Evaluation of matching noise for imputation techniques based on nonparametric regression function

Pier Luigi Conti

Università "La Sapienza", piazzale Aldo Moro 5, 00185 Roma

Daniela Marella *

Università "La Sapienza", piazzale Aldo Moro 5, 00185 Roma

Mauro Scanu

ISTAT, via Cesare Balbo 16, 00184 Roma

Abstract

The aim of the paper is to compare the performance of a class of nonparametric imputation procedures in the simplified context of statistical matching. This class includes both hot deck methods and two procedures based on estimating the regression function between the variables of interest, namely the kNN estimator and the local linear estimator. The regression function is assumed not necessarily linear. Performance is measured by the matching noise given by the discrepancy between the distribution generating genuine data and the distribution generating imputed values.

Key words: statistical matching, missing data, kNN method, local linear regression estimator.

1 Introduction

Let (X, Z) be a bivariate random variable (r.v.) with density function $f(x, z)$, and let A, B be two independent samples of n_A and n_B i.i.d. records from

* Corresponding author.
Email addresses: pierluigi.conti@uniroma1.it (Pier Luigi Conti), daniela.marella@uniroma1.it (Daniela Marella), scanu@istat.it (Mauro Scanu).

(X, Z), respectively, where n_A and n_B are fixed in advance by design. The first n_A records have Z missing while the last n_B records are complete. Hence,

$$\begin{aligned} \left(\mathbf{x}^A\right) &= \left(x_1^A, \ldots, x_{n_A}^A\right) \\ \left(\mathbf{x}^B, \mathbf{z}^B\right) &= \left((x_1^B, z_1^B), \ldots, (x_{n_B}^B, z_{n_B}^B)\right), \end{aligned} \tag{1}$$

are the observed values in A and B, respectively. This is the typical situation in statistical matching where missingness is induced by design. Under this framework, it can be easily proved that the missing data generation process is missing completely at random, see [5].

The missing Z values in A are usually imputed by means of appropriate imputation procedures. The most popular are those based on hot deck, *i.e.* missing Z values are replaced by actually observed values chosen appropriately among the n_B complete records in B. Usually, donor values are selected according to a distance between observed and incomplete records on X [1]. Two of the most popular procedures are distance and random hot deck imputation. Hot deck methods have been largely studied in the statistical literature, see [8] and [9]. By far, distance hot deck is the most used. Generally speaking, hot deck methods have attractive properties: (*i*) they are nonparametric, because they do not need any explicit definition of a parametric data generation model; (*ii*) they impute "live values", *i.e.* actually observed values; (*iii*) they are able to reproduce the marginal and conditional distributions of the variable to impute quite well (at least for large samples).

As a matter of fact, hot deck methods are not the only nonparametric procedures that it is possible to use for imputing missing values. In this paper we investigate the performance of some procedures based on the nonparametric estimation of the joint distribution of observed and missing variables. Such procedures are compared by means of their matching noise given by the discrepancy between the distribution generating genuine data and the distribution generating imputed data [14]. If these two distributions coincide, the imputed data set can be analyzed as if it was a completely observed data set generated by the distribution generating genuine data (the joint distribution of (X, Z)). Otherwise estimators based on the complete synthetic data set could be inappropriate for inferring properties of the model underlying data. An example of study of the matching noise for a class of nonparametric imputation procedures based on kNN methods (including distance hot deck) is in [12]. In that case, comparisons are performed when $f(x, z)$ is a bivariate normal density. The result was that mean kNN plus random residual method (Section 2.3.1) performs better in reconstructing the marginal distribution of Z. When the interest is in the conditional distribution of $Z \mid X$ distance and random hot deck seem to perform quite well.

In this paper we consider several different nonparametric imputation techniques: distance hot deck, random kNN and two stochastic imputation procedures based on estimating the regression function of Z given X, namely the kNN estimator and the local linear regression estimator. The matching noise of such procedures is evaluated by means of appropriate simulations under different models for (X, Z), each characterized by different regression functions of Z on X and homoschedasticity. In the sequel imputed data is denoted with the r.v. $(X, \widetilde{Z})$.

The paper is organized as follows. In Section 2 a class of nonparametric imputation procedures both deterministic and stochastic are described. In Section 3 the matching noise is formally evaluated. Finally, in Section 4 a simulation study is implemented.

2 Nonparametric imputation procedures

In order to appropriately impute missing data, the model that generates imputations should equal the data generating model: the distribution of $(X, \widetilde{Z})$ should coincide with the distribution of (X, Z). Either implicitly or explicitly, the model that generates imputations is estimated from the observed data. In the case of the data set (1), the model should be estimated from the donor file B.

In the sequel a short description of widely used imputation methods is given. Formally, for each $a = 1, \ldots, n_A$, let $\mathbf{b}(a) = (b_1(a), \ldots, b_k(a))$ be the labels of the $k \geq 1$ nearest neighbours of x_a^A in B, such that

$$d(x_a^A, x_{b_1(a)}^B) \leq \cdots \leq d(x_a^A, x_{b_k(a)}^B), \quad d(x_a^A, x_{b_k(a)}^B) \leq d(x_a^A, x_b^B)$$

$\forall\, b \notin \{b_1(a), \ldots, b_k(a)\}$, where $d(.,.)$ is the Euclidean distance. Let $\mathbf{x}_{\mathbf{b}(a)}^B = (x_{b_1(a)}^B, x_{b_2(a)}^B, \ldots, x_{b_k(a)}^B)$ and $\mathbf{z}_{\mathbf{b}(a)}^B = (z_{b_1(a)}^B, z_{b_2(a)}^B, \ldots, z_{b_k(a)}^B)$ the vectors of corresponding X and Z values, respectively.

2.1 kNN random hot deck

Once the k nearest neighbours of x_a^A, $x_{\mathbf{b}(a)}^B$, are obtained, one could impute the missing z_a^A by randomly choosing a label $\tilde{b}(a)$ among $b_j(a)$, $j = 1, \ldots, k$, and in taking imputed values

$$\tilde{z}_a^A = z_{\tilde{b}(a)}^B, \qquad a = 1, \ldots, n_A. \tag{2}$$

A generalized version of this approach is in [1]. A value is taken at random assuming different probabilities of selection for the donor records: observations close to x_a^A have higher probabilities than those further away.

2.2 Distance hot deck

When $k = 1$, the imputation method described in Section 2.1 reduces to distance hot deck. Imputed data are obtained as:

$$\tilde{z}_a^A = z_{b_1(a)}^B, \qquad a = 1, \ldots, n_A. \tag{3}$$

In other words, each record in A is matched with the closest record in B.

2.3 Methods based on nonparametric regression function

Since [18], when X and Z are continuous a very important role has been played by the regression function of Z on X. More precisely, a linear regression function is assumed. A simple (and natural, as well) idea to impute missing data consists in using a nonparametric estimator of the (not necessarily linear) regression function of Z on X (see, for instance, [13]). Suppose that X and Z are related through the relationship

$$Z = m(X) + \epsilon \tag{4}$$

where $m(x) = E[Z\,|X = x]$ is the regression function of Z given X and $\epsilon = Z - m(X)$ is the error term, such that $E[\epsilon\,|X = x] = 0$ for every x. For the sake of simplicity, in the sequel we will further assume that the errors are homoschedastic, *i.e.* $V[\epsilon\,|X = x] = \sigma^2$ independent of x. A simple idea to impute missing data $\mathbf{z}^A$ in the sample A could consist of the following steps.

