SORT 35 (1) January-June 2011, 39-62

Imputation of numerical data under linear edit restrictions

Wieger Coutinho[1], Ton de Waal[2] and Marco Remmerswaal[3]

Abstract

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 generally do not take such edit restrictions into account. In the present article we describe two general approaches for imputation of missing numerical data that do take the edit restrictions into account. The first approach imputes the missing values by means of an imputation method and afterwards adjusts the imputed values so they satisfy the edit restrictions. The second approach sequentially imputes the missing data. It uses Fourier-Motzkin elimination to determine appropriate intervals for each variable to be imputed. Both approaches are not based on a specific imputation model, but allow one to specify an imputation model. To illustrate the two approaches we assume that the data are approximately multivariately normally distributed. To assess the performance of the imputation approaches an evaluation study is carried out.

MSC: 62-02, 62-07, 90C90

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

[1] Tarwekamp 172, 2592 XN The Hague, The Netherlands

[2] Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands (tel. +31 70 337 4930, e-mail: t.dewaal@cbs.nl or tonwaal@planet.nl).

[3] Rijswijk University of Professional Technical Education, Lange Kleiweg 80, 2288 GK Rijswijk, The Netherlands
Received: June 2009
Accepted: November 2010

respond altogether. This is called unit non-response. Unit non-response is not considered in this article. 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 article 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 article we develop two general approaches for imputation of missing numerical data that ensure that edits are satisfied, while at the same time allowing one to specify a statistical imputation model. 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) hardly any research on general approaches to imputation under edit restrictions has been carried out. An exception is imputation based on a truncated multivariate normal model (see, e.g., Geweke, 1991, and Tempelman, 2007). Imputation based on a truncated multivariate normal model can take the edit restrictions we consider in this article into account. Using this model has two drawbacks, however. First of all, the truncated multivariate model is computationally very demanding and complex to implement in a software program. Second, it is obviously only suited for data that (approximately) follow a truncated multivariate normal distribution, not for data that follow other distributions. Some software packages developed by NSIs, such as GEIS (Kovar and Whitridge, 1990), SPEER (Winkler and Draper, 1997), SLICE (De Waal, 2001) and Banff (Banff Support Team, 2008), also ensure that edits are satisfied after imputation. However, these packages only apply relatively simple imputation models, whereas our approaches allow more complicated imputation models.

Both approaches we describe in this article allow one to separate the imputation model from how the edits are handled. In other words, the two approaches described are not based on a specific imputation model, but allow one to specify an imputation model. For both approaches a broad class of imputation models can be applied.

To illustrate the two approaches we will assume in this article that the data are approximately multivariately normally distributed. In fact, in our calculations we will treat the unknown distribution of the data as being 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 approaches examined in this article. We only use the (approximate) multivariate normal model to illustrate how our general approaches can actually be applied in practice. We have selected the (approximate) multivariate normal model for computational convenience. We certainly do not want to suggest that this model is the most appropriate one for the data sets we have used in our evaluation study. Another

computationally convenient choice would have been to use hot-deck imputation instead of the (approximate) multivariate normal model.

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 remainder of this article is organised as follows. Section 2 first discusses the kind of linear edits on which we will focus in this article. Section 3 describes an adjustment approach where imputed records are later adjusted so they satisfy the specified edits. A second imputation approach is described in Section 5. A fundamental role in this approach is played by Fourier-Motzkin elimination. We refer to this imputation approach as the FM approach. The Fourier-Motzkin elimination technique itself is explained in Section 4. Section 6 illustrates the FM approach by means of an example. An evaluation study and its results for the (approximate) multivariate normal model are described in Section 7. In that section we compare the results of the adjustment approach with the FM approach for the multivariate normal imputation model. Finally, Section 8 concludes the article with a short discussion.

2. Linear edit restrictions

In this article 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 themselves by x_i $(i = 1, \ldots, n)$. We assume that edit j $(j = 1, \ldots, J)$ can be written in either of the two following forms:

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

or

$$a_{1j}x_1 + \ldots + a_{nj}x_n + b_j \geq 0. \tag{2}$$

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

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

$$T = P + C, \tag{3}$$

where T is the turnover of an enterprise, P its profit, and C its costs. Edit (3) 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 (3) can be written in the form (1) as $T - P - C = 0$.

Edits of type (2) are referred to as inequality edits. An example is

$$T \geq 0, \tag{4}$$

expressing that the turnover of an enterprise should be non-negative. An inequality edit such as (4), expressing that the value of a variable should be non-negative, is also referred to as a non-negativity edit.

