UNIVERSITÀ DELLA CALABRIA

Dipartimento di Economia e Statistica
Ponte Pietro Bucci, Cubo 0/C
87036 Arcavacata di Rende (Cosenza)
Italy
http://www.ecostat.unical.it/

Working Paper n. 15 - 2010

MISSING-VALUES ADJUSTMENT FOR MIXED-TYPE DATA

Agostino Tarsitano
Dipartimento di Economia e Statistica
Università della Calabria
Ponte Pietro Bucci, Cubo 0/C
Tel.: +39 0984 492465
Fax: +39 0984 492421
e-mail: agotar@unical.it

Marianna Falcone

Collaboratrice di ricerca

e-mail: maryfalcon@libero.it

Agosto 2010

Missing-values adjustment for mixed-type data[☆]

Agostino Tarsitano[*,a], Marianna Falcone

[a] *Dipartimento di Economia e Statistica - Università della Calabria*
Via Pietro Bucci, cubo 1C, 87036 Rende (CS) - Italy

Abstract

In this paper we propose a new method of single imputation, reconstruction, and estimation of non-reported, incorrect or excluded values both in the target and in the auxiliary variables where the first is on ratio or interval scale and the last are heterogeneous in measurement scale. 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 partial distance matrices computed for the various types of variables. In particular, we address the problem of proper weighting 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. We have demonstrated, through analysis of simulated and actual data sets, the appropriateness of this approach. Our main contribution has been to show that mixed data may optimally be combined to allow accurate reconstruction of missing values in the target variable even in the absence of some data in the other fields of the record.

Key words: hot-deck imputation, nearest neighbor, general distance coefficient, least power mean.

[☆]This document is a report of the results of research that have emerged during the preparation of the thesis of M. Falcone.

[*]Corresponding author

Email addresses: agotar@unical.it (Agostino Tarsitano), maryfalcon@libero.it (Marianna Falcone)

1. Introduction

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

There are various strategies to handle the problems posed by the missing observations. These include: additional data collection; application of likelihood-based procedure that allows to model incomplete data; deductive reconstruction; using only part of the available data; weighting records; revise the data set in an attempt to replace the missing data with plausible values (imputation). The present paper concerns the latter method.

Imputation techniques have been extensively studied during the last few decades and a number of approaches have been proposed. For an overview of the methods, see for example Little & Rubin, (2002) and Kalton & Kasprzyk, (1982). Some of these methods are nowadays available in standard statistical software (or could easily be implemented), although there is little consensus as to the most appropriate technique to use for a particular situation. In the present paper, we did not perform an exhaustive review of data imputation methods, but instead discuss only the nearest neighbor hot deck imputation (NNHDI) that has been used for a number years and enjoys a high prestige for both theoretical and computational work.

NNHDI is a based on a non-random sample without replacement from the current data set (it is due to this the term "hot deck" in the name of the method). More specifically, NNHDI looks for the nearest subset of records most similar to the records having missing values, where neareness is specified in terms of minimizing the distances between the formers and the latters. To this end, a general distance measure for the comparison of two records that share some, but not necessarily all, the 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 numeric, ordinal, binary and categorical 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 restricted to non-missing components for each type of variable:

binary (symmetrical and asymmetrical), categorical, ordinal, interval/ratio. Then a compromise distance can be achieved by using a convex combination of all the partial distances ("partial" because each of them is linked to a specific type of variables and not to the globality of the issues reported in the records). 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 nearest neighbor hot deck imputation method. 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. The objective is to analyze the performance of the NNHDI as an imputation method. Experiments performed on real data sets demonstrate the ability of the proposed method to compensate for missing values when several types of variables occur in the same data set. 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 $(X_1, X_2, \cdots, X_m)$. Without loss of generality, 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 the data set S_n. In practical applications, many factors influence which value is missing and which is not, so that the lacunae in the data are usually not confined to particular fields but can be in any position within the record. As a consequence, one or more auxiliary variables may be missing, although we have excluded from the set of usable data, the records that have a missing value for all the auxiliary variables.

For each donee record or receptor $R_i = (y_i, x_{i,1}, x_{i,2}, \ldots, x_{i,m})$, $i > \nu$ and y_i missing, we select a pool (or reference set) $\mathbf{J}_i = \{j_{1,i}, j_{2,i}, \ldots, j_{k,i}\} \subset S_\nu$ of k similar records ("donors") where k 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.

Bankier *et al.*, (1994), Bankier *et al.*, (1995) pointed out that the imputed values of a recipient record should come from a single record donor if possible rather than two or more donors. Welniak & Coder, (1980) noted that, if $k = 1$ then all missing information is imputed from the same donor favoring the preservation of the interrelationships between variables. As a confirmation of this strategy, Jöreskog & Sörbom, (1993) introduced a technique identical to the NNHDI attempting to impute values missing in one case from another case with similar observed values (if it exists) and doing so using a minimization criterion on a set of matching variables.

It must be said that when the attributes characterizing the records include a large number of qualitative, ordinal and quantitative variables, at the same time or when there are auxiliary variables with many distinct values, it is extremely difficult to obtain a perfect retrieval *i.e.* to find a single donor record that precisely matches the recipient record in any field, particularly if there can be the impact of a missing value in more than one variables in the same record. In contrast, the idea of the NNHDI algorithm is that, to predict whether a record will show a certain type of pattern, implies an assumption that the intended recipient is not a singleton, but belongs to a certain type of cluster and will therefore show the certain kind of pattern. In facts, NNHDI first collects records similar to the receptor making use of additional information provided by the auxiliary variables and then integrate 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, (1982) suggested that this may be a source of some concern, but one must take into account the fact that the best donor for a segment of the record could be different from the best donor for another segment when incompleteness also affects auxiliary variables. The NNHDI method, however, does not use an explicit model relating y and $\mathbf{x}$ and, hence, it is expected to be more robust against model violations than methods based on explicit models, such as ratio imputation and regression imputation. Chen & Shao, (2000).

The possibility of reducing bias by NNHDI may be improved if unreported values are characterized by a missing at random (MAR) mechanism. In the phraseology of the field this means that missing values on the target variable follow a pattern that does not depend on the unreported data in the target variable, but only on observed data. 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 fill 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, they will also be small 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, (1997).

Let $\boldsymbol{\psi}$ a vector of missingness indicator for y

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

Under a MAR dynamic, the selection probabilities verify the condition

$$Pr\left(\boldsymbol{\psi}|\mathbf{y}_\nu, \mathbf{y}_{n-\nu}\right) = Pr\left(\boldsymbol{\psi}|\mathbf{y}_\nu\right). \tag{2}$$

This is equivalent to saying that, given the observed data, the inability to observe a realization from y does not depend on the data that are not observed. Unfortunately, validation of MAR assumption is difficult because there is usually a scarce amount of information on the unobserved data. However, the more relevant and related are the auxiliary variables with the target variable or with the propensity to give a y response, the more likely is the MAR hypothesis.

2.1. Formation of the reference sets

In general, NNHDI 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 $R_{\nu+1}, R_{\nu+2}, \ldots, R_n$ according to the rule: record $R_s \in S_\nu$ is added to $\mathbf{J}_i$ if $|\mathbf{J}_i| < k$ or if

$$\max_{j \in \mathbf{J}_i} \delta\left(\mathbf{x}_i, \mathbf{x}_j\right) \leq \delta\left(\mathbf{x}_i, \mathbf{x}_s\right); \; i = \nu+1, \nu+2, \ldots, n; \; s = 1, 2, \ldots, n_\nu \tag{3}$$

where $\delta(.)$ is the distance between two records in the space of 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 (non-random) samples of fixed size k. We admit

that the $\mathbf{J}_i$'s are artificial and that they may be non-representative samples of the target variable population because each $\mathbf{J}_i$ is a subset of a subset of the observed records (specifically those having an effective value for y). It must 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, (1960)[pp. 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 corrected and changed later (see, in this respect, Ghosh, (2007) and references therein); this gives a certain impression since there is no guarantee that each receptor belongs to a compact and homogeneous cluster formed by at least k records to be validly employed as donors; in addition, the fact that NNHDI inexorably finds k donors even if none of them is actually near the receptor, inflates the risk of irrelevant records in the reference set.

Although there are some rules which link k to the number n of records of the data set, the size of the reference set remains somewhat arbitrary. The value of k should be kept small enough to improve the speed of the imputation process and to bring into the imputation process values derived by the most similar records. On the other hand, if k is very small and the donors are not nearby due to data sparseness, the imputed value tends to be very poor. The robustness of 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 out of proportion the variability of imputed values. Moreover, as k increases, the mean distance between the receptor and the donors gets larger. Eventually, as k approaches $n - \nu$, the NNHDI algorithm converges to ordinary mean imputation where units far from the receptor are forced to be equally informative as the nearest records. In this situation, it is likely that implausible imputed values will result.

Literature have tested only small k values $(3, 5, 10, 15, 20)$; after all, single imputation is very much alike the estimation of a univariate mean, which does not require large samples. In Little & Rubin, (2002) is suggested that $k = 3$ to $k = 5$ will suffice. Wettschereck & Dietterich, (1995) chose k to optimize the leave-one-out cross-validation performance of the imputation algorithm (see Section 5.2) by trying all possible of k in a vast range of values and broking ties in favor of the smaller value of k. Friedman *et al.*, (1977) found empirically that values ranging from $k = 8$ to $k = 16$ work well for nearest neighbor searching. To determine k in our experiments, we have used the Sturge's rule $k = 1 + \log_2(n)$ discussed, for example, in Hyndman, (1995).

