Calibrated imputation of numerical data under linear edit restrictions

Jeroen Pannekoek, Natalie Shlomo and Ton de Waal

Discussion paper (09016)

Statistics Netherlands

The Hague/Heerlen, 2009

Explanation of symbols

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

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

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Facility Services

Cover
TelDesign, Rotterdam

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

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

Internet
www.cbs.nl

ISSN: 1572-0314

6008309016 X-10

Calibrated imputation of numerical data under linear edit restrictions

Jeroen Pannekoek, Natalie Shlomo[1] and Ton de Waal

Summary: A common problem faced by statistical offices is that data may be missing from collected data sets. The typical way to overcome this problem is to impute the missing data. The problem of imputing missing data is complicated by the fact that statistical data often have to satisfy certain edit rules 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 present 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. To assess the performance of the imputation methods an evaluation study is carried out.

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 some of the items may be missing. Persons may, for instance, refuse to provide information on their income or on their sexual habits, while at the same time giving answers to other, less sensitive questions on the questionnaire. Enterprises may not provide answers to certain questions, because they may consider it too complicated or too time-consuming to answer these specific questions. Missing items of otherwise responding units is called item non-response. Whenever we refer to missing data in this paper we will mean item non-response.

Missing data is a well-known problem that has to be faced by basically all institutes that collect data on persons or enterprises. In the statistical literature ample attention is hence paid to missing data. The most common solution to handle missing data in data sets is imputation, where missing values are estimated and filled in. An important problem of imputation is to preserve the statistical distribution of the data set. This is a complicated problem, especially for high-dimensional data. For more

[1] University of Southampton

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. 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.

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. 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 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. An evaluation study is described in Section 4. Finally, Section 5 ends the paper 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. An inequality edit such as (3), expressing that the value of a variable should be non-negative, is also referred to as a non-negativity edit.

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. The data for this case are as shown in table 1.

Table 1. Three variables with known totals.

Variable	x_1	x_2	x_3
record $i=1$	x_{11}	x_{12}	x_{13}
$i=2$	x_{21}	x_{22}	x_{23}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i=r$	x_{r1}	x_{r2}	x_{r3}
Total	X_1	X_2	X_3

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. $X_1 + X_2 = X_3$.

Using (4) we can eliminate x_{i3}, and substitute $x_{i3} = x_{i1} + x_{i2}$ into the other edits (5) to (7). In this way we have only inequalities restrictions.

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. 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:

$$l_{ij} \leq x_{ij} \leq u_{ij} . \qquad (9)$$

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

Te 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:

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

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 ordinary least squares (OLS) estimators of the parameters, $\hat{\beta}_0$ and $\hat{\beta}$ say, we obtain predictions for the missing values in $\mathbf{x}_t$ using

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

where $\mathbf{x}_{p.mis}$ contains the 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 $\tilde{\mathbf{x}}_t$ consists of the observed values and the predicted values filled in for the missing values $\tilde{\mathbf{x}}_t = (\mathbf{x}_{t.obs}^T, \hat{\mathbf{x}}_{t.mis}^T)^T$, 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 totals of the missing data for the target variable as well as the predictor. These totals can be calculated as $X_{p.mis} = X_p - \sum_i x_{p.obs,i}$ and $X_{t.mis} = X_t - \sum_i x_{t.obs,i}$, respectively. 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

$$\begin{aligned} \mathbf{x}_{t.obs} &= b_0 + b\mathbf{x}_{p.obs} + e \\ X_{t.mis} &= b_1 m + bX_{p.mis} \end{aligned} \qquad (10)$$

with *m* the number of records with missing values. We apply OLS to estimate the parameters of the model (10) which will then be used to predict and impute the missing values in $\mathbf{x}_t$, i.e.

$$\hat{\mathbf{x}}_{t.mis} = \hat{b}_1 + \hat{b}\mathbf{x}_{p.mis}, \qquad (11)$$

and so the sum of the predicted values will equal

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

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

$$\begin{bmatrix} \mathbf{x}_{t.obs} \\ X_{t.mis} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{x}_{p.obs} \\ 0 & m & X_{p.mis} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon} \\ 0 \end{bmatrix}$$

or

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

If this model is estimated by OLS, 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 the predictions $\hat{x}_{t.mis,i}$ are not within their admissible intervals. In this subsection we will consider additive adjustments to the predicted values such that the adjusted predictions satisfy both the sum constraint and the interval constraints and the adjustments are as small as possible. The adjusted predictions can be written as:

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

with $\mathbf{a}_t$ the vector with adjustments to the predictions for the missing values in the target variable $\hat{\mathbf{x}}_{t,mis}$. Since the predictions $\hat{\mathbf{x}}_{t,mis}$ are determined such that the sum constraint is satisfied, it follows that adjustments are constraint to sum to zero. Adjustments which are minimal, in a least squares sense, can therefore be obtained as the solution to the quadratic programming problem:

$$\text{minimize } \mathbf{a}_t^T \mathbf{a}_t$$

$$\text{subject to } \begin{matrix} \mathbf{l}_t \le \hat{\mathbf{x}}_{t,mis} + \mathbf{a}_t \le \mathbf{u}_i \\ \mathbf{1}^T \mathbf{a}_t = \mathbf{0} \end{matrix},$$

with $\mathbf{l}_t$ the vector with lower boundaries and $\mathbf{u}_t$ the vector with upper boundaries for the corresponding elements of $\mathbf{a}_t$. This problem can of course be solved by an algorithm for solving general quadratic programming problems. As a simple alternative we may consider the following heuristic algorithm which alternates between adjusting to satisfy the interval constraints and adjusting to satisfy the sum constraint.

This algorithm starts with $\mathbf{a}_t^{(0)} = 0$ and the predictions (11) that 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

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

or

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

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 added

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 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.