3. An adjustment approach

A straightforward approach to let imputed values satisfy specified edits is to use an adjustment approach consisting of two steps. In the first step the missing data are imputed without taking the edits (1) and (2) 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 article we illustrate our approaches by indeed assuming 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 refer to this as the first imputation step.

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

$$\sum_i w_{\text{adj},i} \left| x_{\text{first},i} - x_{\text{final},i} \right| \tag{5}$$

subject to the condition that the values $x_{\text{final},i}$ $(i = 1,\ldots,n)$ satisfy all edits (1) and (2) and the condition that for all variables x_i that were observed $x_{\text{final},i}$ equals the corresponding observed value. The latter condition means that only the values *imputed* in the first imputation step may be modified. In (5) the $x_{\text{first},i}$ $(i = 1,\ldots,n)$ are known values and the $x_{\text{final},i}$ $(i = 1,\ldots,n)$ are the unknowns. The $w_{\text{adj},i}$ $(i = 1,\ldots,n)$ are non-negative adjustment weights, reflecting how serious one considers a change of a unit in variable x_i to be.

The adjustment weights $w_{\text{adj},i}$ $(i = 1,\ldots,n)$ can be calculated in many ways. In our application we have set $w_{\text{adj},i} = 1/\bar{x}_{\text{first},i}$, where $\bar{x}_{\text{first},i}$ is the average value of the i-th variable. In this way, the objective function (5) takes the relative deviation between $x_{\text{first},i}$ and $x_{\text{final},i}$ rather than the absolute deviation into account. All the weights $w_{\text{adj},i} = 1/\bar{x}_{\text{first},i}$ $(i = 1\ldots,n)$ were indeed non-negative.

The problem of minimising the objective function (5) subject to the condition that the values $x_{\text{final},i}$ $(i = 1,\ldots,n)$ satisfy all edits (1) and (2) can be formulated as a linear

programming problem by introducing additional variables u_i $(i = 1, \ldots, n)$ and adding the constraints

$$u_i \geq x_{\text{first},i} - x_{\text{final},i} \tag{6}$$

and

$$u_i \geq x_{\text{final},i} - x_{\text{first},i}. \tag{7}$$

It is easy to see that the problem of minimising the objective function

$$\sum_i w_{\text{adj},i} u_i \tag{8}$$

subject to (6), (7), the condition that the values $x_{\text{final},i}$ $(i = 1, \ldots, n)$ satisfy all edits (1) and (2) and the condition that for all variables x_i that were observed $x_{\text{final},i}$ equals the corresponding observed value yields the same optimal value for the objective function (8) and the same optimal values for $x_{\text{final},i}$ $(i = 1, \ldots, n)$ as minimising (5) subject to the condition that the values $x_{\text{final},i}$ $(i = 1, \ldots, n)$ satisfy all edits (1) and (2), and the condition that for all variables x_i that were observed $x_{\text{final},i}$ equals the corresponding observed value (see also Chvátal, 1983).

In the problem of minimising (8) subject to (6), (7), the condition that the values $x_{\text{final},i}$ $(i = 1, \ldots, n)$ satisfy all edits (1) and (2), and the condition that for all variables x_i that were observed $x_{\text{final},i}$ equals the corresponding observed value, the $x_{\text{final},i}$ and u_i $(i = 1, \ldots, n)$ are the unknowns. This linear programming problem can, for instance, be solved by means of the well-known simplex algorithm, an interior-point algorithm (see, e.g., Chvátal, 1983, and Nemhauser and Wolsey, 1988) or a generalized reduced gradient method (see, e.g., Lasdon et al., 1978).

The adjustment approach is quite a general and logical approach. In the first step one can apply the imputation method and imputation model that are 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 main strength of the adjustment approach is its simplicity: one does not need to implement complicated algorithms in a computer program or buy special-purpose software. Standard software, such as Excel, suffices to implement the adjustment approach. In our application we have indeed used the solver offered by Excel. To be precise: we have used the generalized reduced gradient method as implemented in the GRG2 code of the Excel solver (see Lasdon et al., 1978, and Fylstra, 1998). We have used the GRG2 code of Excel instead of the implementation of the simplex algorithm in Excel as we noted that the GRG2 method as implemented in Excel resulted in a larger number of records not "lying on the boundary of the feasible region defined by the edits".

In our evaluation study we pay special attention to the number of records "lying on the boundary of the feasible region defined by the edits". In this article we define a

record to "lie on the boundary of the feasible region defined by the edits" if at least one of the inequality edits is satisfied with equality. We are aware that this is a ambiguous definition, and also one that differs from the usual definition of "lying on the boundary" as used in the theory of linear programming. Namely, our definition of "lying on the boundary of the feasible region defined by the edits" is dependent on how the edits are stated, rather than only on the shape of the feasible region. For instance, an edit given by "$x = y + z$" can also be expressed as two inequality edits: "$x \leq y + z$" and "$x \geq y + z$". In the latter case, *all* records will lie on the boundary of the feasible region defined by the edits after imputation according to our definition. In our definition we implicitly assume that edits are stated as balance edits instead of (pairs of) inequality edits whenever possible. In all practical situations occurring at statistical offices we have encountered so far this was always the case.

