LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
INSTITUT FÜR STATISTIK

Humera Razzak, Christian Heumann

Application of Iterative Hybrid MI Approach to Household Survey Data with Complex Dependency Structures

Technical Report Number 231, 2019
Department of Statistics
University of Munich

http://www.statistik.uni-muenchen.de

Application of Iterative Hybrid MI Approach to Household Survey Data with Complex Dependency Structures

Humera Razzak[1], Christian Heumann[2]

[1] Department of Statistics, Ludwig-Maximilians-Universität München. Ludwigstr. 33, D-80539, München, Germany. Humera.Razzak@stat.uni-muenchen.de

[2] Christian Heumann, Department of Statistics, Ludwig-Maximilians Universität München, Ludwigstr. 33 D-80539 München, Germany. christian.heumann@stat.uni-muenchen.de

ABSTRACT

The multiple indicator cluster survey (MICS) is a household survey tool designed to obtain internationally comparable, statistically rigorous data of standardized indicators related to the health situation of children and women. Missing data in a large number of categorical variables are a serious concern for MICS, following complex dependency structures and inconsistency problems that impose severe challenges to the investigators. Despite the popularity of multiple imputation of missing data, its acceptance and application still lag in large-scale studies with complicated data sets such as MICS. We propose interdependent hybrid multiple imputation (HMI) techniques which combines features of existing MI approaches to handle complex missing data in large scale household surveys. The iterative HMI approach is observed to be a good competitor to the existing approaches, with often smaller root mean square errors, empirical standard errors and standard errors. Regardless of any combination, the iterative HMI method is markedly superior to the existing MI methods in terms of computational efficiency. Results from household data example support the capacity of proposed method to handle complex missing data.

Keywords: word; Survey data; hybrid multiple imputation; household data; complex;MICS

1. Introduction

Key indicators or background variables related to the health situation of children and women are measured in complex household surveys e.g. multiple indicator cluster survey (MICS). These indicators enable countries to produce data that can further be used in policies and programs. Datasets of such surveys have mixed type variables that are both multilevel categorical and continuous variables. However, missing data in a large number of variables are a serious concern for household surveys, following complex dependency structures and inconsistency problems that impose severe challenges to the investigators. For example the MICS 2014 house hold data file that we analyze, 26819 only out of 41413 observations have complete data on a set of more than 200 background variables. Respondent's may refuse to provide a requested piece of information based on various reasons, such as unwillingness, lack of capability to answer, reservation on sensitivity of question, confidentiality and privacy etc. This results in the failure to collect complete information. Generally, this non-response behavior is referred to as item non-response (INR). Most typically, high rate of INR occurs for simple demographic variables such as age, sex or marital status however, questions related to income or wealth are often related to high rate of INR (e.g. Riphahn and Serfling 2005; Hawkes and Plewis 2006). Beside INR general reasons for the missing datasets include data entry errors, system failures etc.

Analysis of data for scientific investigations becomes complicated, biased and less efficient in presence of missing information. In recent decades, lots of effort has been made in development of statistical methods to carter missing data. Missing data can be handled by "Multiple Imputation" (MI). MI, first introduced by Rubin (1987), is widely regarded as the "gold standard" approach to handle missing data problem, with many documented advantages over complete case analyses. Multiple random values for the missing data under a statistical

model can be generated to estimate the values multiple times using MI. This results in M >1 multiple complete datasets. MI combines the results which account extra variability caused by the missing data. The complete datasets can be analyzed by using standard statistical procedures or so called "Rubin's inference". Multivariate normal model, the log linear model, or the general location model (Schafer 1997) are examples of MI. Despite the popularity of MI, its acceptance and application still lag in large-scale studies with complicated data sets such as MICS data. Hence, MI is restricted in one or the other way and not dedicated to the complex household survey data.

The paper is organized as follows: First, we provide a description of notations and assumptions of missing mechanisms then briefly describing some fundamentals of missing data and MI. In Section 3 we describe hybrid architectures in detail. In Section 4 we present the simulations studies, the methods used in the analyses and relevant results to evaluate our proposed approach. Section 5 presents the imputation of the household data. We conclude with a discussion in Section 6.

2. Fundamentals of Missing Data and Multiple Imputation (MI)

2.1. Notations and Assumptions of Missing Mechanisms

In general, there are three types of missingness generating mechanisms. Missing categories can be classified into: (i) missing completely at random (MCAR), (ii) missing at random (MAR), (iii) missing not at random (MNAR) (Little and Rubin 2002). Let Y be the data with $n \times p$ dimensions. Assume, y_{ij} refers to the i_{th} value of variable j from Y where $i=1,\ldots, n$ and $j=1,\ldots, p$. Suppose, there are two components of the data set $Y = \{Y^{miss}, Y^{obs}\}$ where, the first component denotes the observed part of the data and the second component is the missing data.

Let H be a response indictor matrix with same dimensions as Y indicating, if an element of Y is missing.

$$H_{ij} = \begin{cases} 0 \;\; if \; y_{ij} \; is \;\; missing \\ 1 \;\; if \; y_{ij} \; is \;\; observed \end{cases}$$

Missing Completely At Random (MCAR): $Pr\left(H|Y^{miss}, Y^{obs}\right) = Pr(H)$.

Missing At Random (MAR): $Pr\left(H|Y^{miss}, Y^{obs}\right) = Pr(H|Y^{obs})$.

Missing Not At Random (MNAR): $\left(H|Y^{miss}, Y^{obs}\right) \neq Pr(H|Y^{obs})$.

The third assumption is also called non-ignorable (NI) (Little and Rubin 2002) and not further used in the paper.

2.2. Rubin's inference

In general any measure of interest Q (e.g. parameter estimates $\hat{\theta}$) is assessed by the average

$$\overline{Q}_M = \frac{1}{M} \sum_{m=1}^{M} \widehat{Q_m} \tag{1}$$

using M estimates $\widehat{Q_m}$ derived from the imputed complete data sets. The total variability of the estimate is given by

$$T_M = \left(1 + \frac{1}{M}\right) B_M + \overline{W}_M \tag{2}$$

where

$$\overline{W}_M = \frac{1}{M} \sum_{m=1}^{M} \widehat{W_m} \tag{3}$$

and

$$B_M = \frac{1}{M-1} \sum_{m=1}^{M} \left(\widehat{Q_m} - \overline{Q}_M \right)^2 \tag{4}$$

are the averages of the within-imputation variances $\widehat{W_m}$ and the between-imputation variance, respectively.

2.3. Literature Review of Existing Studies in Large-Scale Complex Surveys

There are two general approaches for MI. Fully conditional specification (FCS; also known as sequential regression and MI using chained equations (MICE)) and MI based on the joint posterior distribution of incomplete variables, often referred to as joint modelling (JM) (Raghunathan et al. 2001; van Buuren 2007; Schafer 1997; van Buuren et al. 2006).