The main computational drawback of the NNHDI approach is that the algorithm searches the donors through all the data set. To form the pools $\mathbf{J}_i, i = n_\nu+1, n_\nu+2, \ldots, n$, (3) it considers $n_\nu \times (n - n_\nu)$ distances (although, only a part of them have to be kept in memory) and compares $\delta(\mathbf{X}_i, \mathbf{X}_s)$ with the largest element in each pool. 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 too large to be treated within an acceptable time limit, the donors can be searched in a subset of S_ν. Several works that aim to solve this limitation can be found in the literature, *e.g.* Wilson & Martinez, (2000).

NNHDI does not necessarily produce disjoint reference sets: $|\mathbf{J}_r \cap \mathbf{J}_s| > 0$ for $r \neq s$. Moreover, it leaves unused records that do not fit in any neighborhood. To this purpose, an objection to NNHDI is that data from some records could be used many times as donors and other records excluded from "donation" thus depriving the imputation of the benefits that could have been derived there from. According to Sande, (1982) this will increase the variance while possibly reducing the bias of the estimate. Also, this may imply inflating the size of certain subpopulations in the data set. Kaiser, (1983) pointed out that the excessive use of a single donor results in poor estimates; Schieber, (1978) recommended that each complete record was allowed to be a donor only once. If repeated donations and omitted contributions are considered a problem to be alleviated, one can apply the strategy proposed by Colledge *et al.*, (1978) or Giles, (1988).

2.2. Imputing for missing data

Once the reference sets of $R_i \in S_{n-\nu}, i = n_\nu + 1, n_\nu + 2, \ldots, n$ have been formed, the information on y contained in the pool of donors $\mathbf{J}_i$ has to be synthesized into an estimate for the missing value of the target variable $\hat{y}_i$. The operation consists of replacing the missing value y_i in the donee R_i by $\hat{y}_i$ derived from the values of y observed in $\mathbf{J}_i$. These operations must be repeated for each receptor.

Imputation of missing values should be done with care and control since imputed values will be treated as really observed values. If the cardinality $|\mathbf{J}_i| = 1$, then the value y_s of the nearest neighbor 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 $\hat{y}_i$ from y_s. The same procedure can be applied to the missing values in all the auxiliary

variables of the recipient record, provided that they are measured on an interval or ratio scale; for qualitative and binary attributes, the value in the nearest neighbor may be imputed; for ordinal variables *ad hoc* procedures should be applied.

Manzari, (2004) suggested that the recipient record should closely resemble that single donor, but equally good imputation actions, based on available donors, should have similar chances of being selected. One of those imputations actions is randomly selected by giving a better chance to records that are simultaneously closer to a receptor and have a lower number of missing fields. Many other authors (see, for example, Andridge & Little, (2010), Siddique & Belin, (2008)) suggest performing a random draw from the reference set. An advantage of this option is that the imputed value would reflect, at least partially, sampling and imputation uncertainty about the actual datum. On the other hand, random imputation requires the specification of an appropriate probability model even in situations in which the features detectable from what is known of the data do not legitimate any specific distributional assumptions to assist in making imputations.

If $k_i > 1$ then a synthesis of all the evidence acquired on y through the reference set $\mathbf{J}_i$ is needed. A wide variety of estimation methods have been discussed in the literature. The methods differed from one another in the way donor records are employed. A common imputation technique is using a mean $\hat{y}_i$ of the values observed in the reference set. In this paper attention is concentrated on the L_α (least power mean) estimator that minimizes

$$S_\alpha = \sum_{j \in \mathbf{J}_i} |y_j - \hat{y}_i|^\alpha \qquad \alpha \geq 0 \tag{4}$$

with respect to $\hat{y}_i$. See Pennecchi & Callegaro, (2006). The simple mean imputation $\hat{y}_i = E(y|\mathbf{J}_i)$ is obtained for $\alpha = 2$. If we set $\alpha = 1$, $me = [(k_i + 1)/2]$ and $me' = k_i + 1 - me$ we obtain the median imputation:

$$\hat{y}_i = \frac{y_{me} + y_{me'}}{2}. \tag{5}$$

If $\alpha = 1$ and k_i even, the solution of (4) is not unique, but uniqueness is attained using (5). For $\alpha \to \infty$, (4) yields the midrange

$$\hat{y}_i = \frac{y_{\min} + y_{\max}}{2}. \tag{6}$$

For $\alpha \to 0$ we find a reference to the zero norm and $\hat{y}_i$ is the mode (not necessarily unique) of the target variable. It can be easily seen that (4) does not necessarily coincide with one of the values of y observed for the donors. However, by construction, we have

$$\min_{j \in \mathbf{J}_i} \{y_j\} \leq \hat{y}_i \leq \max_{j \in \mathbf{J}_i} \{y_j\} . \tag{7}$$

It follows that, even though $\hat{y}_i$ is a fictitious value, it is not too far from actually observed values. It must be noted that Andridge & Little, (2010) do not consider methods that impute summaries of values for a set of donors, as hot deck methods, although they share some common features. **?**, (**?**) noted that, for point estimation, single random imputation is less efficient than conditional mean imputation because the random imputation mechanism introduces extra noise, although not in the extent to which it corrects the downward bias in variability estimation due to single imputed values.

In our procedure, α is not fixed but must be optimized to fit the observed values of the target variable. To this end we have used the procedure devised by Mineo & Ruggeri, (2005) (see also Jönnson & Wohlin, (2006)) that takes into account the relationship between α and the tail behavior of the exponential power probability density function

$$f(y) = \frac{\exp\left\{-\dfrac{|y-\mu_\alpha|^\alpha}{\alpha\sigma_\alpha^\alpha}\right\}}{2\sigma_\alpha \alpha^{1/\alpha}\Gamma(1+1/\alpha)} \quad \alpha > 0 \tag{8}$$

fitted to y_j for $j \in \mathbf{J}_i$. The symbols $\mu_\alpha, \sigma_\alpha, \alpha$ denote, respectively, the location, scale and shape parameters of the model. The idea underlying (8) is that the actual probability density function of y in the reference set is, at least approximately, symmetrical. If $\alpha \geq 1$ the densities proposed by (8) are "bell"-shaped (leptokurtic if $0 < \alpha < 2$ and platikurtic if $\alpha > 2$). A variable that has this type of distribution has a little or moderate amount of variability, which suggests a relatively homogeneous cluster of donors. Antimodal densities (that is, "U"-shaped), are obtained for $0 < \alpha < 1$. In this case, the variable has a very large amount of variability that could be attributable to two distinct groups of donors. The range of shapes produced in HDNNI should include J-shaped distributions (L-shaped distributions) in which most of the variability could be attributable to many potential donors being almost certain of having (not having) their value selected, while a few have much lower (higher) probabilities. These shapes, however, are not in the

family of curves generated from (8) so that there could be a systematic error in the exponential power approximation when an asymmetric distribution is approximated by a symmetric one.

In an attempt to improve the reliability of our imputation algorithms, we have applied more versatile models such as the beta and the two-sided power distributions that offer simple and flexible curves and enable us to obtain a reasonable fitting in the presence of a broad range of nonnormalities. The results, however, were discouraging in the sense that the accuracy of the imputed values was lower and the computer time required greater than with the exponential power distribution. A partial explanation of this could be that $\mathbf{J}_i$ is a non-probabilistic sample formed using the subset of records most similar to the receptor R_i; as such, their distribution, is not seriously affected by outliers or marked heaviness of the tails.

The index of tail behavior recommended by Mineo & Ruggeri, (2005) is

$$VI_\alpha = \frac{E|y-\mu|}{\sqrt{E|y-\mu|^2}} = \frac{\left[\Gamma\left(1/\alpha\right)\Gamma\left(3/\alpha\right)\right]^{0.5}}{\Gamma\left(2/\alpha\right)}. \tag{9}$$

The *normalp* is an R (R Development Core Team, (2009)) package containing a collection of tools related to the exponential power distribution. In particular a command that estimates the shape parameter α by means of the index of kurtosis (9) and a command that estimates the location parameter $\mu_\alpha = \hat{y}_i$ for fixed α. The procedure is sufficiently efficient and fast to be applied for small, medium and large data sets.

Once a record has been reconstructed, it can be used to repair other records. For example, the target variable can exchange its role with an auxiliary variable and HDNNI can iteratively complete the missing fields of all records. This is a controversial subject. On one side is the fact that using an imputed value as if it were an actual value increases the degree of intervention; in addition, increasing the set of complete records sequentially implies that the construction of the reference set is now dependent on a properly ordering of the variables and on the order in which complete and incomplete records are processed. On the other side is the reiteration of the algorithm which allows to exploit more efficiently the observed data. In this paper, we have chosen to neglect the reuse of imputed records, but further studies should investigate benefits and drawbacks of this option.

It is worth observing that imputation based approaches for handling missing data are frequently implemented in concert with editing rules, that is,

logical or consistency bounds for the missing values that must be incorporated in the imputation process. In the interest of simplicity, our analysis is confined to analytical interrelationships among the variables ignoring edits and consistency checking. Of course, one must be aware that data repairement cannot prevent the investigators from false conclusions due to badly designed instruments, poor data collection, and elusive populations. The analyst has to keep in mind that, although the sample size is increased, imputation does not add new genuine information about the data set.

3. Distance measurement for mixed data

Let $\mathbf{X}$ be an object-by-variable data matrix containing measurements of n objects on 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 that 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 the two situations: if two records have few co-presences in a great number of binary variables considered symmetric then the similarity between them may be judged quite large even if they have very little in common. Conversely, if the variables are considered binary asymmetric, a large number of co-absences could be judged insignificant. Finally, m_5 variables are categorical with three or more categories with potentially different numbers 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, we have $m = m_1 + \ldots + m_5$.

