On the matching noise of some nonparametric imputation procedures

Daniela Marella [1]
Dipartimento di Statistica, Probabilità e Statistiche Applicate
Università di Roma "La Sapienza"

Mauro Scanu
Dipartimento per la Produzione Statistica e il Coordinamento Tecnico Scientifico
Istituto Nazionale di Statistica, Roma

Pier Luigi Conti
Dipartimento di Statistica, Probabilità e Statistiche Applicate
Università di Roma "La Sapienza"

Abstract. The aim of the paper is to evaluate the matching noise produced by nonparametric imputation techniques referring to the kNN method, both with fixed and variable number of donors k. The matching noise is evaluated formally and via a simulation
Keywords. kNN method, missing data, statistical matching.

1 Introduction

Partially observed data sets are ubiquitous in applied statistics. Usually missing data are filled in, and only the final imputed data set is disseminated. Imputation is justified by practical problems, and at the same time its use is controversial.

The practical problem that imputations overcome are different. For instance, if a partially observed data set is analyzed by different departments in the same institute (as usual for economic samples of the national statistical institutes), different treatment of missing values may be the cause of inconsistencies between the disseminated results. Another problem concerns the dissemination of sample data to third parties (research institutes, universities, international organizations) which are unaware of the data production process, that is a useful source of information when missingness occurs. References on the importance of imputing partially observed data set are Titterington et al. (1989) and Haziza (2001).

The controversial issues of imputation mainly concern the statistical characteristics of the imputed data set. An imputed data set is not a real data set, and statistical conclusions drawn from an imputed data set are questionable. Let $D = \{\mathbf{x}_i; i = 1, \ldots, n\}$ be a sample of n i.i.d. k-variate records from a distribution $f(\mathbf{x}|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. Let $\mathbf{x}_{i;obs}$ and $\mathbf{x}_{i;mis}$ be the observed and missing part in the ith record respectively, for every $i = 1, \ldots, n$. Imputation of missing items consists in choosing a substitute $\widetilde{\mathbf{x}}_{i;mis}$ for the missing components of D. The final imputed data set is a complete data set $\widetilde{D}$ on which the usual estimators and tests are applied. The ideal situation is given by an imputed data set $\widetilde{D}$ whose imputations are randomly generated by the (unknown) conditional distribution $f(\mathbf{x}_{mis}|\mathbf{x}_{i;obs}; \boldsymbol{\theta})$ for every i, while the actual imputation generating distribution $g(\mathbf{x}_{mis}|\mathbf{x}_{i;obs})$ may

[1] Dipartimento di Statistica, Probabilità e Statistiche Applicate, Università di Roma "La Sapienza", Piazzale Aldo 5, 00185, Rome, Italy

be different. The appropriateness (or better, the reliability) of $\widetilde{D}$ depends on two distinct aspects: (i) the missing data generating mechanism and (ii) the imputation mechanism. These two aspects interact: discrepancies between f and g for each missing data pattern affect the reliability of the imputed data set, and are especially important for highly probable missing data patterns.

Missing data patterns and their probabilistic relationships with observed data are well described in the statistical literature, since the seminal paper by Rubin (1976). On the contrary, the study of the discrepancies between the data and imputation processes have received just little attention. A remarkable exception is the statistical matching problem, Paass (1985), where such a discrepancy is named matching noise. Statistical matching consists in integrating information coming from two samples A and B of size n_A and n_B respectively, with no common units, a set of commonly observed variables $\mathbf{X}$, and distinctly observed variables $\mathbf{Y}$ in A and $\mathbf{Z}$ in B. Matching of A and B is performed by imputing in A an observed $\mathbf{Z}$ in B by means of the commonly observed variables $\mathbf{X}$. To this purpose, the usual imputation techniques can be applied under the assumption of conditional independence of $\mathbf{Y}$ and $\mathbf{Z}$ given $\mathbf{X}$, usually denoted as the *conditional independence assumption*, CIA for short (for other dependence relationships, see D'Orazio et al. (2006) and references therein).

It can be easily proved that, when A and B are i.i.d. samples from the same distribution and n_A and n_B are fixed by survey design, the missing data generation process is missing completely at random, D'Orazio et al. (2006). As a matter of fact, the evaluation of the reliability of the matched file consists only of the matching noise.

The goal of this paper is to discuss the matching noise produced by a class of nonparametric imputation procedures in the simplified context of statistical matching. This class is based on the kNN nonparametric estimation of the regression function of $\mathbf{Z}$ on $\mathbf{X}$ in B, and includes some of the most popular nonparametric imputation procedures (as distance hot deck). The asymptotic properties of the imputation procedures are formally analyzed, and then studied by simulation.

The paper is organized as follows. In Section 2 the statistical framework for the matching problem is described. In the same section a set of nonparametric imputation procedures based on the kNN method with fixed number k of donors, are discussed, showing when they produce matching noise. All of them are nonparametric techniques, *i.e* their are not based on a parametric model for the data. In Section 3 the matching noise of different procedures is formally evaluated. In particular, in Section 3.1 the matching noise for kNN procedure is computed. The results of Section 3.1 are particularized for the most popular nonparametric imputation procedure, distance hot deck (Section 3.2). In Section 4, a nonparametric imputation procedure with a variable number of donors k (henceforth d_0-Kernel) is described and the corresponding matching noise is formally evaluated. Finally in Section 5 the matching noise of different procedures and its effect on some estimators is shown via simulation.

2 Statistical framework for matching noise

Let $(\mathbf{X}, \mathbf{Z}) = ((X_1, \ldots, X_P), (Z_1, \ldots, Z_R))$ be a $(P+R)$-variate r.v. and denote by $f(\mathbf{x}, \mathbf{z})$ its joint density function (d.f., for short). Let further A and B be two independent samples of size n_A, n_B, respectively, generated by $(\mathbf{X}, \mathbf{Z})$. Finally,

assume that only $\mathbf{X}$ is observed in A, and $(\mathbf{X}, \mathbf{Z})$ is observed in B. Hence, $\mathbf{Z}$ is missing in A. The sample data can be then written as

$$\begin{aligned} \left(\mathbf{x}_a^A\right) &= \left(x_{a1}^A, \ldots, x_{aP}^A\right), \qquad a = 1, \ldots, n_A \\ \left(\mathbf{x}_b^B, \mathbf{z}_b^B\right) &= \left(x_{b1}^B, \ldots, x_{bP}^B, z_{b1}^B, \ldots, z_{bR}^B\right), \qquad b = 1, \ldots, n_B \end{aligned}$$

for samples A, B, respectively.

The construction of a complete synthetic data set containing $(\mathbf{X}, \mathbf{Z})$, with no parametric assumptions on the family of distributions for the variables of interest, is usually faced by means of nonparametric imputation procedures. These procedures are based on filling missing values with observed ones. More formally, these methods consist in completing the records of a file (the recipient file A, say) by means of the records of the other file (the donor file B, say). The final product is a unique, synthetic data file where all the variables of interest are simultaneously recorded.

