Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2011, Article ID 290380, 20 pages
doi:10.1155/2011/290380

Research Article

Missing-Values Adjustment for Mixed-Type Data

Agostino Tarsitano and Marianna Falcone

Dipartimento di Economia e Statistica, Università della Calabria, Via Pietro Bucci, Cubo 1C, 87036 Rende (Cosenza), Italy

Correspondence should be addressed to Agostino Tarsitano, agotar@unical.it

Received 9 December 2010; Revised 25 May 2011; Accepted 1 July 2011

Academic Editor: Murray Clayton

We propose a new method of single imputation, reconstruction, and estimation of nonreported, incorrect, implausible, or excluded values in more than one field of the record. In particular, we will be concerned with data sets involving a mixture of numeric, ordinal, binary, and categorical variables. Our technique is a variation of the popular nearest neighbor hot deck imputation (NNHDI) where "nearest" is defined in terms of a global distance obtained as a convex combination of the distance matrices computed for the various types of variables. We address the problem of proper weighting of the partial distance matrices in order to reflect their significance, reliability, and statistical adequacy. Performance of several weighting schemes is compared under a variety of settings in coordination with imputation of the least power mean of the Box-Cox transformation applied to the values of the donors. Through analysis of simulated and actual data sets, we will show that this approach is appropriate. Our main contribution has been to demonstrate that mixed data may optimally be combined to allow the accurate reconstruction of missing values in the target variable even when some data are absent from the other fields of the record.

1. Introduction

Missing values are pieces of information which are omitted, lost, erroneous, inconsistent, patently absurd, or otherwise not accessible for a statistical unit about which other useful data are available. Failures in data collection are a matter of major concern both because they reduce the number of valid cases for analysis which, in turn, may result in a potential loss of valuable knowledge, and because, when there is a wide difference between complete and incomplete records, they introduce bias into the estimation/prediction process.

The problems posed by observations that are truly missing or considered as such can be handled by using different strategies. These include additional data collection; application of likelihood-based procedures that allow modeling of incomplete data; deductive reconstruction; the use of only part of the available data (listwise or pairwise

deletion); weighting records; imputation, that is, revision of the data set in an attempt to replace the missing data with plausible values. The present paper is centered on the latter method.

Imputation techniques have been extensively studied over the last few decades, and a number of approaches have been proposed. For an overview of the methods, see, for example, Little and Rubin [1] and Kalton and Kasprzyk [2]. Some of these methods are nowadays available in standard statistical software (or can easily be implemented) although there is little consensus as to the most appropriate technique to use for a particular situation. It is not our intention to provide an exhaustive review of data imputation methods; instead, we discuss just the nearest neighbor hot deck imputation (NNHDI) which has been used for a number of years and enjoys high prestige for its theoretical and computational characteristics.

The NNHDI method involves a nonrandom sample without replacement from the current data set (this explains the term "hot deck" in the name of the method). More specifically, the NNHDI looks for the nearest subset of records which are most similar to the record with missing values, where nearness is specified according to the minimization of the distances between the former and the latter. With this aim, a general distance measure for the comparison of two records that share some, but not necessarily all, auxiliary variables has to be derived. Actually, this is only a part of the problem. Another is the fact that real-world data sets frequently involve a mixture of numerical, ordinal, binary, and nominal variables.

To deal with the simultaneous presence of variables with different measurement scales, we take our point of departure from the computation of a distance matrix which is restricted to nonmissing components for each type of variable. In this way, a compromise distance can be achieved using a combination of all the partial distances ("partial" because each of them is linked to a specific type of variable and not to the globality of the issues reported in the data set). We address the problem of specifying differential weights for each type of variable in order to reflect their significance, reliability, and statistical adequacy for the NNHDI procedure.

The remainder of the paper is organized as follows. The next section gives a brief overview of the NNHDI method. Here, we introduce an imputation technique in which the missing value of the target variable is replaced with the least power mean of the Box-Cox transformation applied to the observed values from selected donors. In Section 3, we give a description of the methodology used to compute distances with an emphasis on the measure of distance for mixed data. In Section 4, we devise a compromise distance between records. In Section 5, an application of the various systems of weighting is given, followed by an evaluation of our approach. Finally, in Section 6, we highlight some areas for future work.

2. Nearest Neighbor Hot Deck Imputation

Let S_n be a data set consisting of n records $R_1, R_2, \ldots, R_n$ in which an interval or ratio-scaled variable y (the target variable or variable of interest) is recorded together with other m auxiliary or matching variables $(X_1, X_2, \ldots, X_m)$. We assume that ν of the n records have a valid observation for the target variable forming the set S_ν of the first ν records of S_n. In practical applications, many factors influence which value is missing and which is not, so that *lacunae* in the data are usually not confined to particular fields but can be at any position within the record. As a consequence, one or more auxiliary variables may be missing although records which have a missing value for all the auxiliary variables are excluded from usable data.

For each receptor record $R_i = (y_i, x_{i,1}, x_{i,2}, \ldots, x_{i,m})$, $i > \nu$ and y_i missing, the NNHDI selects a neighborhood or reference set $\mathbf{J}_i = \{j_{1,i}, j_{2,i}, \ldots, j_{k,i}\} \subset S_\nu$ of k similar records or donors where $k \geq 1$ is the fixed size of all the reference sets. To be a donor, the record must have a valid value both for y and for at least one of the auxiliary variables fully present in the receptor.

The donors provide a basis for determining imputed values. If the cardinality of the ith reference set is $|\mathbf{J}_i| = 1$, then the value y_s of the nearest record R_s can simply be copied into R_i, or a transformation from the auxiliary variables in R_s can be applied to the correspondent data on R_i to determine the imputed value $\widehat{y}_i$ from y_s. Bankier et al. [3, 4] pointed out that the imputed values of a recipient record should come from a single record donor if possible rather than from two or more donors. Welniak and Coder [5] noted that if $k = 1$, then all missing information is imputed from the same donor so favoring the preservation of the interrelationships between variables. However, when the attributes include a large number of qualitative and quantitative variables at the same time or when there are auxiliary variables with many distinct values, it is extremely difficult to find a single donor record that precisely matches the recipient record.

The idea of the NNHDI method is that each receptor record is not an isolated case but belongs to a certain cluster and will therefore show a certain pattern. In fact, the NNHDI first collects records that are similar to the receptor by making use of the auxiliary variables and then integrates the data of alternative records into a consistent and logically related reference set. Hence, several donors may be involved in completing a single deficient record. Sande [6] observed that this may be a source of some concern, but such a worry diminishes if one takes into account the fact that the best donor for a segment of the recipient record may be different from the best donor for another segment when incompleteness also affects auxiliary variables. For example, the perfect neighborhood obtained using nominal variables might turn out to be an inadequate cluster for numerical variables. Wettschereck and Dietterich [7] noted that, in noisy data sets, the k-nearest neighbor algorithm performs better than simple nearest neighbor.

The possibility of reducing bias with the NNHDI may be reinforced if unreported values are characterized by a missing at random (MAR) mechanism. In the phraseology of this field, this means that missing values on the target variable follow a pattern that does not depend on the unreported data in y, but only on observed data. Let $\boldsymbol{\psi}$ be a vector of indicator variables for the ith record such that

$$\psi_i = \begin{cases} 1 & \text{if } y_i \text{ is observed,} \\ 0 & \text{if } y_i \text{ is missing,} \end{cases} \qquad i = 1, 2, \ldots, n. \tag{2.1}$$

Under an MAR dynamic, the selection probabilities verify the following condition: $\Pr(\boldsymbol{\psi} \mid \mathbf{y}_\nu, \mathbf{y}_{n-\nu}) = \Pr(\boldsymbol{\psi} \mid \mathbf{y}_\nu)$. This is equivalent to saying that, given the observed data, the inability to observe a realization from y is not a consequence of the data that are not observed. The missingness pattern, however, may depend on auxiliary variables that may be the reason for missingness or are joint causes and can thus contribute to filling the voids. In fact, the values observed for the auxiliary variables both for the donee and for the donors are compared under the tacit assumption that if distances, however defined, are small for the auxiliary variables, the records will also be close to one another for the target variable. Consequently, the existence of strong relationships between target and auxiliary variables has a positive impact on the ability of the NNHDI to determine more compact and homogeneous

reference sets which, as a result, increase the quality of the imputed values. See Abbate [8]. Unfortunately, the validation of the MAR assumption is difficult because there is not usually much information regarding the unobserved data. However, the more relevant and related the auxiliary variables are to the target variable, the more likely the MAR hypothesis is.