FCS is an iterative process which cycles through incomplete variables one at a time and imputes data on a variable-by-variable basis. A conditionally specified imputation model known as MICE, visits sequentially each incomplete variable and draws alternately the imputation parameters and the imputed values. FCS MI approach imputes variables one at a time from a series of univariate conditional distributions (van Buuren et al. 2006). FCS approach requires existence of joint distribution for convergence, which is a major downside of this approach. It is possible to get the joint distribution under rather general conditions (Liu et al. 2014; Zhu and Raghunathan 2015). However, correct specification of conditional distributions can guarantee consistency of inferences based on the imputed data even in the absence joint distribution. In MICE missing values can be present in many variables and user can specifies regression methods according to the types of variables. For example classification and regression tree (CART) (Burgette and Reiter 2010) for categorical variables and predictive mean matching (PMM) (Rubin and Schenker 1986) which is the default imputation technique for continuous data. CART is a nonparametric method. CART uses splitting algorithms to divide the values of a variable into homogeneous subgroups. On the other hand, PMM approach uses predicted value obtained by a

linear regression model to impute an observed value. The predicted value is among the values of donor pool which are closest to the value predicted for the missing one. Software packages implementing MICE includes "mice" (van Buuren and Groothuis-Oudshoorn 2011; van Buuren 2012), "mi" in R (Su et al. 2011) and "IVEware" in SAS (Raghunathan et al. 2002). Despite of many advantages, MICE has few downsides for example, MICE mostly use parametric models. Those models are hard to implement due to lack of compatibility and complex dependencies among variables. Moreover, implementation is difficult due to higher order interactions effects or many nonlinear relations in regression model (see Burgette and Reiter (2010)). Implementation of MICE becomes very time consuming in presence of large number of categorical variables. PMM can be problematic, when sample size is large (van Buuren 2011) and CART can subject to odd behaviors in high dimensions. Another limitation of CART is that the corresponding joint distribution based on conditional models might not exist (Si and Reiter 2013). Moreover, variables with many levels are preferred to variables with few levels in CART, e.g. Breiman et al. (1984) and Kim and Loh (2001).

Joint modeling (JM) draws missing values simultaneously for all incomplete variables using a multivariate distribution (Schafer 1997). Draws from fitted distribution are used to create imputations. Dirichlet Process Mixture of Products of Multinomial Distributions Model (DPMPM) provides a fully Bayesian, non-parametric JM approach to MI for high dimensional categorical data (Manrique-Vallier and Reiter 2015; Si and Reiter 2013). Dunson and Xing (2009) proposed DPMPM for the first time. This approach uses nonparametric Bayesian versions of latent class models to multiply impute high-dimensional categorical data (Vermunt et al. 2008). The DPMPM imputation routines are implemented in the R software package, "NPBayesImputeCat" (Quanli et al. 2018). Softwares "Realcom-impute" (Carpenter and

Kenward 2011), R package “pan” (Schafer and Zhao 2014), R package “jomo” (Quartagno and Carpenter 2015) implement JM approach.

Like many complex models, the effectiveness of DPMPM still lags in capturing the many features of empirical data. It is not possible to implement JM approach in the multilevel context if missingness also occurs in the random slope variable(s) (Carpenter and Kenward 2011). Modeling mixed type variables can make the specification of a joint distribution very difficult. MI approaches described above are available in standard computer packages (SAS, Stata and R). See Horton and Kleinman (2007) for an overview of available MI procedures and packages. FCS and JM MI approaches were originally proposed for dealing with item nonresponse in cross-sectional data sets. Despite of being commonly available in existing softwares, these methods are hard to implement in large scale data sets with many categorical variables and many levels.

In large-scale complex surveys many types of variables with special data situations have to be handled. To do so, several methods have been proposed in the literature over recent years. For example Audigier et al. (2018) deal with quantitative variables. Manrique-Vallier and Reiter (2014, 2015), Audigier et al. (2017) among many deal for qualitative and Audigier et al. (2016) and Murray and Reiter (2016) deal for mixed data. Methods for qualitative and mixed data tend to perform well particularly for small number of observations and dataset having multilevel categorical variables. Moreover, these methods often require less execution time. However, some of these approaches require knowledge of complicated models and other need transformations (or other tricks) for continuous variables or assume missing values in few variables. Categorizing of continuous variables can subject to considerable loss of information (van Buuren and Groothuis-Oudshoorn 2011). Husson et al. (2019) have proposed a MI method based on multilevel singular value decomposition (SVD) for quantitative, categorical, or mixed data. This

method performs SVD on between and within groups variability of the data. Downside of this method is that it does not take into account the uncertainty associated with predicting missing values from observed values. Goßmann (2016) proposed the application of CART in combination with multiple imputation and data augmentation for large-scale survey. Mislevy (1991) presented the idea to combine multiple imputation with latent variables that were used to estimate population characteristics when individual values were missing in complex surveys. A Bayesian approach for flexible handling of missing values is proposed by Aßmann et al. (2016) which handles continuous and categorically scaled background variables in large-scale surveys. Stekhoven and Bühlmann (2012) have presented a machine learning technique based on non-parametric models called random forest models to impute ordinal missing data. It has many desirable properties such that can be applied to a variety of categorical data, a mix of categorical and continuous data. It does not require any specific distributional assumption. It can handle nonlinear relationships among variables (Doove et al. 2014; Shah et al. 2014). Random forest approach to MI is implemented in R packages "mice" (van Buuren and Groothuis-Oudshoorn 2011; van Buuren 2012) and "missForest" (Doove et al. 2014; Stekhoven and Bühlmann 2012). Shah et al. (2014) found that random forest-based MICE tends to perform better than parametric MICE on survival data. Hybrid MI based on dependence models (Razzak and Heumann 2019) is another approach to impute complex household survey data. The dependence models impute continuous covariates using FCS MI given the categorical covariates already imputed using JM MI. The Hybrid MI based on dependence models not only yields better predictive performance of generalized linear models (GLMs) (Nelder and Wedderburn 1972) for binary response (Razzak and Heumann 2019) but are also observed to be a good competitor to the existing approaches, with often smaller root mean square errors and less computational cost. However,

hybrid dependence models do not use the information of continuous covariates for imputing categorical covariates. In this article, we extend the hybrid imputation approach based on dependence models by categorizing continuous variables. We propose two iterative hybrid imputation approaches for mixed data in complex household surveys where missing values in continuous covariates are imputed by using the information of already imputed categorical variables and continuous variables are categorized to impute categorical variables. We review inference in GLMs with binary response and mixed type missing covariates in large scale survey for a proposed and existing methods.

3. Proposed Hybrid Architectures

Consider the motivational question in section one. Performance of JM and FCS approaches to obtain complete information on mixed type covariates in large scale surveys are limited and subject to specific tasks. Moreover, these approaches are generally not equipped to handle a wide range of complexities in large scale data, categorical variables, and different heretical relations. We propose that various features of JM and FCS methods can be combined to obtain complete data with the limitations discussed above. To do so, we propose two easy and simple to implement variants of hybrid architecture that use the idea of categorizing continuous data. In first variant of hybrid architecture, we use the concept of categorizing continuous variables before the imputation of categorical data. Second variant uses initial imputed values. These values are obtained by categorization of continuous data before the imputation of categorical data. Unlike existing approaches, where categorization results in loss of power, proposed approaches restore the continuous variables in their original form. These variants are computational fast and can be applied to both categorical and continuous data in high dimensions.

3.1. Proposed Hybrid Architecture 1

Algorithm 1: Iterative Hybrid MI 1

Require: *P nxp* matrix with incomplete data

$Miss_{cat}$, $Miss_{num}$ ← Division of *p* variables into factor and continuous subsets.

for $z = 1, \ldots, Z$ **do**

for $m = 1, \ldots, M$ **do**

$Imp^{z}_{num\_cat_m}$ ← Categorizing $Miss_{num}$.

$Imp^{z}_{cat_m}$ ← imputation using JM approach for $Miss._{cat} \mid Imp^{z}_{num\_cat_m}$.

