Imputation of numerical data under linear edit restrictions

Discussion paper 7012

Wieger Coutinho, Ton de Waal and Marco Remmerswaal

The authors wish to thank Sander Scholtus for his contributions to this paper and Jeroen Pannekoek, Frank van de Pol and Natalie Shlomo for their comments on earlier drafts of this paper.

Statistics Netherlands

Voorburg/Heerlen, 2007

Explanation of symbols

.	= data not available
*	= provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
0 (0,0)	= less than half of unit concerned
–	= (between two figures) inclusive
blank	= not applicable
2005–2006	= 2005 to 2006 inclusive
2005/2006	= average of 2005 up to and including 2006
2005/'06	= crop year, financial year, school year etc. beginning in 2005 and ending in 2006
2003/'04–2005/'06	= crop year, financial year, etc. 2003/'04 to 2005/'06 inclusive

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

Publisher
Statistics Netherlands
Prinses Beatrixlaan 428
2273 XZ Voorburg

Cover design
WAT ontwerpers, Utrecht

Prepress
Statistics Netherlands - Facility Services

Information
E-mail: infoservice@cbs.nl
Via contact form: www.cbs.nl/infoservice

Where to order
E-mail: verkoop@cbs.nl

Internet
http://www.cbs.nl

ISSN: 1572-0314

6008307012 X-10

Summary: A common problem faced by statistical offices is that data may be missing from collected data sets. The typical way to overcome this problem is to impute the missing data. The problem of imputing missing data is complicated by the fact that statistical data often have to satisfy certain edit rules, which for numerical data usually take the form of linear restrictions. Standard imputation methods for numerical data as described in the literature generally do not take such linear edit restrictions on the data into account. In the present paper we describe two algorithms for imputation of missing numerical data that do take the edit restrictions into account. Both methods assume that the data are approximately multivariately normally distributed. The first method uses the estimated multivariate normal model to impute the missing values and afterwards adjusts the imputed values so they satisfy the edit restrictions. The second method sequentially imputes the missing data in a record. It uses Fourier-Motzkin elimination to determine appropriate intervals for each variable to be imputed. To assess the performance of these two imputation methods an evaluation study is carried out.

Keywords: Fourier-Motzkin elimination, imputation, linear edit restrictions, linear programming

1. Introduction

National statistical institutes (NSIs) publish figures on many aspects of society. To this end, these NSIs collect 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.

Missing data is a well-known problem that has to be faced by basically all institutes that collect data on persons or enterprises. In the statistical literature ample attention is hence 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 Kalton and Kasprzyk (1986), Rubin (1987), Kovar and Whitridge (1995), Schafer (1997), Little and Rubin (2002), and Longford (2005).

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 the profit and the costs of an enterprise have to sum up to its turnover, 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. While imputing a record, we aim to take these edits into account, and thus ensure that the final, imputed record satisfies all edits. The imputation problem at NSIs is hence given by: impute the missing data in the data set under consideration in such a way that the statistical distribution of the data is preserved as well as possible subject to the condition that all edits are satisfied by the imputed data.

For academic statisticians the wish of NSIs to let the data satisfy specified edits may be difficult to understand. Statistically speaking there is indeed hardly a reason to let a data set satisfy edits. However, as Pannekoek and De Waal (2005) explain, NSIs have the responsibility to supply data for many different, both academic and non-academic, users in society. For the majority of these users, inconsistent data are incomprehensible.

They may reject the data as being an invalid source or make adjustments themselves. This hampers the unifying role of NSIs in providing data that are undisputed by different parties such as policy makers in government, opposition, trade unions, employer organizations, etc. As mentioned by Särndal and Lundström (2005, p. 176): "Whatever the imputation method used, the completed data should be subjected to the usual checks for internal consistency. All imputed values should undergo the editing checks normally carried out for the survey".

Simple sequential imputation of the missing data, where edits involving fields that have to be imputed subsequently are not taken into account while imputing a field, may lead to inconsistencies. Consider, for example, a record where the values of two variables, x and y, are missing. Assume these variables have to satisfy three edits saying that x is at least 50, y is at most 100, and y is greater than or equal to x. Now, if x is imputed first without taking the edits involving y into account, one might impute the value 150 for x. The resulting set of edits for y, i.e. y is at most 100 and y is greater than or equal to 150, cannot be satisfied. Conversely, if y is imputed first without taking the edits involving x into account, one might impute the value 40 for y. The resulting set of edits for x, i.e. x is at least 50 and 40 is greater than or equal to x, cannot be satisfied.

In this paper we develop two algorithms for imputation of missing numerical data that do take the edit restrictions into account. Both methods assume that the data are approximately multivariately normally distributed. In fact, in our calculations we will treat this unknown distribution as following a multivariate normal distribution exactly. For data that have to satisfy edits defined by linear inequalities this is surely incorrect, because at best the data could follow a truncated normal distribution but never a regular normal distribution. Our simplification makes it relatively easy to determine marginal and conditional distributions, which are needed for one of the two imputation methods, examined in this paper. In order to estimate the parameters of the multivariate normal distribution, we have used the EM algorithm. As starting values for the EM algorithm we have used the observed means and covariance matrix of the complete cases. Our implementation of the EM algorithm is based on Schafer (1997).