Let p be the number of non empty subsets of variables. The measurement of the dissimilarity through a distance function that depends on the scale of each group of variables can be realized in several ways. To begin with, it is possible to perform a separate distance analysis for each group and then to compare and synthesize imputation results from different sources. A conflict may occasionally emerge because of irreconcilable differences between the patterns discovered in the various distance matrices. The perfect neighborhood obtained by using multistate variables might turn out to be an inadequate cluster for numerical variables. In real applications, it is unlikely that a separate NNHDI will generate compatible results. In this context,

the question arises whether we can avoid the conflict by limiting the scope of the research (see Anderberg, (1973)[pp. 93-94]). Furthermore, the cost of repeated analysis of large data sets could be too high. Colledge *et al.*, (1978) conducted imputation in several phases each corresponding to a group of variables. Each phase replaced missing values in a particular group by the values of donors for which the variables in that group were observed. This method assumes implicitly that the group of variables is conditionally independent, given the matching variables (Little & Rubin, (2002)[p. 70]) and no interactions among the variables of different groups are covered.

The simplest way to deal with a mixture of variable types is to partition it into types and confine the analysis to the dominant type. Even if 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 produced in the wrong scale.

When simultaneously handling nominal, ordinal, binary, *etc.* characteristics, one may be tempted to ignore their difference and use a distance measure suitable for quantitative variables, but incorrect for the other types. Naturally, this is an absurd solution, but in fact it often works.

Another approach is to convert one type of variables to the other, while retaining as much of the original information as possible, and then use a distance function defined on the final type. In a similar vein, Anderberg, (1973)[p. 94] observed that the primary question to be faced is which variable type should be chosen as the single type of the analysis. For instance, multistate variables can be transformed into classes coded with 1's and 0's thus treating them as asymmetric binary variables along with the original binary variables; then, the records now consisting of only numeric variables can be compared using traditional distance functions for quantitative variables (see Franck & Todeschini, (1994)[p. 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 devised for symmetric binary variables. A consequent of this option is that a large number of records will be considered as alike and, hence, the less influence the quantitative variables are likely to have on the distance function. In any events, conversion between scales comports loss of information and knowledge.

3.1. General distance coefficient

The performance of NNHDI critically depend on the distance/dissimilarity used to form the reference sets. A reasonable approach is to process all variable together and perform a single imputation procedure using a coefficient of dissimilarity explicitly designed for mixed data. In this work we have opted for the following measure of global distance:

$$\delta_{i,j} = \sum_{t=1}^{p} \left[h_{i,j}^{(t)} + \left(1 + h_{i,j}^{(t)}\right) \delta_{i,j}^{(t)} \right] \tag{10}$$

with

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

where $\delta_{i,j}^{(t)}$ is the t-partial distance between R_i and R_j. The indicator $h_{s,i,j}$ is zero if the comparison of R_i and R_j is valid with respect of the s-th variable; $h_{s,i,j} = 1$ otherwise. Conventionally, we set $\delta_{i,j}^{(t)} = 1$ if $h_{i,j}^{(t)} = 1$. The transformation in (10) has the merit of defining a distance even when none of the variables of the group is observed for both records. Imputation will fail if there are no donor records with complete data on one or more of the variables on which the receptor record is complete. Consequently, missing values in the auxiliary variables can reduce the pool of donors to the point where imputation becomes impossible at least for a subset of cases Enders, (2010)[p. 50]. In this situation, the receptor should perhaps be excluded from the set of usable records and treated with a different method. Also, $h_{s,i,j} = 1$ if $x_{i,s}$ and/or $x_{j,s}$, for a variety of reason, cannot have a real value such as the number of pregnancy complications for a male respondent in a survey, the number of years since quit smoking for non-former smokers or the leaf size for lichens. Distance computation can then be performed reliably without a possibility that a distance between exactly similar patient cases would not be zero. If $h_{i,j}^{(t)} = 0, t = 1, 2, \ldots, p$ then (10) reduces to the all-purpose measure of dissimilarity proposed by Gower, (1971). See also Kaufman & Rousseeuw, (1990)[p. 19], Di Ciaccio, (1992), Murthy *et al.*, (2003), Seber, (2004)[pp. 357-358].

Usually, the distances in the right-hand side of (10) have the common thread of being scaled to vary in the unit interval:

$$0 \leq \delta_{i,j}^{(t)} \leq 1, \; t = 1, 2, \ldots, p \tag{11}$$

where zero is achieved only when the two records are identical in all non empty fields and one when the two records maximally differ in all fields validly compared. Condition (11) is necessary, otherwise the combined representation would copy the structure of the indicator with the largest distances.

Since donors may have missing values themselves, distances are, of necessity, computed over shared variables and values contained in one record but missing in the other are ignored. Formula (10) is in line with the principle that the reliability of a distance decreases with the reduction of meaningful comparisons. To illustrate the concept, suppose that $x_{i,s} = x_{j,s}, h_{s,i,j} = 0, s = M_{t-1}, \ldots, M_t$ then $\delta_{i,j}^{(t)} = 0$. The contribution of the t-th group to the global distance is the fraction of valid comparisons: $0 \leq h_{i,j}^{(t)} \leq 1$; conversely, if $\delta_{i,j}^{(t)} = 1$ then the contribution becomes $1 \leq 1 + 2h_{i,j}^{(t)} \leq 3$ which increases with the number of missing values in one or in both the records. Thus, records having less valid fields are penalized in order to compensate for their lower usability. This choice has the desirable effect of restraining selection of donors that share too few features with the receptor.

A limitation of (10) is that variables can substitute each other, that is, a higher distance on one variable can compensate for a lower value on another. The influence of X_s can be increased or decreased by rescaling its contribution with 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 in a global matrix of distance. To keep computations at a manageable level, we assign differential weights to the group of the variables, but not to each single variable

$$\mathbf{D} = \sum_{t=1}^{p} w_t \mathbf{D}_t \qquad \text{with } w_t \geq 0; \quad \sum_{t=1}^{p} w_t = 1 \tag{12}$$

with $\mathbf{D}_t = \mathbf{H}_t + (\mathbf{U} + \mathbf{H}_t) \odot \mathbf{\Delta}_t$, $\mathbf{\Delta}_t = \delta_{i,j}^{(t)}$. Here, $\mathbf{U}$ is a matrix of 1's and $\odot$ indicates the Hadamard product between two matrices. The choice (12) reduces the flexibility of the general dissimilarity coefficient, but the search of an optimal weighting system is simplified.

According to Gower, (1971), each $d_{i,j}^{(t)}$ should be a dissimilarity coefficient that generates an Euclidean distance matrix $\mathbf{\Delta}_t$. Pavoine *et al.*, (2009) show that if the $\mathbf{D}_t$'s are Euclidean, then $\mathbf{D}$ is also Euclidean. Nonetheless, the status of being a Euclidean matrix for $\mathbf{D}_t$ could be modified by the transformation (10). If $\mathbf{D}_t$ is not Euclidean, it is possible to determine constants ϕ_1

and ϕ_2 such that a matrix with elements

$$a) \quad \text{Lingoes Condition} \quad \left[\left(d_{i,j}^{(t)}\right)^2 + 2\phi_1\right]^{0.5} \quad i \neq j \tag{13}$$

$$b) \quad \text{Caillez Condition} \quad d_{i,j}^{(t)} + \phi_2 \quad i \neq j \tag{14}$$

is Euclidean (see Gower & Legendre, (1986)[theorem 7]). In the present paper, we have used the second option.

3.2. Distances involved in the general coefficient

For the present paper, we have selected some of commonly used distance functions that have a range of $[0, 1]$, irrespective of the number of variables so that the distances are unaffected by the number of fields

3.2.1. Ratio and interval scale

Euclidean distance.

$$d_{i,j}^{(1)} = \sqrt{\sum_{h_s,i,j=0} \left(\frac{x_{i,s} - x_{j,s}}{m_1 r_s}\right)^2} \tag{15}$$

where r_h is the observed range of the h-th variable. The difference in standardized values may be viewed as the fractional distance relative to the maximum possible distance between two records. This transformation effectively provides more weight to those attributes with a smaller range. Gower & Legendre, (1986) ensure that (15) forms a Euclidean distance matrix.

The choice (15) does not make any direct use of correlation between variables; thus two highly correlated variables will contribute twice as much to the total distance as they should. To overcome this problem one can perform principal component analysis to reduce the m_1 variables to $m_1' < m_1$ uncorrelated factors. Actually, most of the multivariate techniques use the Mahalanobis distance because it takes into account the variability of the values in all dimensions, though this would require, among other things, a number of donors much more greater than m_1.

3.2.2. Ordinal scale

Linear disagreement index.

$$d_{i,j}^{(2)} = \sqrt{\sum_{h=M_1+1}^{M_2} \left[\frac{x_{i,m_1+h} - x_{j,m_1+h}}{m_2 \left(r_h - 1 \right)} \right]^2} \tag{16}$$

where $r_h = \max \{\mathbf{X}_h\}$. The values of the h-th ordinal variables are integers in $[1, r_h]$. Since (16) is just (15) applied to ranks, we can infer that it too generates a Euclidean matrix.

3.2.3. Binary symmetric

The dissimilarity for binary symmetrical variables is defined as the number of such variables on which records have different values divided by the total number of binary variables (simple matching coefficient).

$$d_{i,j}^{(3)} = \sqrt{\frac{\sum_{h=M_2+1}^{M_3} \delta \left(x_{i,h}, x_{j,h} \right)}{m_3}}; \quad \delta \left(x_{i,h}, x_{j,h} \right) = \begin{cases} 1 & \text{if } x_{i,h} \neq x_{j,h} \\ 0 & \text{otherwise} \end{cases} . \tag{17}$$

It ranges in value between zero, when the two records match on any of the M_3 variables of the group to unity, when they fail to match on every variable.