1. Estimate the regression function $m(x)$ by the sample B. From now on, such an estimator will be denoted by $\widehat{m}^B(x)$.

2. Let

$$\widehat{\epsilon}_b^B = z_b^B - \widehat{m}^B(x_b^B), \qquad b = 1, \ldots, n_B, \tag{5}$$

be the corresponding residuals in B.

3. Impute the missing z_a^As by

$$\tilde{z}_a^A = \widehat{m}^B(x_a^A) + \tilde{\epsilon}^B, \qquad a = 1, \ldots, n_A, \tag{6}$$

where $\tilde{\epsilon}^B$ is drawn at random among $\hat{\epsilon}_1^B, \ldots, \hat{\epsilon}_{n_B}^B$.

The rationale of steps 1-3 is simple: at first model (4) is estimated by the complete sample B, and then used to impute the missing data z_a^As in A. According to [2] this is a *stochastic* imputation method. Clearly, if estimated residuals $\tilde{\epsilon}^B$ are omitted in (6), so that

$$\tilde{z}_a^A = \widehat{m}^B(x_a^A), \qquad a = 1, \ldots, n_A, \tag{7}$$

then the imputation method is *deterministic*. In the sequel, a short description of two imputation procedures based on estimating the regression function $m(x)$ through the kNN estimator and the local linear regression estimator is given.

2.3.1 kNN methods

The kNN imputation method consists in estimating the nonparametric regression function $m(x)$ by the kNN method. Formally, the regression function $m(x)$ is estimated by the average of Z corresponding to the k nearest neighbours of x. When $x = x_a^A$:

$$\widehat{m}^B(x_a^A) = \frac{1}{k} \sum_{j=1}^{k} z_{b_j(a)}^B, \qquad a = 1, \ldots, n_A.$$

Deterministic imputation computed from the estimated nonparametric regression function is:

$$\tilde{z}_a^A = \widehat{m}^B(x_a^A), \qquad a = 1, \ldots, n_A. \tag{8}$$

The corresponding stochastic imputation is obtained by

$$\tilde{z}_a^A = \widehat{m}^B(x_a^A) + \tilde{\epsilon}^B, \qquad a = 1, \ldots, n_A, \tag{9}$$

where $\tilde{\epsilon}^B$ is chosen at random from the residuals computed as in (5) on file B. The key point in using the kNN estimator (8) is the choice of the parameter k, that determines the amount of smoothing of z_b^Bs data. It plays a role similar to the bandwidth for kernel smoothers. A software for imputing missing data through the use of different methods based on the selection of k nearest neighbours (including also the procedures in Sections 2.2 and 2.1) has been recently developed [1].

It can be shown ([14], [3]) that distance hot-deck described in Section 2.2 is equivalent to impute missing data through the kNN method, with $k = 1$. Such a procedure seems to be at first sight a deterministic technique, because residuals estimated as in Equation (5) are null. As a matter of fact this method imputes at the same time both the regression function and the residual. However, this does not mean that the matching noise is null. It can be proved (see [12]) that the matching noise still affects this imputation approach for finite n_B, although it becomes negligible for large n_B.

2.3.2 Local polynomial estimator

As an alternative to kNN estimator, the *local polynomial estimators* [6] represent a simple and useful class of estimators of the regression function $m(x)$. Suppose that $m(x)$ possesses $p+1$ derivatives, and denote by $m^{(j)}(x)$ its jth derivative, $j = 1, \ldots, p+1$. The basic idea consists in approximating $m(t)$ locally by a polynomial of order p:

$$\begin{aligned} m(t) &\approx m(x) + m^{(1)}(x)(t-x) + \cdots + \frac{1}{p!} m^{(p)}(x)(t-x)^p \\ &= \beta_0 + \beta_1(t-x) + \cdots + \beta_p(t-x)^p. \end{aligned} \tag{10}$$

Model (10) may be considered as a "usual" polynomial model on a local scale, with parameters $\beta_0, \ldots, \beta_p$ depending on x. They may be estimated by the weighted least squares method, which consists in minimizing the quantity:

$$\sum_{b=1}^{n_B} \left(Z_b^B - \sum_{j=0}^{p} \beta_j (X_b^B - x)^j \right)^2 K_h\left(X_b^B - x\right)$$

where $K(\cdot)$ is a non-negative weight function, $K_h(t) = h^{-1}K(t/h)$, and h (the bandwidth) is a smoothing parameter determining the size of the neighbourhood of x used in estimating $m(x)$.

Local polynomial estimators have been proved as particularly useful, and efficient as well. Their merits are thoroughly discussed in [6]. In particular, when $p = 0$ the local polynomial estimator reduces to the Nadaraya-Watson estimator, that may be written as:

$$\widehat{m}_0^B(x) = \frac{\sum_{b=1}^{n_B} Z_b^B K_h\left(X_b^B - x\right)}{\sum_{b=1}^{n_B} K_h\left(X_b^B - x\right)} \tag{11}$$

When $p = 1$, the local polynomial estimator reduces to the *local linear estimator*, that may be written in the form:

$$\widehat{m}_1^B(x) = \frac{S_2^B(x)\, T_0^B(x) - S_1^B(x)\, T_1^B(x)}{S_2^B(x)\, S_0^B(x) - S_1^B(x)^2} \tag{12}$$

where

$$S_j^B(x) = \sum_{b=1}^{n_B} (x - X_b^B)^j \, K_h\left(X_b^B - x\right)$$
$$T_j^B(x) = \sum_{b=1}^{n_B} Z_b^B \, (x - X_b^B)^j \, K_h\left(X_b^B - x\right)$$

as $j = 0,\ 1,\ 2$.

The local linear estimator (12), if compared to the Nadaraya-Watson estimator (11), does have several advantages. First of all, it does not suffer of the so-called "boundary effect" [6], consisting in being severely inefficient when x is close at the extremes of its range. Secondly, since it is based on a first-order local fit, it does not really need to assume that the variance $V[\epsilon\,|X = x]$ is independent of x, because it is approximately the same in a local neighbourhood of x.

A crucial element, in determining the performance of the local polynomial estimator, is the choice of the bandwidth h. This point will be discussed in the simulation study of Section 4.

3 Evaluation of the matching noise for the imputation procedure based on local polynomial regression estimator

One of the key issues in order to assess the accuracy of imputation procedures is to study the discrepancy between the distribution that generates genuine data (*i.e.* the distribution of (X, Z)) and the distribution that generates imputed data (*i.e.* the distribution of $(X, \widetilde{Z})$). For all the imputation procedures described in Section 2, based on donors selected according to a distance with the recipient x_a^A, and from the independence of different observations, it turns out that the distribution of $(X, \widetilde{Z})$ is given by:

$$f_{X_a^A, \widetilde{Z}_a}(x, z) = \int f_{X_a^A X_{\mathbf{b}(a)}^B \widetilde{Z}_a}(x, \mathbf{t}, z) d\mathbf{t} = f_X(x) \int f_{X_{\mathbf{b}(a)}^B | X_a^A}(\mathbf{t}|x) f_{Z|X_{\mathbf{b}(a)}^B} d\mathbf{t}$$

where $\mathbf{t}$ is a vector of dimension $k \geq 1$. As a matter of fact, if X is categorical, A and B observe all the categories of X, and distance hot deck is considered, the matching noise is null. Generally speaking, a continuous X does not allow the definition of an imputation procedure with a null matching noise. The matching noise will depend on two elements:

- the distance between the recipient x_a^A and the donors $x_{\mathbf{b}(a)}^B$;

- how $\widetilde{Z}$ is defined as a function of the observed nearest records $x^B_{\mathbf{b}(a)}$.