The reason why we pay special attention to the number of records on the boundary of the feasible region defined by the edits is that, when the adjustment approach is applied, a record that does not satisfy the edits after the first imputation step, will often be adjusted in such a way that the final, adjusted, record lies on the boundary of the feasible region defined by the edits.

In the next three sections we describe our second imputation approach, the FM approach. We begin the description of the FM approach 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 (1), we then write x_r as

$$x_r = -\frac{1}{a_{rs}} \left(b_s + \sum_{i \neq r} a_{is} x_i \right) \tag{9}$$

Expression (9) 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 $(i = 1,\ldots,m)$, then the new edits are also consistent as they can be satisfied by u_i $(i = 1,\ldots,m; i \neq r)$. Conversely, note that if the new edits are consistent, say they can be satisfied by values v_i $(i = 1,\ldots,m; i \neq r)$, then the original edits are also consistent as they can be satisfied by the values v_i $(i = 1,\ldots,m)$ where v_r is defined by filling v_i $(i = 1,\ldots,m; i \neq r)$ into (9).

If x_r does not occur in an equality but only in inequalities, we consider all pairs of edits (2) 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

$$x_r \leq -\frac{1}{a_{rs}} \left(b_s + \sum_{i \neq r} a_{is} x_i \right), \tag{10}$$

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

$$x_r \geq -\frac{1}{a_{rt}} \left(b_t + \sum_{i \neq r} a_{it} x_i \right). \tag{11}$$

Edits (10) and (11) 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

$$-\frac{1}{a_{rt}} \left(b_t + \sum_{i \neq r} a_{it} x_i \right) \leq -\frac{1}{a_{rs}} \left(b_s + \sum_{i \neq r} a_{is} x_i \right). \tag{12}$$

The implied edit (12) 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 (12) 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, for instance, when all current edits are inequalities involving x_r 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 $(i = 1,\ldots,m)$, then the new edits are also consistent as they can be satisfied by u_i

$(i = 1,\ldots,m; i \neq r)$. 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 $(i = 1,\ldots,m; i \neq r)$, then the minimum of the right-hand sides of (12) for the v_i $(i = 1,\ldots,m; i \neq r)$ is larger than, or equal to, the maximum of the left-hand sides of (12) for the v_i $(i = 1,\ldots,m; i \neq r)$. This implies that we can find a value v_r such that

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

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. Note that as one only has to consider pairs of edits, the number of implied edits is obviously finite. We illustrate Fourier-Motzkin elimination by means of the example below.

Example: Suppose there are four variables, T (turnover), P (profit), C (costs), and N (number of employees), and that the edits are given by (3), (4),

$$P \leq 0.5T, \tag{13}$$

$$-0.1T \leq P, \tag{14}$$

$$T \leq 550N. \tag{15}$$

If we eliminate variable P, we use equation (3) to express P in terms of T and C. That is, we use $P = T - C$. After Fourier-Motzkin elimination, we obtain the edits (4), (15),

$$T - C \leq 0.5T, \quad (\text{equivalently: } 0.5T \leq C) \tag{16}$$

and

$$-0.1T \leq T - C \quad (\text{equivalently: } C \leq 1.1T). \tag{17}$$

The main property of Fourier-Motzkin elimination says that the original set of edits (3), (4), and (13) to (15) for T, P, C and N can be satisfied if and only if the set of edits (4), and (15) to (17) for T, 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 (4), and (15) to (17) 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 (4), and (15) to (17), we first copy the edits not involving C, i.e. edits (4) and (15). Moreover, we can combine edits (16) and (17) to obtain

$$0.5T \leq 1.1T, \tag{18}$$

which is equivalent to (4). So, eliminating C from (4), and (15) to (17) leads to edits (4) and (15). The main property of Fourier-Motzkin elimination says that the set of edits (4), and (15) to (17) for T, C and N can be satisfied if and only if edits (4) and (15) for T and N can be satisfied. Combining the two results we have found, we conclude that the original set of edits (3), (4), and (13) to (15) for T, P, C and N can be satisfied if and only if edits (4) and (15) for T and N can be satisfied. □

5. An imputation approach based on Fourier-Motzkin elimination

The FM approach consists of the following steps:

0. Assume a statistical imputation model for the data, and – if necessary for the model – estimate the model parameters.

We order the variables to be imputed from the variable with the most missing values to the variable with the least missing values. If two of more variables have the same number of missing values, we order them in an arbitrary way. For each record to be imputed, we apply Steps 1 to 5 below. We repeat this process 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 for the record under consideration.
2. Use Fourier-Motzkin elimination to eliminate the variables to be imputed for the record under consideration from set of edits $E(0)$ in the fixed order described above 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. Set k equal to the number of variables to be imputed for the record under consideration.
3. Draw a value for the k-th variable to be imputed.
4. If the drawn value lies inside the feasible interval $E(k-1)$, accept it and go to Step 5. If it lies outside the feasible interval, reject it and return to Step 3.
5. If $k = 1$, all variables have been imputed and we stop. 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. We update k by $k := k-1$, 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.

In Step 0 one can either assume an implicitly defined statistical imputation model, for instance when one wants to apply hot-deck imputation, or an explicitly defined imputation model, such as the multivariate normal model like we do in this article. In both cases we suggest to draw a value for the variable to be imputed from the conditional distribution of the selected variable given all known values, either observed or already imputed ones.

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 of times that a value for a particular variable may be 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 the 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.

The variables are imputed in reverse order of elimination. Since we have ordered the variables to be imputed from the variable with the most missing values to the variable with the least missing values before applying Steps 1 to 5 of the above algorithm, the variables are imputed in order of increasing number of missing values. That is, the variable with the least missing values is imputed first and the variable with the most missing values last.

As mentioned before, to illustrate our approaches we assume in this article that the data are multivariately normally distributed, and we use the EM algorithm to estimate the model parameters.

It is well known that in the worst case Fourier-Motzkin elimination can be computationally very expensive. However, the imputation problems arising in practice at statistical offices only have a limited number of variables and edits. The largest problems we are aware of have a few hundreds of variables and slightly more than 100 edits. For realistic problems of this limited size, Fourier-Motzkin elimination is generally sufficiently fast. In fact, it has been shown for the related – but computationally much more demanding – error localization problem of the same size in terms of variables and edits that in practical cases arising at statistical offices the computational performance of Fourier-Motzkin elimination is generally acceptable (see De Waal and Coutinho, 2005, and De Waal, 2005). In our application of Fourier-Motzkin elimination in this article to small imputation problems, the computing time of Fourier-Motzkin elimination was negligible, i.e. close to 0 seconds, for all runs. Once the parameters of the multivariate

normal distribution had been determined by means of the EM algorithm, imputing the missing values took only took a few seconds for the entire data sets. Moreover, in the imputation process, the bulk of the computing time for the FM approach was spent on drawing values from the multivariate normal distribution rather than on Fourier-Motzkin elimination

The main reason for developing the FM approach is the fact that promising results have been obtained by so-called sequential imputation methods. Sequential imputation methods are a well-known class of imputation methods, see, e.g., Van Buuren and Oudshoorn (1999 and 2000), Raghunatan et al. (2001) and Rubin (2003). These imputation methods sequentially impute the variables and allow a separate imputation model to be specified for each variable. By imputing all variables containing missing data in turn and iteratively repeating this process several times, the statistical distribution of the imputed data generally converges to an unspecified multivariate distribution. The main strength of sequential imputation is its flexibility: rather than using one multivariate imputation model for all variables simultaneously, which is generally computationally demanding and complex to handle, one can specify a different imputation model for each variable. Sequential imputation methods can be extended to ensure that they satisfy edits. In principle, the FM approach can be implemented as a sequential imputation approach that allows such an extension, although in our illustration we assume a multivariate normal distribution as imputation model rather than separate imputation models for the variables to be imputed (see Tempelman, 2007, and Pannekoek, Shlomo and De Waal, 2008, for other extensions of sequential imputation to ensure that edits are satisfied).

Of course, the adjustment approach may also be used in a sequential imputation approach, namely one may first use a sequential imputation approach and later adjust the imputed values so they satisfy the edits. A fundamental difference between this approach and the FM approach is that in the adjustment approach the imputed values are adjusted simultaneously afterwards, whereas in the FM approach each separate imputed value is immediately adjusted in order to ensure that all edits can be satisfied. Immediately adjusting each imputed value in order to ensure that all edits can be satisfied might improve (or deteriorate) the statistical results as subsequent imputed values may depend on previously imputed values (see also Section 7.3).

6. Illustration of the FM approach

In this section we illustrate the FM approach 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 (3), (4) and (13) to (15).