$Imp^{z}_{num_m}$ ← imputation using FCS approach for $Miss._{num} \mid Imp^{z}_{cat_m}$.

end for

end for

The first variant of proposed hybrid architecture generates a complete data set in three steps. Incomplete data is divided in to two sub groups (i.e. one containing incomplete continuous data ($Miss_{num}$) and other having incomplete categorical data ($Miss_{cat}$)). **Step 1:** variables in $Miss_{num}$ are categorized $Miss_{num.cat}$. **Step 2:** JM technique is applied on $Miss_{cat}$ given additional covariates $Miss_{num.cat}$ to generate complete categorical data. Complete categorical data generated in this step contains complete categorical variables Imp_{cat} and complete *c*ategorized variables $Imp_{num.cat.}$. In first step, categorization allows the information on continuations variables to impute categorical variables. **Step3**: FCS technique is applied to impute missing values in original continuous variables $Miss_{num}$ given additional categorical variables Imp_{cat}. Step 3, allows the information on categorical variables to impute continuous variables. Steps 1 to 3 are repeated *M* times to generate multiple copies of complete data sets. Inference (e.g. mean, regression) can be run on each of the newly created, imputed datasets. Finally, estimates can be combined by using 'Rubins rules'. Algorithm 1 explains the proposed method in detail. Schematic diagram illustrating the proposed hybrid architecture 1can be seen in supplementary file (see Figure S1).

3.2. Proposed Hybrid Architecture 2

Algorithm 2: Iterative Hybrid MI 2

Require: *P nxp* matrix with incomplete data

0. $Miss_{cat}$, $Miss_{num}$ ← Division of p variables into factor and continuous subsets.

1. **Initialization**

(a) Initialize missing values for categorical variables: Imp_{cat_i} ← single imputation using JM approach for $Miss_{.cat}$.

(b) Initialize missing values for continuous variables: Imp_{num_i} ← single imputation using FCS approach for $Miss_{num} \mid Imp_{cat\_i}$.

(c) Initialize categorized values for continuous variables: $Imp^{z}_{num\_cat_i}$ ← Categorizing Imp_{num_i}

for $z = 1, \ldots, Z$ **do**

for $m = 1, \ldots, M$ **do**

2. **Update imputed values**

(a) $Imp^{z}_{cat_m}$ ← imputation using JM approach for $Miss_{.cat} \mid Imp_{num\_cat\_i}$.

(b) $Imp^{z}_{num_m}$ ← imputation using FCS approach for $Miss_{.num} \mid Imp^{z}_{cat_m}$.

(c) $Imp^{z}_{num\_cat_m}$ ← Categorizing $Imp^{z}_{num_m}$.

end for

end for

The second variant of proposed hybrid architecture is a two steps approach. **Step 1**: **(a)** Initialize values for categorical variables (Imp_{cat_i}) by applying JM approach to $Miss_{cat}$. **(b)** Given the initial values for categorical variables, single iteration of the FCS algorithm is run to $Miss_{num}$ for initialization of values for continuous variables Imp_{num_i}. Information on categorical variables is used for the generation of Imp_{num_i} whereas, no information available on continuous variables is used in generation of Imp_{cat_i}. **(c)** Initial values for continuous variables Imp_{num_i} are categorized $Imp_{num.cat\_i}$ to allow usage of information available on continuous variables for imputing categorical variables. **Step 2: (a)** Given the initial categorized variables ($Imp_{num.cat\_i}$) as additional covariates, complete categorical variables with updated values (Imp_{cat}) are

generated by applying JM approach to $Miss_{cat}$. **(b)** Given updated values of additional covariates Imp_{cat}, complete continuous variables (Imp_{num}) with updated values are generated by applying single iteration of FCS approach to $Miss_{num}$. **(c)** Updated values of complete continuous variables are categorized ($Imp_{num.cat}$). Steps 2(a-c) are repeated *M* times with new updated values of Imp_{cat}, Imp_{num} and $Imp_{num.cat}$ to obtain *M* complete data sets. Algorithm 2 explains the proposed method in detail. Schematic diagram illustrating the proposed hybrid architecture 2 is provided in supplementary file (see Figure S2).

4. A Simulation study

To investigate the performance of hybrid architectures via simulation, somewhat large numbers *(X=39)* of mixed type variables are generated. To generate first thirty one binary (X_b) variables a multivariate normal (MVN) distribution is used and correlated random covariates C_i compromising 1000 observations are generated. The marginal distributions are: $C_i \sim N(0, 0.5)$, where $i=\{1,\ldots,31\}$. The correlation structure is given as:

$$R = \begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}.$$

Where $\rho = 0.5$. Random covariates (C_i) are transformed into binary values (X_b) using the following threshold:

$$X_{b_i} = \begin{cases} 0 & if \quad C_i \leq 0, \\ 1 & if \quad C_i > 0. \end{cases}$$

Where $i=\{1,\ldots,31\}$.

In order to generate outcomes for the two multilevel categorical covariates i.e. (X_{m_1} and X_{m_2}), we first generate two random covariates from normal distributions (ND) given as: $C_{32} \sim N(\mu_1; \sqrt{2})$, $C_{33} \sim N(\mu_2; \sqrt{2})$, where μ_1 and μ_2 are described as:

$$\mu_1 = 0.1 + 0.1 \sum_{i=1}^{31} X_{b_i} + 0.1 X_{b_2} X_{b_3} + 0.1 X_{b_5} X_{b_8} + 0.1 X_{b_2} X_{b_{29}} \quad (5)$$

$$\mu_2 = 0.1 + 1.1 \sum_{i=1}^{19} X_{b_i} + 0.1 \sum_{i=20}^{31} X_{b_i} + 0.1 C_{32} + 0.1 X_{b_2} X_{b_3} + 0.1 X_{b_5} X_{b_8} + 1.1 X_{b_2} X_{b_{29}}. \quad (6)$$

.

Further, all observations in $C_{31}$and C_{32} are randomly split into various homogeneous groups and two multilevel categorical variables X_{m_1} and X_{m_2} are formed with four and six categories respectively. To encode complex dependence relationships with higher order interactions, we generate another binary covariate $X_{b_{32}}$ from Bernoulli distribution with probabilities governed by the logistic regression with

$$logit\ Pr(X_{b_{32}}) = 0.001 - 0.01 X_{b_1} - 0.09 X_{b_2} - 0.09 X_{b_3} - 0.09 X_{b_4} + 0.05 X_{b_5} + 0.08 X_{b_6} - 0.02 X_{b_7} + 0.08 X_{b_8} + 0.01 X_{b_9} + 0.01 X_{b_{10}} - 0.02 X_{b_{11}} + 0.01 X_{b_{i12}} - X_{b_{13}} + 0.02 X_{b_{14}} - 0.01 X_{b_{15}} + 0.02 X_{b_{16}} - 0.03 X_{b_{17}} - 0.02 X_{b_{18}} - 0.07 X_{b_{19}} + 0.08 X_{b_{20}} + 0.08 X_{b_{21}} + 0.01 X_{b_{22}} + 0.09 X_{b_{23}} + 0.09 X_{b_{24}} + 0.05 X_{b_{25}} + 0.08 X_{b_{26}} - 0.02 X_{b_{27}} + 0.08 X_{b_{28}} + 0.08 X_{b_{29}} - 0.01 X_{b_{30}} + 0.09 X_{b_{31}} + 0.02 C_{32} + 0.02 C_{33} + 0.02 X_{b_{12}} X_{b_{29}} - 0.02 X_{b_{15}} X_{b_{18}} X_{b_{29}}. \quad (7)$$