The synthetic data set can be used in "genuine" inference procedures only when it can be (at least approximately) considered as a sample generated from the joint distribution of $(\mathbf{X}, \mathbf{Z})$. As a consequence, the discrepancy between the joint probability distribution of the variables of interest (a) in the population, and (b) in the imputed file (*i.e.* the matching noise) is of primary interest. Attempts at evaluating such discrepancy have been performed in the literature on statistical matching, see D'Orazio et al. (2006) and references therein. In Rässler (2002) normative suggestions for evaluating the accuracy of a statistical matching procedure are provided.

In subsequent sections, we study the matching noise produced by a number of different nonparametric imputation techniques. The final output of these procedures is a new data set $\widetilde{A}$ with records $(\mathbf{x}_a^A, \widetilde{\mathbf{z}}_a)$, $a = 1, \ldots, n_A$, where $\widetilde{\mathbf{z}}_a$ is a $\mathbf{z}$-value observed in B associated by the imputation technique to record a in A.

Formally, a family of nonparametric imputation techniques can be described as follows. For every $\mathbf{x}_a^A$ in A, let $\boldsymbol{b}(a) = (b_1(a), \ldots, b_k(a))$ be the labels of its k donor records in B, on the basis of the n_B observations $\mathbf{x}_b^B$, $b = 1, \ldots, n_B$, and let $\mathbf{X}_{\boldsymbol{b}(a)}^B$ be the corresponding vector of r.v.'s $(\mathbf{X}_{b_1(a)}^B, \ldots, \mathbf{X}_{b_k(a)}^B)$. Next, the corresponding $\mathbf{z}$-values $\mathbf{Z}_{\boldsymbol{b}(a)}^B = (\mathbf{Z}_{b_1(a)}^B, \ldots, \mathbf{Z}_{b_k(a)}^B)$ are considered. Finally, the missing value $\mathbf{Z}_a^A$ is imputed by $\widetilde{\mathbf{Z}}_a = g(\mathbf{Z}_{\boldsymbol{b}(a)}^B)$, $g(\cdot)$ being an appropriate function. Common examples are the arithmetic mean of $\mathbf{Z}_{b_j(a)}^B$, $j = 1, \ldots, k$, their median, or a randomly chosen value from $\mathbf{Z}_{b_j(a)}^B$, $j = 1, \ldots, k$.

Since the observed $\mathbf{x}$-values in A are generated from $\mathbf{X}$, the records $(\mathbf{x}_a^A, \widetilde{\mathbf{z}}_a)$ in $\widetilde{A}$ are generated from a r.v. $(\mathbf{X}, \widetilde{\mathbf{Z}})$, say. The donor procedure works appropriately if the distribution of $(\mathbf{X}, \widetilde{\mathbf{Z}})$ coincides with (is "not too far from") the distribution of $(\mathbf{X}, \mathbf{Z})$. The usual factorization rules for d.f.s lead, with obvious notation, to

$$f_{\mathbf{X}_a^A \mathbf{X}_{\boldsymbol{b}(a)}^B \widetilde{\mathbf{Z}}_a}(\mathbf{x}, \mathbf{t}, \mathbf{z}) = f_{\mathbf{X}}(\mathbf{x})\, f_{\mathbf{X}_{\boldsymbol{b}(a)}^B | \mathbf{X}_a^A}(\mathbf{t}|\mathbf{x})\, f_{\widetilde{\mathbf{Z}} | \mathbf{X}_{\boldsymbol{b}(a)}^B \mathbf{X}_a^A}(\mathbf{z}|\mathbf{x}, \mathbf{t}).$$

Once it is known that $\mathbf{X}_{\boldsymbol{b}(a)}^B = \mathbf{t}$, $\widetilde{\mathbf{Z}}_a$ and $\mathbf{X}_a^A$ are independent, hence the following relationship

$$f_{\mathbf{X}_a^A \mathbf{X}_{\boldsymbol{b}(a)}^B \widetilde{\mathbf{Z}}_a}(\mathbf{x}, \mathbf{t}, \mathbf{z}) = f_{\mathbf{X}}(\mathbf{x})\, f_{\mathbf{X}_{\boldsymbol{b}(a)}^B | \mathbf{X}_a^A}(\mathbf{t}|\mathbf{x})\, f_{\widetilde{\mathbf{Z}} | \mathbf{X}}(\mathbf{z}|\mathbf{t})$$

holds. As a consequence, the synthetic sample data $(\mathbf{x}_a^A, \widetilde{\mathbf{z}}_a)$, $a = 1, \ldots, n_A$, can be considered as composed by observations (identically distributed but not generally

independent) generated from:

$$\begin{aligned} f_{\mathbf{X}_a^A \widetilde{\mathbf{Z}}_a}(\mathbf{x}, \mathbf{z}) &= \int f_{\mathbf{X}_a^A \mathbf{X}^B_{b(a)} \widetilde{\mathbf{Z}}_a}(\mathbf{x}, \mathbf{t}, \mathbf{z})\, d\mathbf{t} \\ &= f_{\mathbf{X}}(\mathbf{x}) \int f_{\mathbf{X}^B_{\boldsymbol{b}(a)}|\mathbf{X}_a^A}(\mathbf{t}|\mathbf{x})\, f_{\widetilde{\mathbf{Z}}|\mathbf{X}}(\mathbf{z}|\mathbf{t})\, d\mathbf{t}. \end{aligned} \tag{1}$$

The matching noise is determined by two elements: (i) the presence of the donor distribution $f_{\mathbf{X}^B_{\boldsymbol{b}(a)}|\mathbf{X}_a^A}(\mathbf{t}|\mathbf{x})$; (ii) the combination of the donor values $\widetilde{\mathbf{Z}} = g(\widetilde{\mathbf{Z}}^B_{\boldsymbol{b}(a)})$. It is easy to see that, if $k = 1$ and $g(.)$ is the identity function, the matching noise is null if the r.v.s $\mathbf{X}_a^A$, $\mathbf{X}^B_{b_1(a)}$ coincide almost surely, so that the r.v.'s $(\mathbf{X}_a^A, \widetilde{\mathbf{Z}}_a)$ and $(\mathbf{X}, \mathbf{Z})$ possess the same d.f.. This is possible when $\mathbf{X}$ is categorical and all the categories are observed in both A and B. In all other cases, the two distributions are different.

In the next sections, we will illustrate the influence of the matching noise for different, widely used donor selection procedures. This problem has been addressed by many authors (see Sims (1972), Rodgers (1984), Paass (1986), Rässler (2002) p. 21-22) but an explicit probabilistic evaluation of the matching noise is still missing.

In what follows, the r.v.s $\mathbf{X}$ and $\mathbf{Z}$ will be assumed absolutely continuous.

3 Matching noise for kNN nonparametric imputation techniques

In this section we explicitly evaluate the matching noise for a class of nonparametric imputation procedures that includes some of the most used ones: the distance and random hot deck imputation procedures. This class is defined by assuming that the k donors to a record $a \in A$ are given by the k nearest neighbours of $\mathbf{x}_a$ in B, $a = 1, \ldots, n_A$. Formally, let D be a positive definite matrix, and let $d(\mathbf{x}_a^A, \mathbf{x}_b^B) = ((\mathbf{x}_b^B - \mathbf{x}_a^A)' D (\mathbf{x}_b^B - \mathbf{x}_a^A))^{1/2}$ be the corresponding Euclidean distance. The k nearest neighbours of $\mathbf{x}_a^A$ are the $k \geqslant 1$ observations $\mathbf{x}^B_{\boldsymbol{b}(a)} = (\mathbf{x}^B_{b_1(a)}, \ldots, \mathbf{x}^B_{b_k(a)})$ in B which are closest to $\mathbf{x}_a^A$, according to the distance d. The imputed value $\widetilde{\mathbf{Z}}_a^A$ is then a function $g(\mathbf{Z}^B_{b_1(a)}, \ldots, \mathbf{Z}^B_{b_k(a)})$.