We focus on Steps 1 to 5 of the approach for a specific record. We assume that the data follow a multivariate normal distribution, and assume that the model parameters, means $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, estimated in Step 0 of our approach 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 (3), (4) and (13) to (15) (Step 1 of our approach). We obtain (3), (4), (13), (14) and

$$T \leq 2750, \tag{19}$$

Now, we sequentially eliminate the variables for which the values are missing from the edits. We start by eliminating P from (3), (4), (13), (14) and (19). This leads to the edits (4), (16), (17) and (19). Edits (4), (16), (17) and (19) have to be satisfied by C and T.

We next eliminate variable C, and obtain edits (4), (18) and (19). Edit (18) is equivalent to (4). The edits that have to be satisfied by T are hence given by (4) and (19). The feasible interval for T is therefore given by [0, 2750]. We have now completed Step 2 of our approach.

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 approach). Suppose we draw the value 1200.

We fill in the imputed value for T into the edits for C and T, i.e. edits (4), (16), (17) and (19) (Step 5 of the approach). We obtain

$$1200 \geq 0,$$

$$600 \leq C,$$

$$C \leq 1320,$$

$$1200 \leq 2750.$$

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 approach). 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 (3), (4), (13), (14) and (19) (Step 5 of the approach). We obtain

$$1200 = P + 700,$$

$$1200 \geq 0,$$

$$P \leq 600,$$

$$-120 \leq P,$$

$$1200 \leq 2750.$$

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 data sets R_{all} and R_{ineq} contain raising weights. These raising weights differ across different (strata of) records, and are used in some of our evaluation measures. In data S all raising weights were set to 1. 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, i.e. did not explicitly demand that after imputation the balance edit holds true for all records, 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 approaches used. This is confirmed by our evaluation study (results not reported in this article). Second, we have removed one of the variables, R_4, involved in the balance edit and its associated non-negativity edit from R_{all}. R_{ineq} is the resulting data set. This data set obviously does not have to satisfy any balance edit. The removed variable R_4 does not occur in any of the other edits apart from its associated non-negativity edit.

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 approaches 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 (unweighted) 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

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

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

Variable	Number of missing values	Mean
R_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

variables of data set S in Table 3. The means are taken over all observations in the complete versions of the data sets.

Variable R_8 in data sets R_{all} and R_{ineq} does not contain any missing values and is only used as auxiliary variable.

7.2. Evaluation measures

To measure the performance of our imputation approaches we use several evaluation measures, The first measure we use is the d_{L1} measure proposed by Chambers (2003). This d_{L1} measure is the average distance between the imputed and true values defined as

$$d_{L1} = \frac{\sum_{k\in M} w_k \,|\hat{y}_k - y_k^*|}{\sum_{k\in M} w_k},$$

where $\hat{y}_k$ is the imputed value in record k of the variable under consideration, y_k^* the corresponding true value, M denotes the set of n_{imp} records with imputed values for variable y and w_k is the raising weight for record k.

The second measure we use is the m_1 measure, which has also been proposed by Chambers (2003). This measure, which measures the preservation of the first moment of the empirical distribution of the true values, is defined as

$$m_1 = \left| \sum_{k\in M} w_k(\hat{y}_k - y_k^*) / \sum_{k\in M} w_k \right|.$$

That m_1 measures the preservation of the first moment of the empirical distribution of the true values becomes clear if we rewrite m_1 as

$$m_1 = \left| \sum_{k\in D} w_k(\hat{y}_k - y_k^*) / \sum_{k\in D} w_k \right| \times \left(\frac{\sum_{k\in D} w_k}{\sum_{k\in M} w_k} \right),$$

where D denotes the entire data set. The quantity $\sum_{k\in D} w_k y_k^* / \sum_{k\in D} w_k$ is an estimate for the population mean. So, m_1 is the deviation of the first moment of the empirical distribution from the first moment of the true distribution times the constant factor $\sum_{k\in D} w_k \big/ \sum_{k\in M} w_k$.

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

$$rdm = \frac{\sum_{k\in M} \hat{y}_k - \sum_{k\in M} y_k^*}{\sum_{k\in M} y_k^*}.$$

Smaller absolute values of the above three measures indicate better imputation performance.

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.

The three evaluation measures described so far all measure the deviation of the imputed values from the true values. The next three evaluation measures measure statistical aspects, such as the preservation of the empirical distribution and the preservation of standard errors.

The first of these measures is the percent difference between the standard deviation (STD) of the mean of the imputations to the standard deviation of the mean of the true values:

$$100\frac{(STD_{\text{imp}} - STD_{\text{true}})}{STD_{\text{true}}}$$

A smaller absolute value for the percent difference between the standard deviation of the mean of the imputations to the standard deviation of the mean of the true values indicates better performance.