We then generate outcomes for the two continuous covariates i.e. X_{n_1} and X_{n_2} from normal distributions (ND). Description is as follows

$$X_{n_1} \sim N(\mu_3; \sqrt{0.5}).$$

Where, $\mu_3 = 0.002 + 0.5 X_{b_1} - 0.15 X_{b_2} + 0.25 X_{b_3} - 0.6 X_{b_4} - 0.88 X_{b_5} + 0.11 X_{b_6} + 0.2 X_{b_7} - 0.5 X_{b_8} + 0.1 X_{b_9} - 0.2 X_{b_{10}} + 0.3 X_{b_{11}} + 5 X_{b_{12}} - 0.2 X_{b_{13}} + 0.3 X_{b_{14}} + 0.4 X_{b_{15}} + 0.1 X_{b_{16}} + 0.1 X_{b_{17}} - 0.1 X_{b_{18}} - 0.1 X_{b_{19}} - 0.10 X_{b_{20}} - 0.1 X_{b_{21}} - 0.1 X_{b_{22}} - 0.2 X_{b_{23}} - 0.1 X_{b_{24}} + X_{b_{25}} + X_{b_{26}} + 0.1 X_{b_{27}} + 0.1 X_{b_{28}} + 0.1 X_{b_{29}} + 0.1 X_{b_{30}} + 0.1 X_{b_{31}} + 0.2 C_{32} -$

$$0.1\,C_{33} + 0.5\,X_{b_{32}} + 0.2X_{b_{11}}\,X_{b_{12}}\,X_{b_{13}} - 0.2\,X_{b_{15}}X_{b_{18}} + 0.2X_{b_{12}}\,X_{b_{29}}. \quad (8)$$

$$X_{n_2} \sim N\,(\mu_4;\sqrt{0.5}).$$

Where, $\mu_4 = 3 - 0.5X_{b_1} - 0.2X_{b_2} + 0.05X_{b_3} - 0.6X_{b_4} - 0.08X_{b_5} + 0.01X_{b_6} + 0.2X_{b_7} + 0.2X_{b_8} + 0.1X_{b_9} - 0.1X_{b_{10}} + 0.2X_{b_{11}} + 0.5X_{b_{12}} - 0.2X_{b_{13}} + 0.3X_{b_{14}} + 0.4X_{b_{15}} + 0.1X_{b_{16}} + 0.1X_{b_{17}} - 0.1X_{b_{18}} - 0.1X_{b_{19}} - 0.1X_{b_{20}} - 0.1X_{b_{21}} - 0.1X_{b_{22}} - 0.2X_{b_{23}} - 0.1X_{b_{24}} + 0.1X_{b_{25}} + 0.1X_{b_{26}} + 0.1X_{b_{27}} + 0.1X_{b_{28}} + +0.1X_{b_{29}} + 0.1X_{b_{30}} + 0.1X_{b_{31}} + 0.2C_{32} - 0.1\,C_{33} + 0.5\,X_{b_{32}} + 0.2X_{b_{11}}\,X_{b_{12}}\,X_{b_{13}} - 0.2\,X_{b_{15}}X_{b_{18}} + 0.2X_{b_{12}}\,X_{b_{29}} + X_{n_1}$. (9)

Both continuous covariates are highly positively correlated i.e. $r = 0.9$.

Covariate dependent binary response y is generated from Bernoulli distributions with probabilities governed by the logistic regression with

$$logitPr(y) = -3 - 3X_{b_1} + 3X_{b_2} + 3X_{b_3} + 3X_{b_4} - 3X_{b_5} + 3X_{b_6} - 3X_{b_7} + 3X_{b_8} + 3X_{b_9} + 3X_{b_{10}} + 2X_{b_{11}} + 3X_{b_{12}} - 2X_{b_{13}} + 3X_{b_{14}} + 3X_{b_{15}} + 3X_{b_{16}} - 4X_{b_{17}} - 0.3X_{b_{18}} - 0.3X_{b_{19}} - 0.3X_{b_{20}} - 0.3X_{b_{21}} - 3X_{b_{22}} - 3X_{b_{23}} - 3X_{b_{24}} - 3X_{b_{25}} - 3X_{b_{26}} - 3X_{b_{27}} - 3X_{b_{28}} - 3X_{b_{29}} + 3X_{b_{30}} + 3X_{b_{31}} + 3X_{m_{1_2}} + 3X_{m_{1_3}} + 1X_{m_{1_4}} + 1X_{m_{1_5}} + 1X_{m_{1_6}} + 3X_{m_{2_2}} + 3X_{m_{2_3}} + 3X_{m_{2_4}} - 3X_{b_{32}} + 3X_{n_1} + 3\,X_{n_2} - 3X_{b_9}\,X_{b_{15}} - 3\,X_{b_1}X_{b_{17}} + 3X_{b_{13}}\,X_{b_{30}}$. (10)

Equations 5–10 include high-order interactions to represent the type of complex dependence structures. Imputation approaches based on log-linear models or chained equations may fail to capture these structures. There is no particular importance of the specific values of the coefficients. Nonzero coefficients are specified for higher order interactions for generating complex dependencies. The analysis model of interest is the GLMs with link "logit". The observations in all covariates are missing (at random) with the probabilities based on a logistic probability distribution model. Probabilities for a random covariate *X* are given as:

$$\pi_{X_i} = \frac{e^{(-2-X_j)}}{(1 + e^{(-2-X_j)})}. \quad (11)$$

Where $i=\{1,\ldots,39\}$ and $j \neq i$. Missingness in X_i is attributed solely to other observed variable X_j. This yields 10% of the observations to be MAR.

We use a JM technique called DPMPM MI for categorical variables. DPMPM MI technique is selected due its ability to identify complex dependencies structure among categorical variables and computational efficient qualities in high dimensions. We use a FCS technique called MICE for continuous variables. MICE is selected due to its popularity and applications in wide range of fields. For comparison, two MICE based MI methods namely "$Mice_{CART}$" (classification and regression trees (CART)) and "$Mice_{DEF}$" (which uses logistic regression models for categorical and "PMM" for continuous variables as default) are used. Proposed hybrid architectures are implemented as "H.CART" and "H.DEF". The mixtures of multinomial distributions approach is combined with the MICE algorithms "CART" and "Default" in H.CART" and "H.DEF" respectively. Further, we express "H.CART" as "$H.CART_1$" and "$H.CART_2$" indicating first and second hybrid architectures based on CART. Similarly first and second hybrid architectures based on "default" are expressed as "$H.DEF_1$" and "$H.DEF_2$" respectively. JM technique in hybrid architectures is implemented with prior specifications $a_\alpha = 0.25$, $b_\alpha = 0.25$, and somewhat large number of mixture components i.e. $k=80$. We used R (R Core Team 2018) version 3.0.1 to perform all calculations. The packages "mice" (van Buuren and Groothuis-Oudshoorn 2011), version 2.17 and "NPBayesImputeCat" (Quanli et al. 2018), version 0.6 were used to perform MICE for continuous data and Non-Parametric Bayesian MI for categorical variables, respectively. These blended versions of joint and sequential modeling MI techniques make it possible to obtain complete datasets with information available on both types of variables. The imputation model contains all of the variables from the generated data in order to preserve the relationships between the variables of interest (Schafer 1997; Moons et al. 2006; White et al. 2011; van Buuren 2012). The parameters of interest are estimated using Rubin's aforementioned method on $Z = 1000$ simulation runs. Ten

imputed data sets for each of the proposed and the MICE MI methods are generated for realistic applications (Fichman and Cummings 2003). Table 1 displays the performance of MI methods for simulated data. Graphical comparisons of the imputation methods based on boxplots (White et al. 2011; van Buuren 2012) of standard errors and point estimates across 1000 simulations for regression coefficients are presented in Figures 1 and 2 respectively.