3.2.4. Binary asymmetric

The dissimilarity is measured by the ratio between the number of binary asymmetrical variables in which both records have a positive value to the number of this type of variables.

$$d_{i,j}^{(4)} = \sqrt{\frac{\sum_{h=M_3+1}^{M_4} \delta \left(x_{i,h}, x_{j,h} \right)}{m_4}}; \quad \delta \left(x_{i,h}, x_{j,h} \right) = 1 - \min \left\{ x_{i,h}, x_{j,h} \right\} . \tag{18}$$

This index, known as Russell-Rao dissimilarity measure, is simply the fraction of variables in which both records had the trait of interest. Gower & Legendre, (1986) established that (17) and (18) form Euclidean distance matrices.

3.2.5. Categorical

The distance is given by the sum of the number of states of the politomies in which the two records under comparison have the same state, divided by the total number of states across all the politomies (McQuitty coefficient, Bijnen, (1973), pp. 13-14).

$$d_{i,j}^{(5)} = \sqrt{\frac{\sum_{h=M_4+1}^{M_5} \delta\left(x_{i,h}, x_{j,h}\right)}{m_5}}; \quad \delta\left(x_{i,h}, x_{j,h}\right) = \begin{cases} \dfrac{l_h}{\lambda_5} & \text{if } x_{i,h} \neq x_{j,h} \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

where $\lambda_5 = \sum_{h=M_4+1}^{M_5} l_h$. Each comparison can be scored as λ_5 different dichotomies by setting l_h of these to 1 when R_i and R_j coincide on the h-th politomy or to 0 when R_i and R_j are different. Then $d_{i,j}^{(5)}$ generates an Euclidean distance matrix (see Gower, (1971)).

4. Combining individual distances

To use the combined distance function (12) a user must supply proper weights for the various type of variables. Bankier *et al.*, (2000) hypothesized that the weights should be smaller for variables where it is considered less important that they match or variables considered more likely to be in error or to be affected by missingness. Istat, (2004), in respect of the MAR dynamic of missing values, determined the weights according to the degree of association between the target and the auxiliary variables in the complete records and, consequently, a heavier weight is put on variables more strictly linked with the former For the moment we ignore this weighting scheme, although it is one that must be considered seriously in future research.

4.1. Equal and proportional weighting

The weights $w_h, h = 1, 2, \ldots, p$ could be determined on the basis of an a priori judgment of what is important and what should be prioritized in regard to the partial distance matrices. In other words, investigators give weights to groups based on an intuitive understanding of the data, but if they do not know the context well, their assessment may be inadequate and will introduce biases. Chiodi, (1990) found equal weighting to be more valuable for his data.

$$w_t = \frac{1}{p}, \quad t = 1\,2, \ldots, p. \tag{20}$$

Formula (20) regards all types of variables to be equally important when determining the global distance matrix. This solution may be a perfectly reasonable solution given that we rarely know *a priori* if some types are more important 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 understanding their contribution to the variability of the data set as a whole.

Romesburg, (1984) based the fraction of each type of variable:

$$w_t = \frac{m_t}{m}, \quad t = 1\,2, \ldots, p. \tag{21}$$

If all the auxiliary variables were assumed to be equal in importance (*e.g.* when all the auxiliary variables are strictly associated with the target variables), independently of the scale on which they are measured, then this option would be the right choice. The original version of the Gower's coefficient is a weighted average of three different measures of dissimilarity where the weights are the fractions of each type of variable.

4.2. Equalizing standard deviations of partial distances

The significance of a group of variables to determine the global distance depends, among other things, on the variability of the $d_{i,j}^{(t)}$'s so that types leading to high variance of pairwise distances will thus be more likely to have high influence into the global distance. To ensure the respect of this principle we can use the inverse of a measure of variability of distances. For example

$$w_t = \begin{cases} \dfrac{\dfrac{1}{\sigma\left[d_{i,j}^{(t)}\right]}}{\sum_{r=1}^{p}\left[\dfrac{1}{\sigma\left[d_{i,j}^{(t)}\right]}\right]} & \text{if } \sigma\left[d_{i,j}^{(t)}\right] > 0 \\ 0 & \text{otherwise} \end{cases} \qquad t = 1, 2, \ldots, p \tag{22}$$

where $\sigma\left[d_{i,j}^{(t)}\right]$ is the standard deviation of the distances in the strictly lower triangular part of $\mathbf{D}_t$. If the weight are determined according to (23) then the entries of the lower triangular part of all distance matrices 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 impact more on the calculation of the global distance. Kagie *et al.*, (2008) observe that there is no reason to assume, without reference to particular facts of the problem at hand, that nominal variables result more important than quantitative ones. Therefore, an adaptation is necessary. Following Kagie *et al.*, (2008) and Lee *et al.*, (1978), the distances in the strictly lower triangular part of $\mathbf{D}_t$ can be normalized to have an average value of one in the data set.

$$w_t = \frac{\dfrac{1}{\mu\left[d_{i,j}^{(t)}\right]}}{\sum_{r=1}^{p}\left[\dfrac{1}{\mu\left[d_{i,j}^{(t)}\right]}\right]}; \qquad t = 1, 2, \ldots, p \tag{23}$$

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

4.4. Distatis weighting

To obtain an optimal system of weights, we need an expression for how much a type of variables affects the global distance. This can be derived from the total sum of squares of the elements of the partial distance matrices $\mathbf{D}_t, t = 1, 2, \ldots, p$ where the weights are chosen so that the variance of the elements in $\mathbf{D}$ is maximized which, in turn, naturally leads to the Distatis procedure developed by Abdi *et al.*, (2005) (see also D'Urso & Vichi, (1998)).

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

$$\mathbf{B}_t = \left(\frac{1}{\lambda_{1,t}}\right)\mathbf{S}_t, \quad \mathbf{S}_t = -0.5\mathbf{C}\mathbf{D}_t^2\mathbf{C}^t, \ \mathbf{C} = \mathbf{I}_n - n^{-1}\mathbf{u}_n\mathbf{u}_n^t \tag{24}$$

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 for example Albers *et al.*, (2007)) that $\mathbf{S}_t$ is symmetric, positive semi definite and with zero row-sums.

Let $\boldsymbol{\beta}_t = Vec(\mathbf{B}_t), t = 1, 2, \ldots, p$ be the column vector obtained by stacking the column of $\mathbf{B}_t$ on top of one another. These vectors are organized in a $n^2 \times p$ matrix $\mathbf{Z} = \left[\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_p\right]$. The central step of Distatis is to create the aggregate cross-product matrix $\mathbf{A} = \left(\mathbf{N}^{-0.5}\mathbf{Z}\right)^t \left(\mathbf{Z}\mathbf{N}^{-0.5}\right)$ where $\mathbf{N}^{-0.5}$ is a diagonal matrix whose elements are the square roots of the reciprocal values in the diagonal of $(\mathbf{Z}^\mathbf{t}\mathbf{Z})$. The generic element $a_{r,s}$ of $\mathbf{A}$ is the vectorial correlation coefficient (Escoufier, (1973)) between the cross-product matrix derived from the partial distance matrix $\mathbf{D}_r$ and $\mathbf{D}_s$, $r, s = 1, 2, \ldots, p$.

$$a_{r,s} = \frac{\boldsymbol{\beta}_r^t \boldsymbol{\beta}_s}{\|\boldsymbol{\beta}_r\| \|\boldsymbol{\beta}_s\|}. \tag{25}$$

Naturally, $a_{r,s} = a_{s,r}$ and $a_{r,r} = 1, r = 1, 2, \ldots, p$. Since the scalar product in (25) verifies the relationship $\boldsymbol{\beta}_r^t \boldsymbol{\beta}_s = Trace\left(\mathbf{B}^\mathbf{t}{}_r \mathbf{B}_s\right)$ and since $\mathbf{B}_r$ and $\mathbf{B}_s$ are symmetric and positive semi definite, then

$$Trace\left(\mathbf{B}_r^t \mathbf{B}_s\right) \geq \lambda_{n,r} Trace\left(\mathbf{B}_s\right) \tag{26}$$

where $\lambda_{n,r}$ is the smallest eigenvalue of $\mathbf{B}_r$ (see Fang *et al.*, (1994)); it follows that $0 \leq a_{r,s} \leq 1$.

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

$$\mathbf{B} = \sum_{t=1}^{p} w_t \mathbf{B}_t \quad \text{where} \quad \mathbf{w} = \left(\mathbf{u}_p^t \mathbf{q}\right)^{-1} \mathbf{q}. \tag{27}$$

The weights $\mathbf{w}$ are determined in such a way that the groups of variables who judge similar to the others get higher weights than those who disagree with most of the rest of the variables. The related global distance matrix is

$$\mathbf{D} = \sum_{t=1}^{p} w_t \mathbf{D}_t. \tag{28}$$

Distatis weights are demanding on the computational point of view because they involve an eigen-analysis of potentially very large matrices. However, the computational task can be simplified 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 some numerical experiments to assess how well 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 the donors. Of particular interest is determining which weighting scheme of the partial matrices would have the smallest detrimental effect on the accuracy of data when using imputed values. In this regard, we have used data sets with variables completely known for all units (where necessary, we have removed incomplete records). The main reason for this choice is to have total control over the missing data in each experiment. We then create missing values that follow a MAR mechanism so that it becomes possible to evaluate the performance of NNHDI by computing the extent of the imputation bias in comparison to the results that would be obtained had there been no missing data. All simulations were performed using the freely available *R* (R Development Core Team, (2009)) statistical software, thus allowing all researchers access to any suitable methods identified.