Another evaluation measure is a sign test using paired data. This sign test can be carried out by creating a new variable that is defined as the difference between the original value and the imputed value. The test with the null hypothesis that the median of the difference is equal to zero is equivalent to the test that the medians of the original and imputed values are equal. The sign statistic is defined as

$$S = (n^+ - n^-)/2$$

where n^+ is the number of values greater than 0 and n^- the number of values less than 0. A small p-value means that we reject the null hypothesis of equal medians. We will interpret this as: a larger p-value indicates better performance.

The next evaluation measure we use is the Kolmogorov-Smirnov non-parametric test statistic (*K-S*). This statistic is used to compare the empirical distribution of the original values to the empirical distribution of the imputed values (also proposed by Chambers, 2003). For unweighted data, the empirical distribution of the original values is defined

as: $F_{y^*}(t) = \sum_{i \in M} I(y_i^* \leq t)/n_{\text{imp}}$, where n_{imp} is again the number of imputed values and I the indicator function. $F_{\hat{y}}(t)$ is defined similarly. The *K-S* is defined as:

$$K\text{-}S = \max_j(|F_{y^*}(t_j) - F_{\hat{y}}(t_j)|),$$

where the $\{t_j\}$ values are the $2n_{\text{imp}}$ jointly ordered original and imputed values of y. A smaller value for *K-S* indicates better performance.

The final evaluation measure we consider is the number of records on the boundary of the feasible region defined by the edits. Records lying on the boundary of the feasible region defined by the edits are outliers in some sense. Too many outliers in this sense, and any other sense, could make the imputed data less suited for certain statistical analyses. The number of records on the boundary defined by the edits should preferably be close to the actual number of records of the true data on the boundary of the feasible region defined by the edits. This evaluation measure is a bit less important than the others, which measure the statistical quality of the imputed values more directly.

We use the measures in a relative way, namely to compare the adjustment approach to the FM approach. 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 which of the evaluation measures is considered more important.

7.3. Evaluation results

Both imputation approaches described in this article 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 approaches 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 approach (see Section 5) is set to 160 in our experiments. The value of 160 for N_{draw} is based on a limited explorative, trial-and-error search, aiming to find an optimal trade-off between the quality of the imputations and the required computing time. The results for data set R_{all} are presented in Table 4 for the adjustment approach and Table 5 for the FM approach.

Table 4: *Evaluation results for the adjustment approach on data set R_{all}.*

Variable	d_{L1}	m_1	*rdm*	*percent difference*	*sign*	*K-S*
R_1	2069.21	1145.83	0.15	0.02	0.57	0.03
R_2	226.91	108.27	0.17	0.22	0.90	0.00
R_3	303.79	285.46	−0.21	−0.13	0.12	0.01
R_4	283.05	263.84	1.79	3.71	0.00	0.01
R_5	16.11	13.21	−0.01	0.00	0.00	0.38
R_6	41.00	40.61	2.65	0.10	0.00	0.03
R_7	86.37	75.14	1.42	1.87	0.07	0.00

Table 5: *Evaluation results for the FM approach on data set R_{all}.*

Variable	d_{L1}	m_1	*rdm*	*percent difference*	*sign*	*K-S*
R_1	3108.97	2633.30	0.34	−0.02	0.00	0.00
R_2	290.66	235.89	0.33	0.11	0.00	0.00
R_3	169.68	130.85	−0.04	0.00	0.02	0.17
R_4	183.83	152.04	0.40	0.16	0.00	0.02
R_5	68.29	61.31	0.01	0.00	0.13	0.74
R_6	27.37	26.87	1.83	−0.40	0.00	0.00
R_7	95.44	92.48	2.17	0.87	0.00	0.00

Variable R_8 does not have any missing values, so no evaluation results for R_8 are presented in Tables 4 and 5. The results for data set R_{ineq} are presented in Table 6 for the adjustment approach and Table 7 for the FM approach.

Table 6: *Evaluation results for the adjustment approach on data set R_{ineq}*

Variable	d_{L1}	m_1	*rdm*	*percent difference*	*sign*	*K-S*
R_1	1868.22	256.14	−0.25	−0.36	0.01	0.10
R_2	205.16	34.67	−0.37	−0.42	0.00	0.00
R_3	1490.74	1451.99	−0.99	−0.93	0.00	0.00
R_5	1227.87	541.04	−0.49	−0.44	0.00	0.00
R_6	2783.81	2783.81	592.50	58.43	0.00	0.00
R_7	14.40	12.03	−0.54	−0.47	0.00	0.36

