UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Working Paper M09/17

Methodology

Calibrated Imputation Of Numerical Data Under Linear Edit Restrictions

Jeroen Pannekoek, Natalie Shlomo, Ton De Waal

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 and that values of variables sometimes have to sum up to known totals. Standard imputation methods for numerical data as described in the literature generally do not take such edit rules and totals into account. In the paper we describe algorithms for imputation of missing numerical data that do take edit restrictions into account and that ensure that sums are calibrated to known totals. The methods sequentially impute the missing data, i.e. the variables with missing values are imputed one by one. To assess the performance of the imputation methods a simulation study is carried out as well as an evaluation study based on a real dataset.

CALIBRATED IMPUTATION OF NUMERICAL DATA UNDER LINEAR EDIT RESTRICTIONS

Jeroen Pannekoek[1], Natalie Shlomo[2] and Ton De Waal[1]

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 and that values of variables sometimes have to sum up to known totals. Standard imputation methods for numerical data as described in the literature generally do not take such edit rules and totals into account. In the paper we describe algorithms for imputation of missing numerical data that do take edit restrictions into account and that ensure that sums are calibrated to known totals. The methods sequentially impute the missing data, i.e. the variables with missing values are imputed one by one. To assess the performance of the imputation methods a simulation study is carried out as well as an evaluation study based on a real dataset.

Keywords: imputation, linear edit restrictions, benchmarking

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

[1]Jeroen Pannekoek, Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands

[2]Natalie Shlomo, Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom

[3]Ton de Waal, Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands

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 be referring to 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), Longford (2005) and references therein.

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. 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). Another additional problem is that data sometimes have to sum up to known totals. While imputing a record, we aim to take these edits and known totals into account. As far as we know, imputation methods that ensure that edits are satisfied and at the same time ensure that totals are preserved have not yet been developed. Taking known totals into account obviously improves the quality of the imputations, at least with respect to the preservation of totals since the imputed totals exactly equal the known totals.

The problem of imputing missing data in records having to satisfy edits such that at the same time known totals are satisfied can arise in the context of a survey amongst a subpopulation of enterprises. Often large enterprises, i.e. enterprises with a number of employees exceeding a certain threshold value, are integrally observed. Some of those enterprises may, however, not provide answers to all questions, and some may even not answer any question at all. Totals corresponding to this subpopulation of enterprises may be known from other sources, e.g. from available register data, or may already have been estimated from other sources. As data of enterprises usually have to satisfy edits, imputation of such a dataset then naturally leads to the problem we consider in the present paper.

The remainder of this paper is organized as follows. Section 2 introduces the edit restrictions we consider in this paper. Section 3 develops a number of imputation algorithms for our problem. Section 4 describes evaluation measures that will be used to compare the imputation algorithms. A simulation study is described in Section 5 and an application on a real dataset is described in Section 6. Finally, Section 7 concludes with a brief 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_j (j=1,…,n). We assume that edit k (k=1,...,K) can be written in either of the two following forms:

$$a_{1k}x_1 + \ldots + a_{nk}x_n + b_k = 0, \tag{1a}$$

or

$$a_{1k}x_1 + \ldots + a_{nk}x_n + b_k \geq 0. \tag{1b}$$

Here the a_{jk} and the b_k are certain constants, which define the edit.

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

$$T = P + C, \tag{2}$$

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

Edits of type (1b) are referred to as inequality edits. An example is

$$T \geq 0 , \tag{3}$$

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

3. Imputation Algorithms Satisfying Edits and Totals

To illustrate how to deal with edit restrictions and (population) totals, we consider a case where we have r records with only three variables as shown in Table 1.

[PLACE TABLE 1 HERE]

These columns contain missing values that require imputation. Just as the observed data, the imputed data have to satisfy the following edit restrictions:

$$x_{i1} + x_{i2} = x_{i3} \tag{4}$$

$$x_{i1} \geq x_{i2} \tag{5}$$

$$x_{i3} \geq 3x_{i2} \tag{6}$$

$$x_{ij} \geq 0 \ (j=1,2,3), \tag{7}$$

in addition the following (population) total restrictions have to be satisfied

$$\sum_{i=1}^{r} x_{ij} = X_j \ (j=1,2,3), \tag{8}$$

where we assume that the population totals are given and consistent with each other, i.e. the totals X_j (j=1,2,3) satisfy the edits (4) to (7).

