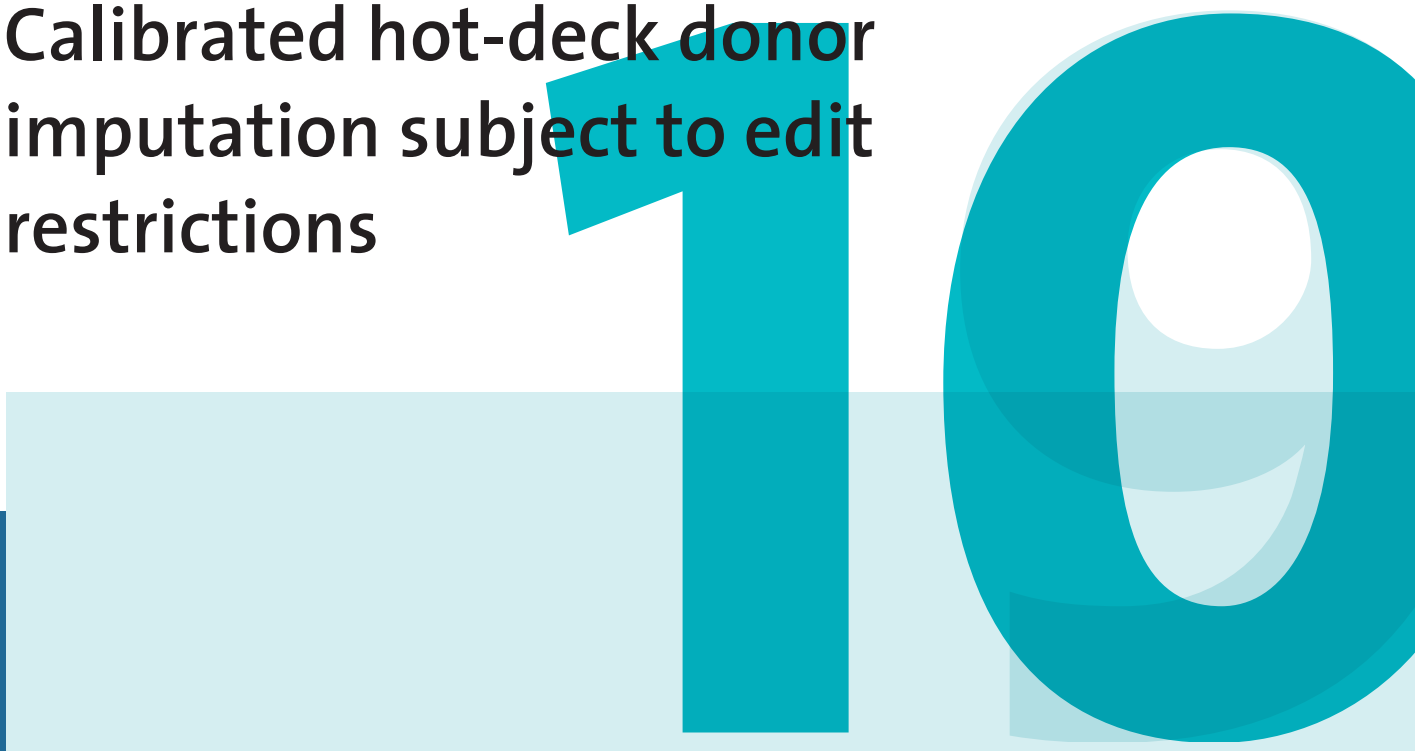
Calibrated hot-deck donor imputation subject to edit restrictions



Wieger Coutinho, Ton de Waal and Natalie Shlomo

Discussion paper (201016)

Statistics Netherlands

The Hague/Heerlen, 2010

Explanation of symbols

.	= data not available
*	= provisional figure
**	= revised provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2008–2009	= 2008 to 2009 inclusive
2008/2009	= average of 2008 up to and including 2009
2008/'09	= crop year, financial year, school year etc. beginning in 2008 and ending in 2009
2006/'07–2008/'09	= crop year, financial year, etc. 2006/'07 to 2008/'09 inclusive

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

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Grafimedia

Cover
TelDesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

60083201016 X-10

Calibrated hot-deck donor imputation subject to edit restrictions

Wieger Coutinho, Ton de Waal and Natalie Shlomo

Summary: A major problem that has to be faced by basically all institutes that collect statistical data on persons or enterprises is that data may be missing in the observed data sets. The most common solution to handle missing data is imputation. At national statistical institutes and other statistical institutes, the imputation problem is further complicated owing to the existence of constraints in the form of edit restrictions that have to be satisfied by the data. Examples of such edit restrictions are that someone who is less than 16 years old cannot be married in the Netherlands, and that someone whose marital status is unmarried cannot be the spouse of the head of household. Records that do not satisfy these edits are inconsistent, and are hence considered incorrect. Another additional problem for categorical data is that the frequencies of certain categories are sometimes known from other sources or have already been estimated. In this paper we develop imputation methods for categorical data that take these edits and known frequencies into account while imputing a record.

Keywords: categorical data, edit restrictions, imputation, population frequencies

1. Introduction

National statistical institutes (NSIs) publish figures on many aspects of society. To this end, these NSIs collect and process data on persons, households, enterprises, public bodies, etc. A major problem that has to be faced is that data may be missing from the collected data sets. Some units that are selected for data collection cannot be contacted or may refuse to respond altogether. This is called unit non-response. Unit non-response is not considered in this paper. For many records, i.e. the data of individual respondents, data on some of the items may be 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. Missing items of otherwise responding units is called item non-response. Whenever we refer to missing data in this paper we will mean item non-response.

In the statistical literature ample attention is paid to missing data. The most common solution to handle missing data in data sets is imputation, where missing values are estimated and filled in. An important problem of imputation is to preserve the

statistical distribution of the data set. This is a complicated problem, especially for high-dimensional data. For more on this aspect of imputation and on imputation in general we refer to several excellent articles and books on imputation, such as Kalton and Kasprzyk (1986), Rubin (1987), Schafer (1997), Little and Rubin (2002), and Longford (2005). Imputation methods can be divided into two broad classes: methods for categorical data and methods for numerical data. In the present paper we focus on imputation of missing categorical data. We assume that the main aim of the NSI is to estimate population frequencies of categories.