Table 7: *Evaluation results for the FM approach on data set R_{ineq}*

Variable	d_{L1}	m_1	*rdm*	*percent difference*	*sign*	*K-S*
R_1	3105.74	2719.82	0.33	−0.01	0.00	0.00
R_2	278.66	225.06	0.30	0.10	0.00	0.00
R_3	359.48	277.00	−0.09	0.00	0.00	0.01
R_5	1844.58	1762.83	0.14	−0.01	0.00	0.00
R_6	27.07	26.66	1.78	−0.39	0.00	0.00
R_7	85.50	82.26	1.80	0.60	0.00	0.00

Table 8: *Evaluation results for the adjustment approach on data set S*

Variable	d_{L1}	m_1	*rdm*	*percent difference*	*sign*	*K-S*
R_1	13943.12	13916.90	142.57	863.15	0.20	0.00
S_2	17440.92	8066.39	0.05	0.17	0.10	0.06
S_3	9941.38	9767.14	13.14	68.89	0.00	0.00
S_4	32672.09	31633.86	0.19	0.37	0.00	0.11
S_5	11404.99	5274.79	−0.04	−0.02	0.00	0.35
S_6	2221.02	1430.56	0.18	0.37	0.00	0.00
S_7	3472.59	1405.63	0.16	0.51	0.00	0.01
S_8	5062.49	4818.50	3.63	11.52	0.00	0.00
S_9	5715.68	3569.85	0.02	0.00	0.87	0.95
S_{10}	28261.21	28064.01	7.22	20.89	0.00	0.00

Table 9: *Evaluation results for the FM approach on data set S.*

Variable	d_{L1}	m_1	*rdm*	*percent difference*	*sign*	*K-S*
R_1	62.39	50.19	0.51	5.81	0.01	0.00
S_2	6754.16	2204.84	−0.01	−0.01	0.00	0.11
S_3	3413.06	3268.38	4.40	28.46	0.00	0.00
S_4	4594.46	3229.51	0.02	0.00	0.13	0.66
S_5	35442.70	28136.41	−0.19	0.01	0.56	0.00
S_6	3600.36	2597.57	−0.33	0.59	0.00	0.00
S_7	15202.73	10779.74	1.21	8.49	0.79	0.00
S_8	21984.15	21247.71	16.01	81.38	0.13	0.00
S_9	3959.69	1940.22	0.01	0.00	0.87	0.95
S_{10}	2223.89	1289.30	0.33	3.76	0.20	0.05

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

It is hard to draw conclusions from Tables 4 to 9. For some variables the adjustment approach leads to better results than the FM approach. For other variables the opposite happens. This is not very surprising as both approaches rely on the same statistical model for drawing imputation values, which fails to capture all distributional aspects of the data. In order to draw some conclusions we examine how often one approach leads to better results than the other, where "better" is defined as "closer to zero" for all evaluation measures considered in Tables 4 to 9 except for the sign test. For the sign test "better" is defined in the opposite way, i.e. the larger the *p*-value, the better the performance. For data set R_{all}, the results for the adjustment approach in Table 4 are in 19 cases better than those for the FM approach in Table 5. The opposite happens in 16 cases. For data set R_{ineq}, the results for the adjustment approach in Table 6 are in 13 cases better than those for the FM approach in Table 7. The opposite happens in 15 cases. For data set *S*, the results for the adjustment approach in Table 8 are in 20 cases better than those for the FM approach in Table 9. The opposite happens in 31 cases. From this we conclude that for data sets R_{all} and R_{ineq} the results for the six evaluation measures of the adjustment approach are comparable to the results for the FM approach. The inclusion or exclusion of the balance edit in R_{all}, respectively R_{ineq} does not seem to affect the results much. For the more complicated data set *S* the FM approach leads to slightly better results than the adjustment approach. This is probably caused by the fact that in the FM approach 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 approach 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.

In Table 10 the average number of records on the boundary of the feasible region over 10 evaluation experiments for the adjustment approach and the FM approach on data sets R_{all}, R_{ineq}, and *S* are presented. For the FM approach 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 the six evaluation measures considered before for $N_{\text{draw}} = 1$ and $N_{\text{draw}} = 1000$ (not presented here) are comparable to the results presented in Tables 5, 7, and 9, where $N_{\text{draw}} = 160$. 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 approach (1)	Average number for FM approach (160)	Average number for FM approach (1000)	Average number for the adjustment approach	Actual number for complete data
Data set R_{all}	499.4	468.2	468.0	499.8	495
Data set R_{ineq}	435.8	397.4	397.0	394.1	424
Data set S	200.5	186.6	186.8	185.5	2

Table 10 shows that the result for data set R_{ineq} for the FM approach 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 adjustment approach for any of the three values of N_{draw}. For data set R_{all} it depends of the value of N_{draw} which approach 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 adjustment approach are slightly closer to the actual number of records on the boundary for the complete data than the FM approach for any of the three values of N_{draw}. The difference between the results for the adjustment approach and the FM approach for $N_{\text{draw}} = 160$ are, however, negligible.