In this paper we sequentially impute the variables with missing data. Suppose we impute variable j. In order to impute a certain missing field x_{ij}, we first fill in the observed and previously imputed values for the other variables in record i into the edits. This leads to a reduced set of edits involving only the variables to be imputed. For instance, if in the above example (4) to (7) the observed values of variable x_1

in record i equals 10 and the values of variables x_2 and x_3 are missing, then the reduced set of edits is given by

$$10 + x_{i2} = x_{i3},$$

$$10 \geq x_{i2},$$

$$x_{i3} \geq 3x_{i2},$$

$$x_{ij} \geq 0 \ (j=2,3).$$

Once the reduced set of edits has been determined for a record i, we eliminate all equations from this reduced set of edits. That is, we sequentially select an equation and one of the variables x involved in this equation. We then express x in terms of the other variables in the selected equation, and substitute this expression for x into the other edits in which x is involved. For instance, assuming that all values are missing for a certain record i in the above example (4) to (7), we can eliminate x_{i3} by substituting the expression $x_{i3} = x_{i1} + x_{i2}$ into the other edits (5) to (7). In this way we obtain a set of edits involving only inequalities restrictions for the remaining variables. Later, once we have obtained imputation values for the variables involved in the set of inequalities, we can find values for the variables we have eliminated by back-substitution. For instance, in our example where we have eliminated x_{i3} from the edits (4) to (7), once we have obtained imputation values for x_{i1} and x_{i2} we can obtain a consistent value for x_{i3}, i.e. a value satisfying all edits, by filling in the values for x_{i1} and x_{i2} in (4).

Next, we eliminate any remaining variables except x_{ij} itself from the set of edits by means of Fourier-Motzkin elimination (see, e.g., De Waal and Coutinho, 2005). The edits for x_{ij} can then be expressed as interval constraints:

$$l_{ij} \leq x_{ij} \leq u_{ij} \,. \tag{9}$$

The problem for variable j now is to fill in the missing values with imputations, such that the sum constraint (8) and the interval constraints (9) are satisfied.

Below we present three different approaches to solving this problem. The first two approaches are based on standard regression imputation techniques, but with (slight) adjustments to the imputed values such

that they satisfy the constraints (8) and (9). The third approach is an extension of MCMC algorithms described in the literature, which generates imputations that directly satisfy the constraints (8) and (9).

3.1 Adjusted Predicted Mean Imputation

The idea of this algorithm is to obtain predicted mean imputations that satisfy the sum constraint and then adjust these imputations such that they also satisfy the interval constraints. To illustrate this idea we use a simple regression model with one predictor but generalisation to multiple regression models is straightforward.

3.1.1 Introducing some notation by the example of standard regression imputation

Suppose that we want to impute a target column $\mathbf{x}_t$ using as a predictor a column $\mathbf{x}_p$. The standard regression imputation approach is based on the model:

$$\mathbf{x}_t = \beta_0 \mathbf{1} + \beta \mathbf{x}_p + \boldsymbol{\varepsilon},$$

where $\mathbf{1}$ is the vector with ones in every entry, i.e. (1,1,…,1) and $\boldsymbol{\varepsilon}$ is a vector with random residuals.

We assume that the predictor is either completely observed or already imputed, so there are no missing values in the predictor anymore. There are of course missing values in $\mathbf{x}_t$ and to estimate the model we can only use the records for which both $\mathbf{x}_t$ and $\mathbf{x}_p$ are observed. The data matrix for estimation consists of the columns $\mathbf{x}_{t.obs}, \mathbf{x}_{p.obs}$, where *obs* denote the records with $\mathbf{x}_t$ observed (and *mis* will denote the opposite). With the OLS estimators of the parameters, $\hat{\beta}_0$ and $\hat{\beta}$ we obtain predictions for the missing values in $\mathbf{x}_t$ using

$$\hat{\mathbf{x}}_{t.mis} = \hat{\beta}_0 \mathbf{1} + \hat{\beta} \mathbf{x}_{p.mis},$$

where $\mathbf{x}_{p.mis}$ contains the $\mathbf{x}_p$-values for the records with $\mathbf{x}_t$ missing and $\hat{\mathbf{x}}_{t.mis}$ are the predictions for the missing $\mathbf{x}_t$-values in those records. The imputed column $\widetilde{\mathbf{x}}_t$ consists of the observed values and the predicted values filled in for the missing values $\widetilde{\mathbf{x}}_t = (\mathbf{x}_{t.obs}^T, \hat{\mathbf{x}}_{t.mis}^T)^T$, where T denotes the transpose.