In [12] the matching noise that affects kNN method is determined. It is proved that $X^B_{\mathbf{b}(a)}|X^A_a$ converges in distribution to k-dimensional vector whose elements are equal to x^A_a. Hence, stochastic kNN (9), distance hot deck (3) and selection of a random element from the k nearest neighbours (2) tend asymptotically to be matching noise free, while deterministic kNN (8) is unavoidably biased. In the sequel, we will prove that a similar result holds for imputation techniques based on local linear regression estimator.

Proposition 1 *Assume that the model (4) holds, with* (Z_b, X_b), $b = 1, \ldots, n_B$ *i.i.d. random variables. Assume further that* $\widehat{m}(\cdot)$ *tends in probability to* $m(\cdot)$ *as* n_B *goes to infinity, and that* $F_\epsilon(x) = Pr(\epsilon \leq x)$ *is continuous. If*

$$\hat{F}_{n_B}(x) = \frac{1}{n_B} \sum_{b=1}^{n_B} I_{(\hat{\epsilon}_b \leq x)} \tag{13}$$

is the empirical distribution function (e.d.f.) based on the residuals $\hat{\epsilon}_b$*s, then*

$$sup_x |\widehat{F}_{n_B}(x) - F_\epsilon(x)| \tag{14}$$

converges in probability to zero as n_B *goes to infinity.*

Proof Let $F_{n_B}(x)$ be the empirical distribution function based on the errors ϵ_bs:

$$F_{n_B}(x) = \frac{1}{n_B} \sum_{b=1}^{n_B} I_{(\epsilon_b \leq x)}. \tag{15}$$

We first show that

$$|\widehat{F}_{n_B}(x) - F_{n_B}(x)| \tag{16}$$

converges in probability to zero pointwise as n_B goes to infinity. First of all, it is not difficult to see that

$$E[|\hat{F}_{n_B}(x) - F_{n_B}(x)|] = E\left[\frac{1}{n_B}\left|\sum_{b=1}^{n_B}(I_{(\hat{\epsilon}_b \leq x)} - I_{(\epsilon_b \leq x)})\right|\right]$$

$$\leq \frac{1}{n_B} \sum_{b=1}^{n_B} E|(I_{(\hat{\epsilon}_b \leq x)} - I_{(\epsilon_b \leq x)})|$$

$$= \frac{1}{n_B} \sum_{b=1}^{n_B} Pr(|(I_{(\hat{\epsilon}_b \leq x)} - I_{(\epsilon_b \leq x)})| = 1)$$

$$= \frac{1}{n_B} \sum_{b=1}^{n_B} [Pr(\hat{\epsilon}_b \leq x, \epsilon_b > x) + Pr(\hat{\epsilon}_b > x, \epsilon_b \leq x)] \tag{17}$$

where $|(I_{(\hat{\epsilon}_b \leq x)} - I_{(\epsilon_b \leq x)})| = 1$ if either $(\epsilon_b \leq x, \hat{\epsilon}_b > x)$ or $(\epsilon_b > x, \hat{\epsilon}_b \leq x)$, and $|(I_{(\hat{\epsilon}_b \leq x)} - I_{(\epsilon_b \leq x)})| = 0$ otherwise. Next, we may write

$$\begin{aligned}
Pr(\hat{\epsilon}_b \leq x, \epsilon_b > x) &= Pr(z_b - \widehat{m}(x_b) \leq x, z_b - m(x_b) > x) \\
&= Pr(z_b \leq \widehat{m}(x_b) + x, z_b > m(x_b) + x) \\
&= Pr(m(x_b) + x < z_b \leq \widehat{m}(x_b) + x) \\
&= E_{x_b}[Pr(m(x_b) + x < z_b \leq \widehat{m}(x_b) + x | x_b)] \\
&= E_{x_b}[Pr(m(x_b) + x < z_b \leq \widehat{m}(x_b) + x, |\widehat{m}(x_b) - m(x_b)| < \delta | x_b)] \\
&\quad + E_{x_b}[Pr(m(x_b) + x < z_b \leq \widehat{m}(x_b) + x, |\widehat{m}(x_b) - m(x_b)| \geq \delta | x_b)] \\
&\leq E_{x_b}[Pr(m(x_b) + x < z_b \leq \widehat{m}(x_b) + x, |\widehat{m}(x_b) - m(x_b)| < \delta | x_b)] \\
&\quad + E_{x_b}[Pr(|\widehat{m}(x_b) - m(x_b)| \geq \delta | x_b)] \qquad (18)
\end{aligned}$$

for every $\delta > 0$. Since $\widehat{m}(\cdot)$ is a consistent estimator of $m(\cdot)$, the second expected value in (18) goes to zero as n_B goes to infinity. As far as the first one is concerned, from the inequality

$$\begin{aligned}
&E_{x_b}[Pr(m(x_b) + x < z_b \leq \widehat{m}(x_b) + x, |\widehat{m}(x_b) - m(x_b)| < \delta | x_b)] \qquad (19) \\
&= E_{x_b}[Pr(m(x_b) + x < z_b \leq \widehat{m}(x_b) + x, m(x_b) - \delta < \widehat{m}(x_b) < m(x_b) + \delta | x_b)] \\
&\leq E_{x_b}[Pr(m(x_b) + x < z_b < m(x_b) + x + \delta | x_b)] \\
&= Pr(x < \epsilon_b \leq x + \delta) < \tau
\end{aligned}$$

and from the continuity of $F_\epsilon(x)$, it is seen that for every $\tau > 0$ there exists n_{B_0} "large enough" such that (19) is smaller than τ for any $n_B \geq n_{B_0}$. The same consideration holds for $Pr(\hat{\epsilon}_b > x, \epsilon_b \leq x)$ in (17). This clearly implies that, for each fixed x, the quantity

$$|\widehat{F}_{n_B}(x) - F_{n_B}(x)| \tag{20}$$

converges in probability to zero as n_B goes to infinity. From the Glivenko-Cantelli theorem we know that

$$sup_x|F_{n_B}(x) - F_\epsilon(x)| \tag{21}$$

converges almost surely to zero as n_B goes to infinity. Since $F_\epsilon(x)$ is continuous, this is enough to conclude that (14) holds. That is, from (20) and (21) it is immediate to see that

$$sup_x|\widehat{F}_{n_B}(x) - F_\epsilon(x)| \tag{22}$$

converges in probability to zero as n_B goes to infinity, *i.e.* the e.d.f. of residuals tends to reproduce the d.f. of ϵ_bs. As a consequence of Proposition 1, it is now easy to conclude that matching based on local linear regression estimator is asymptotically "noise-free".

Proposition 2 *Under the same assumptions of Proposition 1*

$$\tilde{z}_a^A = \widehat{m}^B(x_a^A) + \tilde{\epsilon}^B \tag{23}$$

possesses, as n_B goes to infinity, the same distribution as

$$z_a^A = m(x_a^A) + \epsilon \tag{24}$$

4 A simulation study

In this section we perform a simulation experiment to evaluate the matching noise produced by the nonparametric imputation techniques described in Section 2. It is necessary to resort to simulation procedures because it is not always possible to compute explicitly the matching noise associated to a given imputation technique. The simulation study has been carried out by using the software R ([15]).