In the sequel we will denote by $\boldsymbol{\Psi}_b$ the quantities $\mathbf{X}_b^B - \mathbf{x}_a^A$, by W_b the quantity $\boldsymbol{\Psi}_b' D \boldsymbol{\Psi}_b$, by $\boldsymbol{\Psi}_{n_B:1} \leqslant \cdots \leqslant \boldsymbol{\Psi}_{n_B:n_B}$ the ordered $\boldsymbol{\Psi}_b$s, and by $f_{\boldsymbol{\Psi}}$ the conditional d.f. of $\boldsymbol{\Psi}_b$ given $\mathbf{X}_a^A = \mathbf{x}_a^A$.

3.1 Formal evaluation of the matching noise

In order to evaluate the matching noise for the kNN nonparametric imputation procedures, let $\boldsymbol{\Gamma} = (\Gamma_1, \ldots, \Gamma_k)$ be the r.v. taking the value $\boldsymbol{b}(a) = (b_1(a), \ldots, b_k(a))$ for every observed sample, i.e. the k nearest neighbour labels of each record in the sample. Consider next the joint probability

$$P(\boldsymbol{\Gamma} = \boldsymbol{b}(a), \boldsymbol{\Psi}_{n_B:j} \leqslant \boldsymbol{\psi}_j, j \leqslant k) = \frac{1}{D_{n_B,k}} P(\boldsymbol{\Psi}_{n_B:j} \leqslant \boldsymbol{\psi}_j, j \leqslant k | \boldsymbol{\Gamma} = \boldsymbol{b}(a)) \tag{2}$$

where $D_{n_B,k} = n_B\,(n_B - 1) \cdots (n_B - k + 1)$ and let $\mathbb{S}_k = \{(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, .., \boldsymbol{\psi}_k) : \boldsymbol{\psi}_1^{'} D \boldsymbol{\psi}_1 \leqslant \boldsymbol{\psi}_2^{'} D \boldsymbol{\psi}_2 \leqslant ... \leqslant \boldsymbol{\psi}_k^{'} D \boldsymbol{\psi}_k\}$ be a k-dimensional subset of $\mathbb{R}^P$, and $(-\infty, \boldsymbol{\psi}_j] = \{\boldsymbol{a} \in$

$\mathbb{R}^P : \boldsymbol{a} \leqslant \boldsymbol{\psi}_j\}$ the orthant with (upper) vertex $\boldsymbol{\psi}_j$, $j = 1, 2, .., k$. Since

$$
\begin{aligned}
& P(\boldsymbol{\Psi}_{n_B:j} \leqslant \boldsymbol{\psi}_j,\ j \leqslant k | \boldsymbol{\Gamma} = \boldsymbol{b}(a)) \\
& = P(\boldsymbol{\Psi}_{b_j(a)} \leqslant \boldsymbol{\psi}_j,\ j \leqslant k,\ W_t \geqslant W_{b_k(a)},\ t \notin \boldsymbol{b}(a) | \boldsymbol{\Gamma} = \boldsymbol{b}(a)) \\
& = \int_{\mathbb{S}_k} P\left(\boldsymbol{\Psi}_{b_j(a)} \leqslant \boldsymbol{\psi}_j,\ j \leqslant k,\ W_t \geqslant W_{b_k(a)},\ t \notin \boldsymbol{b}(a) | \boldsymbol{\Gamma} = \boldsymbol{b}(a),\right. \\
& \qquad \left. \boldsymbol{\Psi}_{b_j(a)} = \mathbf{x}_j,\ j \leqslant k\right) \prod_{j=1}^{k} f_{\boldsymbol{\Psi}}(\mathbf{x}_j) d\mathbf{x}_j \\
& = \int_{\mathbb{S}_k \cap (\bigcap_{j=1}^{k} (-\infty, \boldsymbol{\psi}_j])} P(W_t \geqslant \mathbf{x}_k^{'} D \mathbf{x}_k\ \forall t\ \notin \boldsymbol{b}(a)) \prod_{j=1}^{k} f_{\boldsymbol{\Psi}}(\mathbf{x}_j)\, d\mathbf{x}_j \\
& = \int_{\mathbb{S}_k \cap (\bigcap_{j=1}^{k} (-\infty, \boldsymbol{\psi}_j])} \prod_{t \notin \boldsymbol{b}(a)} P(W_t \geqslant \mathbf{x}_k^{'} D \mathbf{x}_k) \prod_{j=1}^{k} f_{\boldsymbol{\Psi}}(\mathbf{x}_j) d\mathbf{x}_j \\
& = \int_{\mathbb{S}_k \cap (\bigcap_{j=1}^{k} (-\infty, \boldsymbol{\psi}_j])} P(W \geqslant \mathbf{x}_k^{'} D \mathbf{x}_k)^{n_B - k} \prod_{j=1}^{k} f_{\boldsymbol{\Psi}}(\mathbf{x}_j) d\mathbf{x}_j
\end{aligned}
$$

it is seen that (2) is equal to

$$
\frac{1}{D_{n_B,k}} \int_{\mathbb{S}_k \cap (\bigcap_{j=1}^{k} (-\infty, \boldsymbol{\psi}_j])} P(W \geqslant \mathbf{x}_k^{'} D \mathbf{x}_k)^{n_B - k} \prod_{j=1}^{k} f_{\boldsymbol{\Psi}}(\mathbf{x}_j) d\mathbf{x}_j. \tag{3}
$$

Hence, the marginal d.f. of $(\boldsymbol{\Psi}_{n_B:1}, \boldsymbol{\Psi}_{n_B:2}, ..., \boldsymbol{\Psi}_{n_B:k})$ is given by

$$
f_{\boldsymbol{\Psi}_{n_B:1}\, \boldsymbol{\Psi}_{n_B:2} \ldots \boldsymbol{\Psi}_{n_B:k}}(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, ..., \boldsymbol{\psi}_k) = P(W \geqslant \boldsymbol{\psi}_k^{'} D \boldsymbol{\psi}_k)^{n_B - k} \prod_{j=1}^{k} f_{\boldsymbol{\Psi}}(\boldsymbol{\psi}_j). \tag{4}
$$

Finally, taking into account that $\mathbf{X}^B_{\boldsymbol{b}(a)} = (\mathbf{X}^B_{b_1(a)}, \ldots, \mathbf{X}^B_{b_k(a)})$ coincides with $(\boldsymbol{\Psi}_{n_B:1} + \mathbf{x}^A_a, \ldots, \boldsymbol{\Psi}_{n_B:k} + \mathbf{x}^A_a)$, we have proved the following result.