At NSIs the imputation problem is further complicated owing to the existence of constraints in the form of edit restrictions, or edits for short, that have to be satisfied by the data. Examples of such edits are that someone who is less than 16 years old cannot be married in the Netherlands, and that someone whose marital status is unmarried cannot be the spouse of the head of household. Records that do not satisfy these edits are inconsistent, and are hence considered incorrect. The problem of missing categorical data having to satisfy edits is examined by Winkler (2003).

Another additional problem for categorical data is that the frequencies of certain categories are sometimes known from other sources or have already been estimated. A population frequency of a category may, for instance, be known from an available related register. Alternatively, previously estimated frequencies may be known, and assumed fixed. In the Dutch Social Statistical Database estimated frequencies are fixed and later used to calibrate estimates of other quantities (see Houbiers, 2004, and Knottnerus and Van Duin, 2006). In fact, this strategy of fixing frequencies and later using these fixed frequencies to calibrate other quantities to be estimated forms the basis of the so-called repeated weighting method: a weighting method designed to obtain unified estimates when combining data from different sources. In this paper we do not use a weighting approach, but instead aim to take these edits and known frequencies into account while imputing a record.

The related problem of imputation of missing numerical data having to satisfy edits and to preserve totals is examined in Pannekoek, Shlomo and De Waal (2008).

The remainder of this paper is organized as follows. Section 2 introduces the edit restrictions we consider in this paper. Section 3 describes the imputation algorithms we have developed for our problem. An evaluation study is described in Section 4. Finally, Section 5 ends the paper with a brief discussion.

2. Edits and frequencies for categorical data

2.1 Edits for categorical data

We denote the number of variables by n. Furthermore, we denote the domain, i.e. the set of all allowed values of a variable i by D_i. In the case of categorical data, an edit j is usually written in so-called *normal form*, i.e. as a collection of sets F_j^i

(i=1,2,...,n):

$$F_j^1 \times F_j^2 \times ... \times F_j^n ,$$

meaning that if for a record with values (v_1, v_2, ..., v_n) we have $v_i \in F_j^i$ for all i=1,2,...,n, then the record fails edit j, otherwise the record satisfies edit j.

For instance, suppose we have three variables: *Marital Status*, *Age* and *Relation to Head of Household*. The possible values of *Marital Status* are "Married", "Unmarried", "Divorced" and "Widowed", of *Age* "< 16 years" and "≥ 16 years", and of *Relation to Head of Household* "Spouse", "Child", and "Other". Suppose we have two edits. The first edit saying that someone who is less than 16 years cannot be married, and the second one that someone who is not married cannot be the spouse of the head of household. In normal form the first edit can be written as

({Married}, {< 16 years}, {Spouse, Child, Other}), (1)

and the second one as

({Unmarried, Divorced, Widowed}, {< 16 years, ≥ 16 years}, {Spouse}). (2)

2.2 Frequencies for categorical data

When a frequency for categorical data is known, for instance because it has already been estimated, this simply means that one knows how many units in the data set should have a specific value for a certain variable. For instance, one may know how many people in the data set have a certain age and how many people in the data set are married, even though some values of the variable *Age* and the variable *Marital Status* are missing in an observed, but incomplete data set. In this paper we assume that for several categories such frequencies are know, and our aim is to obtain a fully imputed data set that preserves these frequencies. In this paper we will also refer to frequencies as totals.

3. The imputation methods

3.1 The basic idea

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

Usually, hot-deck donor imputation is applied multivariately, i.e. several missing values in a record are imputed simultaneously, using the same donor record. For our problem that approach is less suited. If an imputed record fails the edits, all one can do is reject the donor record and 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 a relatively complicated set of edits one may not even be able to find a donor record for a certain recipient record such that the resulting imputed record satisfies all edits.

Even if we were able to find single donor records for all records requiring imputation, this would then solve only part of our problem as the totals would only be preserved in very rare cases.

We therefore apply sequential univariate hot-deck donor imputation, where for each missing value in a record requiring imputation in principle a different donor record may be selected. The variables with missing values are imputed sequentially. For each variable, the records for which the value of the variable under consideration is missing are imputed one by one. Once all records for the variable under consideration have been imputed, the next variable with missing values is considered. The univariate hot-deck imputation methods we apply are described in Subsection 3.2. These univariate hot-deck imputation methods are used to order the possible values for a certain missing field. Whether a value is actually used to impute the missing field depends on whether the edits can be satisfied and the totals can be preserved.

While imputing a missing value, care is taken to ensure that the record can satisfy all edits. Only values of donor records that can result in a consistent record, i.e. a record that satisfies all edits, are eligible to be used. In Subsection 3.3. we explain how we determine whether a value is eligible to be used for imputation. For each record we make a list of eligible values for imputation for the variable under consideration.

An eligible value is only used for actual imputation if the total can be preserved. Before an eligible value is actually used to impute a value, we first check whether the corresponding total can be preserved. If this total cannot be preserved, the value is rejected and the next value on the list of eligible values is selected. This process goes on until we find an eligible value such that the corresponding total can be preserved.

3.2 Univariate hot-deck imputation methods

In this paper we apply two univariate hot-deck donor imputation methods: a nearest-neighbour approach and a random hot-deck approach.

3.2.1 Nearest-neighbour hot-deck imputation

Suppose we want to impute a certain variable v in a record r_0. In the nearest-neighbour approach we calculate for each other record r for which the value of v is not missing (records for which the value of the target variable v is missing are not used as donor records) a distance given by

$$D(r_0, r) = \sum_i w_i(c_i^0, c_i^r),$$

where the sum is taken over all variables except variable *v*, c_i^0 denotes the original value of the *i*-th variable in record r_0, c_i^r the original value of the *i*-th variable in record *r*, and $0 \leq w_i(c_i^0, c_i^r) \leq 1$ a user-specified weight expressing how serious one considers a difference between c_i^0 and c_i^r to be. The weight $w_i(c_i^0, c_i^r)$ equals zero, if $c_i^0 = c_i^r$. The weight $w_i(c_i^0, c_i^r)$ is large if one considers the difference between c_i^0 and c_i^r to be important, and small if one consider the difference to be unimportant. The original value of the *i*-th variable in record r_0, c_i^0, or the original value of the *i*-th variable in record *r*, c_i^r, may be missing. If c_i^0 or c_i^r is missing, we set $w_i(c_i^0, c_i^r)$ to 1.