In more detail, 500 i.i.d. records from a normal distribution X with mean 1 and variance 5 have been generated. Four regression functions, plotted in Figure 1 and listed below in (25), have been used to model the relationship between the predictor X and the response variable Z.

$$
\begin{aligned}
m_1(x) &= 0.4\left(\tfrac{x+5.7}{13.4}\right)+1 \\
m_2(x) &= 0.3+4\left(\tfrac{x+5.7}{13.4}\right)-3\left(\tfrac{x+5.7}{13.4}\right)^2 \\
m_3(x) &= 4\left(\tfrac{x+5.7}{13.4}\right)-2+2exp\left\{-16\left[4\left(\tfrac{x+5.7}{13.4}\right)-2\right]^2\right\} \\
m_4(x) &= -132\left(\tfrac{x+5.7}{13.4}\right)^4+258\left(\tfrac{x+5.7}{13.4}\right)^3-170\left(\tfrac{x+5.7}{13.4}\right)^2+47\left(\tfrac{x+5.7}{13.4}\right)+1
\end{aligned}
\tag{25}
$$

The function $m_1(x)$ is linear in x, while the second and fourth functions are quadratic and quartic functions respectively. The third function is bump shaped. Normal random errors have been used for all test functions, $\epsilon \sim N(0,\sigma^2)$ for different values of σ^2. More specifically, the values $\sigma^2=(0.3)^2, \sigma^2=(0.5)^2$ have been used. We begin the simulation study setting $\sigma^2=(0.3)^2$.

Let the recipient file A consist of these 500 observations, with Z dropped. For each regression function $m_i(x)=E[Z\,|X=x]$ the simulation involves the following steps:

(1) A donor sample B composed by n_B i.i.d. records has been generated exactly as A, except that the Z values are not dropped. Different values of n_B = 800-2000(200) have been used.
(2) The missing Zs have been imputed by the imputation techniques described in Section 2.
(3) Steps 1 to 2 have been repeated 400 times.

In order to evaluate the closeness between the data generating model and the imputation generating model, a divergence measure based on the Kolmogorov-Smirnov distance (KS) has been used. At first, the matching noise for the marginal distribution of Z has been evaluated. Formally speaking, for each donor sample v (for $v=1,2,\ldots,400$), KS distance compares the empirical distribution function (edf) of imputed values $\widetilde{Z}$ in A ($\hat{F}_{\widetilde{Z},v}(z)$) with the edf of true values ($\hat{F}_{Z,A}(z)$). A mean of such values over the 400 donor files is then taken as a global divergence measure, namely:

$$
KS_Z=\frac{1}{400}\sum_{v=1}^{400}KS_Z(v)=\frac{1}{400}\sum_{v=1}^{400}\left[\sup_{-\infty<z<\infty}\mid \hat{F}_{Z,A}(z)-\hat{F}_{\widetilde{Z},v}(z)\mid\right] \tag{26}
$$

In Figure 2 we report the matching noise produced by distance hot deck, random kNN, mean kNN and mean kNN plus random residual for each regression

function. In accordance with the variance-bias trade off, k has to be defined as a function of sample size n_B such that $k/n_B \to 0$, as $n_B \to \infty$. The value of $k = \sqrt{n_B}$ has been chosen, according to [17].

As Figure 2 shows, the mean kNN is the worst method. This imputation technique underestimates variability, since 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 $Z \mid X$. In fact, the mean kNN plus random residual seems to perform better for all regression functions. Figure 2 also suggests that the mean kNN plus random residual works better when the population regression function is "complex".

The differences between matching noises associated to mean kNN plus random residual, distance hot deck and random kNN respectively have been checked by performing the usual difference of means test. In Tables 1 and 2 we report the test statistic (denoted by $\tau(m_i)$) for each regression function and for different values of n_B.

As a matter of fact, as n_B increases the test statistic values decrease towards the test acceptance region for all test functions. In particular, for donor file sizes large enough (*e.g.* $n_B \geq 1800$) and for simple regression functions (linear, quadratic) the matching noise differences between random kNN and mean kNN plus random residual are not significant. As a consequence, performance of the imputation techniques depends on both data generating model complexity and donor file size. However, for n_B large enough the methods have a similar behavior in reconstructing the marginal distribution of Z.

We now proceed to examine the performance of an alternative imputation method based on the local linear regression estimator of $m_i(x) = E[Z \,|X = x]$ as described in Section 2.3.2. More specifically, two versions of the method have been implemented: (i) deterministic imputation (7); (ii) stochastic imputation (6).

As shown in Section 2.3.2, the local linear regression estimator $\widehat{m}(x)$ is obtained by fitting local straight lines in a neighborhood of x, with weights given by a kernel function K. In the simulation we have used a Gaussian kernel. An important point is the bandwidth selection, whose magnitude influences the amount of local smoothing. In the sequel we consider three selection rules of the smoothing parameter:

- The "Rule of Thumb" (*Rot*) bandwidth selection;
- The "Generalized Cross Validation" (*Gcv*) bandwidth selection;
- The bandwidth selection rule (*Plug*) given by [16];

Rot is a crude bandwidth selector but requires little programming effort. Besides it is so little time consuming that other methods are hard to compete

with. Essentially, the *Rot* bandwidth selector estimates the unknown quantities $(\sigma_i^2, m_i^{''})$ appearing in the asymptotically optimal constant bandwidth fitting a polynomial of order 4 to $m_i(x)$ [7].

Gcv is a simplification of the ordinary cross-validation bandwidth selection rule having the advantage to be less computationally intensive, since it does not require to fit the model n times, one for each delete-one data ([10]). In the *Gcv* procedure the prediction error is estimated in a grid of points defined as $h_j = Ch_{j-1}, j \geq 1$ with $C = 0.1$. We start from $h = h_0 = h_{min}$ and we stop when $h > h_{max}$, where $h_{min} = (\max x - \min x)/n_B$ and $h_{max} = (\max x - \min x)/2$.

Plug is an adaptation of a plug-in bandwidth selector, see [16]. The basic idea is to estimate the unknown quantity $(\sigma_i^2, m_i^{''})$ in the asymptotically optimal constant bandwidth by partitioning the range of X into N blocks, and by fitting a quartic polynomial in each block. The number N is chosen by Mallows's approach [11].

Let LRot, LGcv, LPlug be the deterministic imputation methods (7) based on the local linear regression estimators coming from the bandwidth selectors mentioned above. Figure 3 shows that the corresponding matching noises possess the same order of magnitude than the mean kNN procedure. All methods are deterministic and do not improve the mean kNN plus random residual performance. In order to recover a part of the data variability, their stochastic version has been considered, where the residual is drawn at random from the residuals distribution obtained through the implementation of the same method on the donor file B. For instance, in the LRot plus random residual the imputation value for the Z variable is given by $\tilde{z}_a^A = \widehat{m}_i^B(x_a^A) + \tilde{\epsilon}^B$, where $\tilde{\epsilon}^B$ is drawn at random from the residual $\widehat{\epsilon}_b^B = z_b^B - \widehat{m}_i^B(x_b^B)$ computed on the donor file through a local linear regression with the Rot bandwidth selector. The results are reported in Figure 4.

For all test functions, the stochastic imputation techniques based on both kNN estimator and local linear estimator seem to perform better than their deterministic counterparts : adjusting the regressed values in order to account for the residual variability reduces the matching noise. Such a behaviour is more evident for complex regression functions (*i.e.* functions 3 and 4). The differences between such imputation methods checked by performing the usual difference of means test are not significant for $n_B \geq 1200$ and *p-value* $= 0.01$. As a consequence, the preference will be given to the mean kNN plus random residual since it is computationally easier and does not require any bandwidth selection.