2.1. Formation of the Reference Sets

The NNHDI method is implemented as a two-stage procedure. In the first stage, the data set S_ν is searched to form the neighborhood or reference set $\mathbf{J}_i$ for each receptor in $S_{n-\nu}$. In the next stage, the values of y observed in the reference set are used to compute the replacement value. The reference set is built simultaneously for $R_{\nu+1}, R_{\nu+2}, \ldots, R_n$ following the rule: record $R_s \in S_\nu$ is added to $\mathbf{J}_i$ if $|\mathbf{J}_i| < k$ or if

$$\max_{j\in\mathbf{J}_i} \delta(\mathbf{x}_i, \mathbf{x}_j) \leq \delta(\mathbf{x}_i, \mathbf{x}_s); \quad i = \nu+1, \nu+2, \ldots, n;\ s = 1, 2, \ldots, n_\nu, \tag{2.2}$$

where $\delta(\cdot)$ is the distance between two records in terms of the auxiliary variables. At the end of the process, the records corresponding to the first k distances become the neighborhood $\mathbf{J}_i$ of R_i. The solutions $\mathbf{J}_i$, $i = \nu+1, \nu+2, \ldots, n$ form mathematical (nonrandom) samples of fixed size k. We admit that the $\mathbf{J}_i$s are artificial and that they may be nonrepresentative samples of the target variable population because each $\mathbf{J}_i$ is a subset of a subset of the observed records (to be precise, records similar to R_i with an effective value for y). It should be emphasized, however, that imputation of missing values is one of those circumstances in which a biased selection may be preferable to probability sampling. See Deming [9, pages 32-33].

A peculiar characteristic of k-NNHDI is the restriction to a predefined cardinality k of $\mathbf{J}_i$, $i = \nu+1, \nu+2, \ldots, n$ that cannot be changed later (see [10]). Such a constraint may cause perplexity since *a priori* there can be no firm experimental evidence that each receptor belongs to a compact and homogeneous cluster formed by at least k records which are to be validly employed as donors; in addition, the fact that NNHDI inexorably finds k donors, even if none of them is actually near to the receptor, increases the risk of irrelevant records in the reference set.

Although there are some rules which link k to the size n of the data set, the cardinality $|\mathbf{J}_i|$ remains somewhat arbitrary. The value of k should be kept small enough to improve the speed of the algorithm and to bring process values derived by the most similar records into the imputation. On the other hand, if k is very small and the donors are not found nearby due to data sparseness, the imputed value tends to be very poor. The robustness of the NNHDI to noisy data is expected to improve with the number of donors used for imputation, but too many donors increase the computational cost and may enlarge the variability of imputed values out of proportion. Moreover, as k increases, the mean distance between the receptor and the donors becomes larger, especially if a significant proportion of the data is missing. In the literature, only small k values, $(3, 5, 10, 15, 20)$ have been tested. Wettschereck and Dietterich [7] chose k on the basis of the performance of the imputation algorithm by trying all possible k over a vast range and broking ties in favor of the smaller value of k. Friedman et al. [11] found empirically that values ranging from $k = 8$ to $k = 16$ work well when searching for a nearest neighbor. To determine k in our experiments, we have used Sturge's rule $k = 1 + \log_2(n)$ discussed, for example, in Hyndman [12].

A computational drawback of the NNHDI is that the algorithm searches for donors through the entire data set, and this limitation can be serious in the case of large databases. To

form the neighborhoods $\mathbf{J}_i$, $i = n_\nu+1, n_\nu+2, \ldots, n$, the NNHDI considers $n_\nu \times (n-n_\nu)$ distances (although only a minority of them have to be kept in memory) and compares $\delta(\mathbf{X}_i, \mathbf{X}_s)$, $s \notin \mathbf{J}_i$ with the most distant element in each neighborhood. With the vast improvement in computers, NNHDI methods are not nearly as prohibitive as they used to be. Nevertheless, if the data set is judged to be too large to be treated within an acceptable time limit, the search for the donors can be confined to a subset of S_ν. See, for example, Wilson and Martinez [13].

The NNHDI does not necessarily produce distinct reference sets; $|\mathbf{J}_r \cap \mathbf{J}_s|$ may be greater than zero for $r \neq s$. Moreover, it leaves unused records that do not fit into any neighborhood. For this reason, one objection to the NNHDI is that data from some records could be used many times as donors and other records are excluded from "donation," thus depriving the imputation of the benefits which might have been derived using more information. According to Sande [6], this will increase the variance while possibly reducing the bias of the estimate. Furthermore, it may imply inflating the size of certain subpopulations in the data set. Kaiser [14] pointed out that the excessive use of a single donor leads to poor estimates, and Schieber [15] recommended that each complete record was only allowed to be a donor once. If repeated donations and omitted contributions are a problem to be alleviated, one may apply the strategy proposed by Colledge et al. [16] or Giles [17].

2.2. Imputing for Missing Data

Let $\mathbf{J}_i$ be the reference set of $R_i \in S_{n-\nu}$, $i = n_\nu + 1, n_\nu + 2, \ldots, n$ formed in the first stage. The information on y contained in $\mathbf{J}_i$ has to be synthesized into an estimate $\widehat{y}_i$ for the missing value. This operation should be carried out with care and control since imputed values will be treated as actually observed values, and statistical analysis is carried out using the standard procedures developed for data without any missing observations.

Since $|\mathbf{J}_i| > 1$, then a synthesis of all the evidence acquired is needed. Many proposals for estimating y_i have been advanced. A common imputation technique is the use of a mean $\widehat{y}_i$ of the values observed in the reference set. In this paper, attention is concentrated on the least power mean estimator that minimizes

$$S_\alpha(y) = \left[\sum_{j\in\mathbf{J}_i} |y_j - \widehat{y}_i|^\alpha\right]^{1/\alpha} \qquad \alpha \geq 0, \tag{2.3}$$

with respect to $\widehat{y}_i$. A classic choice is the simple mean imputation $\widehat{y}_i = E(y \mid j \in \mathbf{J}_i)$ obtained for $\alpha = 2$. If we set $\alpha = 1$, me $= [(k+1)/2]$, and me$' = k + 1 -$ me, we obtain the median imputation. For $\alpha \to \infty$, (2.3) yields the midrange. In our procedure, α is not fixed but must be optimized to fit the observed values of the target variable y_j for $j \in \mathbf{J}_i$. With this aim, we have used the procedure developed by Mineo and Ruggieri [18] (see also [19]) centered on the exponential power probability density function

$$f(y) = \frac{\exp\left\{-\left(|y-\mu_\alpha|^\alpha / \alpha\sigma_\alpha^\alpha\right)\right\}}{2\sigma_\alpha \alpha^{1/\alpha}\Gamma(1+1/\alpha)}, \quad \alpha > 0. \tag{2.4}$$

The symbols $\mu_\alpha, \sigma_\alpha$, and α denote, respectively, the location, scale, and shape parameters of the density. If $\alpha \geq 1$, the curves generated by (2.4) are "bell" shaped. In an attempt to enhance

the reliability of (2.4) for asymmetrical empirical distributions, we have applied the Box-Cox power transformation in order to achieve distributional symmetry of the target variable

$$y_j(\lambda) = \begin{cases} \dfrac{y_j^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \log(y_j) & \text{if } \lambda = 0, \end{cases} \quad y_j > 0,\ j \in \mathbf{J}_i. \tag{2.5}$$

The value of λ has been estimated minimizing the standardized third central moment within the compact interval $-3 \leq \lambda \leq 5$. See Taylor [20]. Once found the value of $\widehat{\lambda}$, model (2.4) is fitted to the target variable in the transformed scale $y_j(\lambda)$ taking into account the relationship between α in (2.3) and the tail behavior of the exponential power distribution. See Mineo and Ruggieri [18].