To impute a missing value, we first select the potential donor value from the record with the smallest distance. If the value is allowed according to the edits, we put this value on an ordered list of potential donor values: the list of eligible values. If that value is not allowed according to the edits, we try the second smallest value, et cetera until we find a donor value that is allowed according to the edits. After all potential donor records have been checked for eligible values, we try all values not occurring in the donor records in a random order. In this way we construct an ordered list of potential donor values.

Note that the potential donor records for a certain recipient are ordered in the same way for each variable with missing values. So, if possible, multivariate imputation, using several values from the first potential donor record on this list, will be used. Only if a value of the first potential donor record cannot be used because this would lead to non-preserved totals, a value from another potential donor record is used.

3.2.2 Random hot-deck imputation

When random hot-deck imputation is applied a random donor record is selected, often within certain subgroups defined by auxiliary data. In our case we use random hot-deck to order the possible values for the missing field. Let *k* denote the number of categories of the variable to be imputed, and let *R* be the total number of records with an observed value for the variable to be imputed. For each category c_k (k=1,…,K) we determine the ratio p_k defined by the number of records with an observed value for the variable to be imputed equal to c_k divided by *R*. We then draw categories c_k (k=1,…,K) without replacement with probabilities p_k (k=1,…,K).

To impute a missing value, we first select the potential donor value that was drawn first. If the value is allowed according to the edits, we put this value on an ordered list of potential donor values. If that value is not allowed, we try the potential donor value that was drawn second, et cetera until we find a donor value that is allowed according to the edits. After all potential donor records have been checked for eligible values, we try all values not occurring in the donor records in a random order. As for nearest-neighbour imputation, we thus construct an ordered list of potential donor values.

In our application we order the potential donor records for a recipient in the same way for each variable. Like in the case of nearest-neighbour imputation, this implies that if possible, multivariate imputation, using several values from the first potential donor record on this list, will be used.

3.3 Satisfying edit restrictions

In order to ensure that the set of edits can be satisfied, we derive so-called implied edits. These implied edits are necessary to guarantee that whenever we impute the current variable to be imputed, the variables remaining to be imputed can indeed be imputed consistently, i.e. such that all edits are satisfied.

To illustrate the use of implied edits, we assume that we have a data set with the three variables *Marital Status*, *Age* and *Relation to Head of Household* and their categories defined in Subsection 2.1. We also assume that these variables have to satisfy edits (1) and (2). Now suppose that both *Marital Status* and *Age* in a certain record are missing, and that the value of *Relation to Head of Household* equals "Spouse". Suppose that we first impute *Age* and subsequently *Marital Status*. In this case we cannot simply ignore the edits involving the variable to be imputed later, *Marital Status*, while imputing *Age*. If we were to ignore the edits involving *Marital Status*, i.e. both edits, we could impute the value "less than 16 years" for the missing value of *Age*. In that case there would be no value for *Marital Status* such that all edits are satisfied.

The edits (1) and (2) imply the implied edit

({Married, Unmarried, Divorced, Widowed}, {< 16 years}, {Spouse}) (3)

which expresses that someone who is less than 16 years of age cannot be the spouse of the head of household. If we take this implied edit into account while imputing the missing value for *Age*, we find that we cannot impute the value "< 16 years" and that only "≥ 16 years" is allowed. When "≥ 16 years" is imputed, *Marital Status* can indeed be imputed in a consistent manner.

To determine the set of edits for the remaining variables to be imputed while imputing the current variable, we use the method proposed by Fellegi and Holt (1976) to eliminate a variable.

To eliminate a variable v_t we start by determining all index sets S such that

$$\bigcup_{j \in S} F_t^j = D_t \tag{4}$$

and

$$\bigcap_{j \in S} F_i^j \neq \varnothing \quad \text{for } j \neq t\,. \tag{5}$$

From these index sets we select the *minimal* ones, i.e. the index sets S that obey (4) and (5), but none of their proper subsets obey (4). Given such a minimal index set S we construct the implied edit

$$\bigcap_{j\in S} F_1^j \times \ldots \times \bigcap_{j\in S} F_{t-1}^j \times D_t \times \bigcap_{j\in S} F_{t+1}^j \times \ldots \times \bigcap_{j\in S} F_n^j \ .$$

For example, if we eliminate variable *Marital Status* from the edits (1) and (2), we obtain the implied edit (3).

By adding the implied edits resulting from all minimal sets S to the current set of edits, and then removing all edits involving the eliminated variable, one obtains a set of edits for the remaining variables. It can be shown that if and only if this set of edits for the remaining variables can be satisfied, a value for the eliminated variable exists such that the original set of edits can be satisfied. We call this the lifting principle: a property – namely that the corresponding set of edits can be satisfied – for a certain number of variables is "lifted" to a higher number of variables. The idea of the proof of the lifting principle is that if a value for the eliminated variable such that the original set of edits can be satisfied did not exist, one would be able to construct a violated implied edit, which would be a contradiction (see Fellegi and Holt, 1976, and De Waal and Quere, 2003, for details of the proof).

For records where multiple values are missing, we now order these variables in some order that we describe later in Subsection 3.5. Next, we eliminate the variables according to this order. Let us assume that, say, the values of variables v_1 to v_m are missing. We first substitute the values of the remaining variables into the original set of edits. This gives a set of edits E_0 that have to be satisfied by variables v_1 to v_m. We then eliminate variable v_1 from E_0 and obtain a set of edits E_1 that have to be satisfied by variables v_2 to v_m. Next, we eliminate variable v_2 from E_1 and obtain a set of edits E_2 that have to be satisfied by variables v_3 to v_m. We continue this process until we eliminate v_{m-1} from E_{m-2}, and obtain a set of edits E_{m-1} for variable v_m. For a single variable, edits simply define a set of allowed values for that variable. So, for variable v_m we now know which values are eligible for imputation. By a repeated application of the lifting principle it can be shown that the original set of edits can be satisfied if and only if v_m satisfies E_{m-1}.

