Discussion Paper

Imputation of Numerical Data under Edit Restrictions: The Vertices Approach

2017 | 02

Ton de Waal
Wieger Coutinho

Content

Summary

Imputation is a frequently used way to treat missing data due to item-nonresponse in surveys. In some cases, the imputed values have to satisfy certain edit restrictions, i.e. constraints on the values of (combinations of) variables. Examples of edit restrictions are that the profit of an enterprise equals its turnover minus its costs, and that the turnover of an enterprise should be at least zero. In this paper we develop new imputation methods satisfying edit restrictions. These imputation methods start with the feasible region defined by the edit restrictions on which we build a statistical distribution with support on that region. Any point drawn from the constructed distribution will then automatically satisfy all edit restrictions. The main questions we aim to answer in this paper are whether we can really develop an imputation based on this reverse idea, and, if so, how would such an imputation method look like?

Keywords

Edit restrictions, imputation, linear programming, missing data, polytopes, vertices

1. Introduction

Imputation is a frequently used way to treat missing data due to item-nonresponse in surveys. Finding suitable imputations that preserve the statistical distribution of the complete data as well as possible is a complicated problem and many imputation methods have been developed in order to overcome this problem. Such imputation methods are discussed in a large number of articles and books, such as Andridge and Little (2010), Kalton and Kasprzyk (1986), Rubin (1987), Schafer (1997), Little and Rubin (2002), De Waal, Pannekoek and Scholtus (2011) and Van Buuren (2012).
In some cases, especially at National Statistical Institutes, the imputation process is further complicated due to the existence of so-called edit rules, or edits for short. Edits are constraints on the values of (combinations of) variables. Examples of edits for numerical data are that the profit of an enterprise equals its turnover minus its costs, and that the turnover of an enterprise should be at least zero.
Several imputation methods satisfying edits have been developed. Generally speaking, one can use either an approach where the variables with missing items are imputed sequentially or an approach where all variables with missing data are imputed simultaneously. Sequential imputation methods satisfying edits have been developed by Raghunathan, Solenberger and Van Hoewyk, (2002), Coutinho, De Waal and Remmerswaal (2011), Coutinho and De Waal (2012), Pannekoek, Shlomo and De Waal (2013) and De Waal, Coutinho and Shlomo (forthcoming). A theoretical drawback of a sequential approach based on an imputation model for each variable with missing values is that a joint model compatible the parametric models may not exist (see, e.g., Rubin 2003). A drawback of a non-iterative sequential approach, where missing items are imputed only once in a single iteration, is that the quality of the variables that are imputed later on in the imputation process is generally lower than the quality of the variables that are imputed at the start. This is due to the edit restrictions, which limit the imputation options for the later variables.
Geweke (1991), Tempelman (2007) and Kim (2013) have developed imputation methods satisfying edits that impute all variables with missing data simultaneously. These methods use a statistical distribution as a starting point. Next, that statistical distribution is somehow truncated to the feasible region described by the edits. Geweke (1991) and Tempelman (2007) have proposed to use the truncated multivariate normal distribution. Kim et al. (2014) have developed a Bayesian multiple imputation approach, using a Dirichlet process mixture of multivariate normal distributions as the base imputation engine.
In the current paper we will explore the reverse idea, where we start with the feasible region defined by the edits and then build a statistical distribution with support on that region. Any point drawn from the constructed distribution will then automatically satisfy all edits. The main questions we aim to answer in this paper are whether we can really develop an imputation based on this reverse idea, and, if so, how would such an imputation method look like?
We will examine a parametric version based on a posited statistical distribution for the data and a non-parametric version where instead of using a statistical distribution we use a hot deck-like approach. Like the approaches Geweke (1991), Tempelman

(2007) and Kim at al. (2014), the imputation methods we propose impute all variables with missing data simultaneously.
Tempelman (2007) has also proposed to use a model based on the Dirichlet distribution. However, that imputation method can only be used for a very specific situation, namely when all variables are non-negative, sum up to a known constant, and no other edit restrictions are defined for these variables. In our parametric imputation method, which can be applied to more general edits and variables than Tempelman's method, the Dirichlet distribution also plays an important role.
In the current paper we focus on numerical data. Throughout the paper we assume that the missing data mechanism is either Missing Completely At Random (MCAR) or Missing At Random (MAR). Informally, in the case of MCAR there is no relation between the missing data pattern and the values of the data. In the case of MAR there is a relation between the missing data pattern and the values of the observed data, but not between the missing data pattern and the values of the missing data. Using the values of the observed data, one can then, in principle, correct for the relation between the missing data pattern and the values of the observed data.
The remainder of this paper is organized as follows. Section 2 discusses edit restrictions and their relation with the vertices of a polytope. Section 3 describes our imputation methods. Section 4 gives an evaluation study, and Section 5 concludes the paper with a short discussion.

2. Edits and vertices of a polytope

2.1 Edits

Edits for numerical data are generally either linear equations or linear inequalities. That is, edit e $(e = 1, \dots, E)$ can be written in either of the two following forms:

$$a_{1e}x_1 + \cdots + a_{ne}x_n + b_e = 0 \tag{1a}$$

or

$$a_{1e}x_1 + \cdots + a_{ne}x_n + b_e \geq 0. \tag{1b}$$

Here the a_{je} and the b_e are constants and the x_j $(j = 1, \dots, n)$ are the variables. Edits of type (1a) are referred to as balance edits. An example of such an edit is

$$P = T - C \tag{2}$$

where T is the turnover of an enterprise, P its profit, and C its costs. Edit (2) expresses that the profit of an enterprise equals its turnover minus its costs. Edit (2) can be written in the form (1a) as $T - C - P = 0$.