Missing values must be imputed on the original scale, while (2.5) provides a replacement value in the Box-Cox scale so that the estimate must be transformed back to the original. However, when transforming back, the least power mean undergoes a bias unless $\lambda = 1$ (because the transformation is linear) or unless $\alpha = 1$ because the median is invariant under monotone transformation. Based on preliminary work, not reported here, a heuristic bias-correction factor has been devised as follows:

$$\overline{y}_j = \widetilde{y}_j \left(\frac{\mathrm{GM}}{\mathrm{AM}} \right)^{\tau}, \tag{2.6}$$

where GM and AM are the geometric and the arithmetic mean of the target variable in the original scale and τ is a tuning constant. The factor in (2.6) is a monotone decreasing function of its exponent: the larger τ, is the smaller the factor is, and, hence, the greater the biascorrection is. In our experiments, we tried (2.6) with τ in a discrete interval (τ_1, τ_2) and selected the one whose relative mean error of imputation in the first S_ν records was the smallest.

3. Distance Measurement for Mixed Data

Let $\mathbf{X}$ be an object-by-variable data matrix containing measurements of n objects of a mixture of variables types. Without loss of generality, we may assume that m_1 variables are interval or ratio scaled; m_2 are ordinal variables which rank records in terms of degree without establishing the numeric difference between data points; m_3 variables are binary symmetric (0-0 and 1-1 matches are treated as equally indicative of similarity); m_4 are binary asymmetric (0-0 or 1-1 matches are not regarded as indicative of similarity since 1 is used to indicate the presence of some feature and 0 its absence). It is important to distinguish between the two situations: if two records have few copresences in a large number of binary variables which are considered symmetric, then the similarity between them may be judged as quite large even if they have very little in common. Conversely, if the variables are considered binary asymmetric, then a large number of coabsences could be judged insignificant. Finally, m_5 variables are nominal with three or more categories and a potentially different number of states l_h, $h = 1, 2, \ldots, m_5$. Of course, some of the groups may be empty, and some others may be split into groups of variables of the same type. In each case, $m = m_1 + \cdots + m_5$.

Let p be the number of nonempty subsets of variables. The measurement of the dissimilarity through a distance function that considers all types of variables can be achieved in many ways. To begin with, it is possible to perform a separate analysis for each group and then to compare and synthesize imputation results from alternative sources. A conflict may occasionally emerge because of irreconcilable differences between the patterns discovered in the different distance matrices. In real applications, it is unlikely that separate imputations will generate compatible results; furthermore, the cost of repeated analysis of large data sets may be too high.

The simplest way to deal with a mixture of variable types is to divide the variables into types and confine the analysis to the dominant type. Even though it would be easy to judge which type is "dominant," this practice cannot be recommended because it discards data that may be correct and relevant but is produced in the wrong scale. When simultaneously handling nominal, ordinal, binary, and so forth characteristics, one may be tempted to ignore their differences and use a distance measure which is suitable for quantitative variables but is inappropriate for other types. Naturally, this is an absurd solution, but in practice, it often works.

Another approach is to convert one type of variable to another, while retaining as much of the original information as possible, and, then, to use a distance function suitable for the selected type. Anderberg [21, page 94] argued that the primary question to be faced is which variable type should be chosen. For instance, nominal variables can be transformed into classes which are coded with 1s and 0s thus treating them as asymmetric binary variables along with the original binary variables; subsequently, the records which now only consist of numerical variables can be compared using traditional distance functions for quantitative variables (see [22, page 92]). An evident drawback is the use of a large number of binary variables that are highly interdependent because they imply a choice between mutually exclusive possibilities. Alternatively, quantitative variables could be dichotomized at a fixed level so that the new values can be treated using distance functions that are devised for binary variables. A consequence is that a large number of records will be considered alike, thus reducing the influence of the quantitative variables. In any event, conversion between scales involves a loss of information and knowledge.

3.1. General Distance Coefficient

The performance of the NNHDI method depends critically on the distance used to form the reference sets. A particularly promising approach consists of processing all the variables together and performing a single imputation procedure based on a coefficient of dissimilarity explicitly designed for mixed data. In this work, we have adopted the following measure of global distance:

$$\delta_{i,j} = \sum_{t=1}^{p} \left[h_{i,j}^{(t)} + \left(1 + h_{i,j}^{(t)}\right) \delta_{i,j}^{(t)} \right], \tag{3.1}$$

with

$$h_{i,j}^{(t)} = \frac{\sum_{s=M_{t-1}}^{M_t} h_{s,i,j}}{m_t}, \qquad M_t = \sum_{s=1}^{t} m_s, \qquad M_0 = 0, \tag{3.2}$$

where $\delta_{i,j}^{(t)}$ is the tth partial distance between the records R_i and R_j. Usually, the distances are scaled to vary in the unit interval

$$0 \leq \delta_{i,j}^{(t)} \leq 1, \quad t = 1, 2, \ldots, p, \tag{3.3}$$

where zero is achieved only when the two records are identical in all nonempty fields, and one is achieved when the two records differ maximally in all of the fields validly compared. Condition (3.3) is necessary, otherwise a combined representation of the distances would copy the structure of the indicator with the greatest distances.

Since donors may have missing values themselves, distances are, necessarily, computed for variables which have complete information for both records, while values contained in one record but missing from the other are ignored. The indicator $h_{s,i,j}$ in the expression (3.1) is zero if the comparison of R_i and R_j is valid with respect to the sth variable, whereas $h_{s,i,j} = 1$ if either of the fields is empty. Conventionally, we set $\delta_{i,j}^{(t)} = 1$ if $h_{i,j}^{(t)} = 1$. If $h_{i,j}^{(t)} = 0,\ t = 1, 2, \ldots, p$, then (3.1) becomes the all-purpose measure of dissimilarity proposed by Gower [23]. See also Kaufman and Rousseeuw [24, page 19], Di Ciaccio [25], Murthy et al. [26], and Seber [27, pages 357-358].

Imputation will fail if there is not at least one donor with a valid datum regarding at least one variable which is not missing in the receptor. In passing, we note that missing values in the auxiliary variables can reduce the neighborhood of donors to the point where imputation becomes impossible, at least for a subset of cases. See Enders [28, page 50]. In this situation, the receptor should perhaps be excluded from the set of usable records and treated using a different method.

The global distance (3.1) is in line with the principle that the reliability of a distance decreases with the reduction of meaningful comparisons. Consequently, records having less valid fields are penalized in order to compensate for their lower usability. This choice has in addition the desirable effect of deterring the selection of donors that share too few features with the receptor.

A limitation of (3.1) is that variables can substitute each other, that is, a higher distance in one variable can compensate for a lower value in another. The influence of X_s can be increased or decreased by rescaling its contribution proportionally to w_s and grading it in the range $[0, w_s]$. If the number of variables is high, however, a very complex process is needed to amalgamate all the partial distance matrices into a global matrix of distance. To keep computations to a manageable level, we decided to assign distinct weights to the groups of the variables, but not to each single variable

$$\mathbf{D} = \sum_{t=1}^{p} w_t \mathbf{D}_t \quad \text{with } w_t \geq 0;\ \sum_{t=1}^{p} w_t = 1, \tag{3.4}$$

where $\mathbf{D}_t = \mathbf{H}_t + (\mathbf{U} + \mathbf{H}_t) \odot \mathbf{\Delta}_t$, $\mathbf{\Delta}_t = \delta_{i,j}^{(t)}$. Here, $\mathbf{U}$ is a matrix of 1s, and $\odot$ indicates the Hadamard product between two matrices. The choice of (3.4) reduces the flexibility of the general dissimilarity coefficient, but the search for an optimal weighting system is simplified. It is important to emphasize here that if the partial distance matrices have a similar structure, they generate overlapping reference sets, and no advantage can be gained by combining them as (3.4).