Figures (2), (3) and (4) evaluate the ability of the imputation methods to reproduce the marginal distribution of Z in the synthetic data set. In order to

get information on the closeness of the two distributions $f_{X\widetilde{Z}}(x,z)$ (the distribution generating the genuine data) and $f_{XZ}(x,z)$ (the distribution generating imputed data), the KS distance has been computed between the conditional distribution of $Z \mid X = x_a^A$, $a = 1, \ldots, 500$ ($F_{Z|x_a^A}(z)$) and the conditional empirical distribution $\hat{F}_{\widetilde{Z}|x_a^A}(z)$. To get a synthetic measure, the average over the $n_A = 500$ values has been computed:

$$E[KS_Z^X] \approx \frac{1}{500}\sum_{a=1}^{500} KS_Z(x_a^A) = \frac{1}{500}\sum_{a=1}^{500}\left[\sup_{-\infty<z<\infty} \mid F_{Z|x_a^A}(z) - \hat{F}_{\widetilde{Z}|x_a^A}(z) \mid\right] \quad (27)$$

In Figure 5 we report the discrepancy measure (27) for all the stochastic nonparametric imputation techniques described in Section 2. For each test function the results obtained are described below

- *Test function* $m_1(x)$: distance hot deck, random kNN, mean kNN plus random residual and LRot plus random residual seem to perform better. The methods give equivalent results since the corresponding $E[KS_Z^X]$ are not significantly different for $n_B \geq 800$ and *p-value* $= 0.02$.
- *Test function* $m_2(x)$: distance hot deck, LRot plus random residual and LGcv plus random residual seem to be the best methods. The corresponding $E[KS_Z^X]$ are not significantly different for $n_B \geq 800$ and *p-value* $= 0.01$.
- *Test function* $m_3(x)$: distance hot deck and LRot plus random residual seem to perform better. The corresponding $E[KS_Z^X]$ are not significantly different for $n_B \geq 800$ and *p-value* $= 0.03$.
- *Test function* $m_4(x)$: distance hot deck and LRot plus random residual seem to perform better. The corresponding $E[KS_Z^X]$ are not significantly different for $n_B \geq 800$ and *p-value* $= 0.03$.

As previously stressed, (27) is a crude measure of divergence between the conditional distribution of $Z \mid X = x_a^A$, $F_{Z|x_a^A}(z)$, and the conditional empirical distribution $\hat{F}_{\widetilde{Z}|x_a^A}(z)$. In order to get additional information about the performance of the imputation techniques, in Figure 6 the KS distance $KS_Z(x_a^A)$ is reported for different values of x_a^A. As known, the boundary effect is more evident for the kNN than for the local linear estimator of the regression function. When x_a^A is close to the boundaries, the kNN estimator is based on the computation of averages in an asymmetric region of x_a^A, consisting of a fixed number of k points. Hence, the matching noise of the stochastic imputation method based on the kNN estimator could be severely high when x_a^A is close at the extremes of its observational range. As it appears from Figure 6, when the regression function is very steep at the boundaries the kNN estimator is expected to be more biased (see, for instance, $m_4(x)$), and hence the corresponding imputation method is affected by a severe matching noise.

As a matter of fact, when comparing the conditional distribution of $Z \mid X$ a

slight preference could be given to distance hot deck since it is more easily implemented and surely less computationally intensive.

Suppose now to perform the simulation study assuming that errors have common variance $\sigma^2 = (0.5)^2$. The results regarding the marginal distribution of Z and the conditional distribution of $Z \mid X$ are reported in Figure 7 and 8, respectively. Mean kNN plus random residual seems to have the best performance in recovering the marginal distribution of Z together with the stochastic imputation procedures based on the local linear regression estimator (see, for instance, $m_4(x)$). With regard to the conditional distribution of $Z \mid X$, the results obtained are reported below

- *Test function* $m_1(x)$: distance hot deck, mean kNN plus random residual and LRot plus random residual seem to have the best performance. Such methods give equivalent results since the corresponding $E[KS_Z^X]$ are not significantly different for $n_B \geq 800$ and $p\text{-}value = 0.11$.
- *Test function* $m_2(x)$: distance hot deck, LRot plus random residual and LGcv plus random residual seem to be the best methods. The corresponding $E[KS_Z^X]$ are not significantly different for $n_B \geq 800$ and $p\text{-}value = 0.12$.
- *Test function* $m_3(x)$: distance hot deck and LRot plus random residual seem to perform better. The corresponding $E[KS_Z^X]$ are not significantly different for $n_B \geq 800$ and $p\text{-}value = 0.03$.
- *Test function* $m_4(x)$: distance hot deck, LRot plus random residual and LGcv plus random residual seem to perform better. The corresponding $E[KS_Z^X]$ are not significantly different for $n_B \geq 800$ and $p\text{-}value = 0.02$.

In conclusion, since in a given survey the construction of a complete synthetic data set containing (X, Z) aims at getting information about both the marginal distribution of Z and the full distribution of (X, Z), the stochastic imputation method based on the local linear regression estimator (LRot, LGcv) seems to be the best. Such a method is "almost" as good as the mean kNN plus random residual for the reconstruction of the marginal distribution of Z, and as the distance hot deck when the interest is in the conditional distribution of $Z \mid X$. Besides, the more complex is the functional relationship between the variable of interest, the better seems to be its performance.

References

[1] Aluja-Banet, T., Daunis-i-Estadella, J. and Pellicer, D., (2007) GRAFT, a complete system for data fusion. *Journal of Computational Statistics and Data Analysis*, to appear.

[2] Brick, J.M. and Kalton, G., (1996). Handling missing data in survey research. *Statistical Methods in Medicine*, **5**, 215–238.

[3] 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.

[4] Chung, C.K. and Cheng, P.E., (1995). Nonparametric regression estimation with mossing data. *Journal of Statistical Planning and Inference*, **48**, 85–99.

[5] D'Orazio, M., Di Zio, M. and Scanu, M., (2006). *Statistical Matching: Theory and Practice*. Wiley, Chichester.

[6] Fan, J. and Gijbels, I., (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London.

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

[8] Kalton, G. and Kasprzyk, D., (1986). The Treatment of Missing Survey Data. *Survey Methodology*, **12**, 1–16.

[9] Little, J. and Rubin, D., (1987). *Statistical Analysis with Missing data*. Wiley, New York.

[10] Loader, C., (2004). Smoothing Local Regression Techniques. In: Gentle, G., Härdle, W., Mori, Y. (Ed.), *Handbook of Computational Statistics*, Springer-Verlag..

[11] Mallows, C.P., (1973). Some Comments on C_p. *Technometrics*, **15**, 661–675.

[12] Marella, D., Scanu, M. and Conti P.L., (2006). On the matching noise of some nonparametric imputation procedures. *Technical Report n.5, DSPSA, Università di Roma "La Sapienza"*, 2007. Submitted.

[13] Nielsen, S.F., (2001). Nonparametric conditional mean imputation. *Journal of Statistical Planning and Inference*, **99**, 129-150.

[14] 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.

[15] R Development Core Team (2004). *R: A language and Enmvironment for Statistical Computing*. Vienna. R Foundation for Statistical Computing.

[16] Ruppert, D., Sheather, J. and Wand, M.P.,(1995). An Effective Bandwidth Selector for Local Least Squares Regression. *Journal of the American Statistical Association*, **90**, 1257–1270.