4.1. Evaluation Criteria

The quality of MI methods is evaluated based on two error-based measurements i.e. root mean square error (RMSE) and empirical standard errors (ESE) (Akande 2017; Armina et al. 2017). RMSE is computed as a combination of the bias and variance of the estimate (Burton et.al 2006). ESEs can be considered to access the between imputation variations. The smaller values for RMSEs and ESEs indicate better performance (Oba et al. 2003). RMSE and ESE are calculated using the following formulas:

$$\text{Root mean square error (RMSE}_{\bar{q}_m}) = \sqrt{\frac{\sum_{z=1}^{Z}(\bar{q}_M^z - \beta)^2}{Z}}, \quad (12)$$

$$\text{Empirical standard errors (ESE}_{\bar{q}_m}) = \sqrt{\frac{\sum_{z=1}^{Z}(\bar{q}_M^z - \bar{q})^2}{Z}}, \quad (13)$$

where $\bar{q}_M^z$ denote the estimated parameter pooled over M imputed data sets and Z simulation runs and β denote original parameters.

4.2. Results

There seem to be similarities in structure among all MI methods i.e. all methods are upward biased for binary covariates e.g. X_{b_1}, whereas, the average point estimates based on default and H.DEF methods are closer to the corresponding true values as compared to other methods. CART and hybrid methods are slightly downward biased for multilevel covariate with six levels e.g. $X_{m_{1\_5}} X_{b_1}$. The average point estimates for multilevel covariate with six levels based on

CART and H.CART methods are closer to the corresponding true values as compared to H.DEF methods. All methods are downward biased for the interaction terms e.g. $X_{b_{13}} X_{b_{30}}$, whereas, the average point estimates based on default, CART, H.DEF methods and $H.CART_2$ method are closer to the corresponding true values as compared to $H.CART_1$ (Figure 1). Hybrid and CART methods tend to have smaller standard errors as compared to default method for all covariates, whereas the hybrid methods tend to have similar standard errors as compared to CART for most of the cases (Figure 2). The estimated ESEs for the all hybrid methods are smaller for all types of covariates except the binary covariate. H.DEF methods and $H.CART_2$ show similar or slightly higher ESEs as compared to default and CART methods for the binary covariate. The estimated ESEs for the $H.CART_1$ are smallest for the multilevel covariate with six levels and $H.DEF_2$ has smallest ESEs for the interaction terms. All hybrid methods tend to have smaller estimated RMSEs for binary covariate where $H.DEF_2$ has smallest RMSEs as compared to all methods. The estimated RMSEs for all hybrid methods are slimier to default and CART methods for the multilevel covariate with six levels whereas the $H.CART_1$ has the smallest RMSEs among others. Similarly for interaction term, all hybrid methods tend to have smaller RMSEs for most of the cases where $H.DEF_2$ shows smallest RMSE among the remaining methods (Table 1). The estimated ESEs(RMSEs) and averages of point estimates(standard errors) for all coefficients under hybrid architecture 1 and 2 are provided in supplementary file (Tables S1-S4). Boxplots for point estimates(standard errors) for all coefficients under hybrid architecture 1 and 2 are given in supplementary file (Figures S3-S18).

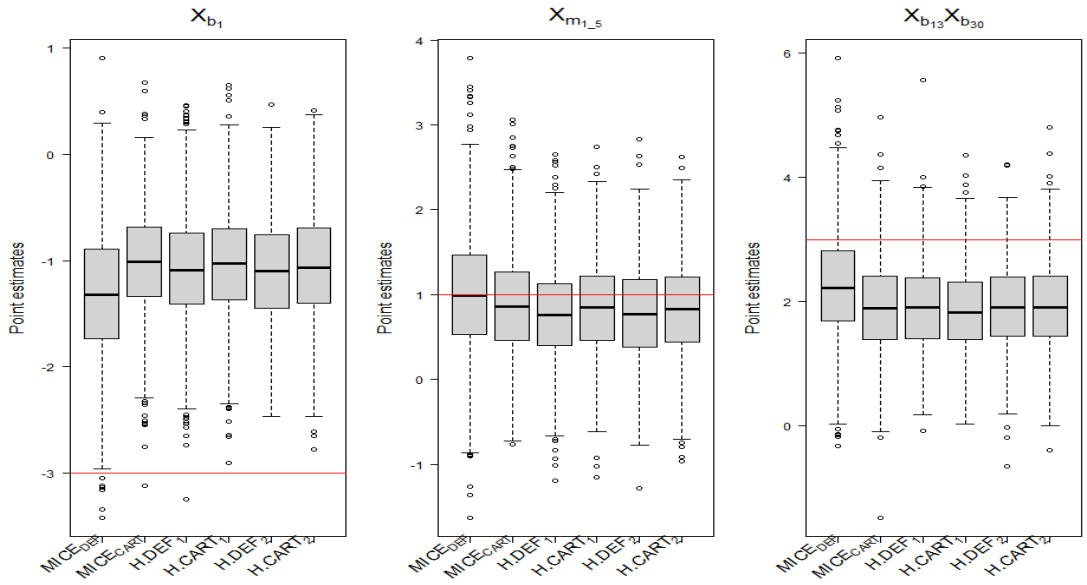


Figure1. Simulated data: Boxplots for the point estimates across 1000 simulations by imputation methods under Missing at Random (MAR) and ten imputations with 10% of missing data. Point estimates are shown for only three regression coefficients, i.e. for variables X_{b_1}, $X_{m_{1\_5}}$, $X_{b_{13}}$ $X_{b_{30}}$ .The horizontal red lines indicate the respective "true" values.

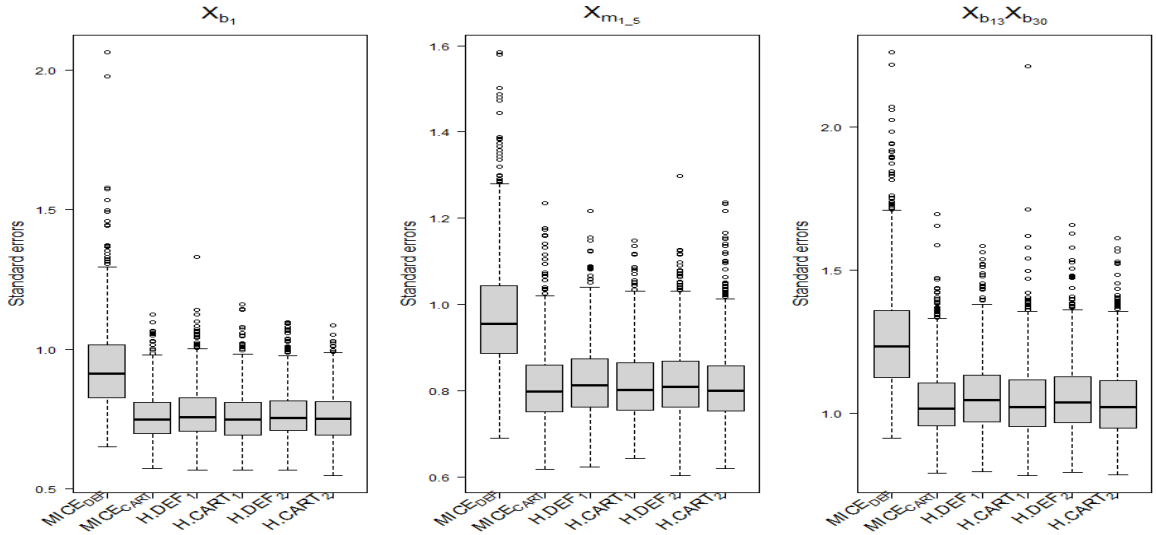


Figure2. Simulated data: Boxplots for the standard errors across 1000 simulations by imputation methods under Missing at Random (MAR) and ten imputations with 10% of missing data. Standard errors are shown for only three regression coefficients, i.e. for variables X_{b_1}, $X_{m_{1\_5}}$, $X_{b_{13}}$ $X_{b_{30}}$.