Gower [23] requires that each $d_{i,j}^{(t)}$ is a dissimilarity coefficient which generates a Euclidean distance matrix $\boldsymbol{\Delta}_t$. Pavoine et al. [29] show that if the $\mathbf{D}_t$s are Euclidean, then $\mathbf{D}$ is also Euclidean. Nonetheless, the status of being a Euclidean matrix could be modified by (3.1). If $\mathbf{D}_t$ is not Euclidean, it is possible to determine constants ϕ_1 such that a matrix with elements $d_{r,s}^{(j)} + \phi_1$, $r \neq s$, is Euclidean. See Gower and Legendere [30, Theorem 7].

3.2. Distances Involved in the General Coefficient

In general, it is not possible to design a coefficient that can be recommended given a particular type of variable because, besides the intrinsic properties of the coefficients, the number and the values of auxiliary variables have a determinant role. For the present paper, we have selected some commonly used distance functions which have a range of $[0, 1]$, irrespective of the number of variables so that distances are unaffected by the number of fields.

(1) Ratio and Interval Scale. Euclidean distance

$$d_{i,j}^{(1)} = \sqrt{(m_1)^{-1} \sum_{\substack{h_{s,i,j}=0,\\ s\in[M_0,M_1]}} \left(\frac{x_{i,s} - x_{j,s}}{r_s}\right)^2}, \tag{3.5}$$

where r_s is the observed range of the sth variable.

(2) Ordinal Scale. Linear disagreement index

$$d_{i,j}^{(2)} = \sqrt{(m_2)^{-1} \sum_{\substack{h_{s,i,j}=0,\\ s\in[M_1+1,M_2]}} \left[\frac{x_{i,s} - x_{j,s}}{r_s - 1}\right]^2}, \tag{3.6}$$

where $r_h = \max\{\mathbf{X}_h\}$. The values of the hth ordinal variables are integers in $[1, r_h]$.

(3) Binary Symmetric. Number of such variables in which records have different values divided by the total number of binary symmetric variables:

$$d_{i,j}^{(3)} = \sqrt{\frac{\sum_{\substack{h_{s,i,j}=0,\\ s\in[M_2+1,M_3]}} \delta(x_{i,s}, x_{j,s})}{m_3}}, \quad \delta(x_{i,s}, x_{j,s}) = \begin{cases} 1 & \text{if } x_{i,s} \neq x_{j,s}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7}$$

(4) Binary Asymmetric. Number of binary asymmetrical variables in which both records have a positive value to the total number of these variables

$$d_{i,j}^{(4)} = \sqrt{\frac{\sum_{\substack{h_{s,i,j}=0,\\ s\in[M_3+1,M_4]}} \delta(x_{i,s}, x_{j,s})}{m_4}}, \quad \delta(x_{i,s}, x_{j,s}) = 1 - \min\{x_{i,s}, x_{j,s}\}. \tag{3.8}$$

(5) Nominal. Number of states of the polytomies in which the two records under comparison have the same state, divided by the total number of states across all the politomies

$$d_{i,j}^{(5)} = \sqrt{\frac{\sum_{\substack{h_{s,i,j}=0,\\ s\in[M_4+1,M_5]}} \delta\left(x_{i,s}, x_{j,s}\right)}{m_5}}, \quad \delta\left(x_{i,s}, x_{j,s}\right) = \begin{cases} \dfrac{l_s}{\lambda_5} & \text{if } x_{i,s} \neq x_{j,s}, \\ 0 & \text{otherwise,} \end{cases} \tag{3.9}$$

where $\lambda_5 = \sum_{s=M_4+1}^{M_5} l_s$. Each comparison can be scored as λ_5 different dichotomies by setting l_s of these to 1 when R_i and R_j coincide on the sth politomy or to 0 when R_i and R_j are different.

All the proposed indices generate a Euclidean distance matrix.

4. Combining Distances

To use the combined distance function (3.4), a user must supply proper weights for the various types of variables. Bankier et al. [31] hypothesized that weights should be smaller for variables when it is considered less important that they match or variables considered more likely to be in error or to be affected by missingness. Istat [32] determined, with respect to the MAR dynamic of missing values, the weights according to the degree of association between the target and the auxiliary variables in complete records. For the moment, we ignore these weighting schemes although they should be considered seriously in future research.

4.1. Equal and Proportional Weighting

The weights w_t, $t = 1, 2, \ldots, p$ could be determined on the basis of an *a priori* judgment of what is important and what should be prioritized with regard to the partial distance matrices. In other words, investigators give weights to types based on an intuitive understanding of the data, but if they do not know the context well, the assessment may be inadequate and introduce bias. Chiodi [33] found equal weighting to be more valuable for his data: $w_t = 1/p$, $t = 1, 2, \ldots, p$. This formula regards all the types of variables to be equally effective when determining the global distance matrix. Such a solution may be perfectly acceptable given that we rarely know *a priori* if some types are more helpful than others. In fact, equal weighting appears to be a valid practice when there are no theoretical or empirical grounds for choosing a different scheme. However, it is inadvisable to equate the weight of the variables without further studying their contribution to the variability of the data set as a whole.

Romesburg [34] based the weights on the proportion of each type of variable: $w_t = m_t/m$, $t = 1, 2, \ldots, p$. If all the auxiliary variables were assumed to have equal efficacy, independently of the scale on which they are measured, then this option would be the right choice. An obvious example is when all the auxiliary variables are strongly associated with the target variables. The original version of the Gower's coefficient is a weighted average of three different measures of dissimilarity where the weights are the proportions of each type of variable.

4.2. Equalizing Standard Deviations of Partial Distances

The significance of a type of variable in determining the global distance depends, among other things, on the variability of $d_{i,j}^{(t)}$, $t = 1, 2, \ldots, p$ so that types leading to a high variance of pairwise distances will, thus, be more likely to have great influence into the global distance. To ensure the respect for this principle, we can use the reciprocal of a measure of "variability for the partial distances." For example,

$$w_t = \begin{cases} \dfrac{1/\left(\sigma\left[d_{i,j}^{(t)}\right]\right)}{\sum_{r=1}^{p}\left[1/\left(\sigma\left[d_{i,j}^{(t)}\right]\right)\right]} & \text{if } \sigma\left[d_{i,j}^{(t)}\right] > 0, \\ 0 & \text{otherwise,} \end{cases} \qquad t = 1, 2, \ldots, p, \tag{4.1}$$

where $\sigma[d_{i,j}^{(t)}]$ is the standard deviation of the distances in the strictly lower triangular part of $\mathbf{D}_t$. If $\mathbf{w}$ is based on (4.1), then the elements of the lower triangular part of $\mathbf{D}_t$ are normalized to have a standard deviation of one.

4.3. Equalizing Mean of Partial Distances

Clear-cut variables such as binary and categorical variables tend to have more of an impact on the calculation of the global distance. Kagie et al. [35] observe that there is no reason to assume, without reference to specific aspects of the problem at hand, that nominal variables are more important than quantitative ones. Therefore, an adaptation is necessary. Following Kagie et al. [35] and Lee et al. [36], the distances in the strictly lower triangular part of $\mathbf{D}_t$ can be normalized to have an average value of one

$$w_t = \frac{1/\left(\mu\left[d_{i,j}^{(t)}\right]\right)}{\sum_{r=1}^{p}\left[1/\left(\mu\left[d_{i,j}^{(t)}\right]\right)\right]}, \quad t = 1, 2, \ldots, p, \tag{4.2}$$

where $\mu[d_{i,j}^{(t)}]$ is the average distance in the strictly lower triangular part of the tth partial distance matrix.