These imputed values will not satisfy the sum constraint but a slightly modified regression approach can ensure that they do and will be described next.

3.1.2 Extending the standard regression imputation to satisfy the sum-constraint

This approach adds to the observed data the known totals of the missing data for the target variable as well as the predictor. These totals are $X_{p.mis} = X_p - \sum_i x_{p.obs,i}$ and $X_{t.mis} = X_t - \sum_i x_{t.obs,i}$, respectively, where the summation is over the records with observed values for the target variable. The total $X_{t.mis}$ is added to the column $\mathbf{x}_{t.obs}$ and the total $X_{p.mis}$ is added to the column $\mathbf{x}_{p.obs}$. Furthermore, the regression model is extended with a separate constant term for the record with the totals of the missing data. The model for these observed data can then be written as

$$\begin{aligned} \mathbf{x}_{t.obs} &= \beta_0 \mathbf{1} + \beta \mathbf{x}_{p.obs} + \boldsymbol{\varepsilon} \\ X_{t.mis} &= \beta_1 m + \beta X_{p.mis} \end{aligned} \tag{10}$$

with m the number of records with missing values for target variable $\mathbf{x}_t$. We apply OLS to estimate the model parameters which will be used to predict and impute the missing values in $\mathbf{x}_t$,. In particular, we impute missing values in $\mathbf{x}_t$ by

$$\hat{\mathbf{x}}_{t.mis} = \hat{\beta}_1 \mathbf{1} + \hat{\beta} \mathbf{x}_{p.mis}, \tag{11}$$

and so the sum of the predicted values over the records with missing values for the target variable will equal

$$\hat{X}_{t.mis} = \sum_i \hat{x}_{t.mis,i} = m\hat{\beta}_1 + \hat{\beta} X_{p.mis}$$

In order to demonstrate the property of this model that the imputed values will sum up to the known total, we re-express the model for the observed data with the known totals added as

$$\begin{bmatrix} \mathbf{x}_{t.obs} \\ X_{t.mis} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{x}_{p.obs} \\ 0 & m & X_{p.mis} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon} \\ 0 \end{bmatrix}$$

or

$$\begin{bmatrix} \mathbf{x}_{t.obs} \\ X_{t.mis} \end{bmatrix} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

If this model is estimated by ordinary least squares (OLS) estimation, the residuals are orthogonal to each of the columns of the model matrix **Z**. Thus, for the second column we obtain $m(X_{t.mis} - \hat{X}_{t.mis}) = 0$

and hence $\hat{X}_{t.mis} = \sum_i \hat{x}_{t.mis,i} = X_{t.mis}$ which implies that the sum of the imputed values equals the known value of this total.

3.1.3 Adjusting the regression imputations to satisfy the sum constraint and the interval constraints

Since the interval constraints have not been considered in obtaining the predicted values, it can be expected that a number of these predictions are not within their admissible intervals. One way to remedy this situation is to calculate adjusted predicted values defined by

$$\hat{\mathbf{x}}_{t.mis}^{adj} = \hat{\mathbf{x}}_{t.mis} + \mathbf{a}_t \,, \tag{12}$$

such that the adjusted predictions satisfy both the sum constraint (which is equivalent to $\sum_i a_{t,i} = 0$) and the interval constraints and the adjustments are as small as possible. One way to find such a value for $\mathbf{a}_t$ is to solve the quadratic programming problem

minimise $\mathbf{a}_t^T \mathbf{a}_t$, subject to $\mathbf{1}^T \mathbf{a}_t = 0$ and $\mathbf{l}_t \le \hat{\mathbf{x}}_{t.mis} + \mathbf{a}_t \le \mathbf{u}_t$,

or, we can minimize the sum of the absolute values of the $a_{t,i}$ instead and solve the resulting linear programming problem.

As a simple alternative we may consider the following algorithm which alternates between adjusting to satisfy the interval constraints and adjusting to satisfy the sum constraint.