Proposition 1 *The conditional d.f. of* $\mathbf{X}^B_{\boldsymbol{b}(a)}$, *given* $\mathbf{X}^A_a = \mathbf{x}^A_a$, *is equal to*

$$
f_{\mathbf{X}^B_{\boldsymbol{b}(a)} | \mathbf{X}^A_a}(\mathbf{x}_1, \ldots, \mathbf{x}_k) = f_{\boldsymbol{\Psi}_{n_B:1} \ldots \boldsymbol{\Psi}_{n_B:k}}(\mathbf{x}_1 - \mathbf{x}^A_a, \mathbf{x}_2 - \mathbf{x}^A_a, ..., \mathbf{x}_k - \mathbf{x}^A_a), \tag{5}
$$

where $f_{\boldsymbol{\Psi}_{n_B:1} \ldots \boldsymbol{\Psi}_{n_B:k}}$ *is given by* (4).

The behaviour of the k donors as n_B increases is studied in Proposition 2.

Proposition 2 *Let* $\boldsymbol{\epsilon}$ *be a* P*-dimensional vector with all components equal to* ϵ. *Using the same notation as in Proposition 1 , and writing* $\mathbf{X}^B_{\boldsymbol{b}(a)} \in (\mathbf{x}^A_a - \boldsymbol{\epsilon}, \mathbf{x}^A_a + \boldsymbol{\epsilon})$ *if and only if* $\mathbf{X}^B_{\boldsymbol{b}_j(a)} \in (\mathbf{x}^A_a - \boldsymbol{\epsilon}, \mathbf{x}^A_a + \boldsymbol{\epsilon})$ *for every* $j = 1, \ldots, k$, *we have:*

$$
\lim_{n_B \to \infty} P(\mathbf{X}^B_{\boldsymbol{b}(a)} \notin (\mathbf{x}^A_a - \boldsymbol{\epsilon}, \mathbf{x}^A_a + \boldsymbol{\epsilon}) \,|\mathbf{X}^A_a = \mathbf{x}^A_a) = 0 \quad \forall\, \epsilon > 0.
$$

Proposition 2 implies that all k components of $\mathbf{X}^B_{\boldsymbol{b}(a)}$ are "close" to $\mathbf{X} = \mathbf{x}^A_a$. Hence, the conditional d.f. of each $\mathbf{Z}^B_{b_j(a)}$, given $\mathbf{X}^B_{\boldsymbol{b}(a)}$, is close to the conditional d.f. of $\mathbf{Z}$, given $\mathbf{X} = \mathbf{x}^A_a$, $j = 1, \ldots, k$. This does not imply that the conditional d.f. of $g(\mathbf{Z}^B_{b_1(a)}, \ldots, \mathbf{Z}^B_{b_k(a)})$, given $\mathbf{X}^B_{\boldsymbol{b}(a)}$ is close to the d.f. of $\mathbf{Z}$ given $\mathbf{X}$.

Example 1 *If g is the mean of $\mathbf{Z}^B_{b_j(a)}$s, $g(\mathbf{Z}^B_{b_1(a)}, \ldots, \mathbf{Z}^B_{b_k(a)})$ tends to the distribution of the sample mean of k i.i.d. copies of $\mathbf{Z}$, given $\mathbf{X}$.*

Example 2 *If g is a random draw from the k nearest neighbours $\mathbf{Z}^B_{b_j(a)}$, $j = 1, \ldots, k$, $g(\mathbf{Z}^B_{b_1(a)}, \ldots, \mathbf{Z}^B_{b_k(a)})$ tends to the distribution of $\mathbf{Z}$, given $\mathbf{X}$, i.e. the matching noise is null.*

Example 3 *An alternative imputation procedure, again based on kNN, could be the following. Assume a (multivariate) nonparametric regression model $\mathbf{Z} = h(\mathbf{X}) + \mathbf{U}$, with $E(\mathbf{U}|\mathbf{X}) = \mathbf{0}$, $E(\mathbf{U}\mathbf{U}'|\mathbf{X}) = \sigma^2\mathbf{I}$, $\mathbf{I}$ being the identity matrix. A simple idea consists in (i) estimating first $h(\mathbf{X})$ by a nonparametric estimator $\widehat{h}$; (ii) defining the residuals $\mathbf{e}^B_b = \mathbf{z}^B_b - \widehat{h}(\mathbf{x}^B_b)$, $b = 1, \ldots, n_B$; (iii) drawing at random a residual $\mathbf{e}^B_{b^*}$; (iv) imputing $\widetilde{z}^A_a = \widehat{h}(\mathbf{x}^A_a) + e^B_{b^*}$. If $\widehat{h}(\cdot)$ is a consistent estimator of $h(\cdot)$ than it is not difficult to see that the matching noise of this procedure vanishes as n_B increases. In this way we have defined a (nonparametric) class of consistent imputation methods. In section 5, the performance of this imputation technique is compared to the hot deck one in the special case of $\widehat{h} = kNN$ estimator of h.*

3.2 An important special case: distance hot-deck

Distance hot-deck is probably the most widely used imputation technique for matching. Each record in the recipient file A is matched with the closest record in the donor file B. Formally speaking, it consists in selecting, for each $a = 1, \ldots, n_A$, the donor $b_1(a) \in B$ such that

$$d(\mathbf{x}^A_a, \mathbf{x}^B_{b_1(a)}) = \min_{b \in B} d(\mathbf{x}^A_a, \mathbf{x}^B_b)$$

It can be shown (Paass (1985), Cohen (1991)) that distance hot-deck is equivalent to impute missing data through the conditional expectation of $\mathbf{Z}$ given $\mathbf{X}$ estimated by the (nonparametric) kNN nearest neighbour method, with $k = 1$. This is actually the most important theoretical justification of distance hot-deck.

Its main properties can be obtained by specializing the results in Section 3.1. More precisely, using the same notation as in Section 3.1, it is immediate to prove the following proposition, that allows the evaluation of the matching noise for distance hot-deck.

Proposition 3 *The conditional d.f. of $\mathbf{X}^B_{b_1(a)}$, given $\mathbf{X}^A_a = \mathbf{x}^A_a$, is equal to*

$$f_{\mathbf{X}^B_{b_1(a)}|\mathbf{x}^A_a}(\mathbf{x}) = f_{\mathbf{\Psi}_{n_B:1}}(\mathbf{x} - \mathbf{x}^A_a),$$

Distance hot-deck exhibits an important feature: its matching noise decreases as n_B increases. In fact, by particularizing Proposition 2, it is immediate to prove the following statement.

Proposition 4 *Using the same notation as in Proposition 2, we have:*

$$\lim_{n_B \to \infty} P(\mathbf{X}^B_{b_1(a)} \notin (\mathbf{x}^A_a - \boldsymbol{\epsilon},\, \mathbf{x}^A_a + \boldsymbol{\epsilon}) \,|\mathbf{X}^A_a = \mathbf{x}^A_a) = 0 \quad \forall\, \epsilon > 0.$$

Despite its similarity with Proposition 2, there is a fundamental difference. In fact, Proposition 4 tells us that, if n_B is large enough, then the matching noise is negligible because $\mathbf{X}^B_{b_1(a)}$ is "close" to $\mathbf{X} = \mathbf{x}^A_a$ with high probability, and hence the conditional distribution of $\widetilde{\mathbf{Z}}$, given $\mathbf{X}^B_{b_1(a)}$, is close to the conditional d.f. of $\mathbf{Z}$, given $\mathbf{X} = \mathbf{x}^A_a$. As discussed in Section 3.1, kNN method with $k > 1$ does not generally possess the same property.