Table1. Simulated data: The performance of methods for MI based on RMSEs, ESEs (top), means of Rubin's estimates i.e. Est(point estimates) and SE(standard errors) (middle) and amount of bias (bottom) under Missing at Random (MAR) with 10% of missing data. Estimated bias is simply a difference between root mean square error and empirical standard error. All results are based on 10 imputations and 1000 simulations. Estimates are shown for only three regression coefficients (Coef.) i.e. for variables X_{b_1}, $X_{m_{1\_5}}$, $X_{b_{13}} X_{b_{30}}$. Bold figures indicate the smallest mean root mean square errors, mean empirical standard errors and amount of bias among various imputation variants.

	Coef.	$MICE_{DEF}$	$MICE_{CART}$	$H.DEF_1$	$H.CART_1$	$H.DEF_2$	$H.CART_2$
ESE_S (RMSEs)	X_{b_1}	0.51(2.04)	**0.51**(2.04)	0.53(**1.99**)	0.52(2.03)	**0.51**(**1.96**)	0.54(2.01)
	$X_{m_{1\_5}}$	0.59(0.60)	0.59(0.60)	0.57(0.61)	**0.55**(**0.58**)	0.57(0.61)	0.57(0.60)
	$X_{b_{13}} X_{b_{30}}$	0.75(1.34)	0.75(1.34)	0.72(1.31)	0.71(1.35)	**0.68**(**1.27**)	0.70(1.29)
Est(SE)	X_{b_1}	-1.329(0.935)	-1.029(0.760)	-1.084(0.773)	-1.037(0.759)	-1.106(0.768)	-1.061(**0.758**)
	$X_{m_{1\_5}}$	1(0.976)	0.876(**0.810**)	0.772(0.825)	0.835(0.814)	0.785(0.820)	0.833(0.813)
	$X_{b_{13}} X_{b_{30}}$	2.258(1.260)	1.893(**1.040**)	1.904(1.061)	1.8498(1.043)	1.927 (1.058)	1.920(1.041)
Bias	X_{b_1}	1.53	1.53	1.46	1.51	**1.45**	1.47
	$X_{m_{1\_5}}$	**0.01**	**0.01**	0.04	0.03	0.04	0.03
	$X_{b_{13}} X_{b_{30}}$	**0.59**	**0.59**	**0.59**	0.64	**0.59**	**0.59**

5. Motivation

Multiple Indicator Cluster Survey (MICS) is an international household survey tool. MICS is developed by the United Nations Children's Fund (UNICEF) to obtain internationally comparable, statistically rigorous data of standardised indicators related to the health situation of children and women. MICS household questionnaire contains information of following dimensions of household head life: education, household characteristics, water and sanitation, salt iodization, hand washing facilities, water quality testing and results etc. Such background variables are important for data analysis, modeling, and policy research.

National study like Government of Pakistan Economic survey (2008) highlighted that nearly 50 million individuals are deprived from safe drinking water in Pakistan. Our motivation stems from data obtained from MICS Punjab, 2014. MICS in Punjab was conducted in the

Punjab province of Pakistan with joint collaboration of the Bureau of Statistics (BOS) Punjab and the United Nations Children's Fund (UNICEF). Final and key findings report, survey plan, list of indicators, questionnaires and training agenda of MICS Punjab 2014 is available for download via a dedicated BOS Punjab website (www.bos.gop.pk). MICS Punjab questionnaire for household contains more than two hundred indicators on variety of household's conditions. For example indicators on house conditions (e.g. number of rooms used for sleeping, main material of floor and roof etc.), access to general facilities (e.g. electricity, radio, television, non-mobile phone, refrigerator etc.), source of drinking water (e.g. main source of drinking water and other purposes, location of the water source, duration to get water and come back, person collecting water, treatment for water to make safer for drinking etc.), sanitation facilities (e.g. type of toilet facility, water available at the place for hand washing, soap or detergent present at place of hand washing etc.). Binary logistic regressions models can be fitted to describe household trends in access to improved water sources and sanitation facilities. Associated factors like location, demographic and socio-economic etc. can be further use for prediction. Information based indicators described above can prove to be very useful in policy making in order to improve quality of drinking water and sanitation in Punjab.

5.1. Imputation of MICS Household Data

We use a secondary household data from the Punjab Multiple Indicator Cluster Survey in 2014 and use a GLM with a logit link is used to describe associations between access to water and sanitation, and geographic, demographic, and socio-economic factors. Most of the background variables related to geographic, demographic, and socio-economic characteristics in MICS data for household are categorical with many categories having complex data structures and large amount of missingness. For example geographical region of Punjab is divided into 36 districts.

Living area has two levels i.e. urban or rural. Statistical models based on survey data sets contain both, continuous and categorical variables and it can be tedious for MICE to specify imputation models and interaction terms in presence of such complications (Van Buuren, and Oudshoorn 1999). Therefore for the proper comparisons, multiple categories for categorical variables were reduced by merging them and a sub-sample of fifty seven variables is selected which contains information on water and sanitization, hand washing and household characteristics. For the sake of keeping the analysis comparable and challenging at the same time, variable "Main material of exterior walls" is included in the sub-sample which has fifteen levels. Among all these variables, forty nine variables are categorical with multiple categories and remaining are continuous, only two variables are fully observed. The missing data rates in most items were moderate. Items carrying great substantive importance, such as "Person collecting water", 83% values were missing; "Energy use for cooking" indicator was missing at approximately 68%; the indicator on whether the child needed to be physically punished to be brought up properly was missing at approximately 37% (see supplementary file (Tables S5-S6)). We assume items are MAR in data under consideration. The R package "VIM" (Templ et al. 2012) is utilized for exploring data and the pattern of missing values. Graphics for the all variables in sub sample are provided in a supplementary file (see Figures S19-S25).

5.2. Logistic Regression Model

To identify key determinants of water quality, we use a dichotomous variable indicating whether the household do anything to the water to make it safer to drink. That is,

$$WT = \begin{cases} 0 \;\; if \;\; household\ do\ not\ do\ anything\ to\ the\ water\ to\ make\ it\ safer\ to\ drink\ , \\ 1 \;\; if \;\; household\ do\ anything\ to\ the\ water\ to\ make\ it\ safer\ to\ drink\ . \end{cases}$$

where WT denotes water treatment status.

We determine two explanatory variables associated with the binary response "WT".

We then used a Logistic regression model, given by

$$logit(p) = ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \quad (14)$$

where X_1, X_2 are the predictor variables, “type of area (rural or urban)” and “soap/other material available for washing hands (yes or no)”, respectively and p denoted the probability that the household do not do anything to the water to make it safer to drink. The binary predictor “soap_avilb_wash_hand ” has the highest amount of missing values (i.e. about 9%) while the amount is rather small in the other two variables (i.e. less than 8% for response “treat_water_make_safe” and less than 6% for predictor “area”). See supplementary file for summary of all variables. Since there are no true values to compare for real data example, we calculated complete case (CC) estimates for comparison purpose. The CC analysis uses only the complete cases (i.e. n = 26819). The point estimates of GLM for “type of area” and “soap/other material available for washing hands” are 1.361 and 1.111 respectively. Whereas, standard errors for “type of area” and “soap/other material available for washing hands” are 0.106 and 0.052 respectively. Similar to simulation study, point estimates and standards for *M=10* completed data sets across 50 simulations are calculated for real data (see Figures 3-4). ESEs and means of point estimates (standard errors) and computational time for various MI methods are shown in Tables 2 and 3 respectively.