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 an MCMC 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. Our algorithm is sketched below.

An MCMC approach, similar to data swapping for categorical data (see Dalenius and Reis, 1982), is the following.

0. Start with a pre-imputed, consistent dataset, i.e. a dataset that satisfies both edits and totals.
1. Randomly select two records.
2. We select a variable. First we note that we know the sum of these 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. We determine the intervals for the two values. For the intervals, we start for each value with the interval that can be derived from the corresponding record. As the two values have to sum up to a known total, the lower bound of an interval may influence the upper bound of the other, and vice versa the upper bound of an interval may influence the lower bound of the other. This leads to two adjusted intervals, which may be narrower than the intervals we started with.
    b. We then draw from a posterior predictive distribution implied by a linear regression model under an uninformative prior one value conditional on the fact that it has to lie inside the corresponding interval. The other value then immediately follows. 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 imputation two values summing up to a total their variances should be 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.

This MCMC approach clarifies why we eliminate all equations from the set of edits before we apply imputation algorithms. 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 study

An evaluation study was carried out on a simulated dataset where variables X_1, X_2 and a predictor *P* were generated using a normal distribution, non-response was added in and the methods applied. In addition, we carried out the algorithms on a real dataset from the 2005 Israel Income Survey and obtained similar results. In this paper we present the results from the real dataset.

We study an unbenchmarked simple predictive mean imputation described in Section 3.1.1 with the adjustments to the imputations that satisfy interval constraints as described in Section 3.1.3; the benchmarked predictive mean imputation as described in Section 3.1.2 also with the adjustments to the imputations that satisfy interval constraints (Section 3.1.3), and the MCMC approach described in Section 3.3. The benchmarked predictive mean imputation with adjustments was used as the pre-imputed dataset for the MCMC approach.

4.1 Evaluation dataset

31. 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 1000 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*:

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

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

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

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

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 deleted, 20% (2,381 records) were selected randomly and their tax variable deleted while 10% of those records (1,191 records) were in common with the missing net income variable, i.e. both their net income and tax variables were 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 groups, 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. Therefore the algorithm as described in Section III.A.2 had to be adjusted 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 the constant term of the regression to constrain the original totals.

4.2 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

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

  where $\hat{y}_i$ is the imputed value in record *i* and y_i^* is the original value of the variable under consideration, *M* denotes the set of *m* records with imputed values for variable *y* and w_i is the raising weight for record *i*. For the purpose of this study $w_i = 1$. 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 inequality edit is satisfied with equality. We denote this number by Z. 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.

- *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: $F_{y^*}(t) = \sum_{i \in M} I(y_i^* \le t)/m$ and similarly $F_{\hat{y}}(t)$ where *I* is the indicator function. The *K-S* is defined as:

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

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

- 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. In addition we can calculate a Wilcoxon signed rank test statistic based on ranks.

- 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 7 ordered bands for the net variable and 5 ordered bands for the tax variable. 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), and P_E is the sum of the multiplied marginal probabilities $p_{i.} = n_{i.}/m$ and $p_{.i} = n_{.i}/m$ and $n_{i.}$ 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.

4.3 Evaluation results

Table 2 contains the results of the evaluation measures as described in Section 4.2 for the three 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) and the MCMC approach (*MCMC*).

Table 2: Results of Evaluation Measures for the three imputation methods

Evaluation Measures	Net Variable			Tax Variable		
	UPMA	*BPMA*	*MCMC*	*UPMA*	*BPMA*	*MCMC*
d_{L1}	2266.1	2132.6	4304.8	786.8	821.7	1393.7
Z (the true value is 0)	204	11	1	123	12	0
K-S	3.535	5.129	9.100	3.521	9.129	11.158
Sign Test	0.0147	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
κ	0.161	0.178	0.117	0.418	0.421	0.226

From the results it is clear that the MCMC approach is doing worse than the other methods on all criteria, except *Z*. The d_{L1} statistic is consistently higher and the other

measures suggest that the distribution of the imputed values is more distorted. However, drawing from a posterior predictive distribution as opposed to the predictive mean imputation will undoubtedly produce better variance estimates and the dataset should be multiply imputed to take into account the uncertainty in the imputation. This will be carried out in future research. The measures when benchmarking the totals (*BPMA*) appear to be slightly better compared to not benchmarking (*UPMA*) except for the *K-S* statistics. The number of records that lie on the boundary *Z* for the unbenchmarked method *UPMA* is a cause for concern. The *MCMC* approach is doing slightly better than the *BPMA* approach in this respect.

5. Discussion

Further research will involve improving the MCMC approach and carrying out a multiple imputation as well as comparing it to the regression approach with added residuals as described in Section 3.2. In order to ensure the normality of the income variables, a log transformation was carried out. Since the sum of the transformed variables is preserved using the benchmarking technique, this does not guarantee that the sum of the original variables is preserved and therefore a correction factor was introduced into the constant term of the regression model. A better approach to this problem will be further investigated.

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 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 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.

6. 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.

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

Kovar, J. en P. Whitridge (1995), Imputation of Business Survey Data. In: *Business Survey Methods* (ed. Cox, Binder, Chinnappa, Christianson & Kott), John Wiley & Sons, New York, pp. 403-423.

Little, R.J.A. and D.B. Rubin (2002), *Statistical Analysis with Missing Data (second edition)*. John Wiley & Sons, New York.

Longford, N.T. (2005), *Missing Data and Small-Area Estimation*. Springer, New York.

Pannekoek, J. and T. De Waal (2005), Automatic Edit and Imputation for Business Surveys: the Dutch Contribution to the EUREDIT Project. *Journal of Official Statistics 21*, pp. 257-286.

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.

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.

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