Once we have determined the edit sets E_k (k=0,…,m-1), we can impute the variables in reverse order. That is, we can impute v_m by drawing values by means of one of our hot-deck imputation methods (see Subsection 3.2) until we have selected an eligible value that can also preserve the total for this variable (see Section 3.4). We fill in this value for v_m into the edits in E_{m-2}. This gives us a set of eligible values for variable v_{m-1}. We continue this procedure until we have imputed all variables. What is important here is that whenever we want to impute a certain variable in a certain record, we know the set of eligible values for that variable in this record. We will use this property to preserve totals (see Subsection 3.4).

For instance, suppose that in our example we order the variables as follows: *Marital Status* and then *Age*. We substitute the value of *Relation to Head of Household* into the edits, and obtain the edits

$$(\{\text{Married}\}, \{< 16 \text{ years}\}) \qquad (6)$$

and

$$(\{\text{Unmarried, Divorced, Widowed}\}, \{< 16 \text{ years}, \geq 16 \text{ years}\}) \quad (7)$$

for *Marital Status* and *Age*. In this very simple case we now only have to eliminate one variable, *Marital Status*, and obtain the edit

$$(\{< 16 \text{ years}\}) \quad (8)$$

that has to be satisfied by *Age*. Edit (8) defines the set of eligible values for *Age*: in this case only the value "≥ 16 years" is allowed. If we impute "≥ 16 years" for the missing value of *Age*, we can be sure that a value for *Marital Status* exists such that all edits are satisfied. Imputing the value "≥ 16 years" for *Age* and substituting this value into edits (6) and (7), gives the edit

$$(\{\text{Unmarried, Divorced, Widowed}\}) \quad (9)$$

for *Marital Status*. The set of allowed values for *Marital Status* hence consists of the value "Married" only.

Implied edits are often used to automatically identify erroneous fields in a data set (see Fellegi and Holt, 1976). It is well-known that in that case the number of implied edits may be immense. In our case the number of implied edits is much less, however. In order to identify erroneous fields automatically, one basically has to generate implied edits for every possible subset of the variables. In our case one only has to consider a limited number of possible subsets, because the variables are eliminated in a fixed order. For instance, if there are 5 variables, one, in principle, has to consider 32 subsets (ranging from eliminating no variables to eliminating all 5 variables) for automatic error localisation, and only 6 subsets (ranging from eliminating no variable, eliminating variable 1, eliminating variables 1 and 2, etc., till eliminating variables 1, 2, 3, 4 and 5).

3.4 Preserving totals

In the previous subsection we have explained that whenever we want to impute a certain variable in a record we know the set of eligible values. For every record we now construct such a set of eligible values for the variable to be imputed. Suppose the variable to be imputed has K categories c_1 to c_K. We can then summarise the situation in a table such as shown below.

Table 1. Illustration of the sets of eligible values

	Cat. c_1	Cat. c_2	…	Cat. c_K
Record 1	*	0	…	*
Record 2	1	0	…	0
Record 3	0	*	…	*
…	…	…	…	…
Record n	*	*	…	0
	t_1	t_2		t_K

Here a 0 denotes that the category is not eligible for imputation, a "*" that the category is eligible for imputation and a 1 that this value occurs in the corresponding

record (i.e. the value of the variable under consideration is not missing in that record). The t_k (k=1,…,K) denote the known totals.

Now, we impute the variable under consideration record by record. We select a value from the set of eligible values for the variable to be imputed for record 1. As explained in Section 3.2 the list of eligible values has been constructed using one of our hot-deck approaches. After a category c_x has been drawn, we perform the following checks:

1. Do we not have too many records with the selected category c_x? If so, we reject the selected category c_x and select a new one. If not, we perform the second check.
2. Will it be possible to preserve the totals involving this variable if we accept the selected category c_x? If so, we accept this value, and go to the next record to be imputed. If not, we reject the selected category c_x and select a new one, which is again subjected to the same checks.

Checking whether we will not have too many records with the selected category c_x is trivial: we simply check whether the number of records so far with the value c_x exceeds the total t_x or not.

Checking whether the total can be satisfied if we accept the selected category c_x is a well-known problem from combinatorial mathematics. It is called the Harem Problem. In the Harem Problem several men (the categories in our case) have to choose a specified number (the t_k in our case) of wives (the records in our case) into their harem. The men all specify which women they would like to have in their harem and which women they do not want in their harem (the *'s and the 0's in our case). The 1's in our case correspond to women these men already have in their harem.

Conditions, and a constructive algorithm, for solving this problem are given in Anderson (1989). The underlying idea of the algorithm is to assign records to categories in a simple manner until one gets stuck. Once that happens a specific algorithm (see Anderson, 1989, for details) is applied with the aim to assign one more record to the categories by reshuffling the assignments of records to categories. This algorithm is repeatedly applied until either all records are assigned to categories, or until one again gets stuck.

In the first case we have constructed a solution to this instance of the Harem problem, and we have shown that it is possible to preserve the totals if we accept the selected category c_x. In the second case a solution to the Harem problem to this instance of the Harem problem is not possible.

Note that if, for a certain variable to be imputed, the first record with a missing value has a solution to the Harem problem, by construction all records to be imputed for that variable can also be imputed.

Obviously, for the first record it is generally easier to find solutions to the Harem problem than for later records. That is, for later records it is generally harder to find

suitable imputation values. To avoid that for the same records it is hard to find suitable imputation values for different variables, we first randomize the records each time before we start imputing a new variable.

We illustrate the Harem problem and our approach to the imputation problem by means of a simple example. Suppose that for a certain variable to be imputed we have the Harem problem summarised in Table 2 below.

Table 2. An example of the Harem problem

	Cat. c_1	Cat. c_2	Cat. c_3
Record 1	0	*	*
Record 2	*	*	*
Record 3	0	0	*
Record 4	*	*	*
Record 5	*	0	*
	3	1	1

Now if we select category c_3 for the first record, the Harem problem for the remaining records turns out be infeasible. This is easy to see: to categories c_1 and c_2 in total all 4 records have to be assigned in some way. However, record 3 cannot be assigned to either of these categories. This means that category c_3 is rejected for record 1, and we have to impute category c_2 for this record. The Harem problem for the remaining records is then feasible. In fact, in this case there is only one solution: assign record 1 to category c_2, record 3 to category c_3, and records 2, 4 and 5 to category c_1.