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 multistate. 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 several cases we have not used the entire data set, but a subset obtained by drawing without replacement a random sample starting from $set.seed(820723)$ in *R*-code for each data set.

(a) Heart dataset (UC-Irvine repository) Frank & Asuncion, (2010). The variables summarizing the medical symptoms considered as important indicators of the patient's condition were classified as metric: $(1, 2, 8, 10)$,

ordinal: $(11, 12)$, binary symmetric: $(2, 6)$, binary asymmetric: $(9, 14)$, multistate: $(3, c11, 12)$; the target variable y is supposed to be the serum cholesterol in mg/dl. The original data set contains 270 records, but we have analyzed a random subset of $n = 30$ records.

(b) Diabetes dataset (UC-Irvine repository). The data consist of 16 variables on subjects who were interviewed in a study to understand the prevalence of obesity, diabetes, and other cardiovascular risk factors in central Virginia for African Americans. For more information about this study see Willems *et al.*, (1997). Only a part of the subjects were the ones who were actually screened for diabetes. Moreover, some cases contained missing values, therefore they were removed. Of the 366 remaining cases, we have used a random subset of $n = 60$. The variables were used as: metric: $(1 : 5, 7, 9, 10, 12 : 15)$, ordinal: (11), binary symmetric: (8), binary asymmetric: (6); the target variable is $y =$ total cholesterol.

(c) Cars data set (package MASS of R). Data from $n = 93$ cars on sale in the USA in 1993. The variables were metric: $(4, 6 : 8, 12 : 15, 17, 19 : 23)$, ordinal: $(9, 11, 18)$, binary symmetric: (16), binary asymmetric: (24), multistate: $(1 : 3, 10, 25)$; the target variable is $y =$ maximum price. 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. Duplicate models were listed at most once. Further description can be found in Lock, (1993).

(d) Demand for medical care (NMES1988, package AER of R). Cross-section data originating from the US National Medical Expenditure Survey (NMES) conducted in 1987 and 1988. 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 are referred to individuals ages 66 and over, all of whom are covered by Medicare (a public insurance program providing substantial protection against health-care cost). To keep the data volume at a manageable level we have randomly selected $n = 150$ records. Variables were used as: metric: $(1 : 6, 11, 15)$, ordinal: $(7, 8)$, binary symmetric: $(9, 13, 14)$, binary asymmetric: $(12, 17 : 19)$; multistate: $(7, 8)$. The income of the individual was chosen as target variable. Details are given by Cameron & Trivedi, (1998).

(e) German Breast Cancer Study Group. This file contains data on 686 women with breast cancer. We have used $m = 15$ variables (two variables were removed because of they showed a constant value across records).

A detailed description of the study is given in Schumacher *et al.*, (1994). The classification of the variables is metric: $(1, 7 : 10, 12, 15)$, ordinal: (4), binary symmetric: $(2, 11, 13)$, binary asymmetric: $(5, 6, 14)$. The target variable was the tumor size (in mm). The algorithms of HDNNI were applied to a random subset of $n = 220$ complete records.

(f) Fatalities in package AER of R. US traffic fatalities panel data annually for 1982 through 1988. 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: $(14 : 16)$, multistate: $(1, 2)$. The number of alcohol-involved vehicle fatalities is used as target variable. See Stock & Watson, (2007) for more details.

(g) Health Care Reform. The variables have been considered as: metric: $(3, 4)$, ordinal: $(6, 8)$, binary symmetric: $(5, 7, 9, 10, 11)$, binary asymmetric: $(12, 14 : 19)$; the target feature chosen for prediction was $y =$ logarithm of monthly gross income. Since the data set is too large, a Monte Carlo data set is substituted in its place; in particular, we have sampled $n = 450$ complete records. Additional information on the data can be found in Winkelmann, (2004).

(h) House prices (package AER of R). Sales prices of houses sold in the city of Windsor, Canada. The data consist of $n = 546$ complete records. We have classified the variables as metric: $(2, 3, 11)$, binary symmetric: (12), binary asymmetric: $(6, 7, 8, 9, 10)$, multistate: $(4, 5)$. The role of target variable has been played by the sale price of a house. For more information about this data set see Anglin & Gencay, (1996).

(i) Assessing consumer credit applications (UC-Irvine repository) Frank & Asuncion, (2010). Each case concerns an application for credit card facilities described by 16 attribute. Metric: $(3, 8, 11, 14, 15)$, ordinal: $(2, 9, 10)$, binary symmetric: $(1, 16)$, binary asymmetric: $(9, 10, 12)$; multistate: $(4 : 7, 13)$. The first continuous variable was selected as the target variable. The data frame contains 690 observations but 37 cases (5%) have one or more missing values and consequently they were removed from further analysis. Hence, the effective number of records is $n = 653$. More information can be found in Quinlan, (1987).

(j) Household Portfolios. The data used in this case are from the Survey of Consumer Finances 2004. We used the data aggregated and transformed by Miniaci & Pastorello, (2010). This file contains 4214 observations

for 15 variables classified as metric: $(2, 3, 13 : 15)$, binary symmetric: $(5, 7, 11)$, binary asymmetric: $(6, 8 : 10, 12)$, multistate: (1). The age in years (multiplied by 100) was selected as target variable. For this data set we have studied a random subset of $n = 850$ records.

A brief description of the characteristics of each data set appears in Table 1. The number of donors k has been established using the Sturge's rule.

Table 1: Information about the data sets used in the paper.

Data sets	n	k	m	Metric	Ordinal	b. symm.	b. asym.	Multistate
Heart	30	6	13	4	2	2	2	3
Diabetes	60	7	15	12	1	1	1	0
Cars	93	8	24	14	3	1	1	5
Medi Care	150	9	19	8	2	3	4	2
Gbsc	220	9	14	7	1	3	3	0
Fatalities	335	10	37	26	3	3	3	2
Health Reform	450	10	16	2	2	5	7	0
House Price	546	11	11	3	0	1	5	2
Credit Approval	653	11	18	5	3	2	3	5
Family Portfolios	850	11	12	5	0	3	3	1

It can be seen that metric variables are, by far, the most prevalent fields in the records and that not all groups are always represented on the various data sets.

5.2. Experiment 1: leave-one-out cross-validation

In this experiment we have used one record to be the part of the data set with a missing value on the target variable, whereas $n_\nu = n - 1$ records were used as the part without missing information. This procedure is repeated until every record in the data set has played the role of receptor. To assess the accuray of the reconstruction process we have calculated the relative mean error between actual and imputed values:

$$E_1^{(q)} = n^{-1} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i^{(q)}}{y_i} \right| 100 \text{ for } y_i \neq 0,\ q = 1, 2, \ldots, 5 \tag{29}$$

where $\hat{y}_i^{(q)}$ is the value obtained from one of the 5 NNHDI procedures compared to the artificially induced missing value y_i. Note that when the original

values of the target variable have r decimal digits, the imputed values were also rounded to the r-th digit.

It must be pointed out that the deletion mechanism adopted here is of the missing completely at random (MCAR) type, whereas HDNNI methods rely on the missingness being of the MAR type. Therefore the results are likely to underestimate both the magnitude and the statistical significance of the imputed values.

Table 2: Leaving-one-out experiment. Relative mean error.

Data sets	Uniform	Proportional	Equal. Mu	Equal. Sd	Distatis
Heart	12.38	12.99	14.02	13.28	12.51
Diabetes	19.41	16.57	18.53	16.95	18.70
Cars	16.92	13.51	17.22	20.23	16.63
Medi Care	112.00	98.75	99.63	124.38	109.29
Gbsc	42.58	43.17	42.77	44.15	43.58
Fatalities	35.92	27.04	43.42	25.73	36.86
Health Care	4.19	4.21	4.26	4.31	4.19
House Price	23.20	23.35	21.51	33.27	24.61
Credit Approvals	28.20	28.03	24.86	27.67	28.54
Family Portfolios	28.21	27.92	24.81	28.19	28.36
Mean rank	**2.9**	**2.3**	**2.8**	**4.0**	**3.0**

The findings in Table 2 for what concern $E_1^{(q)}$ reveals that only negligible difference can be observed among the alternative implementations of HDNNI, particularly for moderate or large data sets. It turns also out that the performance of an aprioristic system of weights (proportional) is remarkably good in terms of relative mean error of imputation. If one is reluctant to use weights that are not data driven, then the mean rank across all imputations and the ten data sets can instead be considered. The last row in Table 2 indicates that equalizing the mean of the nondiagonal entries of the partial distance matrices is a good practice for weighting the partial distances, particularly for larger data sets. The cumbersome technique of Distatis weighting is not likely to lead to high quality of missing value estimates.

Another interesting result in Table 2 is the tendency of the rate of error that decreases as the size of the data set increases. This is in line with the idea that larger data sets permit, as a general rule, the building of a reference set in a region of relatively small volume around the receptor, so that sufficiently good resolution in the estimates of the different conditional densities can be obtained.

5.3. Experiment 2: random percentage of missing in the target variable

This experiment involves random deletion of prefixed percentages of entries in the target variable: $(15, 25, 40)$ for each of the ten data sets described in Section 5.1. If, for a given data set, there were a strict relationship between target and auxiliary variables, it could be expected that the exclusion of the values of the target variables in a simple random sample without replacement, would automatically produce lacunae in y according to a MAR dynamic. This is coherent with the fact that all the variables which are correlated to the target variable are included; variables that have negligible or null impact on y should be ignored because of the useless (or worse than useless) dominance they could exert on the selection of the donors. Conversely, if the target variable has no specific links with auxiliary variables, then the MAR dynamic has to be artificially induced; in other words, we have to remove data in such a way that excluded values of the target variable depend on one or more or all auxiliary variables.