4.4. Distatis Weighting

To obtain an optimal system of weights, we need an expression for how much a certain type of variable affects the global distance. This can be derived from the total sum of the squares of the elements in the partial distance matrices $\mathbf{D}_t$, $t = 1, 2, \ldots, p$ choosing the weights so as to maximize the variance of the elements in the global distance matrix $\mathbf{D}$ which, in turn, leads to the Distatis procedure as developed by Abdi et al. [37] (see also [38]).

In the first step of Distatis, each $\mathbf{D}_t$ is transformed into the cross-product matrix $\mathbf{S}_t$ which has the same information content as $\mathbf{D}_t$ but is more suitable to the eigen-decomposition. Let $\mathbf{u}_n$ be an $n \times 1$ vector of 1s, and let $\mathbf{I}_n$ be the identity matrix of order n. A normalized cross-product matrix for $\mathbf{D}_t$ is

$$\mathbf{B}_t = \left(\frac{1}{\lambda_{1,t}}\right)\mathbf{S}_t, \quad \mathbf{S}_t = -0.5\mathbf{C}\mathbf{D}_t^2\mathbf{C}^t, \quad \mathbf{C} = \mathbf{I}_n - n^{-1}\mathbf{u}_n\mathbf{u}_n^t, \tag{4.3}$$

where $\mathbf{D}_t^2$ is the matrix whose (i,j)th element is the square of $d_{i,j}^{(t)}$, and $\lambda_{1,t} > 0$ indicates the largest eigenvalue of $\mathbf{S}_t$. It is well known (see, e.g., [39]) that $\mathbf{S}_t$ is symmetric, positive semidefinite and has zero row sums.

Let $\boldsymbol{\beta}_t = \text{Vec}(\mathbf{B}_t)$, $t = 1, 2, \ldots, p$ be the column vector obtained by stacking the column of $\mathbf{B}_t$ on top of one another and organizing them into an $n^2 \times p$ matrix $\mathbf{Z} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_p]$. The central step of Distatis is to create the matrix $\mathbf{A} = (\mathbf{N}^{-0.5}\mathbf{Z})^t(\mathbf{Z}\mathbf{N}^{-0.5})$, where $\mathbf{N}^{-0.5}$ is a diagonal matrix whose elements are the square roots of the reciprocal values in the diagonal of $(\mathbf{Z}^t\mathbf{Z})$. The generic element $a_{r,s}$ of $\mathbf{A}$ is the vectorial correlation coefficient [40] between the cross-product matrix derived from the partial distance matrices $\mathbf{D}_r$ and $\mathbf{D}_s$, $r, s = 1, 2, \ldots, p$,

$$a_{r,s} = \frac{\boldsymbol{\beta}_r^t \boldsymbol{\beta}_s}{\|\boldsymbol{\beta}_r\| \|\boldsymbol{\beta}_s\|}. \tag{4.4}$$

Naturally, $a_{r,s} = a_{s,r}$ and $a_{r,r} = 1$, $r = 1, 2, \ldots, p$. Since (4.4) verifies the relationship $\boldsymbol{\beta}_r^t \boldsymbol{\beta}_s = \text{Trace}(\mathbf{B}_r^t \mathbf{B}_s)$ and since $\mathbf{B}_r$ and $\mathbf{B}_s$ are symmetric and positive semidefinite, then $\text{Trace}(\mathbf{B}_r^t \mathbf{B}_s) \geq \lambda_{n,r}\text{Trace}(\mathbf{B}_s)$, where $\lambda_{n,r}$ is the smallest eigenvalue of $\mathbf{B}_r$ (see [41]); it follows that $0 \leq a_{r,s} \leq 1$.

The scope of Distatis is to find a convex linear combination $\boldsymbol{\beta} = \text{Vec}(\mathbf{D})$ of the vectors in $\mathbf{Z}$ that explains the maximum amount of variance of $\mathbf{Z}$ possible. In this sense, Distatis is simply the principal component analysis of $\mathbf{A}$, that is, $\mathbf{A} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^t$ with $\mathbf{Q}^t\mathbf{Q} = \mathbf{I}_p$. Since $\mathbf{A}$ is positive or, at least, nonnegative irreducible, then the Perron-Frobenius theorem (see, e.g., [42]) ensures that there is a single eigenvalue, say θ_1, that is, positive and greater than or equal to all other eigenvalues in modulus and that there is a strictly positive eigenvector $\mathbf{q}$ corresponding to θ_1. The global cross-product matrix can now be found using

$$\mathbf{B} = \sum_{t=1}^{p} w_t \mathbf{B}_t, \quad \text{where } \mathbf{w} = \left(\mathbf{u}_p^t \mathbf{q}\right)^{-1} \mathbf{q}. \tag{4.5}$$

The weights $\mathbf{w}$ are such that the types of variables which are more similar to the others get higher weights than those which disagree with most of the rest of the variables. The related global distance matrix is

$$\mathbf{D} = \sum_{t=1}^{p} w_t \mathbf{D}_t. \tag{4.6}$$

Distatis weights are demanding from a computational point of view because they involve an eigenanalysis of potentially very large matrices. (Only if $P = 2$ are the weights (4.6) easy to compute: 0.5 and 0.5). However, the computational task can be simplified by performing the one-time determination of the weighs on a sufficiently large random sample of complete records.

5. Experimental Results

In this section, we present results of a numerical experiment to test how well the NNHDI reconstructs missing values. More specifically, we examine the five different weighting methods discussed in Section 4 to obtain a global distance matrix in connection with the imputing of the least power mean of donors. It is of particular interest to determine which

weighting scheme for partial matrices would have the smallest detrimental effect on the accuracy of data when using imputed values. In this sense, we have used data sets with variables which are completely known for all units (where necessary, we have removed incomplete records). We then simulated missing values according to a MAR mechanism to evaluate the bias of our NNHDI algorithm. All simulations were performed using *R* [43].

5.1. Description of Data Sets Used in the Experiments

Our experiments were carried out using ten data sets to represent different sized files that are commonly encountered. These data sets are interesting because they exhibit a wide variety of characteristics and have a good mix of attributes: continuous, binary, and nominal. In our study, we have employed actual data sets because of their realism and the ease with which they could be adapted to the experimental setting. In some cases, we have not used the entire data set, but a subset obtained by drawing without replacement a random sample from the data set.

(a) *Heart Disease Cleveland Database (UC-Irvine Repository) Frank and Asuncion [44].* The publicly available Cleveland heart-disease database was used to test the kernel methods. It consists of 303 cases where the disorder is one of four types of heart-disease or its absence. This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence from absence of disease. The variables summarizing the medical symptoms considered as important indicators of a patient's condition were classified as metric: $(1, 2, 8, 10)$, ordinal: $(11, 12)$, binary symmetric: $(2, 6)$, binary asymmetric: $(9, 14)$, and nominal: $(3, 11, 12)$. The "goal" field or outcome refers to the presence of heart disease in the patient. The problem is to predict the outcome given the other 13 values of the characteristics. To make things more apt to our purposes, we changed the problem slightly and used the serum cholesterol in mg/dL as the target variable. The original data set contains 270 complete records, but we have analyzed a random subset of $n = 30$ records.