This algorithm starts with $\mathbf{a}_t^{(0)} = \mathbf{0}$ and the predictions (12) satisfy the sum constraint but not necessarily the interval constraints. Each prediction outside its admissible interval will then be moved to the closest boundary value by an appropriate adjustment, which is the smallest possible adjustment to satisfy the interval constraints, i.e.

$$a_{t,i}^{(1)} = l_{t,i} - \hat{x}_{t.mis,i} \qquad \text{if} \qquad \hat{x}_{t.mis,i} < l_{t,i} \tag{13a}$$

$$a_{t,i}^{(1)} = u_{t,i} - \hat{x}_{t.mis,i} \qquad \text{if} \qquad \hat{x}_{t.mis,i} > u_{t,i} \tag{13b}$$

$$a_{t,i}^{(1)} = 0 \qquad \text{if} \qquad l_{t,i} \le \hat{x}_{t.mis,i} \le u_{t,i} \tag{13c}$$

The adjusted values $\hat{\mathbf{x}}_{t.mis}^{adj}$ will now satisfy the interval constraints but almost surely not the sum constraint, which is equivalent to saying that the $a_{t,i}^{(1)}$ do not sum to zero. To obtain adjustments that also preserve the sum constraint, we divide the *m* units in three set; L_t, U_t, O_t, with numbers of elements m_L, m_U, m_O, according to whether the current adjusted value $\hat{\mathbf{x}}_{t.mis}^{adj}$ is on the lower boundary, upper boundary or neither boundary. Let the current sum of the $a_{t,i}^{(1)}$ be $S_t^{(1)}$, then sum-to-zero adjustments can be obtained as

$$a_{t,i}^{(2)} = a_{t,i}^{(1)} - S_t^{(1)}/(m_U + m_O) \text{ for all } i \in U_t \cup O_t \text{ if } S_t^{(1)} > 0 \tag{14a}$$

or

$$a_{t,i}^{(2)} = a_{t,i}^{(1)} + S_t^{(1)}/(m_L + m_O) \text{ for all } i \in L_t \cup O_t \text{ if } S_t^{(1)} < 0 \tag{14b}$$

Thus, we add or subtract a constant to the $a_{t,i}^{(1)}$ to make them sum to zero, thereby taking care not to subtract anything from $a_{t,i}^{(1)}$'s that already set the $\hat{\mathbf{x}}_{t.mis}^{adj}$ on their lower boundary and not to add anything to $a_{t,i}^{(1)}$'s that already set the $\hat{\mathbf{x}}_{t.mis}^{adj}$ on their upper boundary. After this step it may be that some of the $a_{t,i}^{(2)}$ cause their corresponding $\hat{\mathbf{x}}_{t.mis}^{adj}$ to cross their interval boundaries. In that case both steps (13) and (14) must be repeated.

3.2 Regression Imputation with Random Residuals

It is well known that in general predictive mean imputations show less variability than the true values that they are replacing. In order to better preserve the variance of the true data, random residuals can be added to the predicted means. The adjusted predictive mean imputations considered in the previous section will also be hampered by this drawback because these adjustments are intended to be as close as possible to the predicted means and not to reflect the variance of the original data.

In order to better preserve the variance of the true data we start with the predicted values $\hat{\mathbf{x}}_{t.mis}$ obtained from (11) that already satisfy the sum constraint, and our purpose is to add random residuals to these predicted means such that the distribution of the data is better preserved and in addition both the interval and sum constraints are satisfied. These residuals serve the same purpose (satisfying the constraints) as the adjustments $a_{t,i}$ but in contrast to the $a_{t,i}$, they are not as close as possible to the predicted means; they are intended to also reflect the true variability around these predicted means

A simple way to obtain residuals is to draw each of the *m* residuals by Acceptance/Rejection (AR) sampling (see, e.g., Robert and Casella, 1999, for more on AR sampling) from a normal distribution with mean zero and variance equal to the residual variance of the regression model. This means, by repeatedly drawing from this normal distribution until a residual is drawn that satisfies the interval constraint.

The residuals obtained by this AR-sampling may not sum to zero so that the imputed values do not satisfy the sum constraint. We may then adjust these residuals to sum to zero by the "shift" operation according to (14) after which it may be necessary to again adjust some of the residuals to also satisfy the interval constraint by means of (13).

Instead of this somewhat ad hoc approach we next consider a more sophisticated alternative to the adjusted predictive mean imputation.

3.3 MCMC Approach

The third imputation algorithm we describe is based on a Monte Carlo Markov Chain (MCMC; see, e.g., Robert and Casella, 1999; Liu, 2001, for more on MCMC in general) approach. This MCMC approach is an extended version of similar approaches by Raghunatan et al. (2001), Rubin (2003) and Tempelman (2007; Chapter 6). Raghunatan et al. (2001) and Rubin (2003) do not take edits or totals into account in their MCMC approaches. The MCMC approach of Tempelman (2007) does take edits into account, but not totals. The approach starts with a fully imputed, consistent dataset, for instance obtained by means of the methods of Sections 3.2 and 3.3. Subsequently, we try to improve the imputed values so they preserve

the statistical distribution of the data better. Our algorithm, which is similar to data swapping for categorical data (see Dalenius and Reis, 1982), is sketched below.