3.5 Order of imputing variables

As noted in the previous subsection, with respect to satisfying edits and totals, there may be a problem only for the first record to be imputed for each new variable. If we cannot find a suitable imputation value for that record, we would have to back-track. That is, we would have to return to a previously imputed variable, and impute one or more missing values for that variable in another way. This would lead to an extremely complicated process, because it is hard to specify beforehand which missing values would have to be imputed in another way and how they would have to be imputed. We would more or less have to explore all possibilities, which is obviously very time-consuming.

In an attempt to avoid this situation from happening, at the start of the imputation process we try to fill in values that de-activate edits, i.e. lead to edits that cannot be violated, by variables to be imputed later. If edits cannot be de-activated because this would lead to an unsolvable instance of the Harem problem for the variable under consideration or simply because no value exists that de-activates edits, we follow the approach described earlier.

4. Evaluation study

In this section we describe an evaluation study we have carried out to evaluate our imputation approaches.

4.1 Evaluation data

The evaluation data set was collected from 2001 UK Census tables. The data set included 1,000 randomly selected households from one area. In the data set we have one record per person in the selected households. In total the data set contained 2,447 records. Each record contained six variables: *Age*, *Ethnicity*, *Employment Status*, *Sex*, *Marital Status* and *Relation to Head of Household*.

For this data set three explicit categorical edits were defined:

- Someone whose age is less than 16 years cannot be employed;
- Someone whose age is less than 16 years cannot be married;
- Someone whose relation to the head of household is husband or wife has to be married.

The collected data set for the 2001 UK Census tables did not contain any missing values. In the evaluation measures this data set is indicated by the subscript *orig*. In this data set we randomly introduced fixed percentages of missing values, using a so-called Missing Completely At Random (MCAR) mechanism (see, for instance, Little and Rubin, 2002, and Schafer, 1997, for details about MCAR mechanisms). In each variable we created exactly the same percentages of missing values. We created 6 data sets with varying percentages of missing values per variable: 1%, 2%, 5%, 10%, 20% and 90%. These data sets were imputed, using the imputation methods described in Section 3. The resulting imputed data set were subsequently compared to the true data. The evaluation measures used for this comparison are discussed in Subsection 4.3. In the evaluation measures each of the imputed data sets is indicated by the subscript *imp*.

4.2 The evaluated methods

We have evaluated three different imputation methods: one based on random hot-deck donor imputation and two based on nearest-neighbour hot-deck imputation. Below we will refer to the method based on random hot-deck donor imputation as the random method.

For both versions based on nearest-neighbour imputation $w_i(c_i^0, c_i^r) = 0$ if $c_i^0 = c_i^r$ and $w_i(c_i^0, c_i^r) = 1$ if $c_i^0 \neq c_i^r$ for all variables except *Age*. The two versions based on nearest-neighbour hot-deck imputation differ with respect to the weights used in the distance function for variable *Age*.

In the distance function the values of *Age* are subdivided into four age groups. In one version of the method based on nearest-neighbour hot-deck

imputation $w_i(c_i^0, c_i^r) = 0$ if c_i^0 is the same age group as c_i^r and $w_i(c_i^0, c_i^r) = 1$ if c_i^0 is a different age group than c_i^r. Below we will refer to this method as the equal nearest neighbour method.

In the other version of the method based on nearest-neighbour hot-deck imputation $w_i(c_i^0, c_i^r) = 0$ if c_i^0 is the same age group as c_i^r, $w_i(c_i^0, c_i^r) = 0.25$ if c_i^0 and c_i^r differ by only one age group, $w_i(c_i^0, c_i^r) = 0.5$ if c_i^0 and c_i^r differ by two age groups, and $w_i(c_i^0, c_i^r) = 0.75$ if c_i^0 and c_i^r differ by three age groups. Below we will refer to this method as the unequal nearest neighbour method.

4.3 Evaluation measures

The imputation methods will be compared using the quality measures described below.

Let *D* represent a frequency distribution for a two-way table produced from the data and let *D(r,c)* be the frequency in the cell in row *r* and column *c.*

Distance metric: We use the Hellinger's Distance defined as:

$$HD(D_{orig}, D_{imp}) = \left\{ 0.5 \sum_{r,c} (\sqrt{D_{orig}(r,c)} - \sqrt{D_{imp}(r,c)})^2 \right\}^{1/2}$$

with *orig* and *imp* referring to the original and imputed tables respectively. The *HD* provides a measure of similarity between two probability distributions typically used for positive or zero counts.

Impact on measure of association:

The first measure is defined as the percent difference in the Cramer's V statistic as:

$$RCV(D_{orig}, D_{imp}) = \frac{100 \times \{CV(D_{imp}) - CV(D_{orig})\}}{CV(D_{orig})}$$

where $CV(D) = \sqrt{\frac{\chi^2}{\min(R-1, C-1)}}$ is the Cramer's V measure of association defined in terms of χ^2, the usual Pearson chi-squared statistic for testing independence in the two-way table, *R* is the number of rows and *C* is the number of columns. The *RCV* provides a measure of attenuation of the association in the table.

The second measure of association is defined as the percent difference in the variance of the cell counts:

$$RV(D_{orig}, D_{imp}) = \frac{100 \times \{V(D_{imp}) - V(D_{orig})\}}{V(D_{orig})}$$

where $V(D) = \frac{\sum_{r,c}(D(r,c) - \overline{D})^2}{RC - 1}$. The *RV* provides a measure of attenuation of the counts in the table indicating whether the cell counts are "flattening" as a result of the imputation.

Impact on an ANOVA analysis: Another form of bivariate analysis consists of comparing proportions in a category of a column (outcome) variable between categories of a row (explanatory) variable. Let $P^c(r) = \frac{D(r,c)}{\sum_c D(r,c)}$ be the proportion in column *c* for row *r* and define the between-row variance of this proportion by:

$BV(P^c) = \frac{\sum_r (P^c(r) - P^c)^2}{R - 1}$ where $P^c = \frac{\sum_r D(r,c)}{\sum_{r,c} D(r,c)}$. The measure is defined as:

$$BVR(P^c_{orig}, P^c_{imp}) = \frac{100 \times \{BV(P^c_{imp}) - BV(P^c_{orig})\}}{BV(P^c_{orig})}$$

The *BVR* provides a measure of attenuation of between group differences in an ANOVA analysis and whether the group proportions are tending towards the overall proportion.