The innovative aspect of our work is the way we ensure that edits are satisfied. Despite the fact that much research on imputation techniques has been carried out, imputation under edits is still a rather neglected area. As far as we are aware, apart from some research at NSIs (see, e.g., Tempelman, 2007) no or hardly any research on this particular aspect of imputation has been carried out.

The remainder of this paper is organised as follows. Section 2 first discusses the kind of linear edits on which we will focus in this paper. Section 3 describes an adjustment method where imputed records are later adjusted so they satisfy the specified edits. A second imputation method is described in Section 5.

A fundamental role in this algorithm is played by Fourier-Motzkin elimination. We refer to this imputation method as the FM method. The Fourier-Motzkin elimination technique itself is explained in Section 4. Section 6 illustrates the FM method by means of an example. An evaluation study and its results are described in Section 7. In that section we compare the results of the adjustment method with the FM method. Finally, Section 8 concludes the paper with a short discussion.

2. Linear edit restrictions

In this paper we focus on linear edits for numerical data. Linear edits are either linear equations or linear inequalities. We denote the number of continuous variables by n, and the variables in a certain record by x_i (i=1,…,n). We assume that edit j (j=1,...,J) can be written in either of the two following forms:

$$a_{1j}x_1 + \ldots + a_{nj}x_n + b_j = 0\,, \tag{2.1a}$$

or

$$a_{1j}x_1 + \ldots + a_{nj}x_n + b_j \geq 0\,. \tag{2.1b}$$

Here the a_{ij} and the b_j are certain constants, which define the edit.

Edits of type (2.1a) are referred to as balance edits. An example of such an edit is

$$T = P + C, \tag{2.2}$$

where T is the turnover of an enterprise, P its profit, and C its costs. Edit (2.2) expresses that the profit and the costs of an enterprise should sum up to its turnover. A record not satisfying this edit is obviously incorrect. Edit (2.2) can be written in the form (2.1a) as $T - P - C = 0$.

Edits of type (2.1b) are referred to as inequality edits. An example is

$$T \geq 0\,, \tag{2.3}$$

expressing that the turnover of an enterprise should be non-negative. An inequality edit such as (2.3), expressing that the value of a variable should be non-negative, is also referred to as a non-negativity edit.

3. An adjustment method

A straightforward approach to let imputed values satisfy specified edits is to use an adjustment method consisting of two steps. In the first step the missing data are imputed without taking the edits (2.1) into account. These missing data can, for instance, be imputed by assuming that the data follow a multivariate normal distribution, and use a standard imputation method for this situation (see, e.g., Little and Rubin, 2002, and Schafer, 1997). As already mentioned, in this paper we indeed assume that the data follow a multivariate normal distribution, and impute the missing data of a record by drawing values from the appropriate estimated conditional distribution for the missing data given the observed values.

We denote the values after the first imputation step for the record under consideration by $x_{\text{first},i}$ (i=1,…,n). In the second step, the adjustment step, the final values in the record under consideration, $x_{\text{final},i}$ (i=1,…,n), are determined by minimising the objective function

$$\sum_i w_i \left| x_{\text{first},i} - x_{\text{final},i} \right| \tag{3.1}$$

subject to the condition that the values $x_{\text{final},i}$ (i=1,…,n) satisfy all edits (2.1) and the condition that for all variables k that were observed $x_{\text{final},k}$ equals the corresponding observed value. The latter condition means that only the values *imputed* in the first imputation step may be modified. The w_i (i=1,…,n) are non-negative weights, reflecting how serious one considers a change of a unit in variable i to be. The problem of minimising the objective function (3.1) subject to the condition that the values $x_{\text{final},i}$ (i=1,…,n) satisfy all edits (2.1) is a linear programming problem, and can, for instance, be solved by means of the well-known simplex algorithm (see, e.g., Chvátal, 1983).

The adjustment method is quite a general and logical approach. In the first step one can apply the imputation method that is best from a statistical point of view for the data under consideration. In the second step the imputed values are (hopefully only slightly) adjusted so they satisfy the specified edits. The method has one important drawback, however. Namely, if after the first step an imputed record does not satisfy the edits, the final, adjusted, record will lie on the boundary of the feasible region, i.e. at least one of the inequality edits will be satisfied with equality. In other words, the number of records that lie on the boundary of the feasible region is completely determined by the first imputation step. Apart from using another imputation method in the first step, one has no way of influencing the number of records that lie on the boundary of the feasible region.

In the next three sections we describe our second imputation method, the FM method. We begin the description of the FM method by explaining Fourier-Motzkin elimination.

4. Eliminating variables by means of Fourier-Motzkin elimination

Fourier-Motzkin elimination (see, e.g., Duffin, 1974, and De Waal and Coutinho, 2005) is a technique to project a set of linear constraints involving m variables onto a set of linear constraints involving m-1 variables. The original set of constraints involving m variables can be satisfied if and only if the corresponding, projected set of constraints involving m-1 variables can be satisfied. The standard version of Fourier-Motzkin elimination handles only inequalities as constraints. We use an extended version of Fourier-Motzkin elimination that can also handle equations. In our application of Fourier-Motzkin elimination the constraints are defined by the edits.