(b) *Car Data Set (Package MASS of R).* Data from $n = 93$ cars on sale in the USA in 1993. Cars were selected at random from among 1993 passenger car models that were listed in both the consumer reports issue and the PACE buying guide. Pickup trucks and sport/utility vehicles were eliminated due to incomplete information in the consumer reports source. Duplicate models (e.g., Dodge Shadow and Plymouth Sundance) were listed at most once. The variables have been classified as metric: $(4, 6 : 8, 12 : 15, 17, 19 : 23)$, ordinal: $(9, 11, 18)$, binary symmetric: (16), binary asymmetric: (24), and nominal: $(1 : 3, 10, 25)$; the target variable is y = maximum price. Further description can be found in Lock [45].

(c) *Demand for Medical Care (NMES 1988, Package AER of R).* Cross-section data originating from the US National Medical Expenditure Survey (NMES) conducted in 1987 and 1988 to provide a comprehensive picture of how Americans use and pay for health services. The NMES is based upon a representative, national probability sample of the civilian population and individuals admitted to long-term care facilities during 1987. The data refer to individuals of ages 66 and over, all of whom are covered by Medicare (a public insurance program providing

substantial protection against healthcare cost). These data were verified by cross-checking information provided by survey respondents with providers of healthcare services. In addition to healthcare data, NMES provides information on health status, employment, sociodemographic characteristics, and economic status. The version of the data set used in our experiment contains $m = 19$ variables which we have classified as follows: metric: $(1 : 6, 11, 15)$, ordinal: $(7, 8)$, binary symmetric: $(9, 13, 14)$, binary asymmetric: $(12, 17 : 19)$, and nominal: $(7, 8)$. The income of the individual was chosen as the target variable y. To keep the volume of data within reasonable limits, we have randomly selected $n = 150$ records. Details are given by Cameron and Trivedi [46].

(d) *Fatalities, in Package AER of R.* US traffic fatalities panel data annually from 1982 to 1988 for the 48 contiguous states (excluding Alaska, Hawaii, and the District of Columbia). Researchers using this data set investigate the impact of beer taxes and a variety of alcohol-control policies on motor vehicle fatality rates. Special attention is paid to bias resulting from failing to adequately control for grassroots efforts to reduce drunk driving, the enactment of other laws which simultaneously operate to reduce highway fatalities, and the economic conditions existing at the time the legislation is passed. Among the most interesting characteristics of traffic fatalities found in the data set are the drinking age variable, a factor indicating whether the legal drinking age is 18, 19, or 20; two binary punishment variables which describe the state's minimum sentencing requirements for an initial drunk driving conviction; per capita death rates from accidents occurring at night and for 18 to 20 year olds. The data set comprises 336 observations on $m = 34$ variables. However, record 28 has been excluded because of a lack of valid observations in two fields. All in all $n = 335$ records without missing values are considered. The classification of the variables is metric: $(3 : 9, 11 : 13, 17 : 25, 27 : 34)$, ordinal: $(10, 32, 33)$, binary symmetric: $(2, 11, 13)$, binary asymmetric: $(14 : 16)$, and nominal: $(1, 2)$. The number of alcohol-involved vehicle fatalities is used as target variable. See Stock and Watson [47] for more details.

(e) *Australian Credit Data Set (UC-Irvine Repository) Frank and Asuncion [44].* This file concerns credit card applications. Each case concerns an application for credit card facilities described by 16 attributes. Metric: $(3, 8, 11, 14, 15)$, ordinal: $(2, 9, 10)$, binary symmetric: $(1, 16)$, binary asymmetric: $(9, 10, 12)$, and nominal: $(4 : 7, 13)$. The first metric variable was selected as the target variable. The data frame contains 690 observations but 37 cases (5%) have one or more missing values, and they were consequently removed from further analysis. Hence, the effective number of records is $n = 653$. This data set is considered both scanty and noisy and, probably because of this, has become very popular in the testing algorithms for building classification trees. From our point of view, this dataset is interesting because there is a good mix of attributes: continuous, nominal with small numbers of values, and nominal with larger numbers of values. There are also a few missing values. More information can be found in Quinlan [48].

5.2. Experiment: Random Percentage of Missing Values in All the Variables

In a first phase, we omitted a fraction $\pi_{x_j} \in \{10, 20\}$, $j = 1, 2, \ldots, m$ ensuring that no row $\mathbf{x}$ contains more than $(m - 2)$ incomplete fields. The π_{x_j} records in which the value of the

auxiliary variables has to be removed were chosen by drawing a simple random sample without replacement from the first $\lfloor n/2 \rfloor$ records, once the data set had been sorted in ascending order of y. This step was repeated for each auxiliary variable. In this way, a mild association could be observed between $\mathbf{y}$ and $\mathbf{X}$ because the largest half of y could rely on a richer source of information in terms of auxiliary variables. In the second phase, we omitted a proportion $\pi_y \in \{15, 25, 40\}$ of the target variable, after randomly permuting the auxiliary variables. Records were resorted in ascending order of X_1 allocating the missing values of X_1 in the last positions. Records with a missing y were chosen to be a random sample without replacement of size $\lfloor n\pi_y/m \rfloor + b_j$, where

$$b_j = \begin{cases} 1 & \text{if } \lfloor n\pi_y \rfloor \bmod m \neq 0, \\ 0 & \text{otherwise,} \end{cases} \qquad j = 1, 2, \ldots, m. \tag{5.1}$$

The omission continued until the total number of missing values is reached. The sample of y was drawn with the Midzuno's method. The factors of proportionality were i^{-1}, $i = 1, 2, \ldots, n$ which are inversely related to the order of appearance of the records in the current arrangement of the data set. Consequently, the values of the target variable are more likely to be missing for cases with low values of the auxiliary variable. The above process is iterated for a number of auxiliary variables $X_2, X_3, \ldots$, sufficient to complete the number of missing values to be injected into the target variable. By operating in this manner, the missing values in y have a probabilistic link with the maximum possible number of auxiliary variables. Furthermore, the simultaneous presence of missing values in $\mathbf{y}$ and $\mathbf{X}$ is deterred. The double relationship between y and $\mathbf{x}$ makes the existence of an MAR mechanism in the simulated patterns of missingness plausible. Each combination of $\pi_y \times \pi_{x_i}$ has been repeated $N = 100$ times to reduce irregular variations. The efficiency of the reconstruction process of the NNHDI algorithms is measured by the mean relative error (MRE)

$$E_2^{(q)} = (n - \nu)^{-1} \sum_{i=\nu+1}^{n} \left| \frac{y_i - \widehat{y}_i^{(q)}}{y_i} \right| \quad \text{for } y_i \neq 0,\ q = 1, 2, \ldots, 5, \tag{5.2}$$

where $\widehat{y}_i^{(q)}$ is the value obtained from one of the five NNHDI procedures in which the value of the target variable was considered missing. The index $E_2^{(q)}$ is related to the size of the discrepancies between predicted and observed values. Clearly, a small value of $E_2^{(q)}$ represents a successful method. The results are reported in Table 1.

No strong evidence has been found in the experiment in favor of or against one of the five weighting systems of the partial matrices discussed in Section 4 when both the target and the auxiliary variables are affected by missingness. This implies that practices of using aprioristic system of weights to combine the partial distance matrices do not, apparently, penalize the accuracy of imputed values too greatly, especially for larger data sets. One potential reason for this finding is that the impact of each single variable on the global distance is rather small compared with the effect of the other variables in the same category or with the effect of a dominant group. In these cases, the flexibility in the choice of weights is not very helpful in comparing two records, and Monte Carlo experiments confirm that no algorithm is significantly more efficient than others under these data conditions. However, in measuring the overall quality of missing value estimates produced by a given NNHDI

Table 1: Missing values in all the variables: mean relative error.