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

$$T \geq 0$$

expressing that the turnover of an enterprise should be non-negative.
Together the edits define a feasible region of allowed vectors. We assume that this feasible region is bounded. It is then a polytope (see, e.g. Chvátal 1983 and Schrijver 1986), which we will refer to as the "edit polytope". For statistical data based on observations of actual phenomena this assumption is always satisfied, because we can always find upper and lower bounds for individual variables.

2.2 Vertices of polytopes

Both imputation methods we propose in this paper use the vertices of the edit polytope as starting point. It is well-known that any point in a polytope can be described as a convex combination of its vertices (see, e.g. Chvátal 1983 and Schrijver 1986). Suppose the vertices of the edit polytope are given by $\boldsymbol{x}_1^0$ to $\boldsymbol{x}_r^0$. We use the superscript "0" to indicate that these $\boldsymbol{x}_1^0$ to $\boldsymbol{x}_r^0$ are fixed and known vectors. Any point $\boldsymbol{x}$ in the edit polytope can then be written as

$$\boldsymbol{x} = \sum_{i=1}^{r} \lambda_i \boldsymbol{x}_i^0, \tag{3}$$

for some scalars λ_i $(i = 1, \dots, r)$ satisfying

$$\lambda_i \geq 0 \ (i = 1, \dots, r) \tag{4}$$

and

$$\sum_{i=1}^{r} \lambda_i = 1 \tag{5}$$

The vertices of the edit polytope can be determined by means of several different approaches. For our evaluation study, we used an approach developed by Duffin (1974), which is similar to an approach developed by Chernikova (1964, 1965). For some alternatives approaches we refer to Rubin (1975, 1977).

Example: We will illustrate our parametric imputation method by means of an example. The edits of our example are given below. We have already introduced variables T, P and C in Section 2.1. Variable N denotes the number of employees of an enterprise. (For convenience, we combined upper and lower bound on a variable, which are actually two edits of type (1b), into one expression).

$$P = T - C \tag{6}$$
$$0.5T \leq C \leq 1.1T \tag{7}$$
$$0 \leq T \leq 400N \tag{8}$$
$$0 \leq T \leq 1{,}000{,}000 \tag{9}$$
$$0 \leq C \leq 1{,}000{,}000 \tag{10}$$

$$0 \leq N \leq 1{,}000 \quad (11)$$
$$-500{,}000 \leq P \leq 500{,}000 \quad (12)$$

Note that the lower and upper bounds in (9) to (12) ensure that the polytope is indeed bounded.
In this example there turn out to be (only) five vertices. These are described in Table 1.

Table 1. Vertices in the example

Vertex	T	C	P	N
$\boldsymbol{x}_1^0$	0	0	0	0
$\boldsymbol{x}_2^0$	400,000	400,000	0	1,000
$\boldsymbol{x}_3^0$	400,000	200,000	200,000	1,000
$\boldsymbol{x}_4^0$	400,000	440,000	-40,000	1,000
$\boldsymbol{x}_5^0$	0	0	0	1,000

All points satisfying (6) to (12) can be written as $\sum_{i=1}^{5} \lambda_i \boldsymbol{x}_i^0$, where the λ_i $(i = 1, \dots, 5)$ satisfy (4) and (5). In other words, any vector (T, C, P, N) satisfying (6) to (12) can be written as

$$T = 400{,}000\lambda_2 + 400{,}000\lambda_3 + 400{,}000\lambda_4 \quad (13)$$
$$C = 400{,}000\lambda_2 + 200{,}000\lambda_3 + 440{,}000\lambda_4 \quad (14)$$
$$P = 200{,}000\lambda_3 - 40{,}000\lambda_4 \quad (15)$$
$$N = 1{,}000\lambda_2 + 1{,}000\lambda_3 + 1{,}000\lambda_4 + 1{,}000\lambda_5 \quad (16)$$

with λ_1 to λ_5 satisfying

$$\lambda_i \geq 0 \text{ for } i = 1, \dots, 5$$

and

$$\sum_{i=1}^{5} \lambda_i = 1.$$

3. Imputation methods

3.1 The parametric approach

3.1.1 Main idea

In our parametric approach we specify a probability distribution $\Pr(\lambda_1, \dots, \lambda_r; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of model parameters, for the λ_i $(i = 1, \dots, r)$ such that (4) and (5) are always satisfied. Any point $\boldsymbol{x}$ defined by (3) using λ_i $(i = 1, \dots, r)$ drawn from $\Pr(\lambda_1, \dots, \lambda_r; \boldsymbol{\theta})$ then automatically satisfies all edits. The statistical problem of

finding a suitable probability distribution is hereby shifted from the $\boldsymbol{x}$-space to the $\boldsymbol{\lambda}$-space.