In order to eliminate a variable x_r from the set of current edits by means of Fourier-Motzkin elimination, we start by copying all edits not involving this variable from the set of current edits to a new set of edits Ψ.

If variable x_r occurs in an equation, we express x_r in terms of the other variables. Say, x_r occurs in edit s of type (2.1a), we then write x_r as

$$x_r = -\frac{1}{a_{rs}}\left(b_s + \sum_{i \neq r} a_{is} x_i\right) \qquad (4.1)$$

Expression (4.1) is used to eliminate x_r from the other edits involving x_r. These other edits are hereby transformed into new edits, not involving x_r, that are logically implied by the old ones. These new edits are added to our new set of edits Ψ. Note that if the original edits are consistent, i.e. can be satisfied by certain values u_i (i=1,…,m), then the new edits are also consistent as they can be satisfied by u_i (i=1,…,m; $i \neq r$). Conversely, note that if the new edits are consistent, say they can be satisfied by the values v_i (i=1,…,m; $i \neq r$), then the original edits are also consistent as they can be satisfied by the values v_i (i=1,…,m) where v_r is defined by filling v_i (i=1,…,m; $i \neq r$) into (4.1).

If x_r does not occur in an equality but only in inequalities, we consider all pairs of edits (2.1b) involving x_r. Suppose we consider the pair consisting of edit s and edit t. We first check whether the coefficients of x_r in those inequalities have opposite signs, i.e. we check whether $a_{rs} \times a_{rt} < 0$.

If this is not the case, we do not consider this particular combination (*s*,*t*) anymore. If the coefficients of x_r do have opposite signs, one of the edits, say edit *s*, can be written as an upper bound on x_r, i.e. as

$$x_r \leq -\frac{1}{a_{rs}}\left(b_s + \sum_{i \neq r} a_{is} x_i\right), \tag{4.2}$$

and the other edit, edit *t*, as a lower bound on x_r, i.e. as

$$x_r \geq -\frac{1}{a_{rt}}\left(b_t + \sum_{i \neq r} a_{it} x_i\right). \tag{4.3}$$

Edits (4.2) and (4.3) can be combined into

$$-\frac{1}{a_{rt}}\left(b_t + \sum_{i \neq r} a_{it} x_i\right) \leq x_r \leq -\frac{1}{a_{rs}}\left(b_s + \sum_{i \neq r} a_{is} x_i\right),$$

which yields an implied edit not involving x_r given by

$$-\frac{1}{a_{rt}}\left(b_t + \sum_{i \neq r} a_{it} x_i\right) \leq -\frac{1}{a_{rs}}\left(b_s + \sum_{i \neq r} a_{is} x_i\right). \tag{4.4}$$

The implied edit (4.4) is added to our new set of edits Ψ. After all possible pairs of edits involving x_r have been considered and all implied edits given by (4.4) have been generated and added to Ψ, we delete the original edits involving x_r that we started with. In this way we obtain a new set of edits Ψ not involving variable x_r. This set of edits Ψ may be empty. This occurs when all current edits involving x_r are inequalities and the coefficients of x_r in all those inequalities have the same sign. Note that if the original edits are consistent, say they can be satisfied by certain values u_i (i=1,…,m), then the new edits are also consistent as they can be satisfied by u_i (i=1,…,m; $i \neq r$). This is by definition also true if the new set of edits is empty. Conversely, note that if the new edits are consistent, say they can be satisfied by certain values v_i (i=1,…,m; $i \neq r$), then the minimum of the right-hand sides of (4.4) for the v_i (i=1,…,m; $i \neq r$) is larger than, or equal to, the maximum of the left-hand sides of (4.4) for the v_i (i=1,…,m; $i \neq r$). This implies that we can find a value v_r such that

$$-\frac{1}{a_{rt}}\left(b_t + \sum_{i \neq r} a_{it} v_i\right) \leq v_r \leq -\frac{1}{a_{rs}}\left(b_s + \sum_{i \neq r} a_{is} v_i\right) \text{ for all pairs } s \text{ and } t,$$

which in turn implies that the original edits are consistent. We have demonstrated the main property of Fourier-Motzkin elimination: a set of edits is consistent if and only if the set of edits after elimination of a variable is consistent.

We illustrate Fourier-Motzkin elimination by means of the example below.

Note that as one only has to consider pairs of edits, the number of implied edits is obviously finite.

Example: Suppose there are four variables, T (turnover), P (profit), C (costs), and N (number of employees), and that the edits are given by (2.2), (2.3),

$$P \leq 0.5T\,, \tag{4.5}$$

$$-0.1T \leq P\,, \tag{4.6}$$

$$T \leq 550N\,. \tag{4.7}$$

If we eliminate variable P, we use equation (2.2) to express P in terms of T and C. That is, we use $P = T - C$. After Fourier-Motzkin elimination, we obtain the edits (2.3), (4.7),

$$T - C \leq 0.5T\,, \quad \text{(equivalently:} \quad 0.5T \leq C\,) \tag{4.8}$$

and

$$-0.1T \leq T - C \quad \text{(equivalently:} \quad C \leq 1.1T\,) \tag{4.9}$$

The main property of Fourier-Motzkin elimination says that the set of edits (2.3), and (4.7) to (4.9) for T, C and N can be satisfied if and only if the original set of edits (2.2), (2.3), and (4.7) to (4.9) for T, P, C and N can be satisfied.