Data set	τ	Weights	T15A10	T25A10	T40A10	T15A20	T25A20	T40A20
Hearts	8	Distatis	14.98	15.91	16.26	13.87	16.05	17.66
		Proportional	14.85	17.12	16.66	13.40	17.13	17.81
		Uniform	14.23	16.37	15.38	14.75	16.47	18.41
		Eq. μ	15.50	17.28	16.20	13.62	17.12	17.75
		Eq. σ	14.14	14.82	15.98	11.52	14.31	16.14
Cars	12	Distatis	24.20	24.69	26.27	39.70	37.31	37.40
		Proportional	24.84	24.40	25.45	31.63	33.61	33.03
		Uniform	23.28	25.59	25.86	39.84	37.46	37.12
		Eq. μ	25.53	24.86	26.78	38.40	37.19	36.82
		Eq. σ	38.41	36.55	37.03	39.51	40.00	37.22
Medicare	7	Distatis	64.95	71.32	71.75	71.15	69.09	71.53
		Proportional	65.23	68.03	70.44	66.74	65.20	66.27
		Uniform	68.23	72.10	74.59	71.63	70.11	75.81
		Eq. μ	66.71	71.98	74.24	71.81	69.40	75.24
		Eq. σ	68.60	71.67	73.25	73.22	69.22	69.97
Fatalities	20	Distatis	108.02	123.86	122.08	145.22	142.89	128.83
		Proportional	161.37	170.39	156.02	171.66	168.77	152.75
		Uniform	122.32	133.57	131.89	151.60	144.03	132.22
		Eq. μ	133.87	145.28	139.10	153.47	148.96	136.95
		Eq. σ	87.40	83.19	80.90	95.88	94.80	91.38
Credit appr.	5	Distatis	25.42	25.96	26.38	23.5	23.63	24.59
		Proportional	26.36	26.62	27.08	23.19	23.87	24.39
		Uniform	25.46	25.82	26.26	23.27	23.65	24.68
		Eq. μ	25.68	25.96	26.25	23.18	23.57	24.65
		Eq. σ	25.76	25.78	26.28	23.39	23.81	24.78

variation, the weighting proportional to the number of variables in the groups and the equalization of the mean partial distance performed slightly better than the other NNHDI algorithms.

In general, one would expect that with the increase in the proportion of incomplete records, or in the number of missing values in a record, or both, the quality of estimates would diminish due to the reduction of potentially useful information. Indeed, this seems to be confirmed by the high values of $E_2^{(q)}$. Nonetheless, the values of the index for experiments in which the percentage of missing values in the $\mathbf{x}$ is 20% are not much higher, and in many cases are lower, than for experiments in which the percentage is 10%.

Table 2 shows the average of $E_2^{(q)}$ over all weighting systems for the nine combinations of missing percentages in the target and in the auxiliary variables. The trends of the total mean relative error are not in line with what would be a logical response to the reduced quantity of information, that is, effectively handled. This can be explained in a number of ways. It is plausible that the impact of the missing values in the auxiliary variables is reduced because they are less present in records with a missing value in y. It is also plausible that the $y, \mathbf{x}$ interactions have a solid MAR mechanism so that the increase in the percentage of missing values in $\mathbf{x}$ has moderate detrimental effects on the accuracy of the reconstruction process; on the other hand, high percentages of missing values in y and $\mathbf{x}$ may even strengthen the

Table 2: Missing values in all the variables: mean relative error across the weightings.

Data set	T15A00	T15A10	T15A20	T25A00	T25A10	T25A20	T40A00	T40A10	T40A20
Hearts	8.57	14.74	13.43	8.42	16.30	16.22	8.47	16.10	17.55
Cars	15.65	27.25	37.82	16.22	27.22	37.11	19.26	28.28	36.32
Medicare	72.53	66.74	70.91	71.78	71.02	68.60	70.83	72.85	71.76
Fatalities	36.84	122.60	143.57	39.61	131.26	139.89	53.17	126.00	128.43
Credit approvals	25.77	25.74	23.31	25.82	26.03	23.71	26.44	26.45	24.62

cohesion of the available values in the records concerned. In the light of our experiment, the proportion of missing data does not appear to be the major determinant of imputation errors.

6. Conclusions and Suggestions for Future Research

Missing values often occur in real-world applications and represent a significant challenge in the field of data quality, particularly when the data set variables have mixed types. In this paper, we have conducted an extensive study of the nearest neighbor hot deck imputation (NNHDI) methodology where, for each recipient record with incomplete data for the target variables, a set of donors is selected so that the donors are similar to their recipient with regards to the auxiliary variable. The known values of the donors are then used to derive a value for the missing data by computing the least power mean (together with the Box-Cox transformation) of the target variable in the set of donors.

The particular focus of this paper is on "problematic" data sets containing missing values both in the target and the auxiliary variables and involving a mixture of numeric, ordinal, binary, and nominal variables. It has become increasingly apparent that efficacy and effectiveness of the NNHDI hinges crucially on how the distance between records is measured and different ways of measuring distances lead to different solutions. In this work, we have devised a new global distance function based on the partial distance matrices obtained from the various types of variables appearing in (or missing from) the records of the data set. The separate distance matrices are combined as a weighted average, and the resulting global distance matrix is then used in the search for donors. The contribution of each group of variables to the global distance is scaled with a weight which contracts/expands the influence of the group. In this study, we have compared the performance of five weighting schemes.

To judge the accuracy of the reconstruction process, we have considered a performance indicator which is related to the size of the discrepancies between predicted and observed values. More specifically, the relative absolute mean error was calculated for each method based on five real data sets in three different experiments: leaving one out, incompleteness in the target variables, and incompleteness in all variables. The missing values in the last two experiments were inserted according to an MAR mechanism.

The empirical findings suggest that data-driven weights for the partial distance matrices are moderately preferable to aprioristic weights although the reasons are more theoretical than objective, as the experiments presented in this work give little evidence in support of a specific weighting system. This is mostly due to the strong relationships among the variables of the tested data sets; also, the low variability showed by the least power mean used for imputing missing values might have given a nonmarginal contribution to the insufficient degree of discernability of the weighting systems. On the other hand,

the investigations carried out using the NNHDI demonstrate the ability of this method to compensate for missing values when several type of variables occur in the same data set, even if some of the records have *lacunae* in the auxiliary variables. The key benefit of the NNHDI combined with the new global distance and the least power mean estimator is that the good results achievable in terms of low reconstruction error should not be compensated by strong distributional assumptions or sophisticated modeling.

Our results have also shown that the choice of weights does not significantly affect the quality of imputed values, but we cannot exclude that alternative weights might achieve a superior performance. Perhaps, it is not insignificant that the highest impact of the missing values was in data sets where the metric variables were overrepresented. We plan to study weighting systems based on the degree of interdependency between the target and the auxiliary variables, so we can understand better the implications of such differences for diverse NNHDI algorithms.

The quality of the results of the NNHDI is closely related to the specific observations to be analyzed. Therefore, instead of using a fixed value of k over the entire data set, a locally adaptive choice of the cardinality of the reference sets may be more useful in practice. Alternatively, the value of k can be determined by performing a preliminary analysis of some subsets of the data set at hand. An evaluation of alternative strategies of constructing the reference sets might offer a promising line for further research into the NNHDI.

References

[1] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, Wiley Series in Probability and Statistics, John Wiley & Sons, New York, NY, USA, 2nd edition, 2002.

[2] G. Kalton and D. Kasprzyk, "Imputing for missing survey responses," in *Proceedings of the Section on Survey Research Methods*, pp. 22–31, 1982.

[3] M. Bankier, J. M. Fillion, M. Luc, and C. Nadeau, "Imputing numeric and qualitative variables simultaneously," in *Proceedings of the Section on Survey Research Methods*, pp. 242–247, American Statistical Association, 1994.

[4] M. Bankier, M. Luc, C. Nadeau, and P. Newcombe, "Additional details on imputing numeric and qualitative variables simultaneously," in *Proceedings of the Section on Survey Research Methods*, pp. 287–292, American Statistical Association, 1995.

[5] E. J. Welniak and J. F. Coder, "A measure of the bias in the march CPS earning imputation system and results of a simple bias adjustment procedure," Tech. Rep., U.S. Census Bureau, 1980.