The next step in the vertices approach is to specify a suitable class of probability distributions $\Pr(\lambda_1, \ldots, \lambda_r; \boldsymbol{\theta})$ satisfying (4) and (5). In our evaluation study we will draw the λ_i $(i = 1, \ldots, r)$ from a Dirichlet$_r(\boldsymbol{\theta})$ distribution (see, e.g., Ng, Tian and Tang 2011), where $\boldsymbol{\theta}$ is the vector of model parameters. The Dirichlet($\boldsymbol{\theta}$) distribution is given by

$$\text{Dirichlet}_r(\boldsymbol{\lambda}, \boldsymbol{\theta}) = \frac{1}{B(\boldsymbol{\theta})} \prod_{i=1}^{r} \lambda_i^{\theta_i - 1}$$

with $B(\boldsymbol{\theta})$ the Beta function. The Beta function is itself given by

$$B(\boldsymbol{\theta}) = \frac{\prod_{i=1}^{r} \Gamma(\theta_i)}{\Gamma(\sum_{i=1}^{r} \theta_i)}$$

with Γ the Gamma function.

Obviously, one could also use other distributions to draw the λ_i $(i = 1, \ldots, r)$ from (again see, e.g., Ng, Tian and Tang 2011, for examples of such distributions).

The observed data are used to obtain an estimate $\widehat{\boldsymbol{\theta}}$ for the model parameters. How the model parameters are estimated exactly is explained in Subsection 3.1.2 below.

After we have obtained an estimate $\widehat{\boldsymbol{\theta}}$ for the model parameters, and hence a probability distribution $\Pr(\lambda_1, \ldots, \lambda_r; \widehat{\boldsymbol{\theta}})$, we are ready to draw imputations for each record with missing values.

For each record $\boldsymbol{x}$, we re-order the variables so $\boldsymbol{x}$ can be written as $\boldsymbol{x} = (\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$, where $\boldsymbol{x}_{\text{obs}} = \left(x_{\text{obs},1}, \ldots, , x_{\text{obs},p(\boldsymbol{x}),}\right)$ is the vector of observed values for record $\boldsymbol{x}$ and $\boldsymbol{x}_{\text{mis}} = \left(x_{\text{mis},1}, \ldots, , x_{\text{mis},q(\boldsymbol{x})}\right)$ is the vector of missing values for record $\boldsymbol{x}$. We use $p(\boldsymbol{x})$ to denote the number of observed values in record $\boldsymbol{x}$ and $q(\boldsymbol{x})$ to denote the number of missing values in record $\boldsymbol{x}$.

We need to draw values for the missing part, $\boldsymbol{x}_{\text{mis}}$, conditional on $\boldsymbol{x}_{\text{obs}}$. In the λ-space this means that we need to draw from the probability distribution $\Pr(\lambda_1, \ldots, \lambda_r; \widehat{\boldsymbol{\theta}})$ conditional on certain linear restrictions for the λ_i $(i = 1, \ldots, r)$. These linear restrictions are given by

$$\boldsymbol{x}_{\text{obs},j} = \sum_{i=1}^{r} \lambda_i \boldsymbol{x}_{ij}^0 \qquad (17)$$

where $\boldsymbol{x}_{\text{obs}}$ and the $\boldsymbol{x}_i^0$ $(i = 1, \ldots, r)$ are fixed vectors and the λ_i $(i = 1, \ldots, r)$ are stochastic variables.

By shifting from the $\boldsymbol{x}$-space to the $\boldsymbol{\lambda}$-space we solve one problem – satisfying edits – but create new problems when estimating model parameters and when actually drawing imputations as the observed data are only available in the $\boldsymbol{x}$-space. What complicates matters considerably is that there is no one-to-one correspondence between points in the $\boldsymbol{x}$-space and points in the λ-space. In particular, for a point in the $\boldsymbol{x}$-space there does not need to be a unique set of λ_i $(i = 1, \ldots, r)$ satisfying (4) and (5), but there may be several of such sets. Below we will return to these complications.

3.1.2 Estimating a probability distribution on the vertices

Owing to the shift from the $\boldsymbol{x}$-space to the λ-space and the non-uniqueness of the λ_i $(i = 1, \dots, r)$ for a given $\boldsymbol{x}$, estimating the model parameters may be complicated. For instance, maximum likelihood estimation of the model parameters appears to be too complicated to apply for our approach in most practical situations. Instead, we therefore resort to the method of moments estimation. Even this is non-trivial in our case, as the number of parameters is not directly related to the number of variables in the $\boldsymbol{x}$-space. Instead, the number of parameters is related to the number of vertices of the edit polytope. As a result the set of moment equations may be over- of under-specified. In order to arrive at an approximate solution to the set of moment equations in any case, we therefore solve a minimization problem where we aim to satisfy the moment equations as well as possible.

We will minimise the distance between expected moments of the Dirichlet$_r(\boldsymbol{\theta})$ distribution and the observed moments in the data set. We measure the distance by a weighted sum of the squared differences between expected moments and observed moments. That is, we will minimise

$$\sum_{k \in S} \gamma_k \left(M_{\text{exp},k} - M_{\text{obs},k}\right)^2, \tag{18}$$

where S is a set of selected moments that we want to use for estimating the model parameters, $M_{\text{exp},k}$ are the expected moments under the Dirichlet$_r(\boldsymbol{\theta})$ distribution, $M_{\text{obs},k}$ the observed moments, and the γ_k are non-negative weights reflecting how closely one wants to preserve the corresponding observed moment.

In our evaluation study (see Section 4), we have used the first moments of all variables as our set S in (18). We have set the weight corresponding to a certain variable equal to the reciprocal of the observed mean. If one wishes, one could add more terms, corresponding to higher order moments such as variances, to (18).