0. Start with a pre-imputed, consistent dataset, i.e. a dataset that satisfies both edits and totals.

1. Randomly select two records.

2. Select a variable. First note that we know the sum of the two values of this variable for the two records (namely, the total minus the sum of the imputed values for the other records). We then apply the following two steps.

   a. Determine the allowed intervals for the two values. To determine these intervals, we start for each value with the interval that can be derived from the edits by filling in the (observed or imputed) values of the other variables in the corresponding record. We impute the value, say x, in one of the two records and later derive the value of the other variable y in the other record by subtracting the value of x from the known sum. As the two values have to sum up to a known total, the lower bound on y may influence the upper bound on x, and vice versa, the upper bound on y may influence the lower bound on x. This leads to an adjusted interval for x (y), which may be narrower than the interval for x (y) we started with. For example, if x and y have to sum to 100, the original lower bound on x is 50 and the original upper bound on y is 40, the adjusted lower bound on x is 60.

   b. Draw a value for x from a posterior predictive distribution implied by a linear regression model under an uninformative prior, conditional on the fact that this value has to lie inside the interval for x. As already mentioned, the value for y then immediately follows by subtraction of x from the known sum. Note that the variances of the two values (which is the variance of the posterior predictive distribution) are equal. This is a fortunate "coincidence", because for two variables summing up to a total, their variances, conditional on the total, are equal.

   Now, repeat Steps 1 and 2 until "convergence". Note that "convergence" is a difficult concept as we are referring to the convergence of the distribution. We refer to Robert and Casella (1999) and Liu (2001) for more on convergence of MCMC processes.

An important reason why we use a posterior predictive distribution implied by a linear regression model under an uninformative prior is that this, in principle, allows us to extend our approach to multiple imputation. The extension to multiple imputation is not studied in the present paper, however.

This MCMC approach clarifies why we eliminate all equations from the set of edits before we apply our imputation algorithms (see the beginning of Section 3). If any equations from the set of edits had been left, our MCMC approach would be 'stuck' after the pre-imputation step as we would get the same values over and over again.

4. Evaluation Measures

To measure the performance of the imputation methods we use several methods as described below:

- d_{L1} measure as proposed by Chambers (2003) and 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

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

  where $\hat{x}_i$ is the imputed value in record i and x_i^* is the original value of the variable under consideration, M denotes the set of m records with imputed values for variable x and w_i is the raising weight for record i. The smallest measure indicates better imputation performance.
- 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 of the original inequality edits is satisfied with equality. We denote this number by N_b. Records on the boundary of the feasible region defined by the edits are outlying in some sense. The number of outlying records in this sense (and in any other sense) should be close to the true number of outlying records. The number of imputed records on the boundary should be close to the actual number of records on the boundary for the complete version of the file.

In general, the true number of records on the boundary of the feasible region defined by the edits will be close to zero.

- *K-S* Kolmogorov-Smirnov non-parametric test statistic 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: $F_{x^*}(t) = \sum_{i \in M} I(x_i^* \le t)/m$ and similarly $F_{\hat{x}}(t)$ where *I* is the indicator function. The *K-S* is defined as:

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

where the $\{t_j\}$ values are the 2*m* jointly ordered original and imputed values of *x*.

- a sign test using paired data 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

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

where n^+ is the number of values greater than 0 and n^- the number of values less than 0. Under the null hypothesis, the sign test calculates the *p*-value for *S* using a binomial distribution. A small *p*-value means that we reject the null hypothesis of equal medians. In addition we can calculate a Wilcoxon signed rank test statistic based on ranks.

- the percent difference between the standard deviation (STD) of the mean of the imputations to the standard deviation of the mean of the observations:

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

- a Kappa statistic for a 2 dimensional contingency table containing counts of the records spanned by ordered bands of the original values and ordered bands of the imputed values. We used 6 ordered bands. The Kappa statistic compares the agreement against that which might be expected by chance and is defined as:

$$\kappa = (P_D - P_E)/(1 - P_E)$$

where P_D is the sum of the diagonal probabilities defined by $p_{ii} = n_{ii}/m$ and n_{ii} is the number of records in the diagonal (i,i), i.e. the records for which the original values and the imputed values are in the same ordered band, and P_E is the sum of the multiplied marginal probabilities $p_{i.} = n_{i.}/m$ and $p_{.i} = n_{.i}/m$, where $n_{i.}$ is the row total of i and $n_{.i}$ is the column total of i.