4.4 Evaluation results

Figures 1 through 4 present graphs of quality measures for some main distributions in the tables. The unequal nearest neighbour method provided similar results to the equal nearest neighbour method and hence we compare the random method with the equal nearest neighbour method only in the figures.

Figure 1a presents the Hellinger's Distance (*HD*) on a table of counts spanned by *Age* group and *Employment Status*. For all imputation rates, the equal nearest neighbour method has lower Hellinger's distance compared to the random method. Figure 1b presents the Hellinger's Distance for the table spanned by *Age* group and *Relation to Head of Household* showing similar results.

Figure 1a: Hellinger's distance (HD) on the table Age Group and Employment Status

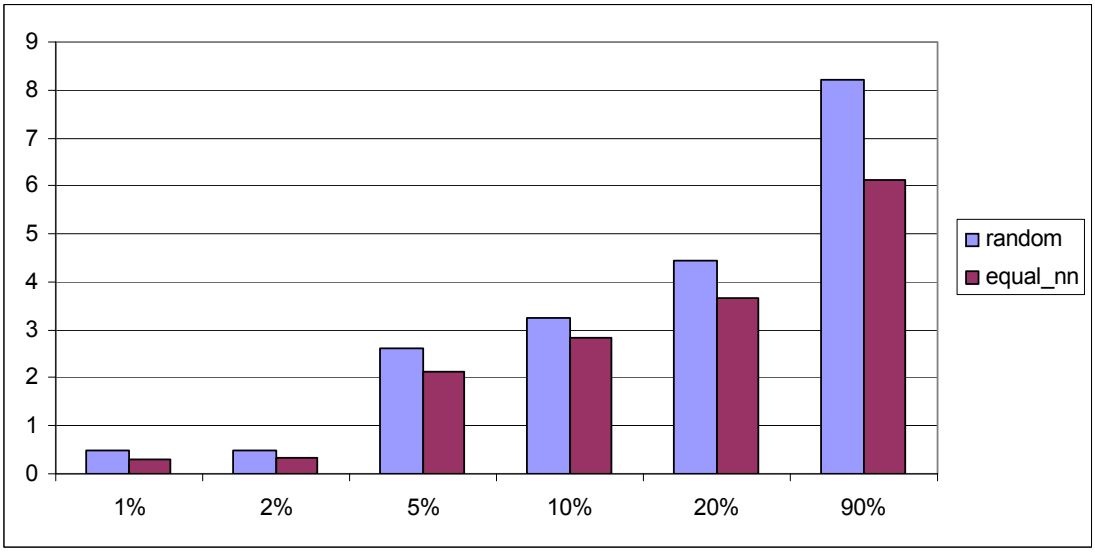


Figure 1b: Hellinger's distance (HD) on the table Age Group and Relation to Head of Household

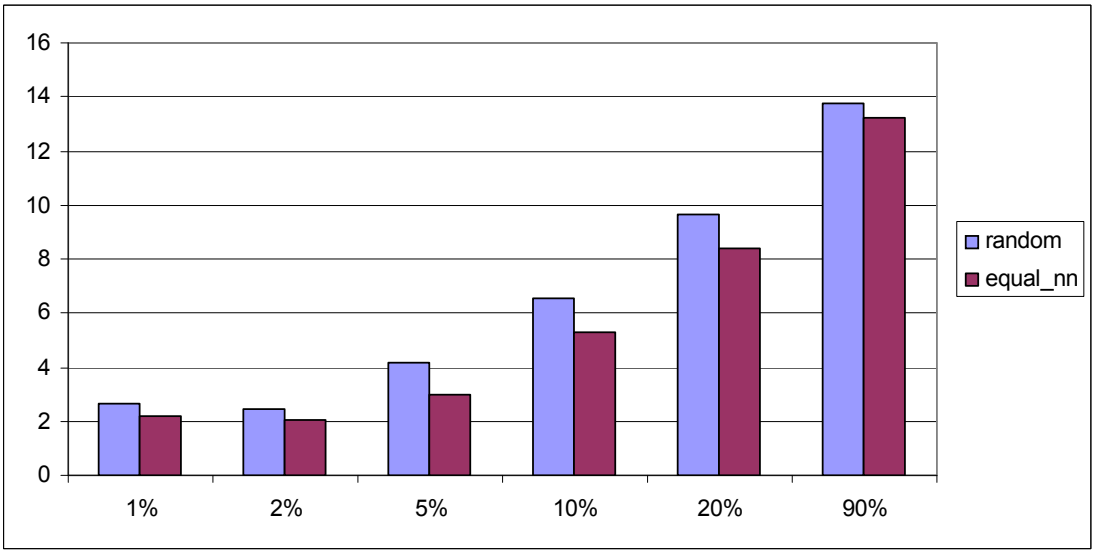


Figure 2a presents the percent relative difference in the variance of the cell counts for the table spanned by *Age* group and *Employment Status*. The negative values of the *RV* measure means that the variance of counts with imputed values is less than the original variance of counts. The cell counts are "flattened" as a result of the imputation, leading to a smaller variance of the counts. The equal nearest neighbour method (as well as the unequal nearest neighbour method) has less change in the variance of the cell counts compared to the random method. Figure 2b presents the *RV* measure for the table spanned by *Age* group and the *Relation to the Head of Household* with similar results.