Example (continued): We continue with our example. We assume that the λ_i $(i = 1, \dots, 5)$ follow a Dirichlet$_5(\boldsymbol{\theta})$ distribution. The expectation of T, C, P and N can then be written as

$$\mathrm{E}T = 400{,}000\left(\frac{\theta_2+\theta_3+\theta_4}{\theta_+}\right)$$

$$\mathrm{E}C = 400{,}000\left(\frac{\theta_2}{\theta_+}\right) + 200{,}000\left(\frac{\theta_3}{\theta_+}\right) + 440{,}000\left(\frac{\theta_4}{\theta_+}\right)$$

$$\mathrm{E}P = 200{,}000\left(\frac{\theta_3}{\theta_+}\right) - 40{,}000\left(\frac{\theta_4}{\theta_+}\right)$$

$$\mathrm{E}N = 1{,}000\left(\frac{\theta_2+\theta_3+\theta_4+\theta_5}{\theta_+}\right)$$

with $\theta_+ = \sum_{i=1}^{5} \theta_i$. Here we have used that $\mathrm{E}\lambda_i = \frac{\theta_i}{\theta_+}$.

We denote the averages for T, C, P and N in the observed data by $\bar{T}_{\text{obs}}$, $\bar{C}_{\text{obs}}$, $\bar{P}_{\text{obs}}$ and $\bar{N}_{\text{obs}}$. Now, we estimate the model parameters θ_i $(i = 1, \dots, 5)$ of the Dirichlet distribution by minimizing

$$\gamma_T(\mathrm{E}T - \bar{T}_{obs})^2 + \gamma_C(\mathrm{E}C - \bar{C}_{obs})^2 + \gamma_P(\mathrm{E}P - \bar{P}_{obs})^2 + \gamma_N(\mathrm{E}N - \bar{N}_{obs})^2,$$

where the γ_T, γ_C, γ_P, and γ_N are user-specified non-negative weights.

3.1.3 Imputing missing data

Once we have estimated the parameters of the probability distribution, the next step is to impute the missing data for each record. As mentioned before, this is a non-trivial task as we need to draw from the probability distribution $\Pr(\lambda_1, \ldots, \lambda_r; \widehat{\boldsymbol{\theta}})$ conditional on certain linear restrictions for the λ_i $(i = 1, \ldots, r)$ given by (17). Only in very exceptional cases, one would be able to derive closed expressions for such a conditional distribution. Instead, one usually has to rely on Monte Carlo methods (see, e.g., Robert and Casella 1999 and Liu 2003 for more on Monte Carlo methods in general). In our case we use a (Metropolised) hit-and-run algorithm (see Chen and Schmeiser 1993 and Romeijn and Smith 1994).

In a hit-and-run algorithm one starts with a point $\boldsymbol{\lambda}^{(0)} = (\lambda_1^{(0)}, \ldots, \lambda_r^{(0)})$ satisfying the linear restrictions for the λ_i $(i = 1, \ldots, r)$ given by (17). Next, another point in the λ-space satisfying (17) is proposed. We accept or reject this proposed point with a certain probability. If the proposed point is accepted, $\boldsymbol{\lambda}^{(1)}$ is set equal to this proposed point, otherwise, $\boldsymbol{\lambda}^{(1)} = \boldsymbol{\lambda}^{(0)}$. Again, we now propose a point, and accept or reject this proposed point with a certain probability, and so on. This defines a Markov chain. The probability of accepting or rejecting a new point in the λ-space is defined in such a way that the stationary distribution of the Markov chain equals the probability distribution $\Pr(\lambda_1, \ldots, \lambda_r; \widehat{\boldsymbol{\theta}})$ conditional on the linear restrictions for the λ_i $(i = 1, \ldots, r)$ given by (17).

In general, there three aspects we need to pay attention to:

- Selecting starting values for the $\lambda_i^{(0)}$ $(i = 1, \ldots, r)$ satisfying (17);
- Proposing a new point $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_r^*)$ in the λ-space satisfying (17) given a current point $\boldsymbol{\lambda}^{(t)} = (\lambda_1^{(t)}, \ldots, \lambda_r^{(t)})$.
- Setting the probability of accepting the new point in the λ-space.

In order to find a starting point inside the polytope we try to solve a linear programming problem given by (5), (17) and

$$\lambda_i \geq \varepsilon \; (i = 1, \ldots, r),$$

where ε is a small positive number, say $\varepsilon = 0.001$. However, for some records this problem may not have a solution. In those cases we accept that some λ_i are equal to zero. For instance, all λ_i $(i = 1, \ldots, r)$ for which $\boldsymbol{x}_{obs,j} = 0$ and $\boldsymbol{x}_{ij}^0 \neq 0$ necessarily have to be equal to zero (see (17)).

For b) we draw a random direction from the current point $\boldsymbol{\lambda}^{(t)} = (\lambda_1^{(t)}, \ldots, \lambda_r^{(t)})$ in the polytope defined by (4), (5) and (17). The maximum step we can take in the selected random direction is determined by the boundary of this polytope. We select a random distance between zero and the maximum step size to propose a new point $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_r^*)$.