Table 10 also shows the effect of the parameter N_{draw} of the FM approach: the higher N_{draw}, the less records will generally lie 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.

If one wants, for the FM approach, 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 for the complete data, 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 approach. The number of imputed records on the boundary of the feasible region is too high for both approaches. By increasing the value of N_{draw} the number of records on the boundary decreases only slowly for the FM approach. 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 approach noted in Section 3 that

the number of records on the boundary of the feasible region for this approach is for a substantial part determined by the first imputation step does not appear to be a major disadvantage in comparison to the FM approach – at least not for our evaluation data – as the results of the adjustment approach are not clearly worse than those of the FM approach in this respect.

8. Discussion

In this article we have described two imputation approaches that lead to imputed data that satisfy specified edits. The main aim of the article was to describe the two general frameworks, which are basically independent of the imputation method or imputation model actually applied. To illustrate how these approaches work in practice we have used a multivariate normal imputation model.

For the data sets in our evaluation study we conclude that, for the multivariate normal imputation model, for 2 of the 3 data sets (R_{all} and R_{ineq}) the FM approach leads to comparable evaluation results as the adjustment approach. For the other data set (data set S) the FM approach leads to (slightly) better than the adjustment approach (see Tables 8 and 9). The FM approach 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 approach. Our study is, however, very limited and more research is necessary before we can draw any definite conclusions.

In our application of the adjustment approach 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 approach may possibly be 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 approach 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 complicated than for the adjustment approach. 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 approach 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 approach and the FM approach. 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 imputation 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 approaches presented in this article, 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 models.

When imputing a missing value in a record in our implementation of the FM approach, 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. 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 approach 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 approaches we have developed in this article 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 approaches. Such models automatically ensure that each 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.

Acknowledgments

The authors would like to thank the reviewers and the editor of the SORT for their useful comments on earlier versions of this article.

References

Banff Support Team (2008). *Functional Description of the Banff System for Edit and Imputation*. Technical Report, Statistics Canada.

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

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

De Waal, T. (2001). SLICE: Generalised software for statistical data editing. In: *Proceedings in Computational Statistics* (ed. J.G. Bethlehem and P.G.M. Van der Heijden), Physica-Verlag, StateNew York, 277–282.

De Waal, T. (2005). Automatic error localisation for categorical, continuous and integer data, *Statistics and Operations Research Transactions*, 29, 57–99.

De Waal, T. and W. Coutinho (2005). Automatic editing for business surveys: an assessment of selected algorithms. *International Statistical Review*, 73, 73–102.

Duffin, R.J. (1974). On Fourier's analysis of linear inequality systems. *Mathematical Programming Studies*, 1, 71–95.

Fylstra, D. L. Lasdon, J. Watson and A. Warren (1998). Design and use of the Microsoft Excel Solver, *Interface*, 28, 29–55.

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

Kalton, G. en D. Kasprzyk (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1–16.

Kovar, J. and P. Whitridge (1990). Generalized edit and imputation system; overview and applications. *Revista Brasileira de Estadistica*, 51, 85–100.

Kovar, J. en P. Whitridge (1995). Imputation of business survey data. In: *Business Survey Methods* (ed. B.G. Cox, D. A. Binder, B.N. Chinnappa, A. Christianson, M. J. Colledge & P.S. Kott), John Wiley & Sons, New York, 403–423.

Lasdon, L.S. and S. Smith (1992). Solving large sparse non-linear programs using GRG, *ORSA Journal on Computing*, 4, 2–15.

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.

Nemhauser, G.L. and L.A. Wolsey (1988). *Integer and Combinatorial Optimisation*. Wiley, 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, 257–286.

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.

Raghunathan, T.E., J.M. Lepkowski, J. Van Hoewyk and P. Solenberger (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27, 85–95.

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

Rubin, D.B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica*, 57, 3–18.

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*. Doctoral thesis, University of Groningen.

Van Buuren, S. and C.G.M., Oudshoorn (1999). *Flexible Multivariate Imputation by MICE*. TNO Preventie en Gezondheid, TNO/PG 99.054, Leiden.

Van Buuren, S. and C.G.M. Oudshoorn C.G.M. (2000). *Multivariate Imputation by Chained Equations: MICE V1.0 User's Manual*. Report PG/VGZ/00.038, TNO Preventie en Gezondheid, Leiden.

Winkler, W.E. and L.A. Draper (1997). The SPEER edit system. *Statistical Data Editing (Volume 2); Methods and Techniques*, United Nations, Geneva.