[6] I. G. Sande, "Imputation in surveys: coping with reality," *The American Statistician*, vol. 36, pp. 145–152, 1982.

[7] D. Wettschereck and T. G. Dietterich, "An experimental comparison of the nearest-neighbor and nearest-hyperrectangle algorithms," *Machine Learning*, vol. 19, no. 1, pp. 5–27, 1995.

[8] C. Abbate, "La completezza delle informazioni e l'imputazione da donatore con distanza mista minima," *Quaderni di Ricerca dell'ISTAT*, vol. 4, pp. 68–102, 1997.

[9] W. E. Deming, *Sample Design in Business Research*, A Wiley Publication in Applied Statistics, John Wiley & Sons, New York, 1960.

[10] A. K. Ghosh, "On nearest neighbor classification using adaptive choice of k," *Journal of Computational and Graphical Statistics*, vol. 16, no. 2, pp. 482–502, 2007.

[11] J. H. Friedman, J. L. Bentley, and R. A. Finkel, "An algorithm for finding best matches in logarithmic expected time," *Association for Computing Machinery Transactions on Mathematical Software*, vol. 3, pp. 209–226, 1977.

[12] R. J. Hyndman, "The problem with Sturges' rule for constructing histograms," Business, July, 1-2, 1995.

[13] D. R. Wilson and T. R. Martinez, "Reduction techniques for instance-based learning algorithms," *Machine Learning*, vol. 38, no. 3, pp. 257–286, 2000.

[14] J. Kaiser, "The effectiveness of hot-deck procedures in small samples," in *Proceedings of the Annual Meeting of the American Statistical Association Javaid Kaiser*, University of Kansas Kalton G.,

Compensating for Missing Survey Data. Ann Arbor, MI: Survey Research Center, University of Michigan, 1983.

[15] S. J. Schieber, "A comparison of three alternative techniques for allocating unreported social security income on the survey of the low-income aged and disabled," in *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1978.

[16] M. J. Colledge, J. H. Johnson, R. Pare, and I. J. Sande, "Large scale imputation of survey data," in *Proceedings of the Section on Survey Research Methods*, pp. 431–436, American Statistical Association, 1978.

[17] P. Giles, "A model for generalized edit and imputation of survey data," *The Canadian Journal of Statistics*, vol. 16, pp. 57–73, 1988.

[18] A. M. Mineo and M. Ruggieri, "A software tool for the exponential power distribution: the normalp package," *Journal of Statistical Software*, vol. 12, pp. 1–24, 2005.

[19] P. Jönsson and C. Wohlin, "Benchmarking k-nearest neighbour imputation with homogeneous Likert data," *Empirical Software Engineering*, vol. 11, no. 3, pp. 463–489, 2006.

[20] J. M. G. Taylor, "Power transformations to symmetry," *Biometrika*, vol. 72, no. 1, pp. 145–152, 1985.

[21] M. R. Anderberg, *Cluster Analysis for Applications*, Academic Press, New York, NY, USA, 1973.

[22] I. E. Franck and R. Todeschini, *The Data Analysis Handbook*, Elsevier, Amsterdam, The Netherlands, 1994.

[23] J. C. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, vol. 27, pp. 623–637, 1971.

[24] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, New York, NY, USA, 1990.

[25] A. Di Ciaccio, "Simultaneous clustering of qualitative and quantitative with missing observations," *Statistica Applicata*, vol. 4, pp. 599–609, 1992.

[26] M. N. Murthy, E. Chacko, R. Penny, and M. Hossain, "Multivariate nearest neighbour imputation," *Journal of Statistics in Transition*, vol. 6, pp. 55–66, 2003.

[27] G. A. F. Seber, *Multivariate Observations*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons, New York, NY, USA, 2004.

[28] C. K. Enders, *Applied Missing Data Analysis*, The Guilford Press, New York, NY, USA, 2010.

[29] S. Pavoine, J. Vallet, A. B. Dufour, S. Gachet, and H. Daniel, "On the challenge of treating various types of variables: application for improving the measurement of functional diversity," *Oikos*, vol. 118, no. 3, pp. 391–402, 2009.

[30] J. C. Gower and P. Legendre, "Metric and Euclidean properties of dissimilarity coefficients," *Journal of Classification*, vol. 3, no. 1, pp. 5–48, 1986.

[31] M. Bankier, M. Lachance, and P. Poirier, "2001 Canadian census minimum change donor imputation methodology," in *Proceedings of the Work Session on Statistical Data Editing, (UN-ECE)*, Cardiff, Wales, 2000.

[32] Istat, *CONCORD V. 1.0: Controllo e Correzione dei Dati. Manuale Utente e Aspetti Metodologici*, Istituto Nazionale di Statistica, Roma, Italy, 2004.

[33] M. Chiodi, "A partition type method for clustering mixed data," *Rivista di Statistica Applicata*, vol. 2, pp. 135–147, 1990.

[34] H. C. Romesburg, *Cluster Analysis for Researchers*, Krieger Publishing, Malabar, Fla, USA, 1984.

[35] M. Kagie, M. van Wezel, and P. J. F. Groenen, "A graphical shopping interface based on product attributes," *Decision Support Systems*, vol. 46, no. 1, pp. 265–276, 2008.

[36] R. C. T. Lee, J. R. Slagle, and C. T. Mong, "Towards automatic auditing of records," *IEEE Transactions on Software Engineering*, vol. 4, no. 5, pp. 441–448, 1978.

[37] H. Abdi, A. J. O'Toole, D. Valentin, and B. Edelman, "DISTATIS: the analysis of multiple distance matrices," in *Proceedings of the the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 42–47, San Diego, Calif, USA, 2005.

[38] P. D'Urso and M. Vichi, "Dissimilarities between trajectories of a three-way longitudinal data set," in *Advances in Data Science and Classification*, A. Rizzi, M. Vichi, and H.-H. Bock, Eds., pp. 585–592, Springer, Berlin, Germany, 1998.

[39] C. J. Albers, F. Critchley, and J. C. Gower, "Group average representations in Euclidean distance cones," in *Selected Contributions in Data Analysis and Classification*, P. Brito, P. Bertrand, G. Cucumel, and F. de Carvalho, Eds., Studies in Classification, Data Analysis, and Knowledge Organization, pp. 445–454, Springer, Berlin, Germany, 2007.

[40] Y. Escoufier, "Le traitement des variables vectorielles," *Biometrics*, vol. 29, pp. 751–760, 1973.
[41] Y. G. Fang, K. A. Loparo, and X. Feng, "Inequalities for the trace of matrix product," *IEEE Transactions on Automatic Control*, vol. 39, no. 12, pp. 2489–2490, 1994.
[42] K. Y. Lin, "An elementary proof of the Perron-Frobenius theorem for non-negative symmetric matrices," *Chinese Journal of Phisics*, vol. 15, pp. 283–285, 1977.
[43] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2009.
[44] A. Frank and A. Asuncion, *UCI Machine Learning Repository*, University of California, School of Information and Computer Sciences, Irvine, Calif, USA, 2010, http://archive.ics.uci.edu/ml.
[45] R. H. Lock, "New car data," *Journal of Statistics Education*, vol. 1, no. 1, 1993, http://www.amstat.org/publications/jse/v1n1/datasets.lock.html.
[46] A. C. Cameron and P. K. Trivedi, *Regression Analysis of Count Data*, vol. 30 of *Econometric Society Monographs*, Cambridge University Press, Cambridge, Mass, USA, 1998.
[47] J. H. Stock and M. W. Watson, *Introduction to Econometrics*, Addison Wesley, Boston, Mass, USA, 2nd edition, 2007.
[48] J. R. Quinlan, "Simplifying decision trees," *International Journal of Man-Machine Studies*, vol. 27, no. 3, pp. 221–234, 1987.