This potential new point is accepted with probability

$$J(\boldsymbol{\lambda}^{(t)}, \boldsymbol{\lambda}^*) = \min\left\{1, \frac{\Pr(\boldsymbol{\lambda}^*; \widehat{\boldsymbol{\theta}})}{\Pr(\boldsymbol{\lambda}^{(t)}; \widehat{\boldsymbol{\theta}})}\right\} \tag{19}$$

where $\Pr(\boldsymbol{\lambda}; \widehat{\boldsymbol{\theta}})$ is the Dirichlet$(\widehat{\boldsymbol{\theta}})$ distribution. We can express $J(\boldsymbol{\lambda}^{(t)}, \boldsymbol{\lambda}^*)$ as

$$J\left(\boldsymbol{\lambda}^{(t)}, \boldsymbol{\lambda}^{*}\right) = \min\left\{1, \frac{\prod_{i=1}^{r}\left(\lambda_i^{*}\right)^{\hat{\theta}_i - 1}}{\prod_{i=1}^{r}\left(\lambda_i^{(t)}\right)^{\hat{\theta}_i - 1}}\right\} \tag{20}$$

The second term in (20) can be calculated as

$$\exp\left(\sum_{i=1}^{r}\left(\hat{\theta}_i - 1\right)\left(\ln(\lambda_i^{*}) - \ln(\lambda_i^{(t)})\right)\right) \tag{21}$$

We draw a large number of random points in the λ-space until convergence. We then draw the final values on which we base our imputations.
Note that we are referring to the convergence of a statistical distribution and that convergence of a distribution is generally hard to establish. In practice, one often monitors the distribution of the generated data over a large number of iterations and uses this to check whether the distribution appears to converge. In our evaluation study we have also applied this pragmatic approach.
For records for which $\lambda_i^{(t)} = 0$ for some $i = 1, \ldots, r$, (20) cannot be computed. We then accept the proposed new point $\boldsymbol{\lambda}^{*}$ if it has less components equal to zero. If $\boldsymbol{\lambda}^{(t)}$ and $\boldsymbol{\lambda}^{*}$ are equal to zero for the same components, we exclude these components while calculating (20) or (21).

Example (continued): Suppose that in a certain record $T = 75{,}000$ and $N = 250$ are observed, and C and P are missing. To impute C and P we need to draw values from

$$\Pr\left(\lambda_1, \ldots, \lambda_5 | T = 75{,}000 \text{ and } N = 250; \hat{\boldsymbol{\theta}}\right),$$

i.e. (see (13) and (16))

$$\Pr\left(\lambda_1, \ldots, \lambda_5 | 400{,}000(\lambda_2 + \lambda_3 + \lambda_4) = 75{,}000 \text{ and } 1{,}000(\lambda_2 + \lambda_3 + \lambda_4 + \lambda_5) = 250; \hat{\boldsymbol{\theta}}\right)$$

The starting λ_i $(i = 1, \ldots, 5)$ have to satisfy (5),

$$400{,}000(\lambda_2 + \lambda_3 + \lambda_4) = 75{,}000 \tag{22}$$
$$1{,}000(\lambda_2 + \lambda_3 + \lambda_4 + \lambda_5) = 250 \tag{23}$$
$$\lambda_i \geq \varepsilon \ (i = 1, \ldots, r), \tag{24}$$

where ε is a sufficiently small positive number, say $\varepsilon = 0.001$.
We can find solutions subject to (5), (22), (23) and (24) by maximizing a linear programming problem with an arbitrary linear target function $f(\lambda_1, \ldots, \lambda_r)$, for instance

$$f(\lambda_1, \ldots, \lambda_r) = \lambda_1 \tag{25}$$

subject to the constraints (5), (22), (23) and (24).
We draw new points until convergence to the distribution. We then draw final values for λ_1 to λ_5. Once we have drawn final values λ_1 to λ_5 we can calculate imputation values for C and P using (14) and (15).

3.2 The non-parametric approach

Our non-parametric approach consists of two steps for each record with missing data. In the first step the nearest complete record is found for each record with missing data. That is, for each record with missing values $\boldsymbol{x}_r$ we find a complete donor record $\boldsymbol{x}_d^*$ that minimizes

$$\sum_{j \in \text{Obs}(\boldsymbol{x}_r)} u_j \left| \boldsymbol{x}_{rj} - \boldsymbol{x}_{dj}^* \right|,$$

where $\text{Obs}(\boldsymbol{x}_r)$ is the set of observed values in record $\boldsymbol{x}_r$ and u_j ($j \in \text{Obs}(\boldsymbol{x}_r)$) are non-negative user-specified weights. In our evaluation study (see Section 4) we have again set u_j equal to the reciprocal of the observed mean for variable x_j, just as we did for the γ_j in (18).

Suppose the donor record $\boldsymbol{x}^*$ for record $\boldsymbol{x}$ can be written as $\boldsymbol{x}^* = \sum_i^r \lambda_i^* \boldsymbol{x}_i^0$ with r again the number of vertices of the edit polytope, $\boldsymbol{x}_i^0$ its vertices, $\lambda_i^* \geq 0$ ($i = 1, \ldots, r$) and $\sum_i^r \lambda_i^* = 1$. The (as yet unknown) imputed version of record $\boldsymbol{x}$ can likewise be written as $\boldsymbol{x} = \sum_i^r \lambda_i \boldsymbol{x}_i^0$ with $\lambda_i \geq 0$ ($i = 1, \ldots, r$) and $\sum_i^r \lambda_i = 1$.