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

[18] Yates, F., (1933). The analysis of replicated experiments when the field results are incomplete. *Emporium J. Exp. Agriculture*, **1**, 129–142.

Table 1
Test statistic between distance hot deck and mean kNN + residual

n_B	$\tau(m_1)$	$\tau(m_2)$	$\tau(m_3)$	$\tau(m_4)$
800	9.73	6.34	9.48	9.45
1200	6.23	3.73	6.17	5.55
1600	4.67	3.44	3.48	3.56
2000	3.10	1.19	3.30	4.19

Table 2
Test statistic between random kNN and mean kNN + residual

n_B	$\tau(m_1)$	$\tau(m_2)$	$\tau(m_3)$	$\tau(m_4)$
800	6.77	2.45	6.83	10.20
1200	4.14	0.88	5.72	6.95
1600	4.40	1.45	5.09	4.22
2000	1.72	−0.47	3.26	3.87

Fig. 1. *Plots of the regression functions* $m_i(x) = E[Z\,|X = x]$.

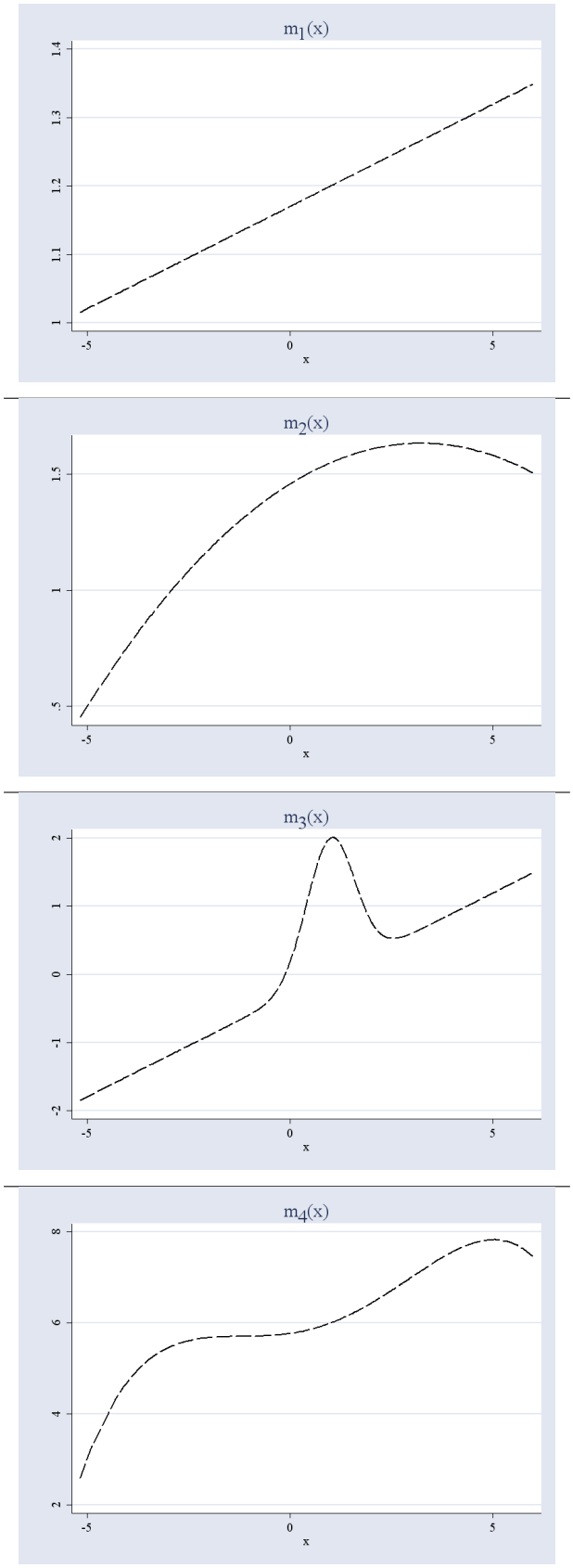

Fig. 2. *KS_Z for distance hot deck, mean kNN, random kNN and mean kNN +residual.*

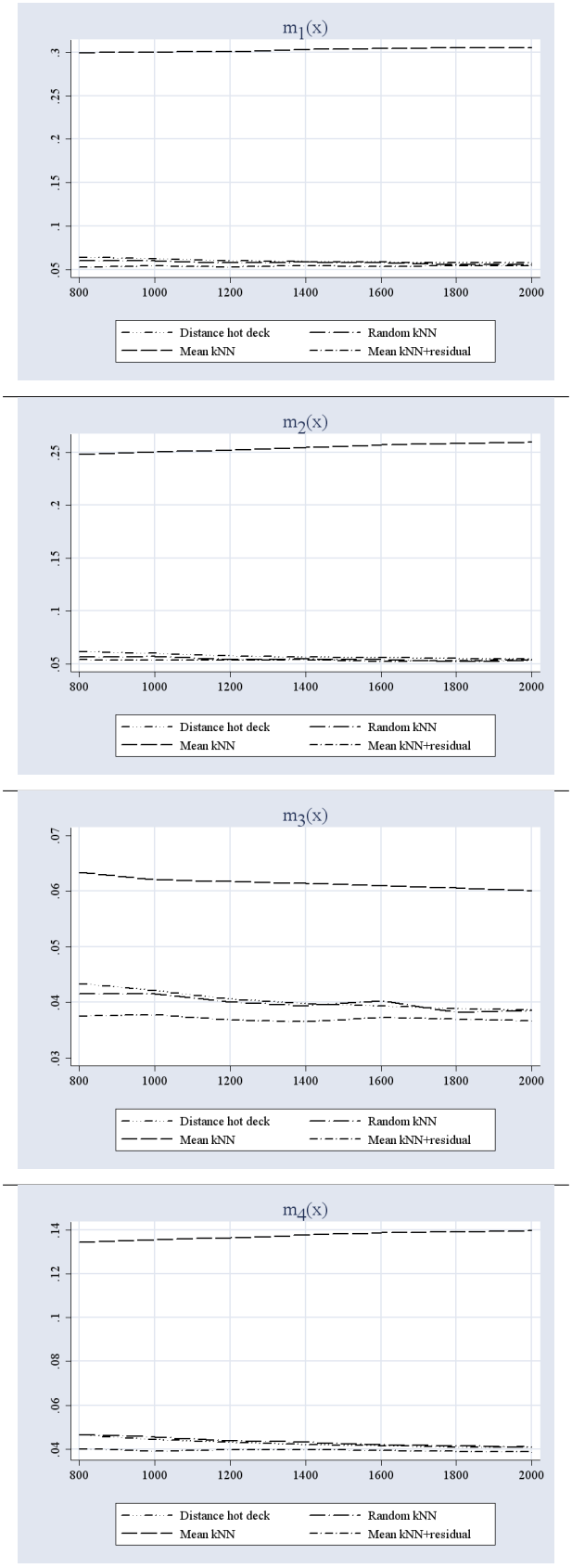

Fig. 3. *KS_Z for distance hot deck, random kNN, mean kNN +residual, local linear regression estimators LRot, LGcv, LPlug .*