4 d_0-Kernel hot-deck

In this section we describe a nonparametric imputation procedure characterized by a variable number of donors k, the d_0-Kernel. For each record in A, the available donor methods described in Section 3 select the k nearest neighbours with fixed k. As a consequence, some donors could be sparse, especially in the tails of the distribution of $\mathbf{X}$. In other words, the kNN method forces units far from the record x^A_a to be equally informative on z^A_a. Since the optimal value of k varies with x^A_a, an obvious extension of the kNN method consists in allowing a possibly different number of donors k for each record x^A_a. In order to accomplish this, we fix a threshold d_0: the records b in B having distance $d(\mathbf{x}^B_b, \mathbf{x}^A_a)$ smaller than d_0 are considered as neighbours of $\mathbf{x}^A_a$, $a = 1, \ldots, n_A$. As a matter of fact, the number $\widetilde{k}$ of neighbours of a has binomial distribution with parameters n_B and $\alpha(d_0) = P(W \leqslant d_0)$. Let $\mathbf{\Gamma}_{\widetilde{k}}$ be the r.v. taking the value $\boldsymbol{b}(a)$ of records $\mathbf{x}^B_b$ such that $d(\mathbf{x}^B_b, \mathbf{x}^A_a) \leqslant d_0$. Then, when $k \geqslant 1$ we have

$$
\begin{aligned}
&P(\widetilde{k} = k, \mathbf{\Gamma}_{\widetilde{k}} = \boldsymbol{b}(a), \mathbf{\Psi}_{n_B:j} \leqslant \boldsymbol{\psi}_j,\, j \leqslant \widetilde{k}) \\
&= \quad P(\widetilde{k} = k)\frac{1}{D_{n_B,k}} P(\mathbf{\Psi}_{n_B:j} \leqslant \boldsymbol{\psi}_j,\, j \leqslant \widetilde{k} | \widetilde{k} = k, \mathbf{\Gamma}_{\widetilde{k}} = \boldsymbol{b}(a)) \\
&= \quad P(\widetilde{k} = k)\frac{1}{D_{n_B,k}} P(\mathbf{\Psi}_{b_j(a)} \leqslant \boldsymbol{\psi}_j,\, j \leqslant k,\, W_t \geqslant W_{b_k(a)}, t \notin \boldsymbol{b}(a) | \\
&\qquad W_{b_j(a)} \leqslant d_0,\, j \leqslant k, W_t > d_0,\, t \notin \boldsymbol{b}(a)) \\
&= \quad P(\widetilde{k} = k)\frac{1}{D_{n_B,k}} P(\mathbf{\Psi}_{b_j(a)} \leqslant \boldsymbol{\psi}_j,\, j \leqslant k | W_{b_j(a)} \leqslant d_0,\, j \leqslant k) \\
&= \quad \frac{P(\widetilde{k} = k)}{D_{n_B,k} P(W \leqslant d_0)^k} \int_{\mathbb{T}_k \cap (\bigcap_{j=1}^k (-\infty, \boldsymbol{\psi}_j])} \prod_{j=1}^k f_{\mathbf{\Psi}}(\mathbf{x}_j) d\mathbf{x}_j \qquad (6)
\end{aligned}
$$

where $\mathbb{T}_k = \{(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, .., \boldsymbol{\psi}_k) : \boldsymbol{\psi}_1' D \boldsymbol{\psi}_1 \leqslant \boldsymbol{\psi}_2' D \boldsymbol{\psi}_2 \leqslant ... \leqslant \boldsymbol{\psi}_k' D \boldsymbol{\psi}_k \leqslant d_0\}$. Taking into account that there are no donors when $\widetilde{k} = 0$, from (6) it is not difficult to compute the following probability

$$
\begin{aligned}
P(\mathbf{\Psi}_{n_B:j} \leqslant \boldsymbol{\psi}_j,\, j \leqslant \widetilde{k} | \widetilde{k} \geqslant 1) &= \sum_{k \geqslant 1} P(\mathbf{\Psi}_{n_B:j} \leqslant \boldsymbol{\psi}_j,\, j \leqslant \widetilde{k}, \widetilde{k} = k | \widetilde{k} \geqslant 1) \\
&= \frac{1}{P(\widetilde{k} \geqslant 1)} \sum_{k \geqslant 1} P(\mathbf{\Psi}_{n_B:j} \leqslant \boldsymbol{\psi}_j,\, j \leqslant \widetilde{k}, \widetilde{k} = k) \qquad (7)
\end{aligned}
$$

where

$$
P(\mathbf{\Psi}_{n_B:j} \leqslant \boldsymbol{\psi}_j,\, j \leqslant \widetilde{k}, \widetilde{k} = k) = \frac{P(\widetilde{k} = k)}{P(W \leqslant d_0)^k} \int_{\mathbb{T}_k \cap (\bigcap_{j=1}^k (-\infty, \boldsymbol{\psi}_j])} \prod_{j=1}^k f_{\mathbf{\Psi}}(\mathbf{x}_j) d\mathbf{x}_j \quad (8)
$$

From (7) it is possible to derive the distribution function of the donors $\mathbf{X}^B_{\boldsymbol{b}(a)}$, and hence the matching noise. The marginal d.f. of $(\boldsymbol{\Psi}_{n_B:1}, \boldsymbol{\Psi}_{n_B:2}, ..., \boldsymbol{\Psi}_{n_B:\widetilde{k}} \mid \widetilde{k} \geqslant 1)$ is given by

$$f_{\boldsymbol{\Psi}_{n_B:j}, j \leqslant \widetilde{k}}(\boldsymbol{\psi}_j, j \leqslant \widetilde{k} \mid \widetilde{k} \geqslant 1) = \frac{1}{P(\widetilde{k} \geqslant 1)} \sum_{k \geqslant 1} \frac{P(\widetilde{k} = k)}{P(W \leqslant d_0)^k} \prod_{j=1}^{k} f_{\boldsymbol{\Psi}}(\boldsymbol{\psi}_j). \tag{9}$$

Finally, taking into account that $\mathbf{X}^B_{\boldsymbol{b}(a)} = (\mathbf{X}^B_{b_1(a)}, \ldots, \mathbf{X}^B_{b_{\widetilde{k}}(a)})$ coincides with $(\boldsymbol{\Psi}_{n_B:1} + \mathbf{x}^A_a, \ldots, \boldsymbol{\Psi}_{n_B:\widetilde{k}} + \mathbf{x}^A_a)$, we have proved the following result.