This was an example of Fourier-Motzkin elimination if the variable to be eliminated is involved in an equation. We now use the resulting set of edits (2.3), and (4.7) to (4.9) for variables T, C and N to give an example of the elimination of a variable involved in inequalities only. If we eliminate variable C from edits (2.3), and (4.7) to (4.9), we copy the edits not involving C, i.e. edits (2.3) and (4.7). Moreover, we can combine edits (4.8) and (4.9) to obtain

$$0.5T \leq 1.1T\,,$$

which is equivalent to (2.3). So, eliminating C from (2.3) and (4.7) to (4.9) leads to edits (2.3) and (4.7). The main property of Fourier-Motzkin elimination says that the set of edits (2.3) and (4.7) for T and N can be satisfied if and only if the set of edits (2.3), and (4.7) to (4.9) for T, C and N can be satisfied. Combining the two results we have found, we conclude that edits (2.3) and (4.7) for T and N can be satisfied if and only if the original set of edits (2.2), (2.3), and (4.7) to (4.9) for T, P, C and N can be satisfied. ■

5. An imputation method based on Fourier-Motzkin elimination

The FM method consists of the following steps:

0. Assume a statistical model for the data, and estimate the model parameters. As mentioned before, in our application we assume that the data are multivariately normally distributed, and we use the EM algorithm to estimate the model parameters.

For each record to be imputed we apply Steps 1 to 5 below until all records have been imputed.

1. Fill in the values of the non-missing data into the edits. This leads to a set of edits $E(0)$ involving only the variables to be imputed.
2. Use Fourier-Motzkin elimination to eliminate the variables to be imputed from these edits until only one variable remains. The set of edits after the i-th variable to be imputed has been eliminated is denoted by $E(i)$. The final set of edits defines a feasible interval for the remaining variable. Select this remaining variable as the current variable to be imputed.
3. Draw a value for the variable currently selected to be imputed from the conditional distribution of the selected variable given all known values; either observed or already imputed ones.
4. If the drawn value lies inside the feasible interval, accept it and go to Step 5. If it lies outside the feasible interval, reject it and return to Step 3.
5. We stop when all variables have been imputed. Otherwise, we fill in the drawn value for the selected variable k into the edits in $E(k-2)$. This defines a feasible interval for the $(k-1)$-th eliminated variable. Select the $(k-1)$-th eliminated variable as the current variable to be imputed, and go to Step 3. ■

Note that the theory developed in Section 4 implies that if the record to be imputed can be imputed consistently, the feasible interval determined in Step 2 or 5 is never empty.

If the feasible interval determined in Step 2 has width 0, there is only one feasible value for the variable under consideration. In this case it is not necessary to draw a value in Step 3. Instead we immediately impute the only feasible value. In some other cases the width of the feasible interval determined in Step 2 may be rather small. In those cases many values may need to be drawn before a value inside the feasible interval is drawn. We therefore set a limit, N_{draw}, on the number times that a value for a particular variable may drawn. If this limit is reached, and no value inside the feasible has been drawn, the last value drawn is set to the nearest value of the feasible interval.

By means of N_{draw} one can indirectly control the number of imputed records on boundary of the feasible region defined by the edits. If N_{draw} is set to a low value, relatively many imputed records will be on this boundary; if N_{draw} is set to a high value, relatively few imputed records will be on the boundary.

In our implementation of the FM method we have eliminated, and hence imputed, the variables in a fixed order. All variables in each data set to be imputed are beforehand randomly assigned a unique number from 1 to the number of variables. In Step 2 of the above algorithm we then select the variable to be imputed according to the assigned numbers, i.e. the variable with the lowest assigned number with a missing value is eliminated first, followed by the variable with the second lowest assigned number with a missing value, and so on. The order in which the variables are imputed is fixed in the sense that – although this order is randomly determined beforehand – the same order is used for all records, and even over all experiments with the same data set.

6. Illustration of the FM method

In this section we illustrate the FM method by means of an example. In our example, we assume that we are given a data set with some missing values, that there are four variables, *T*, *P*, *C* and *N*, and that the edits are given by (2.2), (2.3) and (4.5) to (4.7).

We focus on Steps 1 to 5 of the algorithm for a specific record, and assume that the model parameters, means $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, of the multivariate normal distribution estimated in Step 0 of our algorithm are given by

$$\boldsymbol{\mu} = (1000, 200, 500, 4)$$

and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 13500 & 3000 & 10500 & 60 \\ 3000 & 2500 & 500 & 10 \\ 10500 & 500 & 10000 & 50 \\ 60 & 10 & 50 & 1 \end{pmatrix}.$$

Here the first column/row corresponds to *T*, the second column/row to *P*, the third column/row to *C*, and the fourth column/row to *N*.