We use the measures in a relative way, namely to compare the different methods. The measures are neither necessarily appropriate nor sufficient to measure the impact of imputation on the quality of survey estimates in general. Furthermore, to assess the importance of bias caused by imputation it should be related to other quality aspects, such as sampling variance.

5. Simulation Study

A simulation study was carried out based on variables X_1, X_2 and a predictor P that were generated from a normal distribution using linear transformations to ensure a reasonably realistic degree of correlation between them. The simulated dataset included 5,000 records. Table 2 presents the Pearson correlations between the simulated variables.

[PLACE TABLE 2 HERE]

Edit constraints (4) to (8) are all preserved on the simulated dataset.

Out of the 5,000 records, 1,000 records were randomly selected and their X_1 variable blanked out. Out of those records, 500 were randomly selected and their X_2 variable was also blanked out. An additional 500 records were randomly selected from the remaining 4,000 records and their X_2 variable was blanked out.

We study and compare the following procedures:

- *UPMA* - unbenchmarked simple predictive mean imputation (Section 3.1.1) with adjustments to imputations so they satisfy interval constraints (Section 3.1.3)
- *BPMA* – benchmarked predictive mean imputation (Section 3.1.2) with adjustments to imputations so they satisfy interval constraints (Section 3.1.3)
- *BPMR*- benchmarked predictive mean imputation (Section 3.1.2) with random residuals (Section 3.2)
- *MCMC* – the approach described in Section 3.3. The dataset with *BPMA* was used as the pre-imputed dataset for the *MCMC* approach.

For all methods, the variable X_1 was regressed on the predictor *P*, and X_2 was regressed on the predictor *P* and X_1. After all variables have been imputed once, the next rounds of the sequential procedure uses, for each variable to be re-imputed, all other variables as regressors. Thus after the first round X_1 is regressed on *P* and X_2, and X_2 is regressed on *P* and X_1. Table 3 contains the results of the evaluation measures as described in Section 4 for the four methods averaged across 10 simulations. Note that UPMA and BPMA are deterministic imputations and BPMR and MCMC are stochastic imputations.

[PLACE TABLE 3 HERE]

The results in Table 3 show the similarities between the methods *UPMA* and *BPMA*. These imputations are deterministic and the difference between them is that *BPMA* benchmarks the totals. The simulations show that with benchmarking *BPMA* has less imputed values on the boundaries of the edits as seen for variable X_1 compared to UPMA (the true value for N_b is zero). All other evaluation measures are similar between the two methods. The results in Table 3 also show similarities between *BPMR* based on random residuals and the *MCMC* algorithm. The *K-S* statistic and the relative difference to the standard deviation of the mean is slightly higher for both variables for the *MCMC* approach. The Kappa statistic κ for the *MCMC* is slightly higher for the X_1 but slightly lower for the X_2 variable. Using stochastic imputation methods ensures that we have fewer values on the boundaries of the edits, less impact on the standard

deviation of the mean and smaller *K-S* statistics. Indeed, for the X_1 variable we obtain larger standard deviations of the mean of imputed values than the standard deviations of the observed values. The X_1 variable has a smaller Kappa statistic κ for the stochastic methods compared to the deterministic methods, but a higher Kappa statistic κ for the X_2 variable. The sign test shows that all methods have significantly different medians of the observed values compared to the imputed values under both variables.

Our general conclusion from the simulation study is that, based on the preservation of totals (and edit constraints), preservation of standard deviations and preservation of other distributional properties, we consider *BPMR* and *MCMC* the most promising methods. However, when interest is restricted to preserving the individual values as well as possible, the deterministic methods *UPMA* and *BPMA* perform better.

6. Evaluation Study

6.1. Evaluation Dataset

We use a real dataset from the 2005 Israel Income Survey. The file for the evaluation study contains 11,907 individuals aged 15 and over that responded to all the questions in the questionnaire of the 2005 Israel Income Survey and in addition, earned more than 1,000 Israel Shekels (IS) for their monthly gross income. We focus on three variables from the Income Survey: the gross income from earnings, the net income from earnings and the difference between them (tax). As above, we consider the following edits for each record i:

$$net_i + tax_i = gross_i \tag{15a}$$

$$net_i \geq tax_i \tag{15b}$$

$$gross_i \geq 3 \times tax_i \tag{15c}$$

$$gross_i \geq 0,\ net_i \geq 0,\ tax_i \geq 0 \tag{15d}$$

Item non-response was introduced randomly to the income variables in order to simulate a typical dataset: 20% of the records (2,382 records) were selected randomly and their net income variable blanked out.