Proposition 5 *The conditional d.f. of* $\mathbf{X}^B_{\boldsymbol{b}(a)}$*, given* $\mathbf{X}^A_a = \mathbf{x}^A_a$*, is equal to*

$$f_{\mathbf{X}^B_{\boldsymbol{b}(a)} | \mathbf{X}^A_a}(\mathbf{x}_j, j \leqslant \widetilde{k} \mid \widetilde{k} \geqslant 1) = f_{\boldsymbol{\Psi}_{n_B:j}, j \leqslant \widetilde{k}}(\mathbf{x}_j - \mathbf{x}^A_a, j \leqslant \widetilde{k} \mid \widetilde{k} \geqslant 1), \tag{10}$$

where $f_{\boldsymbol{\Psi}_{n_B:j}, j \leqslant \widetilde{k}}$ *is given by* (9).

Differently from the kNN method the results in Proposition 2 can not be extended to the d_0-Kernel method, unless d_0 goes to zero appropriately as n_B goes to infinity.

5 A simulation study

In the previous sections we have formally evaluated the matching noise for a set of widely used nonparametric imputation techniques. In order to compare the matching noise of these different imputation methods we have performed a simulation experiment. In detail, we have randomly generated 500 i.i.d records from a bivariate normal distribution (X, Z) with means 1, 3, and variances 5, 4, respectively, and covariance 3. Let the recipient file A consist of these 500 observations, with Z dropped. The simulation analysis involves the following steps:

- *Step 1 :* A donor sample B composed by n_B i.i.d. records is drawn from the same bivariate distribution. Different values of n_B have been used, $n_B = 100 - 1000/100$.
- *Step 2 :* the missing Zs have been imputed by the following imputation techniques.
  1. Distance hot deck, with $d(x^A_a, x^B_b) = |x^A_a - x^B_b|$ (Section 3.2);
  2. kNN with $k = \sqrt{n_B}$, and $g(.)$ corresponding to the mean function (mean kNN, see Example 1) and to a random draw (random kNN, see Example 2), respectively . The value of k has been chosen according to Silverman (Silverman (1986), page 19) .
  3. kNN estimator of $h(X) = E(Z|X)$ plus random residual, as defined in Example 3.
  4. d_0-kernel with d_0 chosen to minimize the asymptotic Mean Square Error of the local kernel density function estimator of X (Section 4).
- *Step 3 :* steps 1 to 2 are repeated 400 times.

In order to evaluate the closeness between the data generating model and the imputation generating model a divergence measure between the two distributions $f_{X\widetilde{Z}}(x,z)$ and $f_{XZ}(x,z)$ should be introduced. Such a divergence (matching noise) has been evaluated by the Kolmogorov-Smirnov distance (KS). We begin by evaluating to which extent the imputation procedures are able to recovery the marginal distribution of Z in the synthetic data file. Formally speaking, for each donor sample v (for $v = 1, 2, \ldots, 400$), KS distance compares the empirical distribution of imputed values $\widetilde{Z}$ in A ($\hat{F}_{\widetilde{Z},v}(z)$) with the hypothesized distribution ($F_0(z)$). A mean of such values over the 400 donor files is then taken as a global divergence measure, namely :

$$KS_Z = \frac{1}{400}\sum_{v=1}^{400} KS_Z(v) = \frac{1}{400}\sum_{v=1}^{400}\left[\sup_{-\infty<z<\infty} \mid F_0(z) - \hat{F}_{\widetilde{Z},v}(z) \mid\right] \tag{11}$$

Moreover, in order to analyze the matching noise as n_B increases, (11) has been computed for different donor file sizes n_B (Figure 1).

Note that for all nonparametric imputation techniques, the matching noise decreases as the donor file size n_B increases. Loosely speaking, the mean kNN+ residual technique seems to perform slightly better than other donor methods.

The mean kNN is the worst method. This imputation technique underestimates variability, and this worsens as k increases. In fact, the replacement of the expected value of k nearest neighbors to each missing item implies that the synthetic distribution of $Z \mid X$ is concentrated on the expected value of $Z \mid X$.

Note that (11) evaluates the ability of the donor method to reproduce the marginal distribution of Z in the synthetic data set. In order to get information on the closeness of the two distributions $f_{X\widetilde{Z}}(x,z)$ and $f_{XZ}(x,z)$, the Kolmogorov-Smirnov distance has been computed between the empirical conditional distribution of $Z \mid X = x_a^A$, $a = 1, \ldots, 500$ ($\hat{F}_{\widetilde{Z}|x_a^A}(z)$) and the hypothesized distribution ($F_{0|x_a^A}(z)$). A mean of the $n_A = 500$ values is then taken :

$$E[KS_Z^X] \approx \frac{1}{500}\sum_{a=1}^{500} KS_Z(x_a^A) = \frac{1}{500}\sum_{a=1}^{500}\left[\sup_{-\infty<z<\infty} \mid F_{0|x_a^A}(z) - \hat{F}_{\widetilde{Z}|x_a^A}(z) \mid\right] \tag{12}$$

The results are reported in Figure 2. The distance hot deck and random kNN methods seem to perform better. Furthermore their matching noise is approximately constant as n_B increases. In conclusion Figures 1 and 2 suggest that the imputation techniques give equivalent results with the exception of the mean kNN method. More precisely, in both figures the distances between the matching noise curves associated to distance hot deck, random kNN and mean kNN+residual methods are so small that further analyses are required to evaluate their performance. These should be carried out increasing both the number of samples and their sizes, taking in mind that even the use of a high-speed computer can lead to excessively long calculation times. The same consideration holds for the Figures 3 and 4 regarding the d_0-Kernel method.

The nonparametric imputation methods defined in step 2 depend on a single parameter k, the number of nearest neighbors. These techniques operate given the donation class size. As already mentioned, an obvious extension is to consider a

different number of donors for each record. More precisely, we study the d_0-Kernel hot deck described in Section 4: the main problem is how to choose d_0.

Since the simulation study is carried out to compare the performance of different imputation techniques, we use knowledge on the data generating mechanism to define d_0. Its value has been chosen to minimize the asymptotic Mean Square Error of the local kernel density function estimator of X and it is proportional to $n^{-1/5}$ (Silverman (1986)). Clearly in applied studies the choice of d_0 is more complicated.

The effect of the d_0-Kernel method to avoid far away donors from x_a^A is evident in the tails of the distribution. As a consequence, we expect a better performance of the d_0-Kernel in terms of matching noise, since the method should be able to dose the optimal amount of smoothing across the real line.

By comparing the mean kNN+residual technique both with fixed and variable number of donors k, it appears that the latter performs slightly better than the former. On the other side, the mean imputation method performs worse than the method with a fixed number of donors k. In fact, since the expected value of the random variable $\widetilde{k}$ is bigger than $\sqrt{n_B}$, the d_0-Kernel technique averages on a larger number of neighbours.

Instead of the full distribution of (X, Z), it is possible to measure the performance of the imputation procedures when the aim is the estimation of some statistically meaningful parameters, such as the expectation of Z (μ_Z, say), and the correlation coefficient between the variables of interest ($\rho_{X,Z}$, say). For each nonparametric imputation technique, both with fixed and variable number of donors k, the performance of the sample mean and the sample correlation coefficient has been evaluated in terms of Mean Square Error (Figures 5 and 6).