Now, suppose that for a certain record in our data set we have $N = 5$, and that the values for *T*, *P* and *C* are missing. We first fill in the observed value for *N* into the edits (2.2), (2.3) and (4.5) to (4.7) (Step 1 of our algorithm). We obtain (2.2), (2.3), (4.8), (4.9) and

$$T \leq 2750, \tag{6.1}$$

Now, we sequentially eliminate the variables for which the values are missing from the edits. We start by eliminating P from (2.2), (2.3), (4.8), (4.9) and (6.1). This leads to the edits (2.3), (6.1) and

$$T - C \leq 0.5T \qquad \text{(equivalently: } 0.5T \leq C\text{)}, \qquad (6.2)$$

$$-0.1T \leq T - C \qquad \text{(equivalently: } C \leq 1.1T\text{)}. \qquad (6.3)$$

Edits (2.3) and (6.1) to (6.3) have to be satisfied by C and T.

We next eliminate variable C, and obtain (2.3), (6.1) and

$$0.5T \leq 1.1T\,, \qquad (6.4)$$

Edit (6.4) is equivalent to (2.3). The edits that have to be satisfied by T are hence given by (2.3) and (6.1). The feasible interval for T is therefore given by [0, 2750]. We have now completed Step 2 of our algorithm.

To impute T, we determine the distribution of T, conditional on the value for variable N. The distribution of T turns out to be N(1060, 9900), the normal distribution with mean 1060 and variance 9900. We draw values from this distribution until we draw a value inside the feasible interval (Steps 3 and 4 of the algorithm). Suppose we draw the value 1200.

We fill in the imputed value for T into the edits for C and T, i.e. edits (2.3) and (6.1) to (6.3) (Step 5 of the algorithm). We obtain

$$600 \leq C\,,$$

$$C \leq 1320\,,$$

$$1200 \geq 0\,,$$

$$1200 \leq 2750\,.$$

The feasible interval for C is hence given by [600, 1320]. We determine the distribution of C, conditional on the values for variables N and T. This distribution turns out to be N(656.11, 18181.18). We draw values from this distribution until we draw a value inside the feasible interval (Steps 3 and 4 of the algorithm). Suppose we draw the value 700.

We fill in the imputed values for C and T into the edits that have to be satisfied by C, T and P, i.e. edits (2.2), (2.3), (4.8), (4.9) and (6.1) (Step 5 of the algorithm). We obtain

$$1200 = P + 700\,,$$

$$P \leq 600\,,$$

$$-120 \leq P\,,$$

$$1200 \geq 0\,,$$

$$1200 \leq 2750\,.$$

There is only one feasible value for P, namely 500. The imputed record we obtain is given by $T = 1200$, $C = 700$, P = 500, and $N = 5$. ■

7. Evaluation study

7.1 Evaluation data

For our evaluation study we have used three data sets: a data set with actually observed data from a business survey, data set R^{all}, the same data set but without balance edits, data set R^{ineq}, and a data set with synthetic data, data set S. The main characteristics of these data sets are presented in Table 1.

Table 1. The characteristics of the evaluation data sets.

	Data set R^{all}	Data set R^{ineq}	Data set S
Total number of records	3,096	3,096	500
Number of records with missing values	544	469	490
Total number of variables	8	7	10
Total number of edits	14	12	16
Number of balance edits	1	0	3
Total number of inequality edits	13	12	13
Number of non-negativity edits	8	7	9

The actual values for data set R^{all}, and hence also for data set R^{ineq}, are all known. In the completely observed data set values were deleted by a third party, using a mechanism unknown to us. Data set R^{ineq} was constructed in order to examine the effects of balance edits on the results. In fact, we have removed the balance edit from data set R^{all} in two different ways. First of all, we have only removed the balance edit, but have left all involved variables in the data set. As a consequence, the estimated covariance matrix will be singular and the balance edit will be automatically satisfied by the imputed data, if the parameters of the normal distribution are estimated by means of the EM algorithm using the complete cases to obtain a first estimate for the model parameters as we do in our application. We refer the interested reader to Chapter 4 in Tempelman (2007) for a proof. The evaluation results should hence be the same as for the case where all edits are used, apart from some minor differences due to the stochastic nature of the methods used. This is confirmed by our evaluation study (results not reported in this paper). Second, we have removed the balance edit, one of the variables involved in this balance edit, and its associated non-negativity edit. R^{ineq} is the resulting data set.

The removed variable, R_4, does not occur in any of the other edits apart from its associated non-negativity edit. Removing this variable has the effects that the estimated covariance matrix is non-singular, and that the balance edit will not be automatically satisfied.

Data set S is indirectly based on an observed business survey and its corresponding edits. This observed data was used to estimate the parameters of a multivariate normal model by means of the EM algorithm. Next, data set S was generated by drawing from the estimated multivariate normal model. If a drawn vector did not satisfy all specified edits it was rejected, else it was accepted. In this way 500 vectors were generated. Missing values were generated by randomly deleting for each variable a specified number of values. The number of values deleted was (much) higher than in the actually observed business survey in order to evaluate the performance of our imputation methods for a very complicated situation.