In the second step of our non-parametric imputation approach, we aim to find values for the λ_i ($i = 1, \ldots, r$) that are as close as possible to the corresponding values for λ_i^* ($i = 1, \ldots, r$) of the selected nearest-neighbor donor record $\boldsymbol{x}^*$. We do this by finding λ_i and λ_i^* ($i = 1, \ldots, r$) that minimize

$$\sum_{i=1}^r |\lambda_i - \lambda_i^*| \tag{26}$$

subject to the conditions

$$\sum_i^r \lambda_i^* \boldsymbol{x}_i^0 = \boldsymbol{x}^* \tag{27}$$

$$\sum_i^r \lambda_i \boldsymbol{x}_{ij}^0 = \boldsymbol{x}_{\text{obs},j} \qquad \text{for } j = 1, \ldots, p(\boldsymbol{x}) \tag{28}$$

$$\lambda_i^*, \lambda_i \geq 0 \text{ for } i = 1, \ldots, r \tag{29}$$

$$\sum_i^r \lambda_i^* = 1 \tag{30}$$

$$\sum_i^r \lambda_i = 1 \tag{31}$$

This is a linear programming problem and can, for instance, be solved by means of the simplex algorithm.

A minor technical problem is that many algorithms or software packages for the simplex algorithm cannot minimize a sum of absolute distances directly. We overcome this technical problem by introducing variables μ_i ($i = 1, \ldots, r$) that have to satisfy

$$\mu_i \geq \lambda_i - \lambda_i^* \text{ for } i = 1, \ldots, r \tag{32}$$
$$\mu_i \geq \lambda_i^* - \lambda_i \text{ for } i = 1, \ldots, r \tag{33}$$

Our minimization problem is now given by:

Minimize $\sum_{i=1}^{r} \mu_i$ (34)

subject to (27) to (33).
Since in an optimal solution $\mu_i = \lambda_i - \lambda_i^*$ or $\mu_i = \lambda_i^* - \lambda_i$ $(i = 1, \ldots, r)$, we have that in an optimal solution $\mu_i = |\lambda_i - \lambda_i^*|$ $(i = 1, \ldots, r)$, and the value of (34) will equal the value of (26).

4. Evaluation study

4.1 Evaluation approach

In our evaluation study we have used two data sets. The true values for the data in the two data sets are known. In the completely observed data sets values were deleted by a third party, using a MAR mechanism. For each of our evaluation data sets we thus have two versions available: a version with missing values and a version with complete records. The former version is imputed, without making any use of the complete records. The resulting data set is then compared to the version with complete records.
To evaluate the results of our imputation methods we have compare them to nearest-neighbor hot deck imputation, using the Euclidean distance function, and random hot deck imputation. We will refer to our parametric method as P, to our non-parametric method as NP, to nearest-neighbor hot deck imputation as NN HD and to random hot deck imputation as Random HD.

4.2 Evaluation data

The main characteristics of the two data sets are presented in Table 2.

Table 2. The characteristics of the evaluation data sets

	Data set 1	**Data set 2**
Total number of records	3096	500
Number of records with missing values	287	250
Total number of variables	5	6
Total number of edits	14	17
Number of balance edits	0	1
Total number of inequality edits	14	16
Number of non-negativity edits	5	6

The edit polytope of data set 1 has 21 vertices whereas the edit polytope of data set 2 has 28 vertices.
The numbers of missing values and (unweighted) means of the variables of our data sets are given in Tables 3 and 4. The percentages in brackets are the percentages of

records with a missing value for the corresponding variable out of the total number of 3,096 records for data set 1 and 500 records for data set 2. The means are taken over all observations in the complete version of the data sets.

Table 3. The numbers of missing values and the means in data set 1

Variable	Number of missing values	Mean
R1	0 (0.0 %)	37.4
R2	79 (2.2%)	777.6
R3	76 (4.2%)	11574.8
R4	73 (4.8%)	209.9
R5	67 (2.6%)	169.2

Table 4. The numbers of missing values and the means in data set 2

Variable	Number of missing values	Mean
S1	61 (12.2%)	97.8
S2	86 (17.2%)	175018.3
S3	131 (26.2%)	731.0
S4	61 (12.2%)	175749.3
S5	109 (21.8%)	154286.5
S6	91 (18.2%)	7522.3

4.3 Evaluation measures

In our evaluation study we focus on three different aspects of our imputation approaches: the preservation of individual values, the preservation of totals and means, and the preservation of univariate distributions. We use measures proposed by Chambers (2003) to evaluate these aspects.

To measure the preservation of individual values we use the d_{L1} measure, which calculates the average distance between the imputed and true values. The d_{L1} measure is defined as

$$d_{L1} = \frac{\sum_{i \in M(j)} w_i |\hat{x}_{ij} - x_{ij}^{\text{true}}|}{\sum_{i \in M(j)} w_i}$$

where $\hat{x}_{ij}$ is the imputed value in record i of the variable x_j under consideration and x_{ij}^{true} the corresponding true value. Note that the value of d_{L1} depends on the variable under consideration. The same holds true for the evaluation measures mentioned below.

To measure the preservation of totals and means we use the m_1 measure, which calculates the preservation of the first moment of the empirical distribution of the true values. The m_1 measure is defined as

$$m_1 = \left| \frac{\sum_{i \in M(j)} w_i (\hat{x}_{ij} - x_{ij}^{\text{true}})}{\sum_{i \in M(j)} w_i} \right|$$

To measure the preservation of univariate distributions we use the KS Kolmogorov-Smirnov distance, which compares the empirical distribution of the original values to

the empirical distribution of the imputed values. For weighted data, the empirical distribution of the true values is defined as

$$F_{x_j}(t) = \sum_{i \in M(j)} I\left(w_i x_{ij} \leq t\right) / |M(j)|$$

with $|M(j)|$ the number of records with missing values for the variable x_j under consideration, w_i the survey weight of record i and I the indicator function. Similarly, we define $F_{\hat{x}_j}(t)$. The KS distance is defined as

$$KS = \max_k |F_{x_j}(t_k) - F_{\hat{x}_j}(t_k)|,$$

where the t_k- values are the $2|M(j)|$ jointly ordered true and imputed values. Smaller absolute values of the evaluation measures indicate better imputation performance.
To evaluate to what extent the relationship between different variables are preserved, we also compare the correlations in the (partly) imputed data to the true data. For this we calculate the average absolute difference between the correlations in the true complete data and the imputed data, where we have taken the average over all 10 pairs of variables for data set 1 and all 15 pairs for data set 2.We have also calculated the average of the absolute percent differences (where the percentage is calculated with respect to the correlations in the complete data) over all pairs of variables.