Out of those selected records, 50% (1,191 records) also had their tax variable blanked out. An additional 10% (1,191 records) were selected randomly from the dataset and their tax variable deleted. We assume that the totals of each of the income variables are known.

The variables that were chosen for the predictive mean imputation based on regression modelling were the following: 14 categories of economic branch, 10 categories of occupation, 10 categories of age group, and sex. For each category a dummy variable was created.

In order to ensure the normality of the income variables, a log transformation was carried out. This meant we had to change the algorithm described in Section 3.1.2 slightly since the sum of the log transformed variables which will equal the known log totals will not necessarily mean that the sum of the original variables will equal the known original totals. We used a correction factor to replace the constant term of the regression to constrain the sum of the untransformed, original variables to the original totals. We denote $\mathbf{z} = \log \mathbf{x}$, where the logarithm is taken component-wise, i.e. $\mathbf{z} = (\log(x_1),\ldots,\log(x_r))$, where r is the number of records. From (11), $\hat{\mathbf{z}}_{t.mis} = \hat{\beta}_1 \mathbf{1} + \hat{\beta}\ \mathbf{z}_{p.mis}$ and therefore $\hat{\mathbf{x}}_{t.mis} = \exp(\hat{\beta}_1) \times \exp(\hat{\beta}\ \mathbf{z}_{p.mis})$, where $\exp(\hat{\beta}\ \mathbf{z}_{p.mis})$ is taken component-wise. Summing across the missing values gives: $\hat{X}_{t.mis} = \sum_i \hat{x}_{t.mis,i} = \exp(\hat{\beta}_1) \sum_i \exp(\hat{\beta}\ z_{p.mis,i})$. The correction replaces the constant factor $\exp(\hat{\beta}_1)$ with $\dfrac{\hat{X}_{t,mis}}{\sum_i \exp(\beta\ z_{p,mis,i})}$.

Table 4 contains the results of the evaluation measures as described in Section 4 for the methods: (1) unbenchmarked simple predictive mean imputation with adjustments to the imputations that satisfy interval constraints (*UPMA*), (2) the benchmarked predictive mean imputation with adjustments to the imputations that satisfy interval constraints (*BPMA*), (3) the benchmarked predictive mean imputation with random residuals (*BPMR*) , (4) the MCMC approach (*MCMC*).

[PLACE TABLE 4 HERE]

From the results of Table 4, the *BPMA* approach and the stochastic approaches *BPMR* and the *MCMC* all preserve the totals and edit constraints in the data. The results on the d_{L1} are mixed with the net income variable doing slightly worse for both stochastic approaches but the tax variable showing improvement with the *BPMR* approach. The κ statistics are only slightly lower for both stochastic methods compared to the *BPMA*. As expected, the number of values on the boundary N_b is less for the stochastic approaches and the distribution is preserved better as reflected in the *K-S* statistic, the percent difference in the standard deviation of the mean, and the *p*-value of the sign test. The measures when benchmarking the totals (*BPMA*) appear to be mixed compared to not benchmarking (*UPMA*) depending on the variable. The number of records that lie on the boundary N_b for the unbenchmarked method *UPMA* is a cause for concern.

It is more difficult to draw general conclusions for the real dataset than it was for the simulated dataset, since the results for the real dataset are not univocal across variables. However, based on the fact that the stochastic methods preserve totals (and edit constraints) and preserve standard deviations and other distributional properties better than *UPMA* and *BPMA*, we consider *BPMR* and *MCMC* the most promising methods. The *MCMC* approach would allow for multiply imputing the dataset. In that way we would be able to take the uncertainty in the imputation into account when making inferences. For the other methods, replication approaches (bootstrap or Jacknife) could be employed for this purpose.

6. Discussion

In this paper we have proposed three imputation methods for numerical data that satisfy edit restrictions and preserve totals. As far as we know, such methods have not been developed before. Two of the developed methods are stochastic, aiming to better preserve the variation in the imputed data. One of the developed imputation methods can be easily extended to a multiple imputation approach.