Although we have several reasons to believe that the target variable in each data set interacts with some, if not all, of the auxiliary variables, we find it satisfying to inject some MAR behavior into the process of data exclusion. For a given data set, we assume that the missingness indicator ψ_i introduced in section 2 is a Bernoulli random variable $B(\pi)$ where $0 < \pi < 1$ is a probability expressed as a function of auxiliary variables only. In particular, we assume that the missing values derive from a sequence of n independent Bernoulli random variables with probabilities of success

$$\pi(\mathbf{x}_i) = \frac{1}{1 + e^{-\left(\gamma_0 + \sum_{r=M_0}^{M_4} \gamma_r x_{i,r}\right)}}; \qquad i = 1, 2, \ldots, n \tag{30}$$

where the expression in the exponential includes all the auxiliary variables except multistate variables. See, for example, Ambler & Omar, (2007), Khoshgoftaar & Van Hulse, (2008), Gheyas & Smith, (2009). To render effective the MAR mechanism, we have to determine the parameters $\gamma_r, r = 0, 1, 2, \ldots, M_4$ of (30). For this purpose, we have first obtained the least squares estimates $\hat{\boldsymbol{\gamma}}$ of the parameters in the multiple linear regression model

$$y_i = \gamma_0 + \sum_{r=M_0}^{M_4} \gamma_r x_{i,r}; \qquad i = 1, 2, \ldots, n \tag{31}$$

and then substituted them in the missingness model to compute the probability that y_i should be missing

$$\hat{\pi}\left(\mathbf{x}_i\right) = \frac{1}{1 + e^{-\hat{y}_i / \max(\mathbf{y})}}; \qquad i = 1, 2, \ldots, n \tag{32}$$

where

$$\hat{y}_i = \hat{\gamma}_0 + \sum_{r=M_0}^{M_4} \hat{\gamma}_r x_{i,r}. \tag{33}$$

The records of the data set at hand are sorted according to (32) and divided in two blocks of equal size. The missing data were inserted in y selecting a random sample without replacement of the given size from the first block, with inclusion probabilities proportional to (32). To this end, we have used the Midzuno's method, Midzuno, (1952), in the package *sampling* in *R* (see also the package *samplingbook*).

The accuracy of the reconstruction process has been measured by the mean relative error (MRE)

$$E_2^{(q)} = (n - \nu)^{-1} \sum_{i=\nu+1}^{n} \left| \frac{y_i - \hat{y}_i^{(q)}}{y_i} \right| \text{ for } y_i \neq 0, \ q = 1, 2, \ldots, 5 \tag{34}$$

where $\hat{y}_i^{(q)}$ is the value obtained from one of the 5 NNHDI procedures in which the i-th value of the target variable was considered artificially missing. The index $E_2^{(q)}$ is related to the sizes of the discrepancies between predicted and observed values. Clearly a small value of $E_2^{(q)}$ represents a successful method. Table 3 reports the average error measured for 250 distinct replications of the same scenario to limit erratic fluctuations.

The results shown in Table 3 suggest that the mean relative error of imputed values is not very sensitive to the weighting scheme of the partial distance matrices. Other than a slight preference for weights proportional to the number of features and, limitedly to larger data sets, for weights obtained equalizing the mean of partial distances, there is no strong discernible pattern that allows a clear-cut ranking of the other three techniques.

Intuitively, it seems reasonable to expect that the performance of NNHDI will deteriorate as the fraction of missing values in the target variables increases; this decreased accuracy manifest itself in Table 3, although not to an overwhelming extent.

Table 3: Missing values in the target variable. Mean relative error.

	Uniform	Propor.	Eq. Mean	Eq. sd	Distatis	Uniform	Propor.	Eq. Mean	Eq. sd	Distatis
			Heart					Fatalities		
M15	10.8	10.3	10.3	11.5	11.7	73.3	47.2	80.8	59.1	73.9
M25	11.0	10.6	11.1	11.4	11.1	105.1	59.7	99.9	89.9	98.4
M40	12.1	11.5	13.1	12.0	11.5	206.4	108.0	158.2	176.1	177.5
Mean	11.3	10.8	11.5	11.6	11.4	128.2	71.7	113.0	108.4	116.6
R15	2.9	2.7	2.7	3.3	3.4	3.6	1.4	4.2	2.3	3.5
R25	3.0	2.6	3.0	3.3	3.1	3.9	1.1	3.6	2.9	3.5
R40	3.1	2.5	3.8	3.0	2.6	4.8	1.0	2.4	3.3	3.5
Mean	3.0	2.6	3.2	3.2	3.0	4.1	1.2	3.4	2.8	3.5
			Diabetes					Medi Care		
M15	18.9	16.6	19.0	18.3	18.9	4.7	4.8	4.8	4.9	4.8
M25	19.0	16.9	18.7	18.0	18.5	4.8	4.9	4.8	5.0	5.0
M40	18.9	16.7	19.1	19.0	19.0	5.1	5.3	5.0	5.4	5.5
Mean	18.9	16.7	18.9	18.4	18.8	4.9	5.0	4.8	5.1	5.1
R15	3.2	2.5	3.2	2.9	3.2	2.7	2.9	2.9	3.3	3.0
R25	3.3	2.4	3.2	2.9	3.2	2.5	3.2	2.5	3.3	3.4
R40	3.3	1.9	3.3	3.2	3.3	2.2	3.4	1.7	3.7	4.1
Mean	3.3	2.2	3.3	3.0	3.2	2.5	3.2	2.4	3.5	3.5
			Cars					House Prices		
M15	16.9	13.1	16.8	22.0	17.0	27.1	27.7	24.8	46.3	29.6
M25	20.3	14.5	19.4	24.6	20.9	28.4	29.7	25.9	49.1	31.9
M40	45.6	32.7	38.8	43.7	44.1	33.9	34.6	30.8	56.0	36.9
Mean	27.6	20.1	25.0	30.1	27.3	29.8	30.7	27.1	50.5	32.8
R15	3.0	1.7	2.9	4.3	3.0	2.5	2.7	1.6	5.0	3.3
R25	3.1	1.3	2.9	4.3	3.4	2.3	2.8	1.3	5.0	3.6
R40	3.7	1.7	2.5	3.5	3.5	2.4	2.7	1.1	5.0	3.7
Mean	3.3	1.5	2.8	4.0	3.3	2.4	2.7	1.3	5.0	3.6
			NMES					Cr.Approvals		
M15	187.2	183.2	181.8	190.1	174.1	30.3	29.8	28.1	30.4	30.6
M25	212.2	201.6	201.8	212.2	216.3	31.0	30.9	28.4	31.0	31.1
M40	279.1	228.8	260.5	265.8	278.6	33.9	34.2	30.1	33.5	33.6
Mean	226.2	204.5	214.7	222.7	223.0	31.7	31.6	28.9	31.6	31.8
R15	3.0	2.9	2.8	3.5	2.8	3.2	3.0	2.0	3.4	3.5
R25	3.1	2.8	2.8	3.1	3.2	3.4	3.2	1.6	3.4	3.5
R40	3.6	1.8	2.9	3.1	3.6	3.5	3.6	1.3	3.2	3.3
Mean	3.2	2.5	2.8	3.2	3.2	3.4	3.3	1.6	3.3	3.4
			GBSG					F. Portfolios		
M15	44.8	44.4	44.9	48.7	44.8	32.6	32.6	32.0	32.1	32.4
M25	45.4	44.5	45.3	50.8	45.9	34.4	34.4	34.3	33.7	34.0
M40	47.8	44.8	45.9	50.4	48.2	39.7	39.8	40.1	38.7	39.7
Mean	46.0	44.5	45.4	50.0	46.3	35.6	35.6	35.5	34.8	35.3
R15	3.0	2.8	2.9	3.3	2.9	3.1	3.1	2.8	2.9	3.0
R25	2.8	2.6	2.9	3.8	2.9	3.2	3.2	3.1	2.7	2.9
R40	3.1	2.1	2.5	4.0	3.3	3.1	3.2	3.4	2.3	3.1
Mean	3.0	2.5	2.8	3.7	3.0	3.1	3.2	3.1	2.6	3.0

5.4. Experiment 3: random percentage of missing values in all the variables

In a first phase, we have randomly omitted a percentage $\pi_{x_j} \in \{10, 20\}, j = 1, 2, \ldots, m$ for all auxiliary variables ensuring that no vector $\mathbf{x}$ of values

of the auxiliary variables contains more than $(m-2)$ fields not filled in. The π_{x_j} records in which remove the value of the auxiliary variable were chosen by drawing a simple random sample without replacement for each $(X_1, X_2, \cdots, X_m)$ from the first $\lfloor n/2 \rfloor$ rows, once the data set has been sorted in ascending order of the target variable. In this way, a mild association is observed between $\mathbf{y}$ and $\mathbf{X}$ because the largest half of the target variable can rely on a more rich source of information in terms of auxiliary variables.

In the second phase, we perform a random deletion of prefixed percentages $\pi_y \in \{15, 25, 40\}$ of the target variable. As a necessary premise, the auxiliary variables are randomly permuted. The records of the data set are resorted in ascending order of the first (permuted) auxiliary variable X_1 allocating the missing values of X_1 always in the last positions. At this point, the records with a missing value for y were chosen to be a random sample without replacement of sample size

$$\lfloor n\pi_y/m \rfloor + b_j \quad \text{J} = 1, 2, \ldots, m. \tag{35}$$

The sample is drawn with probability proportional to certain appropriately given numerical quantities (Midzuno's method). Regarding (35), we have $b_j = 0$, but if $\lfloor n\pi_y \rfloor$ is not an exact multiple of m, we set $b_j = 1$ until the total number of missing values is reached. The quantity which determines the proportionality of the Midzuno's method was $i^{-1}, i = 1, 2, \ldots, n$ which is 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 were selectively missing for cases with small values of the auxiliary variable. This process is repeated for a number of auxiliary variables sufficient to complete the number of missing values to be injected in 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; in addition, the simultaneous presence of missing values in $\mathbf{y}$ and $\mathbf{X}$ is discouraged.