4.4 Evaluation results

The evaluation results for data set 1 and data set 2 are presented in Tables 5 and 6, respectively. As variable R1 has no missing values it is not included in Table 5. “Average” is the average of the absolute results over all 4, respectively 6, variables mentioned in the tables.

Table 5. Results for data set 1

	R2	R3	R4	R5	Average
d_{L1}					
P	835	4607	723	19	1237
NP	1009	11206	85	26	2465
NN HD	415	4301	68	24	962
Random HD	1140	8080	81	26	1865
m_1					
P	803	1789	723	5	664
NP	132	4364	16	12	905
NN HD	9	191	12	14	45
Random HD	182	432	9	12	127
KS					
P	0.22	0.29	0.21	0.44	0.23
NP	0.26	0.45	0.35	0.92	0.40
NN HD	0.10	0.13	0.26	1.00	0.30
Random HD	0.32	0.51	0.47	0.92	0.44

Table 6. Results for data set 2

	S1	S2	S3	S4	S5	S6
d_{L1}						
P	103	17372	334	24425	147256	3035
NP	40	155412	277	113634	56501	3424
NN HD	30	48954	544	48413	37922	1786
Random HD	36	118355	548	110335	102786	3338
m_1						
P	103	9615	333	14250	147256	1040
NP	21	149792	81	105745	45943	2283
NN HD	3	141	121	4651	11573	365
Random HD	8	156	32	15369	6.044	495
KS						
P	0.04	0.03	0.07	0.05	0.10	0.06
NP	0.05	0.05	0.05	0.05	0.05	0.05
NN HD	0.03	0.03	0.07	0.05	0.04	0.03
Random HD	0.04	0.09	0.05	0.06	0.06	0.03

The results are rather disappointing for our imputation methods. For both data sets they are outperformed by NN standard for most evaluation measures. An exception is the KS measure for data set 1, for which our parametric methods P performs best. For data set 1 our parametric method P performs slightly better than Random HD, while NP seems to perform worst.

For data set 2 the differences with respect to the KS measure are very small. Method P is better than Random HD with respect to d_{L1}, but worse with respect to m_1. Method NP again seems to be the worst method.
As can be seen in Table 7, NN HD also performs best with respect to the preservation of correlations. The other methods seem to perform about equally well.

Table 7. Average absolute differences between correlations in true data and imputed data

	Data set 1	Data set 2
P	0.0039 (1.4%)	0.1269 (27.1%)
NP	0.0030 (0.9%)	0.1690 (42.2%)
NN HD	0.0006 (0.3%)	0.0410 (21.8%)
Random HD	0.0035 (0.9%)	0.1457 (37.5%)

Finally, Table 8 gives the number of violated edits and the number of violated records, where a record is considered violated if at least one edit is violated for that record.

Table 8. Edit violations

	Data set 1		Data set 2	
	Violated edits	Violated records	Violated edits	Violated records
P	0	0	0	0
NP	0	0	0	0
NN HD	7	7	245	205
Random HD	55	55	250	206

By design our imputation methods P and NP violate no edits and no records. NN HD violates only 7 edits in 7 records, i.e. 1.4% of the total number of records, for data set 1. For data set 2 NN HD violates 245 edits in 205 records, i.e. 6.6% of the total number of records. The results for Random HD were clearly worse for data set 1 and slightly worse for data set 2.
In data set 2 NN standard violates the balance edit 198 times and inequality edits 47 times. For the same data set Random HD violated the balance edit 193 times and inequality edits 57 times.

5. Discussion

In this paper we have explored a new approach to finding imputations for numerical data that have to satisfy certain edit restrictions. In traditional methods one either uses a sequential approach or one starts with a distribution for the complete data and somehow truncates that to the allowed region. We have shown that a reverse approach to the latter approach is also a feasible option. In this reverse approach the region defined by the edit restrictions forms the starting point, not the distribution for the complete data. In particular we have demonstrated how parameters for such

a reverse approach can be estimated, and how missing values can be imputed. Besides a parametric imputation method we have also proposed a non-parametric version.

The results of our imputation methods are a bit disappointing: even our best performing imputation method, the parametric method P, is generally outperformed by standard nearest-neighbor hot deck imputation. An obvious exception is the number of edit violations: by design our imputation methods do not violate any edits, whereas standard imputation methods, such as nearest-neighbor hot deck imputation and random hot deck donor imputation do. So, only when it is considered important that all edits are satisfied, our parametric imputation method should be considered for use in practice.

A possible explanation for the disappointing performance of the proposed imputation methods is that they do not make use of auxiliary information. This in contrast to, for instance, nearest-neighbor hot deck imputation, where auxiliary data are used to find appropriate donors. A possible way to improve the proposed imputation methods is by extending them so auxiliary information can be taken into account.