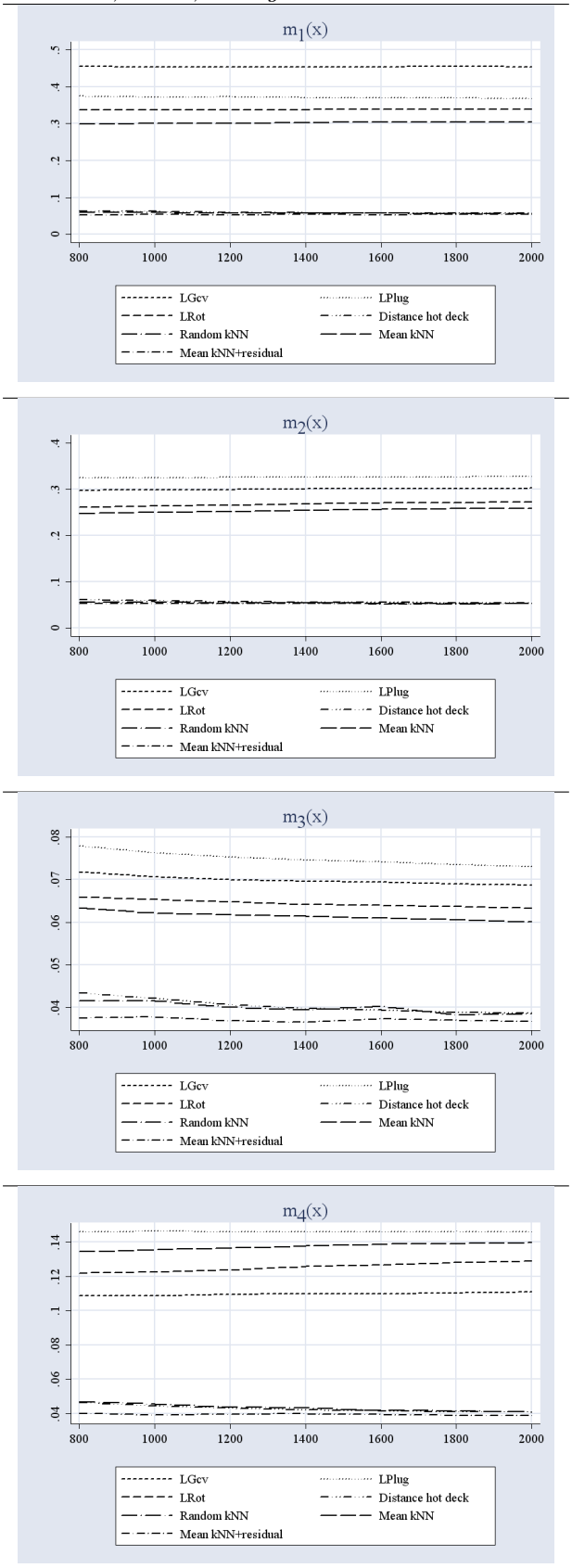

Fig. 4. *KS_Z for distance hot deck, random kNN, mean kNN +residual, LRot+residual, LGcv+residual, LPlug+residual.*

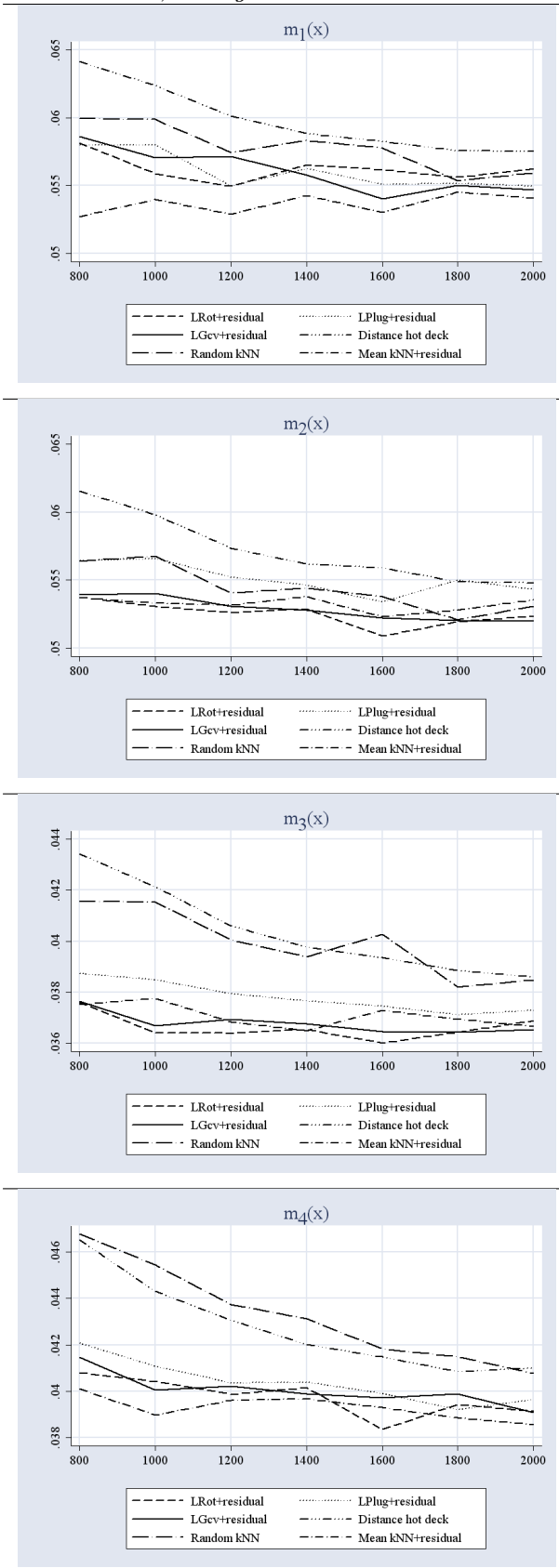

Fig. 5. $E[KS_Z^X]$ *for distance hot deck, random kNN, mean kNN +residual, LRot+residual, LGcv+residual, LPlug+residual.*