The double relationship between the target and the auxiliary variables makes plausible the existence of a MAR mechanism in the simulated patterns of missingness. Each combination of $\pi_y \times \pi_{x_i}$ has been replicated $N = 100$ times to reduce irregular variations. Note that the computation of the partial distance matrices has to be executed for any choice in $\pi_{\mathbf{x}}$. The results are reported in Tables 4.

Table 4: Missing values in all the variables. Mean relative error.

	Uniform	Propor.	Eq.Mu	Eq.sd	Distatis	Uniform	Propor.	Eq.Mu	Eq.sd	Distatis
		Heart					Fatalities			
M15-10	17.6	20.0	18.6	17.7	19.0	219.3	211.4	218.1	205.8	217.2
M25-10	20.4	21.8	21.0	19.2	19.9	224.0	215.9	223.0	222.4	228.8
M40-10	17.3	19.3	19.2	18.1	18.4	208.4	196.9	206.5	218.8	210.2
M15-20	15.7	14.3	14.3	13.6	14.7	226.0	204.3	216.7	208.6	223.5
M25-20	16.8	17.5	16.6	16.8	16.8	224.9	209.5	221.9	222.7	230.5
M40-20	20.6	20.5	19.2	18.9	20.1	210.9	191.8	208.1	215.6	214.1
Mean	18.1	18.9	18.1	17.4	18.2	218.9	205.0	215.7	215.7	220.7
			Diabetes					Medi Care		
M15-10	20.6	18.1	21.2	16.9	20.4	4.7	4.7	4.7	4.8	4.7
M25-10	22.2	19.9	21.9	20.0	22.3	4.7	5.0	4.8	4.8	4.8
M40-10	21.4	21.3	22.3	19.8	22.0	4.6	4.7	4.6	4.7	4.6
M15-20	19.2	21.2	19.2	19.4	20.9	3.9	3.8	3.8	3.9	4.0
M25-20	19.5	22.6	21.0	18.8	20.4	4.2	4.3	4.2	4.4	4.3
M40-20	19.6	20.2	19.8	18.9	21.1	4.3	4.3	4.3	4.4	4.4
Mean	20.4	20.5	20.9	19.0	21.2	4.4	4.5	4.4	4.5	4.5
			Cars					House Prices		
M15-10	52.5	50.0	56.8	46.5	51.2	51.2	46.2	50.4	54.5	50.7
M25-10	54.4	49.9	54.2	45.8	50.5	49.2	46.7	48.7	53.0	49.1
M40-10	49.4	47.2	50.0	46.1	49.3	48.6	46.3	47.7	52.9	48.6
M15-20	53.3	51.9	54.9	47.9	54.3	41.2	38.5	41.5	44.6	41.0
M25-20	54.8	53.6	54.2	50.2	53.5	36.9	33.7	36.9	42.4	37.4
M40-20	54.8	54.6	56.4	50.7	55.0	39.7	36.2	39.5	45.1	40.1
Mean	53.2	51.2	54.4	47.9	52.3	44.5	41.3	44.1	48.7	44.5
			NMES					Cr. Approvals		
M15-10	182.2	175.0	178.4	178.9	172.6	30.7	31.6	31.5	32.9	31.1
M25-10	156.4	159.7	158.2	150.1	156.1	30.8	31.2	31.0	32.3	31.5
M40-10	155.7	166.4	156.2	168.0	162.9	30.5	30.8	30.6	32.0	30.8
M15-20	135.5	133.3	142.2	129.3	136.1	29.8	30.2	29.5	30.6	29.8
M25-20	130.3	137.4	128.1	128.0	130.4	30.0	30.3	29.5	30.5	29.7
M40-20	120.0	127.1	120.1	125.1	122.1	32.2	32.4	32.2	32.7	32.2
Mean	146.7	149.8	147.2	146.6	146.7	30.7	31.1	30.7	31.8	30.8
			GBSG					F. Portfolios		
M15-10	53.0	52.2	51.1	53.0	51.8	35.9	36.2	35.4	36.4	36.0
M25-10	52.3	54.0	53.0	53.3	54.0	37.4	37.4	37.4	36.8	37.7
M40-10	59.6	59.5	58.2	59.9	58.8	38.8	38.9	38.7	38.0	38.9
M15-20	53.0	54.2	52.5	52.7	53.3	32.0	32.2	32.7	31.8	32.0
M25-20	53.8	56.3	54.3	54.2	56.1	31.6	31.2	31.9	30.6	31.3
M40-20	51.3	53.8	52.1	53.8	54.3	33.3	33.4	33.5	32.2	33.1
Mean	53.8	55.0	53.5	54.5	54.7	34.8	34.9	34.9	34.3	34.8

No strong evidence has been found in our third experiment in favor of or against one of the five weighting system 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, apparently, do not penalize the accuracy of imputed values too greatly, especially for the 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 the weights is not very helpful in comparing two records and Monte Carlo experiments confirm that no algorithm is significantly more efficient that others under these data conditions. However, in measuring the overall quality of missing value estimates produced by a given NNHDI variation, the weighting proportional to the cardinality of the group of variables 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 proportion of incomplete records, or of the number of missing values in a record, or both, the quality of estimates would decrease due to the reduction of potentially useful information. In effect, this seems to be confirmed by the values of $E_2^{(q)}$ which are generally higher that those obtained in the previous experiment. Nonetheless, the values of the index (35) for experiments in which the percentage of missing in the auxiliary variables is at 20% are not much higher, and in many cases are lower, than for experiments in which the percentage is at 10%. 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 present rarely in records having a missing value in the target variable. 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 the auxiliary variables has moderate detrimental effects on the accuracy of the reconstruction process; on the contrary, high percentages of missing values in y and $\mathbf{x}$ may even strengthen the cohesion of the values in the records concerned.

6. Conclusions and suggestions for future research

Missing values often occur in real-world applications and represent a relevant challenge in the field of data quality, particularly when the variables of the data set have mixed types. In this paper, we have conducted an extensive study of the nearest neighbor hot deck imputation (NNHDI) methodology where, for each record with incomplete data for the target variables, a set of donors is selected so that the donors are similar to their recipient with respect to the auxiliary variable. The known values of the donors are then

used to derive a value for the missing data computing the least power mean 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 auxiliary variables and involving a mixture of numeric, ordinal, binary and multistate variables. It has become increasingly apparent that efficacy and effectiveness of NNHDI, crucially hinge on how the distance between records is measured and different ways of measuring distance lead to different solutions. We have devised a new global distance function based on the partial distance matrices obtained from the various type of variables appearing (or missing) in the records of the data set. The separate distance matrices are added as a weighted average and the combined distance matrix is then used for classification and ordination. The contribution of each group of variables to the global distance is scaled with a weight which contracts/expands the influence of variables of interest. In this study we have compared the performance of five weighting schemes to define a combined distance matrix for variables of mixed nature.

To judge the accuracy of the reconstruction process, we have considered a performance indicator 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 ten real data sets on three different experiments: leaving-one-out, incompleteness in the target variables and incompleteness in all variables. The missing values in the last two experiments have been artificially inserted following a MAR mechanism.

The empirical findings suggest that data driven weights for the partial distance matrices are 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 specific weighting system. On the other hand, the investigations carried out with 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 a part of the records have lacunae in the auxiliary variables other that the target variable. The main advantage of 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 paid with strong distributional assumptions or sophisticated modeling.

In this work we have compared NNHDI algorithms implementing five different system of weights for the separate distance matrices. The results show that the choice of weights did not significantly affect the estimates.

This is due in most part to the strong relationships among the variables of the tested data sets. Also, the scarce variability of the least power mean used to estimate the missing values might have given a non marginal contribution to the lower distinguishability of the various techniques. However, we have to consider the possibility that alternative weights could achieve superior performance. We plan to study weighting system based on the degree of interdependency between the target and group of auxiliary variables so we can understand better the implications of such differences for diverse NNHDI algorithms.

The quality of the results of NNHDI is closely connected on the specific observations to be classified. 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. The analysis of alternative strategies of constructing the reference set of a receptor might offer a promising line for further research

The imputation technique we have explored is based on the least power mean estimator of the target variables based on the observed values in the reference set of donors. This method does not fully exploit the information generated by the donors. For instance, nearest neighbors may be more informative than distant neighbors, so that a positive score should be assigned to each of the nearest neighbors donors which decreases as the distance from the receptor record increases. This is probably the most adaptive course of action that balances the dual challenges given to NNHDI: to produce reliable replacements and speed of execution. As a point for further research, we would like to test our algorithm for NNHDI in association with other methods of imputation based on weighted mean of donors.

The partial distance between two records is computable if at least one of the auxiliary variables in the group of interest has been validly observed for both the records. If this condition is not met, we are lead to a distance matrix completion problem. In our paper, we have set the missing distance to the its maximum observable value. Although there is no reason to assume that the divergence between two otherwise not comparable records is at the highest level for the given type of variables, this choice has had the desirable effect of restraining selection of donors that share too few features with the receptor. Since distances are not independent from one another, unknown entries in a distance matrix can be reconstructed from a subset of the known distances that share part of the same information. It would be useful for other research to develop solutions to handle missing entries in distance matrices.

References

Abbate, C. (1997). La completezza delle informazioni e limputazione da donatore con distanza mista minima. *Quaderni di Ricerca dell'ISTAT*, **4**, 68–102.