Figure 2a: Percent relative difference in variance of cell counts (RV) on the table Age Group and Employment Status

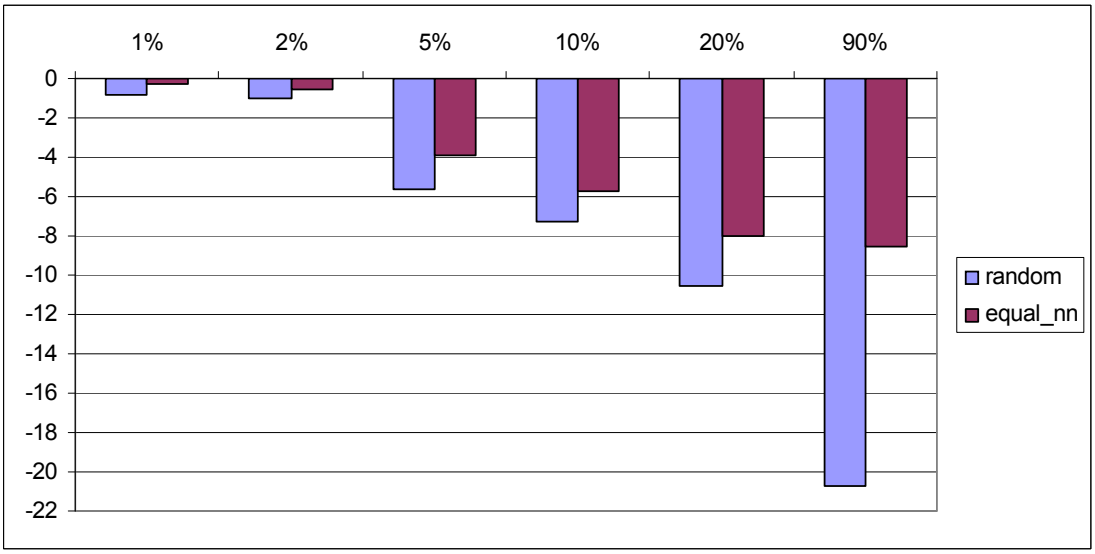


Figure 2b: Percent relative difference in variance of cell counts (RV) on the table Age Group and Relation to Head of Household

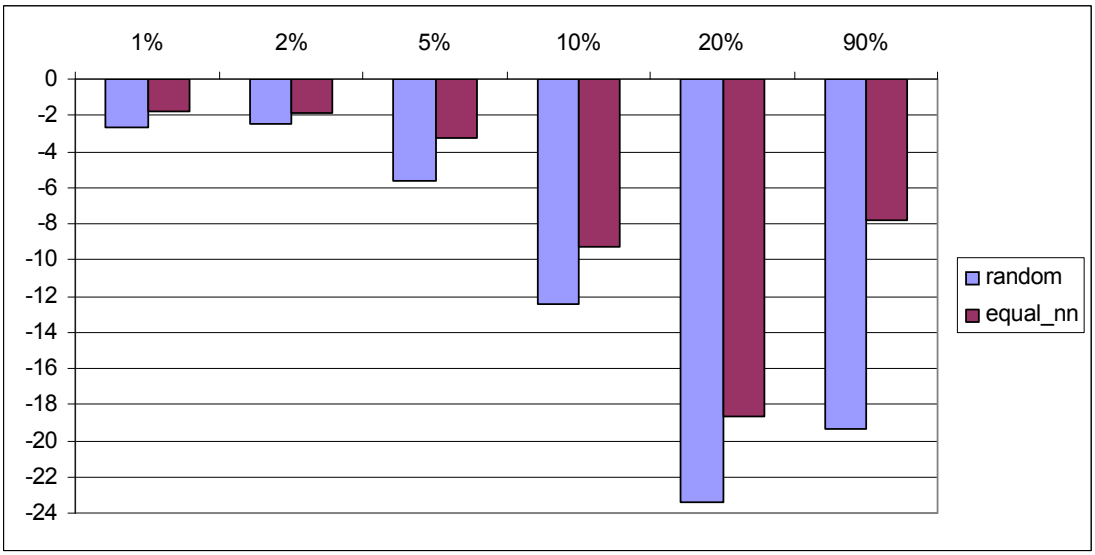


Figure 3 presents the percent relative difference in the between variance of the proportion of employed persons in groups defined by *Sex* and *Age* groups (*BVR*). The negative values of the *BRV* measure means that the between variance of the group proportions of employed persons with imputed values is less than the original between variance. The group proportions are attenuating to the overall proportion as a result of the imputation. Again, equal nearest neighbour method (and the unequal nearest neighbour method) has less change in the *BVR* compared to the random method.

Figure 3: Percent relative difference in between variance (BVR) of proportion of Employed across Sex and Age Groups

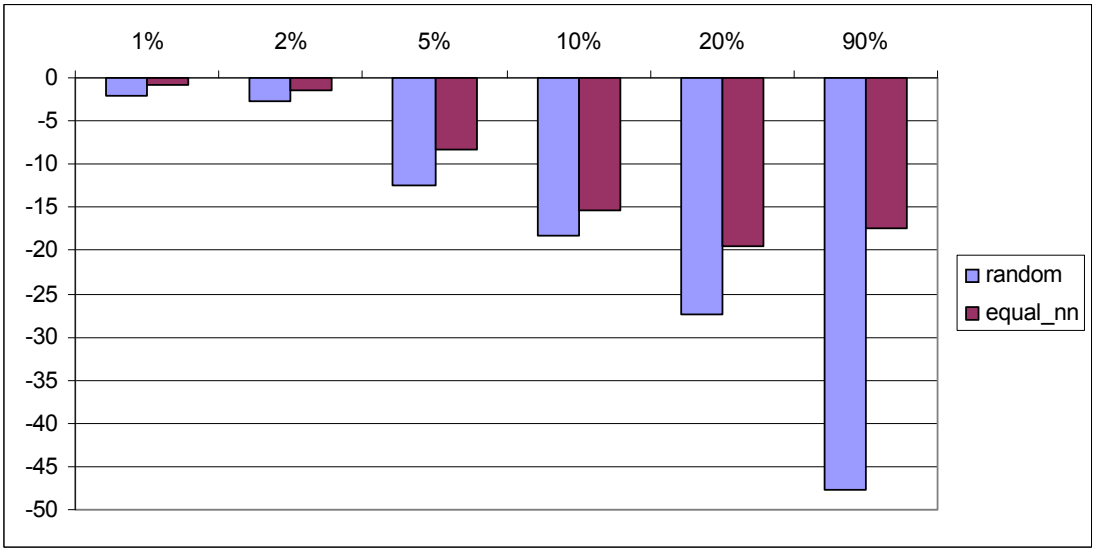


Figure 4a presents the percent relative difference in the Cramer's V statistic of the table spanned by *Age* groups and *Employment Status* (*RCV*). The negative values of the *RCV* measure means that the Cramer's V statistic on the table with imputed values is less than the original Cramer's V statistic. The table of counts is attenuating towards assumptions of independence compared to the original table. For all imputation rates, the equal nearest neighbour method has less change in the Cramer's V statistic than the random method and similarly for the unequal nearest neighbour method. Figure 4b presents the *RCV* measure for the table spanned by *Age* groups and *Relation to Head of Household* with similar results.