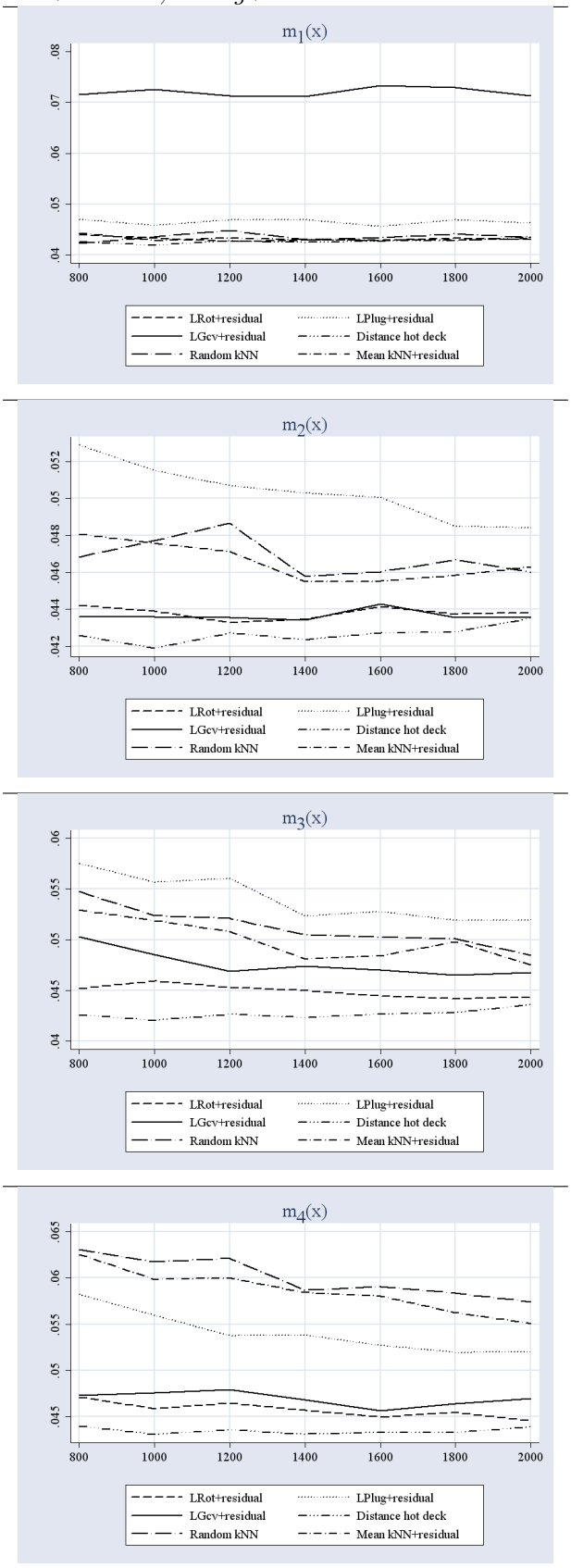

Fig. 6. *$KS_Z(x_a^A)$ for distance hot deck, mean kNN +residual, LRot+residual, LGcv+residual, LPlug+residual.*

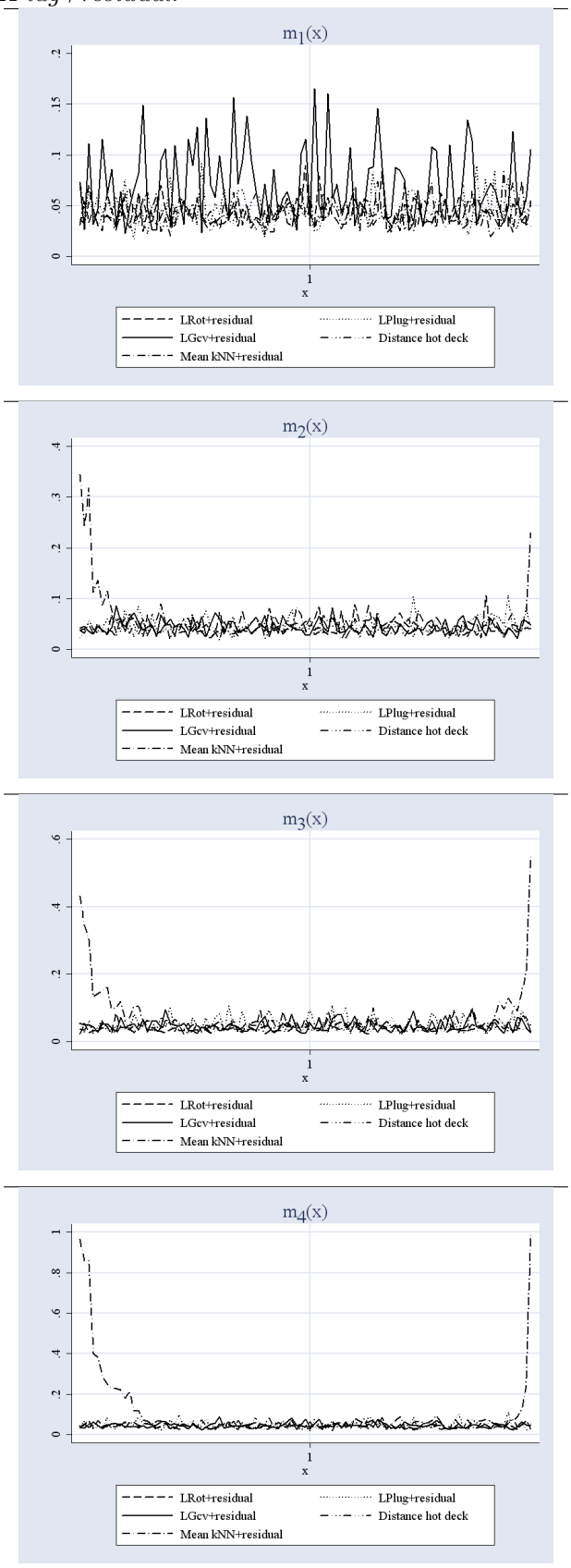

Fig. 7. *KS_Z for distance hot deck, mean kNN, random kNN and mean kNN +residual, LRot+residual, LGcv+residual, LPlug+residual.* ($\sigma^2 = (0.5)^2$)

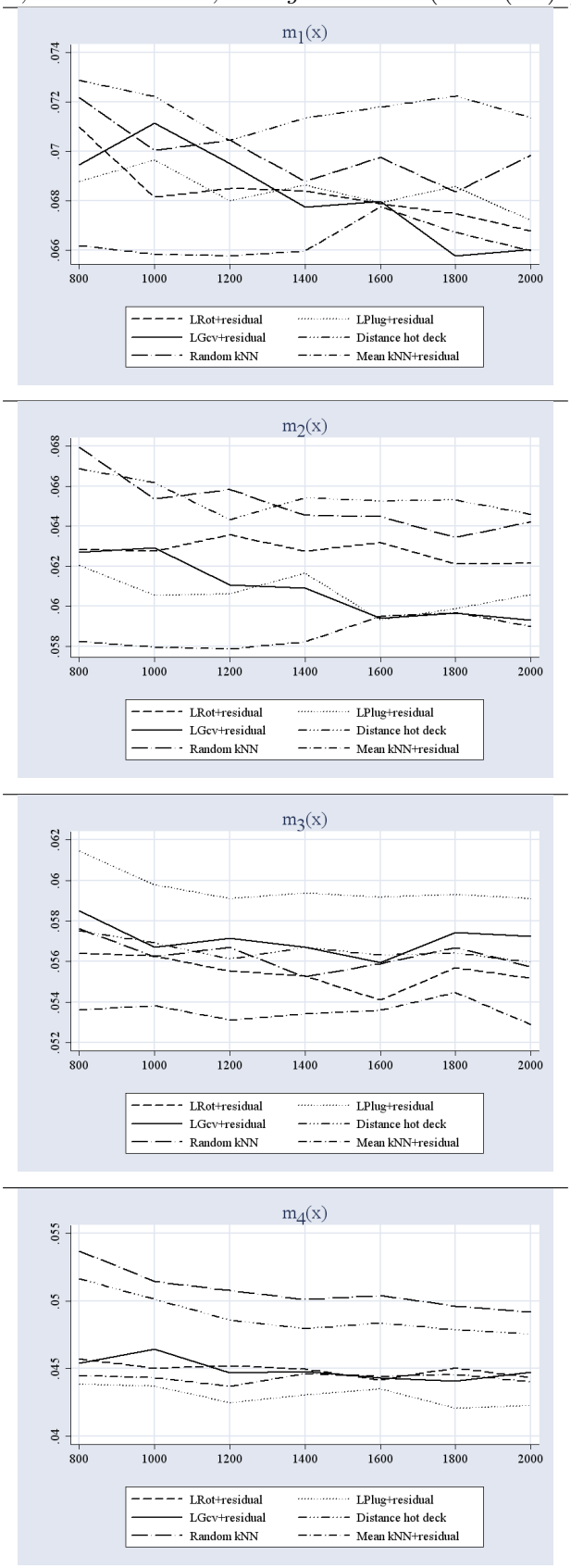

Fig. 8. *$E[KS_Z^X]$ for distance hot deck, mean kNN, random kNN and mean kNN+residual, LRot+residual, LGcv+residual, LPlug+residual.* ($\sigma^2 = (0.5)^2$)