For all three 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 means of the 8, respectively 7, variables of data set R^{all} and data set R^{ineq} are given in Table 2 and those of the 10 variables of data set S in Table 3. The means are taken over all observations in the complete versions of the data sets.

Table 2. The numbers of missing values and the means of the variables of data sets R^{all} and R^{ineq}.

Variable	Number of missing values	Mean
R_1	76	11,574.83
R_2	79	777.56
R_3	130	8,978.70
R_4^*	147	1,034.07
R_5	68	10,012.77
R_6	67	169.24
R_7	73	209.86
R_8	0	37.41

* Data set R^{ineq} does not contain variable R_4.

Variable R_8 does not contain any missing values and is only used as auxiliary variable.

Table 3. The numbers of missing values and the means of the variables of data set S.

Variable	Number of missing values	Mean
S_1	120	97.77
S_2	180	175,018.30
S_3	240	731.03
S_4	120	175,749.33
S_5	180	154,286.53
S_6	180	7,522.34
S_7	180	8,519.65
S_8	180	1,277.04
S_9	120	171,605.57
S_{10}	120	4,143.76

7.2 Evaluation measures

To measure the performance of our imputation methods we use a d_{L1} measure, an m_1 measure, an *rdm* measure, and the number of imputed records on the boundary of the feasible region defined by the edits, i.e. the number of records for which at least one inequality edit is satisfied with equality. The first three criteria have been proposed by Chambers (2003), and have been used in an evaluation study by Pannekoek and De Waal (2005). The d_{L1} measure is the average distance between the imputed and true values defined as

$$d_{L1} = \frac{\sum_{i \in M} w_i \left| \hat{y}_i - y_i^* \right|}{\sum_{i \in M} w_i},$$

where $\hat{y}_i$ is the imputed value in record i of the variable under consideration, M denotes the set of records with imputed values for variable y and w_i is the raising weight for record i.

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

$$m_1 = \left| \sum_{i \in M} w_i (\hat{y}_i - y_i^*) \Big/ \sum_{i \in M} w_i \right|.$$

The *rdm* (relative difference in means) measure is defined as

$$rdm = \frac{\sum_{i \in M} \hat{y}_i - \sum_{i \in M} y_i^*}{\sum_{i \in M} y_i^*}.$$

Smaller values of the above three measures indicate better imputation performance. The number of imputed records on the boundary of the feasible region defined by the edits should be close to the actual number of records on the boundary for the complete versions of the data sets.

To remain consistent with the literature, in particular with the previously published 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.

We use the measures in a relative way, namely to compare the adjustment method to the FM method. The measures are neither necessarily appropriate nor sufficient to measure the impact of imputation on the quality of survey estimates in general. For an actual production process it depends on the intended use of the data whether record level accuracy (d_{L1}) or more aggregate measures of imputation bias like m_1 or *rdm* are more important. Furthermore, to assess the importance of bias caused by imputation it should be related to other quality aspects, such as sampling variance.

7.3 Evaluation results

Both imputation methods described in this paper are of a stochastic nature as they depend on drawing vectors from a probability distribution. To reduce the effects of the stochastic nature of our methods we have repeated each evaluation experiment 10 times, and have calculated the average of these 10 experiments. Unless stated otherwise the value of N_{draw} for the FM method (see Section 5) is set to 160 in our experiments. The results for data set R^{all} are presented in Table 4 for the adjustment method and Table 5 for the FM method.

Table 4. Evaluation results for the adjustment method on data set R^{all}

Variable	d_{L1}	m_1	*rdm*
R_1	2069.20	1145.80	0.15
R_2	226.91	108.27	0.17
R_3	158.79	106.63	-0.04
R_4	532.81	531.39	3.58
R_5	14.81	14.81	-0.01
R_6	41.00	40.617	2.65
R_7	86.37	75.14	1.42

Table 5. Evaluation results for the FM method on data set R^{all}

Variable	d_{L1}	m_1	*rdm*
R_1	3141.27	2593.45	0.34
R_2	277.30	222.28	0.34
R_3	176.55	142.97	-0.05
R_4	189.06	160.18	0.44
R_5	65.59	54.20	0.00
R_6	13.59	13.17	0.90
R_7	83.83	80.40	1.77

Variable R_8 does not have any missing values, so no evaluation results for R_8 are presented.

The results for data set R^{ineq} are presented in Table 6 for the adjustment method and Table 7 for the FM method.

Table 6. Evaluation results for the adjustment method on data set R^{ineq}

Variable	d_{L1}	m_1	*rdm*
R_1	1868.20	256.14	-0.26
R_2	205.16	34.67	-0.38
R_3	1490.70	1452.00	-0.99
R_5	1227.90	541.04	-0.49
R_6	2783.80	2783.80	592.50
R_7	14.40	12.03	-0.54

Table 7. Evaluation results for the FM method on data set R^{ineq}

Variable	d_{L1}	m_1	*rdm*
R_1	3101.88	2717.77	0.33
R_2	271.09	216.54	0.29
R_3	360.10	279.12	-0.09
R_5	1837.67	1756.52	0.14
R_6	13.77	13.37	0.84
R_7	92.65	89.43	1.93

The results for data set S are presented in Table 8 for the adjustment method and Table 9 for the FM method.