5.3. Results

We note that the standard errors for all of the coefficients are smaller compared to their point estimates under all MI methods (see Figures 3-4). The empirical example with real data indicated that the MICE methods and HMI variants yielded differing point estimates. We noticed that point estimates in both default and CART methods are nearer to the estimates in CC analysis for all cases with larger standard errors as compared to hybrid methods (see Table 2). Figure 4 displays smaller standard errors for hybrid variants (i.e. $H.DEF_1$, $H.CART_1$, $H.DEF_2$, $H.CART_2$) as compared to default and CART methods. ESEs and means of standard errors for hybrid variants are also smaller as compared to other methods (see Table 2) whereas these estimates are smaller for $H.DEF_2$ and $H.CART_2$ as compared to $H.DEF_1$ and $H.CART_1$, suggesting better performance over default and CART. Given the results produced by the MI methods, a look at the computation times in Table 3 may be useful for a further comparison. We found that hybrid variants ran quite fast followed by default method whereas, it took almost 5 days by CART to run on standard computers for a small subset of incomplete household data. Surprisingly, this time was reduced to almost half a day when hybrid methods were applied. We also found that hybrid variants also resulted in satisfactory performance when applied the full MICS household data set with hundreds of variables and categories with multiple levels whereas, methods based on MICE were not even able to run this large dataset due to complex structures. Thus, there exist significant differences in terms of the computational efficiency among the MI methods.

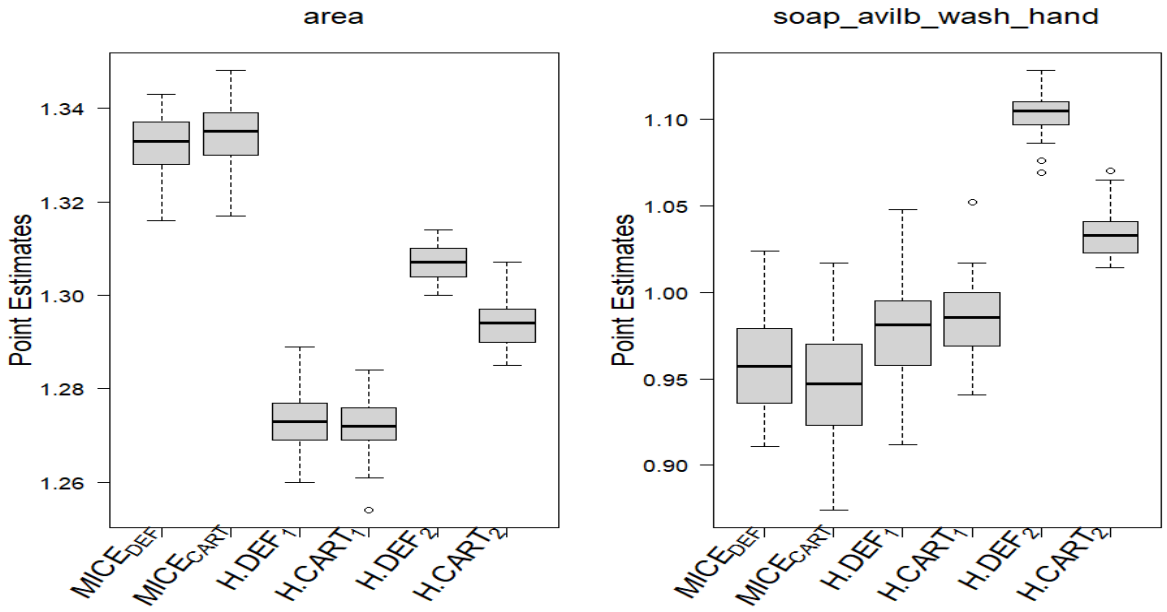


Figure3. Real data: Boxplots for point estimates across 50 simulations by imputation methods under Missing at Random (MAR) and ten imputations.

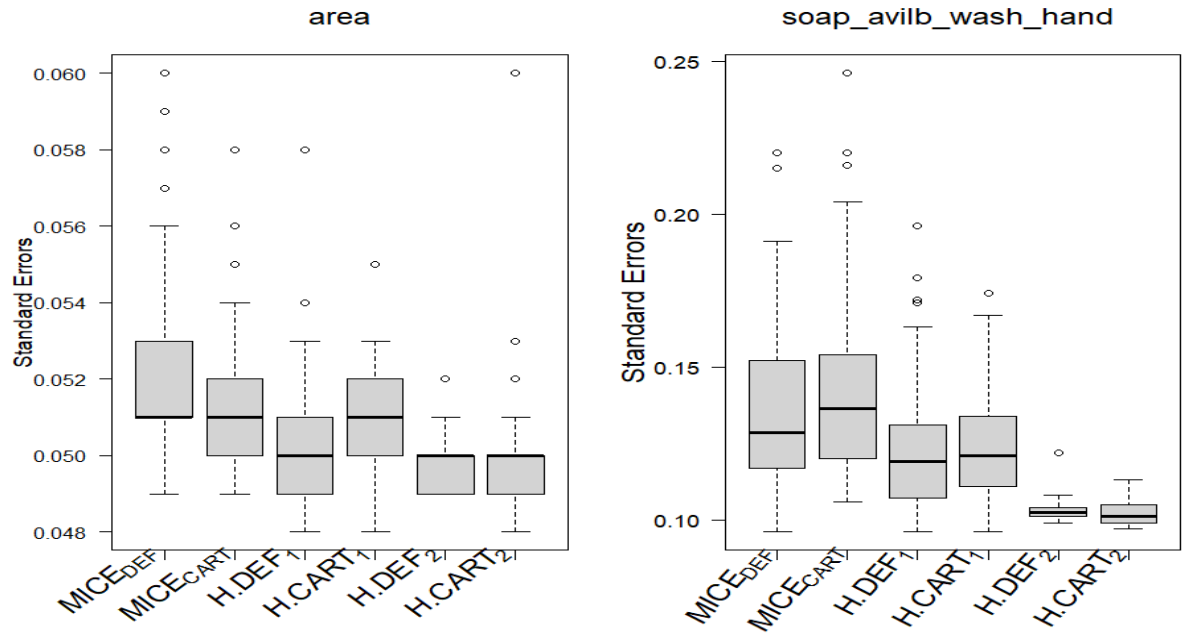


Figure4. Real data: Boxplots for standard errors across 50 simulations by imputation methods under Missing at Random (MAR) and ten imputations.

Table2. Real data: Means of point estimates (standard errors) for two categorical regression coefficients for *M=10* completed data sets across 50 simulations under various MI methods.

Estimates	Methods	Coefficients	
		area	soap_avilb_wash_hand
Means of Est(SE)	$MICE_{DEF}$	1.332(0.052)	0.957(0.137)
	$MICE_{CART}$	1.334(0.051)	0.947(0.143)
	$H.DEF_1$	1.272(**0.050**)	0.976(**0.124**)
	$H.CART_1$	1.271(**0.050**)	0.985(**0.124**)
	$H.DEF_2$	1.307(**0.049**)	1.103(**0.103**)
	$H.CART_2$	1.293(**0.050**)	1.034(**0.102**)
ESEs	$MICE_{DEF}$	0.0061	0.0290
	$MICE_{CART}$	0.0061	0.0350
	$H.DEF_1$	**0.0056**	**0.0286**
	$H.CART_1$	**0.0056**	**0.0209**
	$H.DEF_2$	**0.0032**	**0.0118**
	$H.CART_2$	**0.0045**	**0.0130**

Here Est and SE stand for point estimates and standard errors respectively. Cases where both Hybrid architectures result in minimum standard errors and ESEs as compared to default and CART are highlighted in bold.