Figure 4a: Percent relative difference in Cramer's V on the table of Age Groups and Employment Status

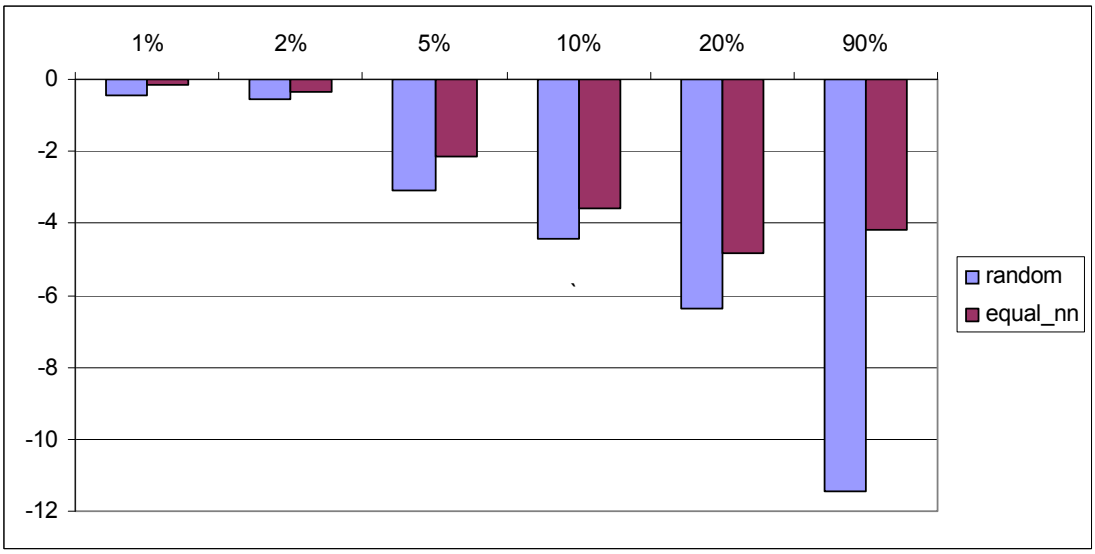
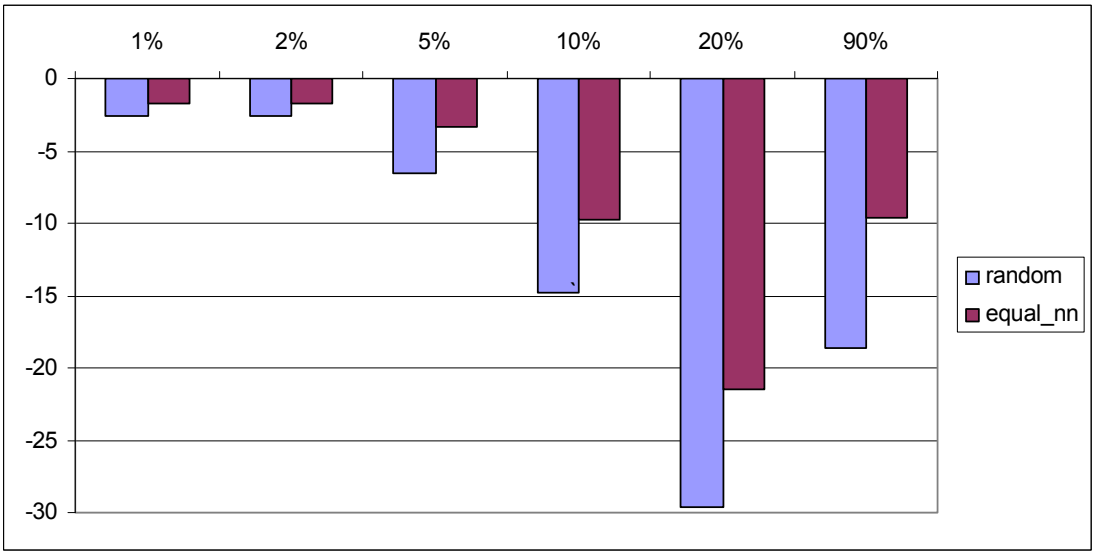


Figure 4b: Percent relative difference in Cramer's V on the table of Age Groups and Relation to the Head of Household



5. Discussion

In this paper we have developed two imputation methods for categorical data that take edits and known frequencies into account while imputing a record. One of the imputation methods proposed in this paper is based on random hot-deck donor imputation and the other on nearest-neighbour donor imputation. Our evaluation study shows the method based on nearest-neighbour imputation to be better than the method based on random imputation. In our evaluation study changing the weights in the distance function of the method based on nearest-neighbour imputation had little or no effect on the outcome of the results.

As far as we are aware the imputation methods proposed in this paper are the first imputation methods for this kind of imputation problem where both edits have to be satisfied and frequencies preserved. As these are the first methods for this kind of problem, many aspects of the developed methods can undoubtedly be improved upon. We encourage the reader to improve the methods proposed in this paper or develop novel methods for this kind of imputation problem himself.

References

Anderson, I. (1989), *A First Course in Combinatorial Mathematics (second edition)*. Oxford University Press, Oxford.

De Waal, T. and Quere, R. (2003), A Fast and Simple Algorithm for Automatic Editing of Mixed Data. *Journal of Official Statistics*, 19, pp. 383-402.

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

Houbiers, M. (2004), Towards a Social Statistical Database and Unified Estimates at Statistics Netherlands, *Journal of Official Statistics*, 20, pp. 55-75.

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

Knottnerus, P. and C. Van Duin (2006), Variances in Repeated Weighting with an Application to the Dutch Labour Force Survey. *Journal of Official Statistics*, 22, pp. 565-584.

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, New York.

Pannekoek, J., N. Shlomo and T. De Waal (2008), *Calibrated Imputation of Numerical Data under Linear Edit Restriction.* UN/ECE Work Session on Statistical Data Editing, Vienna.

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

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

Winkler, W.E. (2003), *Contingency-Table Model for Imputing Data Satisfying Analytic Constraints*. U.S. Bureau of the Census, Washington D.C.