Table 8. Evaluation results for the adjustment method on data set S

Variable	d_{L1}	m_1	*rdm*
S_1	466.17	452.17	4.63
S_2	44304.04	42833.46	-0.26
S_3	32441.33	32332.18	43.50
S_4	28114.87	673.03	0.00
S_5	56780.58	49973.13	-0.33
S_6	33203.49	28916.08	3.72
S_7	22135.07	15792.76	1.77
S_8	75627.02	75118.57	56.62
S_9	127862.42	104736.50	0.60
S_{10}	145238.77	104194.00	-26.82

Table 9. Evaluation results for the FM method on data set S

Variable	d_{L1}	m_1	*rdm*
S_1	13943.12	13916.90	142.57
S_2	17440.92	8066.39	0.05
S_3	9941.38	9767.14	13.14
S_4	32672.09	31633.86	0.19
S_5	11404.99	5274.79	-0.04
S_6	2221.02	1430.56	0.18
S_7	3472.59	1405.63	0.16
S_8	5062.49	4818.50	3.63
S_9	5715.68	3569.85	0.02
S_{10}	28261.21	28064.01	7.22

It is hard to draw conclusions from Tables 4 to 9. For some variables the adjustment method leads to better results than the FM method. For other variables the opposite happens. This is not very surprising as both methods rely on the same statistical model for drawing imputation values. In order to draw some conclusions we examine how often one methods leads to better results than the other, where "better" is defined as "closer to zero". For data set R^{all}, the results for the adjustment method in Table 4 are in 13 cases better than those for the FM method in Table 5. The opposite happens in 8 cases. For data set R^{ineq}, the results for the adjustment method in Table 6 are in 10 cases better than those for the FM method in Table 7. The opposite happens in 8 cases. For data set S, the results for the adjustment method in Table 8 are in 5 cases better than those for the FM method in Table 9. The opposite happens in 25 cases. From this we conclude that for the relatively easy to impute data sets R^{all} and R^{ineq} the results for d_{L1}, m_1, and *rdm* are slightly better for the adjustment method than for the FM method. The inclusion or exclusion of the balance edit in R^{all}, respectively R^{ineq} does not seem to affect the results much. However, for the very complicated data set S the FM method leads to clearly better results than the adjustment method. This is probably caused by the fact that in the FM method the values imputed cannot be too far from their true values as each separately imputed value is at worst on the boundary of its feasible interval. This imputed value is later used as predictor in order to impute other missing values. In the adjustment method the values imputed in the first step may be far from their true values. For the complicated data set S, this is apparently not, or in any case to an insufficient extent, corrected in the adjustment step.

The final evaluation measure we consider is the number of records on the boundary of the feasible region defined by the edits. In Table 10 the average number of records on the boundary of the feasible region over 10 evaluation experiments for the adjustment method and the FM method on data sets R^{all}, R^{ineq}, and S are presented. For the FM method we show the results for three different values of N_{draw}, namely the values 1, 160 and 1000. The value of N_{draw} used is mentioned between brackets. The results for d_{L1}, m_1, and *rdm* for $N_{draw} = 1$ and $N_{draw} = 1000$ (not presented here) are comparable to the results presented in Tables 5, 7, and 9, where $N_{draw} = 160$. In Table 10 we also present the number of records on the boundary of the feasible region for the complete versions of the three mentioned data sets. In almost all cases records of these data sets lie on the boundary of the feasible region because a variable that has to satisfy a non-negativity edit attains the value zero.

Table 10. (Average) number of records on the boundary of the feasible region defined by the edits.

	Average number for FM method (1)	Average number for FM method (160)	Average number for FM method (1000)	Average number for the adjustment method	Actual number for complete data
Data set R^{all}	499.8	467.1	467.0	499.6	495
Data set R^{ineq}	435.7	396.4	396.0	437.3	424
Data set S	178.3	162.4	162.2	178.9	2

Table 10 shows that the result for data set R^{all} for the adjustment method is closer to the actual number of records on the boundary of the feasible region defined by the edits for the complete data than the FM method for any of the three values of N_{draw}. For data set R^{ineq} it depends of the value of N_{draw} which method leads to a result that is the closest to the actual number of records on the boundary for the complete data. For data set S the results of the FM method are closer to the actual number of records on the boundary for the complete data than the adjustment method for any of the three values of N_{draw}.

Table 10 also shows the effect of the parameter N_{draw} of the FM method: the higher N_{draw}, the less records will generally be on the boundary of the feasible region. By means of N_{draw} one can indirectly control the number of records on the boundary of the feasible region. The results of the FM method with $N_{draw} = 1$ in Table 10 are quite close to the results for the adjustment method. This was to be expected as in both methods only one value for each missing value is drawn, using the same underlying statistical model. This drawn value may be later modified. The main difference between the FM method with $N_{draw} = 1$ and the adjustment method is that in the FM method each drawn value may be modified directly after it has been drawn, whereas in the adjustment method all drawn value are later modified simultaneously.