The problem that we have examined in this paper forms part of a more general problem. In this more general problem a non-integral survey amongst the population is held, i.e. only part of the population is

observed. The standard way to use such a sample in order to obtain estimates for population totals is by means of raising weights, which are multiplied with the observed values. Next, these weighted observed values are summed to obtain the desired population estimates. If data are missing and some of the population totals are known, one then has two options: either one first imputes the missing data and then determines raising weights in such a way that the weighted sums equal the known population totals, or one first determines raising weights and then imputes the missing values in such a way the weighted sums equal the known population totals. The former approach is the standard approach. The methods examined in this paper form a first step towards the latter approach. In the present paper all raising weights equal one. In a future paper we plan to extend this to the more general case where the raising weights are not all equal to one.

The methods introduced in this paper can also be used for mass imputation of numerical data. In Houbiers (2004) a statistical database for social data was constructed using so-called repeated weighting based on regression estimators. Whilst benchmarking totals (either based on registers or weighted survey estimates), the method does not preserve edit constraints. The methods in this paper provide an alternative to repeated weighting which can benchmark totals, preserve edit constraints and preserve correlation structures in the data. Initial work in the area of mass imputation for a numerical dataset having the above properties using the methods proposed in the present paper is described in Shlomo, De Waal, and Pannekoek 2009.

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) (available on http://www.cs.york.uk/euredit/).

Dalenius, T. and Reiss, S.P. (1982), Data Swapping: A Technique for Disclosure Control. *Journal of Statistical Planning and Inference 7*. pp. 73-85.

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

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.

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

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.

Liu, J.S. (2001), *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, 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.

Raghunathan, T.E., J.M. Lepkowski, J. Van Hoewyk and P. Solenberger (2001), A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology 27*, pp.85-95.

Robert, C.P. and G. Casella (1999), *Monte Carlo Statistical Methods*. Springer-Verlag, New York.

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*, pp. 3-18.

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

Shlomo, N., De Waal, T. and Pannekoek, J. (2009), Mass Imputation for Building a Numerical Statistical Database. *Presented at the UNECE Statistical Data Editing Workshop, Neuchatel, October, 2009* . http://www.unece.org/stats/documents/ece/ces/ge.44/2009/wp.31.e.pdf

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

Table 1: Illustration of a Dataset

x_{11}	x_{12}	x_{13}
x_{21}	x_{22}	x_{23}
$\vdots$	$\vdots$	$\vdots$
x_{r1}	x_{r2}	x_{r3}
X_1	X_2	X_3

Table 2: Pearson correlation for variables in simulation study

	X_1	X_2	P
X_1	1.000	0.688	0.584
X_2	0.688	1.000	0.396
P	0.584	0.396	1.000

Table 3: Results of Evaluation Measures for the Imputation Methods in the Simulation Study

	X_1				X_2			
	UPMA	*BPMA*	*BPMR*	*MCMC*	*UPMA*	*BPMA*	*BPMR*	*MCMC*
Distance d_{L1}	11.38	11.40	17.35	17.91	12.87	12.82	14.67	14.54
Number on boundary N_b (true value is zero)	153	25	2	2	0	0	0	0
Kolmogorov-Smirnov *K-S*	0.127	0.124	0.048	0.050	0.196	0.190	0.120	0.132
Sign Test (*p*-value)	<.001	<.001	<.001	0.010	<.001	<.001	0.013	<.001
% difference of STD	-32.6%	-32.8%	13.6%	20.4%	-45.7%	-45.7%	-31.1%	-33.8%
Kappa Statistic κ	0.283	0.280	0.148	0.172	0.055	0.053	0.121	0.109

Table 4: Results of Evaluation Measures for the Imputation Methods in the Evaluation Study

Evaluation Measures	Net Income Variable				Tax Variable			
	UPMA	*BPMA*	*BPMR*	*MCMC*	*UPMA*	*BPMA*	*BPMR*	*MCMC*
Distance d_{L1}	2040.4	2132.6	2695.9	2664.2	980.6	821.7	818.6	1154.4
Number on boundary N_b (true value is zero)	163	112	33	15	115	73	39	17
Kolmogorov-Smirnov *K-S*	0.098	0.149	0.049	0.086	0.433	0.323	0.184	0.155
Sign Test (*p*-value)	<.001	<.001	0.035	0.499	<.001	<.001	<.001	0.389
% difference to STD	-41.1%	-37.6%	-11.9%	-19.4%	-3.2%	-4.7%	-3.2%	3.5%
Kappa Statistic κ	0.227	0.215	0.191	0.178	0.228	0.406	0.395	0.240