Abdi, H. and O'Toole, A. J. and Valentin, D. and Edelman, B. (2005). DISTATIS: the analysis of multiple distance matrices. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, CA, (USA), 42–47.

Albers, C., and Critchley, F. and Gower, J. (2007). Group average representations in Euclidean distance cones. In: Brito, P., Bertrand, P., Cucumel, G. and de Carvalho, F. eds. *Selected Contributions in Data Analysis and Classification. Studies in Classification, Data Analysis, and Knowledge Organization*. Springer, Berlin, Germany, 445–454.

Ambler, G. and Omar, R.Z. (2007). A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Statistical Methods in Medical Research* **16**, 277–298.

Anderberg, M. R. (1973). *Cluster Analysis for applications*. Academic Press, New York.

Andridge, R. R and Little, R.J.A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review* **78**, 40–64.

Anglin, P., and Gencay, R. (1996). Semiparametric estimation of a hedonic price function. *Journal of Applied Econometrics*, **11**, 633648. http://www.econ.queensu.ca/jae/1996-v11.6/anglin-gencay/

Bankier, M. and Fillion, J.M. and Luc, M. and Nadeau, C. (1994). Imputing numeric and qualitative variables simultaneously. Proceedings of the Section on Survey Research Methods, American Statistical Association, 242–247.

Bankier, M. and Luc, M. and Nadeau, C. and Newcombe, P. (1995). Additional details on imputing numeric and qualitative variables simultaneously. Proceedings of the Section on Survey Research Methods, American Statistical Association, 287–292.

Bankier, M. and Lachance, M. and Poirier, P. (2000). 2001 Canadian census minimum change donor imputation methodology. Work Session on Statistical Data Editing, UN-ECE, Cardiff.

Bijnen, E. J. (1973). *Cluster analysis. Survey and evaluation of techniques*. Tilburg University Press, Tilburg (NL).

Cameron, A.C. and Trivedi, P.K. (1998). *Regression Analysis of Count Data. Cambridge*, Cambridge University Press, Cambridge. http://www.econ.queensu.ca/jae/1997-v12.3/deb-trivedi/

Chen, J. and Shao, J. (2000). Nearest neighbor imputation for survey data.*Journal of Official Statistics*, **16**, 113–131.

Chiodi, M. (1990). A partition type method for clustering mixed data, *Rivista di statistica applicata*, **2**, 135–147.

Colledge, M. J. and Johnson, J. H. and Pare, R. and Sande, I. J. (1978). Large scale imputation of survey data. Proceedings of the Section on Survey Research Methods, American Statistical Association, 431–436.

Deming, W.E., (1960), *Sample Design in Business Research*, John Wiley & Sons, New York.

Di Ciaccio, A. (1992). Simultaneous clustering of qualitative and quantitative with missing observations. *Statistica Applicata*, **4**, 599–609.

D'Urso, P. and Vichi, M. (1998). Dissimilarities between trajectories of a three-way longitudinal data set. In A. Rizzi, M. Vichi and H.-H. Bock (Eds.) *Advances in data science and classification*, Springer, Berlin, 585–592.

Enders, C.K. (2001). The Performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement*, **61**, 713–740.

Enders C.K. (2010). *Applied missing data analysis*. The Guilford Press, New York.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*, **29**, 751–760.

Fang, Y., and Loparo K. A., and Feng, X. (1994). Inequalities for the trace of matrix product.*IEEE Transactions on Automatic Control*, **39**, 2489–2490.

Frank, A. and Asuncion, A. (2010). UCI Machine Learning Repository, `http://archive.ics.uci.edu/ml`, University of California, Irvine, School of Information and Computer Sciences.

Franck, I.E and Todeschini, R. (1994). *The Data Analysis Handbook*, Elsevier, Amsterdam.

Friedman, J. H. and Bentley, J. L. and Finkel, R. A. (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, **3**, 209–226.

Gheyas, I.A. and Smith, L.S. (2009). A novel nonparametric multiple imputation algorithm for estimating missing data. *Proceedings of the World Congress on Engineering - WCE 2009. Vol II.*, London.

Ghosh, A. K. (2007). On nearest neighbor classification using adaptive choice of k, *Journal of Computational and Graphical Statistics*, **16** 482–502.

Giles, P. (1988). A model for generalized edit and imputation of survey data. *The Canadian Journal of Statistics*, **16**, 57–73.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties, *Biometrics*, **27**, 623–637.

Gower, J. C. and Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients, *Journal of Classification*, **3**, 5–48.

Hyndman, R.J. (1995). The problem with Sturges' rule for constructing histograms *Business*,**Issue: July**, 1–2.

Istat (2004).*CONCORD V. 1.0: Controllo e correzione dei dati. Manuale utente e aspetti metodologici*. Istituto nazionale di statistica, Roma.

Jönnson, P. and Wohlin, C. (2006). Benchmarking k-nearest neighbour imputation with homogeneous Likert data. *Empirical Software Engineering*, **11**, 463–489.

Jöreskog, K.G. and Sörbom, D. (1993) *New Features in PRELIS 2*. Scientific Software, Chicago, IL,

Kagie, M. and van Wezel, M. and Groenen, P. J.F. (2008). A graphical shopping interface based on product attributes. *Decision Support Systems*, **46**, 265–276.

Kaiser, J. (1983). The effectiveness of hot-deck procedures in small samples. In *Proceedings of Annual Meeting of the American Statistical Association* Javaid Kaiser, University of Kansas Kalton G. (1983), *Compensating for Missing Survey Data*. Ann Arbor, MI: Survey Research Center, University of Michigan.

Kalton, G. and Kasprzyk, D. (1982) Imputing for missing survey responses. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 22-31.

Kaufman, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York.

Khoshgoftaar, T.M. and Van Hulse, J. (2008). Imputation techniques for multivariate missingness in software measurement data *Software Quality Journal*, **16**, 563–600

Lee, R. C. T. and Slagle, J. R. and Mong, C. T. (1978). Towards automatic auditing of records. *IEEE Transactions on Software Engineering*, **4**, 441–448.

Lin, K.Y. (1977). An elementary proof of the Perron-Frobenius theorem for non-negative symmetric matrices. *Chinese Journal of Phisics*, **15**, 283–285.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd edition. John Wiley & Sons, New York.

Lock, R. H. (1993). New Car Data. Journal of Statistics Education, **1**. http://www.amstat.org/publications/jse/v1n1/datasets.lock.html

Manzari, A. (2004). Combining editing and imputation methods: an experimental application on population census data. *Journal Of The Royal Statistical Society Series A*, **167**, 295–307.

Mineo, A. and Ruggeri, M. (2005). A software tool for the exponential power distribution: the normalp package. *Journal of Statistical Software*, **12**, 1–24.

Miniaci, R. and Pastorello, S. (2010). Mean-variance econometric analysis of household portfolios, *Journal of Applied Econometrics*, **25**, 481–504.http://www.econ.queensu.ca/jae/2010-v25.3/miniaci-pastorello/

Midzuno, H. (1952). On the sampling system with probability proportional to sum of size. *Annals of the Institute of Statistical Mathematics*, **3**, 99–107.

Murthy, M.N. and Chacko, E.and Penny, R.and and Hossain, M. (2003). Multivariate nearest neighbour imputation. *Journal of Statistics in Transition*, **6** , 55–66.

Pavoine, S. and Vallet, J. and Dufour, A-B and Gachet, S. and Herve, D. (2009). On the challenge of treating various types of variables: application for improving the measurement of functional diversity. *Oikos*, **18**, 391–402.

Pennecchi, F. and Callegaro, L. (2006). Between the mean and the median: the L_p estimator. *Metrologia* **43**, 213–219.

Quinlan, J.R. (1987). Simplifying decision trees, *International Journal of Man-Machine Studies* **27**, 221–234.

R Development Core Team, (2009). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Romesburg, H. C. (1984). *Cluster Analysis for Researchers*. Krieger Publishing, Malabar (FL), USA.

Sande, I.G. (1982). Imputation in surveys: coping with reality. *The American Statistician*, 36, 145-152.

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

Schumacher, M. and Basert, G. and Bojar, H. and Hubner, K. and Olschewski, M. and Sauerbrei, W. and Schmoor, C. and Beyerle, C. and Neumann, R. L. A. and Rauschecker, H. F., for the GBSG: German Breast Cancer Study Group (1994). Randomized 2×2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology*, **12**, 208–2093.

Seber, G.A.F. (2004). *Multivariate Observations*, John Wiley & Sons, New York.

Siddique, J. and Belin, T.R. (2008) Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in Medicine*, **27**, 83–102.

Stock, J. H. and Watson, M. W. (2007). *Introduction to Econometrics, 2nd ed.*, Addison Wesley, Boston.

Welniak, E.J. and Coder, J.F. (1980). A measure of the bias in the march CPS earning imputation system and results of a simple bias adjustment procedure. Technical Report, U.S. Census Bureau.

Wettschereck, D. and Dietterich, T.G. (1995). An experimental comparison of the nearest-neighbor and nearest-hyperrectangle algorithms. *Machine Learning*, **19**, 5–27.

Willems, J.P. and Saunders, J.T. and Hunt, D.E. and Schorling, J.B. (1997). Prevalence of coronary heart disease risk factors among rural blacks: A community-based study. *Southern Medical Journal* **90**, 814–820. http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/diabetes.html

Wilson, D. R. and Martinez T. R. (2000). Reduction techniques for exemplar-based learning algorithms. *Machine Learning*, **38**,257–286.

Winkelmann, R. (2004). Health care reform and the number of doctor visits - An econometric analysis, *Journal of Applied Econometrics*, **19**, 455–472. http://www.econ.queensu.ca/jae/2004-v19.4/winkelmann/