Figure 5 shows that the mean kNN technique better estimates μ_Z, since it generates the best point estimate with regard to a quadratic loss function. This result can be considered as the nonparametric counterpart of results in Buck (1969). Nevertheless, it should not be consider as a good matching method for a general purpose matched file $(X, \widetilde{Z})$. For instance, this is confirmed by the Mean Square Error of the correlation coefficient estimator (Figure 6). The imputation mean method does not preserve the relationship between the variables of interest in the synthetic complete data set. The distance hot deck and random kNN methods preserve the relation between X and Z slightly better than the other methods.

Note that this analysis does not yet allow us to discriminate between the d_0-Kernel and the kNN methods because the results for both the sample mean and the sample correlation coefficient are essentially the same.

There are two practical improvements that we plan to pursue in a future work. The first one is to implement further analyses increasing both the number of samples and their sizes. A second application is to investigate the use of nonparametric regression methods as the local polynomial regression (Härdle (1990)), as an additional method to reconstruct the synthetic data file.

References

Buck, W. (1960). A method of estimation of missing values in multivariate data suitable for use with electronic computer. *Journal of the Royal Statistical Society*, 22, 302-306.

Cohen, M.L. (1991). Statistical matching and microsimulation models. *Improving Information for Social Policy Decisions, the Use of Microsimulation Modeling*,Technical Papers, II, National Academy Press.

D'Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical Matching: Theory and Practice*(John Wiley & Sons, Chichester).

Härdle, W. (1990). *Applied nonparametric regression* (Cambridge University Press, New York).

Haziza, D. (2001). Why do we impute? *The Imputation Bulletin*, 1, 3-4.

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

Paass, G. (1986). Statistical Match: Evaluation of Existing Procedures and Improvements by Using Additional Information. *Microanalytic Simulation Models to Support Social Fiscal Policy (Eds. G.H. Orcutt, J. Merz, H. Quinke)*. Elsevier Science, Amsterdam, 401-420.

Paass, G. (1985). Statistical record linkage methodology, state of the art and future prospects. *Bulletin of the International Statistical Institute, Proceedings of the 45th Session*, LI, Book 2.

Rässler, S. (2002). *Statistical Matching: a Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*(Springer Verlag, New York).

Rodgers, W.L. (1984). An Evaluation of Statistical Matching. *Journal of Business and Economic Statistics*, 2, 91-102.

Rubin, D.B. (1974). Characterizing the Estimation of Parameters in Incomplete-Data Problems. *Journal of the American Statistical Association*, 69, 467-474.

Rubin, D.B. (1976). Inference and missing data.*Biometrika*, 63, 581-592.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*(Chapman & Hall, London).

Sims, C.A. (1972). Comments on: "Constructing a New Data Base From Existing Microdata Sets: the 1966 Merge File", by B.A. Okner. *Annals of Economic and Social Measurements*, 1, 343-345.

Titterington, D.M. and Sedransk J., (1989). Imputation of missing values using density estimation. *Statistics & Probability Letters*, 8, 411-418.

LIST OF FIGURES

Figure 1: *Kolmogorov-Smirnov distance KS_Z for distance hot deck, kNN mean, random kNN and mean kNN +residual with $k = \sqrt{n_B}$.*

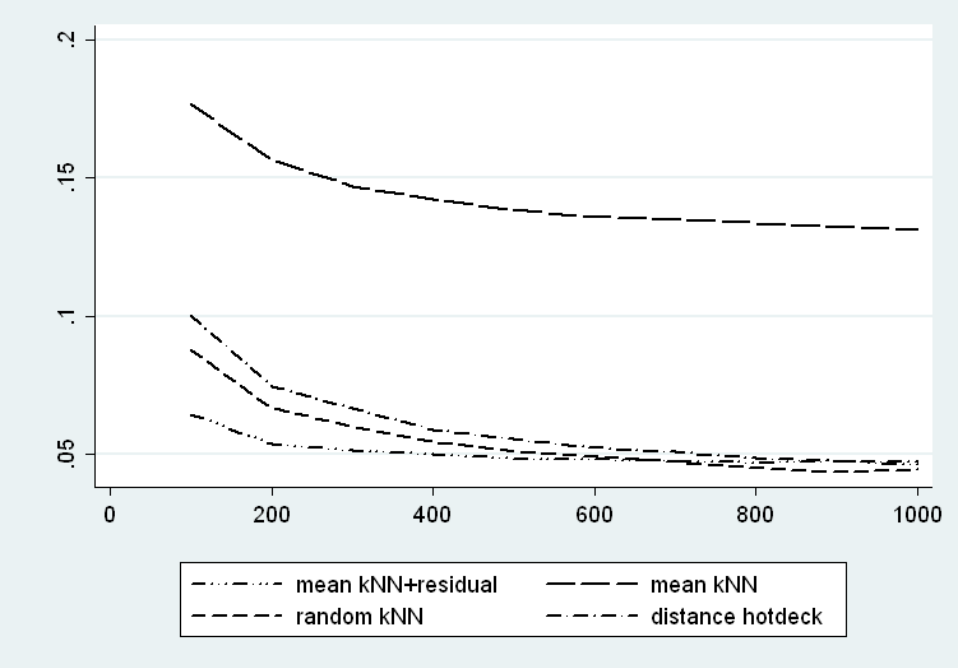

Figure 2: *Kolmogorov-Smirnov distance* $E[KS_Z^X]$ *for distance hot deck, kNN mean, random kNN and mean kNN+residual with* $k = \sqrt{n_B}$.

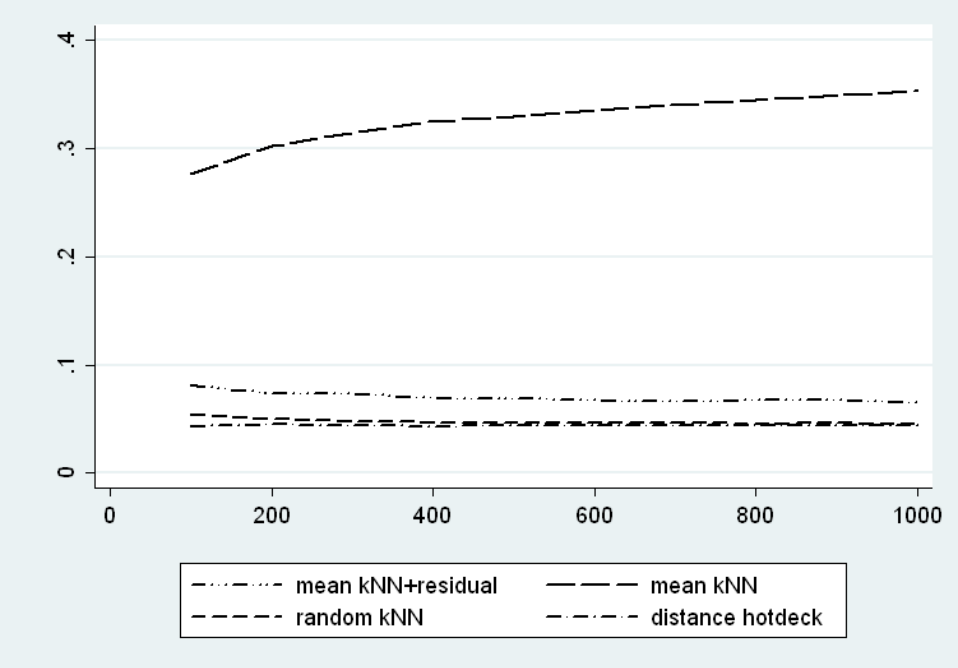

Figure 3: *Kolmogorov-Smirnov distance KS_Z for distance hot deck, kNN mean, random kNN and mean kNN +residual with d_0-Kernel .*