If one wants, for the FM method, the number of imputed records on the boundary of the feasible region defined by the edits to be close to the actual number of records on the boundary, one should choose N_{draw} between 1 and 160 for data sets R^{all} and R^{ineq}. Data set S appears to be too complicated for both the adjustment and the FM method. The number of imputed records on the boundary of the feasible region is too high for both methods. By increasing the value of N_{draw} the number of records on the boundary decreases only slowly for the FM method. Increasing the value of N_{draw} also leads to an increase of the computing time, however. So, although one can influence the number of records on the boundary of the feasible region by changing the value of N_{draw}, the effect of changing the value of N_{draw} is limited, in any case for complicated data sets such as S. The drawback of the adjustment method noted in Section 3 that the number of records on the boundary of the feasible region for this method is completely determined by the first imputation step does not appear to be a major disadvantage in comparison to the FM method – at least not for our evaluation data – as the results of the adjustment method are not clearly worse than the FM method in this respect.

We have also applied a logarithmic transformation to the data before applying our imputation methods. The results were similar to the results presented in this paper.

8. Discussion

In this paper we have described two imputation methods that lead to imputed data that satisfy specified edits. For the data sets in our evaluation study we conclude that in the worst case (data sets R^{all} and R^{ineq}) the FM method leads to comparable or slightly worse evaluation results as the adjustment method. In the best case (data set S) the FM method leads to clearly better results than the adjustment method. The FM method seems to have a built-in mechanism to protect itself from imputing very wrong values. Such a mechanism seems to be lacking from the adjustment method. Our study is, however, very limited and more research is necessary before we can draw any definite conclusions.

In our application of the adjustment method we have used a linear objective function. The main reason for using a linear objective function is that this is easy to implement in a software program. The results of the adjustment method may be possibly improved by using a quadratic objective function instead of our linear one. In any case, for statisticians, minimising a quadratic objective function is more natural and often more logical than minimising a linear objective function.

The FM method has the advantage that one can, indirectly, control the number of records on the boundary of the feasible region defined by the edits. The price that has to be paid for this is that the algorithm is more complex than the algorithm for the adjustment method.

Moreover, the effect of this indirect control over the number of records on the boundary of the feasible region seems limited. From a purely practical point of view, the adjustment method may therefore be a better choice in many cases.

For data set S, far too many records lie on the boundary of the feasible region for both the adjustment method and the FM method. For almost all records on the boundary one or more non-negativity edit is satisfied with equality, i.e. the value of the involved variable equals zero. The fact that far too many non-negativity edits are satisfied with equality strongly indicates that the assumed statistical model, which in our application is assumed to follow a multivariate normal distribution, is incorrect. In order to improve the statistical results of the two imputation methods presented in this paper, the underlying statistical model should be improved. Further research is required to develop such better statistical models as well as computationally tractable methods to handle such methods.

As mentioned in Section 5, in our implementation of the FM method we have imputed the variables in a fixed order that is randomly determined beforehand. The results of the method are, however, likely to be influenced by the order in which the variables are imputed. More research is needed to examine whether the results improve if the data set is multiply imputed, using several different, randomly determined, orders in which the missing data are imputed. More research is also needed in order to determine the best order for the variables to be imputed. A likely candidate for the best order is to impute the missing values according to the accuracy of the imputed values, i.e. impute the missing value that can be imputed most accurately first, followed by the missing value that can be imputed second most accurately, and so on. Future research will have to decide whether this is indeed the best order, or if there are even better imputation orders.

When imputing a missing value in a record in our implementation of the FM method, we use the previously imputed values in this record as auxiliary information. In this way we try to preserve the correlation structure between the imputed values as much as possible. Again, the results of the method are likely to be influenced by the order in which the missing values are imputed. The same questions mentioned in the paragraph above have to be answered.

Using previously imputed values in order to impute a missing value has an obvious drawback: if the stochastic imputation process leads to a bad imputed value, this affects all subsequently imputed values in this record. It remains to be examined if the results of the FM method improve, or deteriorate, if we do not use the previously imputed values as auxiliary information but instead use only the observed data as auxiliary information.

The imputation methods we have developed in this paper can be applied to general linear edit restrictions. If only non-negativity edits are specified, one could possibly also use tobit and logit models instead of our methods. Such models automatically ensure that the variable to be imputed attains a non-negative value. The use of tobit or logit models for imputation subject to non-negativity edits remains to be examined.

We have made the implicit assumption in this paper that apart from values being missing, the data are otherwise correct. In practice, this is rarely true. Data observed in practice generally contain errors. The existence of errors in the data complicates the imputation problem as presented in this paper. It may, in fact, be impossible to impute the missing values in such a way that all edits become satisfied, because the errors in the record inevitably lead to inconsistency. For a data set arising in practice it is therefore recommended to edit the data first, for instance by applying software for automatic editing. An overview of algorithms for automatic editing is given by De Waal and Coutinho (2005).

References

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

Chvátal, V. (1983), *Linear Programming*. W.H. Freeman and Company, New York.

De Waal, T. and W. Coutinho (2005), Automatic Editing for Business Surveys: An Assessment of Selected Algorithms. *International Statistical Review 73*, pp. 73-102.

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

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

Kovar, J. en 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, 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.

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.

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

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