Another somewhat disappointing aspect of our imputation methods is their complexity, especially for the parametric method. When we started the development of our imputation methods we had not fully understood the problems that are caused by the shift from the $\boldsymbol{x}$-space to the $\boldsymbol{\lambda}$-space. This shift considerably complicates the estimation of model parameters, and the actual imputation of missing data. The complexity might be a limiting factor for application of our imputation methods in practice.

The actual imputation for our parametric imputation method can be simplified so it resembles the approach for our non-parametric method. In this simplified approach U records $\boldsymbol{x}_u^*$ $(u = 1, \dots, U)$ are generated by drawing from the estimated Dirichlet distribution on the vertices. Next, for each generated record $\boldsymbol{x}_u^* = \sum_{i=1}^{r} \lambda_{i,u}^* \boldsymbol{x}_i^0$ we find values for the λ_i $(i = 1, \dots, r)$ of a synthetic record $\boldsymbol{x} = \sum_{i=1}^{r} \lambda_i \boldsymbol{x}_i^0$ that are as close as possible to the $\lambda_{i,u}^*$ by minimizing (26) subject to (27) to (31). Finally, we select the synthetic record for which the adjustment given by (26) was the least as our imputed record. In this way the use of the rather complicated Metropolised hit-and-run algorithm of Section 3.1.3 can be avoided, at the expense of having to solve U linear programming problems.

Another limiting factor is that our imputation methods rely on the generation of all vertices of the edit polytope. The number of vertices of a polytope can grow rapidly with the number of variables and/or edits (see, e.g. McMullen 1970 and Avis and Devroye 2000) This means that for problems with a large number of variables and/or edits our imputation will not be computationally feasible in practice.

Despite the shortcomings of our imputation methods, we feel that they might offer an interesting topic for future research. We hope that our paper will inspire other researchers to develop improved versions of our imputation methods that are wider applicable.

Acknowledgement

The authors thank Jeroen Pannekoek for his useful comments on an earlier version of this paper.

References

Andridge R.R. and R.J.A. Little (2010). A Review of Hot Deck Imputation for Survey Nonresponse. International Statistical Review 78, pp. 40-64.

Avis, D. and L. Devroye (2000), Estimating the Number of Vertices of a Polyhedron. Information Processing Letters 73, pp. 137-143.

Chen, M-H. and B.W Schmeiser (1993), Performance of the Gibbs, Hit and Run, and Metropolis Samplers. Journal of Computational and Graphical Statistics 2, pp. 251-272.

Chernikova, N.V. (1964), Algorithm for Finding a General Formula for the Non-Negative Solutions of a System of Linear Equations. USSR Computational Mathematics and Mathematical Physics 4, 151-158.

Chernikova, N.V. (1965), Algorithm for Finding a General Formula for the Non-Negative Solutions of a System of Linear Inequalities. USSR Computational Mathematics and Mathematical Physics, 5, 228-233.

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

Coutinho, W., T. de Waal and M. Remmerswaal (2011), Imputation of Numerical Data under Linear Edit Restrictions. Statistics and Operations Research Transactions 35, pp. 39-62.

Coutinho, W., T. de Waal and N. Shlomo (2013), Calibrated Hot Deck Imputation Subject to Edit Restrictions. Journal of Official Statistics 29, pp. 299-321.

De Waal, T., W. Coutinho and N. Shlomo (forthcoming). Calibrated Hot Deck Imputation for Numerical Data under Edit Restrictions.

De Waal, T., J. Pannekoek and S. Scholtus (2011), Handbook of Statistical Data Editing and Imputation. John Wiley & Sons, New York.

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

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

Kim, H.J., J.P. Reiter, Q. Wang, L.H. Cox and A.F. Karr (2014), Multiple Imputation of Missing or Faulty Values under Linear Constraints. Journal of Business and Economic Statistics 32, pp. 375-386

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.
McMullen, P. (1970), The Maximum Numbers of Faces of a Convex Polytope. Mathematika 17, pp. 179-184.
Pannekoek, J., N. Shlomo and T. de Waal (2013), Calibrated Imputation of Numerical Data under Linear Edit Restrictions, Annals of Applied Statistics 7, pp. 1983-2006.
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.
Romeijn, H.E. and R.L. Smith (1994), Simulated Annealing for Constrained Global Optimization. Journal of Global Optimization 5, pp. 101-126.
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.
Rubin, D.S. (1975), Vertex Generation and Cardinality Constrained Linear Programs. Operations Research 23, pp. 555-565.
Rubin, D.S. (1977), Vertex Generation Methods for Problems with Logical Constraints. Annals of Discrete Mathematics 1, pp. 457-466.
Schafer, J.L. (1997), Analysis of Incomplete Multivariate Data. Chapman & Hall, London.
Schrijver, A. (1986), Theory of Linear and Integer Programming. John Wiley & Sons.
Tempelman, C. (2007), Imputation of Restricted Data. Doctorate thesis, University of Groningen.
Van Buuren, S. (2012), Flexible Imputation of Missing Data. Chapman & Hall/CRC, Boca Raton, Florida.

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2014–2015	2014 to 2015 inclusive
2014/2015	Average for 2014 to 2015 inclusive
2014/'15	Crop year, financial year, school year, etc., beginning in 2014 and ending in 2015
2012/'13–2014/'15	Crop year, financial year, etc., 2012/'13 to 2014/'15 inclusive

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

Colofon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Studio BCO

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbsl.nl/information