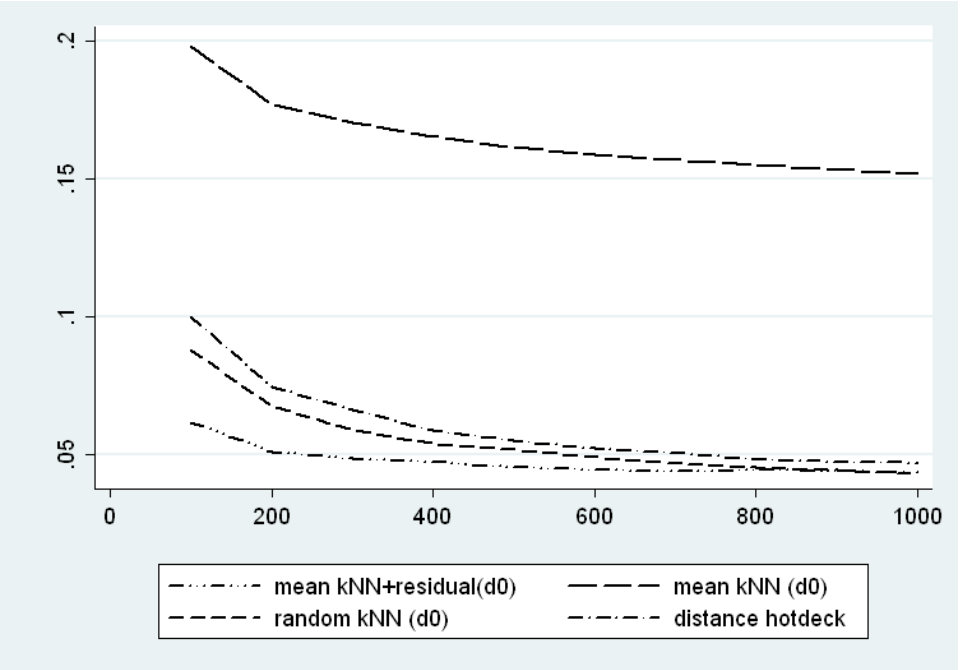

Figure 4: *Kolmogorov-Smirnov distance* $E[KS_Z^X]$ *for distance hot deck, kNN mean, random kNN and mean kNN +residual with* d_0*-Kernel .*

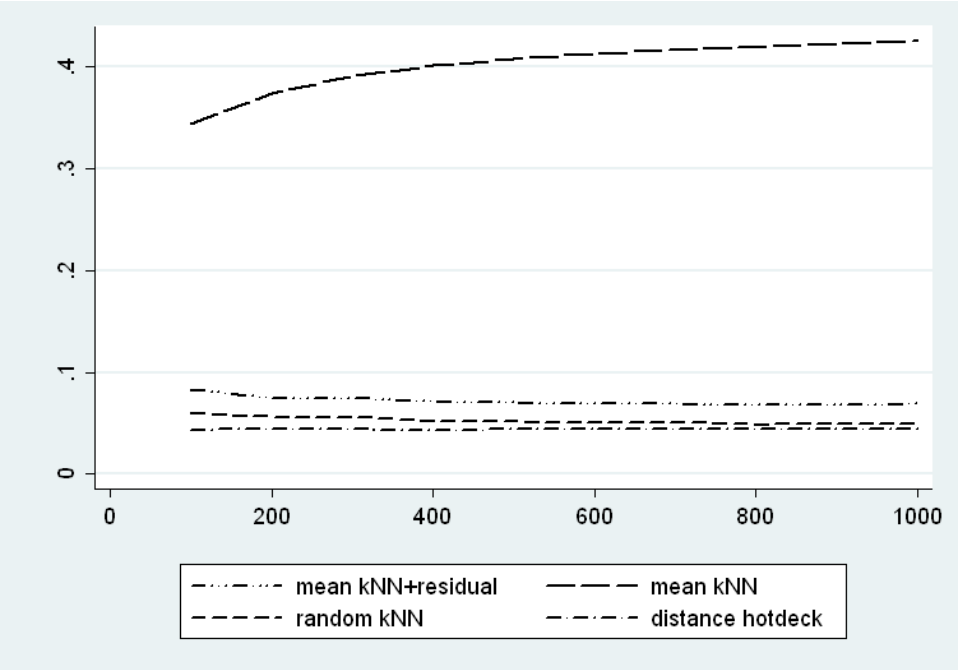

Figure 5: *Mean Square Error of sample mean for distance hot deck, kNN mean, random kNN and mean kNN +residual with* $k = \sqrt{n_B}$.

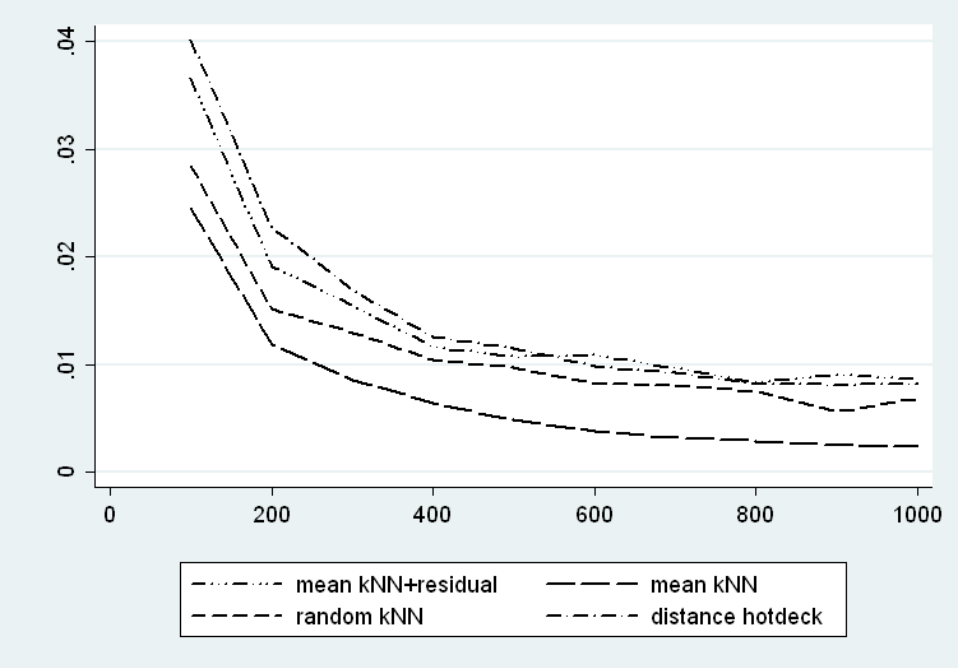

Figure 6: *Mean Square Error of sample correlation coefficient for distance hot deck, kNN mean, random kNN and mean kNN +residual with* $k = \sqrt{n_B}$.