Table3. Real data: Time taken for various MI methods

Method	$MICE_{DEF}$	$MICE_{CART}$	$H._{.}DEF_1$	$H.CART_1$	$H._{.}DEF_2$	$H.CART_2$
Time	2.37_d	4.87_d	$12.48_{h.}$	13.67_h	12.99_h	13.03_h

Note: time = the time to complete 10 multiple imputation by variants of MI across 1000 simulations, h = hours, d = days. The maximum number of iterations is set to 200.

6. Conclusion and future research

This paper describes the mechanisms of two hybrid strategies to handle missing data in large scale survey data with complex dependence structures among categorical variables and high percentage of missing information. After compering the performance of various multiple imputation algorithms, we showed that both proposed hybrid variants of the multiple imputation algorithms were clearly superior to MICE MI methods not only in terms of the accuracy of imputation, but were also markedly superior to the others in terms of the computational

efficiency. Practitioners can easily use our proposed methods to handle complex survey data because our techniques rely mostly on previously implemented algorithms. Our current work is limited to MAR mechanism, however, we believe that the biases due to wrongly assumed missingness mechanism are minimal when the imputation models are kept as rich as possible to the extent where they are estimable. We also believe that a data generating processes considered in simulation study can be generalized to a large number of situations. However, we have no sound grounding to prove that the comparisons we make here will always apply for any data. In particular, we have not yet considered alternative categorizations for continuous variables such as ordinal, unordered or multiple categories. Issues like convergence and appropriate selection of predictors is beyond the scope of the present paper. This study has for the first time provided an overview and a systematic comparison of previous approaches to MI for large scale complex data implemented in conditional models. We propose that the performance of proposed algorithms can be improved by extending the categorization process of continuous variables to ordinal or multiple categories. Since proposed approach requires the covariates to be strongly correlated in order to work properly, further evaluations with diversity of experimental settings will undoubtedly be needed to account for this.

References

Aßmann, C., Gaasch, C., Pohl, S., & Carstensen, C. (2016). Estimation of plausible values considering partially missing backround information: A data augmented MCMC approach. In H.-P. Blossfeld, J. von Maurice, J. Skopek, & M. Bayer (Eds.), *Methological Issues of Longitudinal Surveys* (pp. 505-522).Wiesbaden: Springer.

Audigier, V., Husson, F., & Josse, J. (2016). A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification*,10 (1), 5–26.

Audigier, V., Husson, F., & Josse, J. (2017). Mimca: multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and Computing,* 27 (2), 501–518.

Akande, O., LI, F., Reiter, J., (2017). An empirical comparison of multiple imputation methods for categorical data. *The American Statistician*,71, 162–170.

Armina, R., Zain, A.M., Ali, N.A., & Sallehuddin, R. (2017). A review on missing value estimation using imputation algorithm. *Journal of Physics: Conference Series*, 892(1), 4.

Audigier, V., I. White, I.R., Jolani, S., Debray, T., Quartagno, M., Carpenter, J., S. van Buuren, S., & Resche-Rigon, M. (2018). Multiple imputation for multilevel data with continuous and binary variables. *Statistical Science*, 33(2), 160-183.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. New York: Chapman and Hall.

Burton, A., Altman, D.G., Royston, P., & Holder, R.L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24), 4279–92.

Burgette, L. F., & Reiter, J. P. (2010). Multiple Imputation for Missing Data via Sequential Regression Trees. *American Journal of Epidemiology*, 172(9), 1070-1076.

Carpenter, J.R., & Kenward, M.G. (2011). REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software*, 45(5), 1–14.

Dunson, D. B., & Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104, 1042-1051.

Doove, L., van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72, 92-104.

Fichman, M., & Cummings, J. N. (2003). Multiple Imputation for Missing Data: Making the most of What you Know. *Organizational Research Methods*, 6(3), 282–308.

Government of Pakistan, *Economic Survey of Pakistan*. 2008–09.

Goßmann, S.D. (2016), The application of nonparametric data augmentation and imputation using classification and regression trees within a large-scale panel study, PhD Dissertation Presented to the Faculty for Social Sciences, Economics, and Business Administration at the University of Bamberg.

Hawkes, D., & Plewis, I. (2006). Modelling non-response in the National Child Development Study. *Journal of the Royal Statistical Society Series A,* 169. 479–491.

Horton, N.J., & Kleinman, K.P. (2007). Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *American Statistician*, 61,79–90.

Husson, F., Josse, J., Narasimhan, B., Robin, G., & Traumabase (2019): Imputation of mixed data with multilevel singular value decomposition, *Journal of Computational and Graphical Statistics*, DOI: 10.1080/10618600.2019.1585261

Kim, H., & Loh, W.-Y. (2001). Classification Trees With Unbiased Multiway Splits. *Journal of the American Statistical Association*, 96(454), 589-604.

Little, R. J. A., & Rubin, D. B. (2002). Statistical analysis with missing data. New York, Wiley.

Liu, J., Gelman, A., Hill, J., Su, Y.-S. & Kropko, J. (2014). On the stationary distribution of iterative imputations. *Biometrika*, 101, 155–173.

Mislevy, R.J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika,*56(2), 177-196.

Moons, K.G.M., Donders, R.A.R.T., Stijnen, T., & Harrell, F.E. (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology***,** **5**9(10), 1092–101.

Manrique-Vallier, D., & Reiter, J. P. (2014). Bayesian multiple imputation for large-scale categorical data with structural zeros. *Survey Methodology*, 40,125–134.

Manrique-Vallier, D., & Reiter, J. P. (2015). Bayesian simultaneous edit and imputation for multivariate categorical data. Technical Report. Dept. of Statistics, Duke University.

Murray, J. S. and Reiter, J. P. (2016). Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. *Journal of the American Statistical Association,* 111, 1466–1479.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, Series A 135, 370-384.

Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K., & Ishii, S. (2003), A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*,19, 2088 –2096.

Quartagno, M., & Carpenter, J. (2015). Multiple imputation for IPD meta-analysis: allowing for heterogeneity and studies with missing covariates. *Statistics in Medicine,*35(17), 2938–54.

Quanli, W., Danial, M.V., Reiter, J.P., & Jigchen, H. (2018). NPBayesImputeCat: Non-Parametric Bayesian Multiple Imputation for Categorical Data. R package version 0.1, https://CRAN.R-project.org/package=NPBayesImputeCat.

Rubin, D. B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. Journal of the American Statistical Association, 81, 366–374.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys.New York: John Wiley.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. S*urvey methodology*, 27(1), 85–96.

Raghunathan, T. E., Solenberger, P., & Van Hoewyk, J. (2002), IVEware: imputation and variance estimation software user guide. *Survey Research Center, Institute for Social Research*, University of Michigan.

Riphahn, R. T. & Serfling, O. (2005). Item Non-response on Income and Wealth Questions. *Empirical Economics,* 30(2), 521-538.

R Core Team (2018). R: A Language and Environment for Statistical Computing, Vienna, Austria: R Foundation for Statistical Computing.

Razzak, H., & Heumann, C. (2019). Predictive performance of a hybrid technique for the multiple imputation of survey data. Paper presented at NTTS 2019. Available at: https://coms.events/ntts2019/data/abstracts/en/abstract_0108.html.

Razzak, H., & Heumann, C. (2019). Hybrid multiple imputation in a large scale complex survey. *Statistics in Transition new series,* forthcoming.

Schafer, J.L. (1997). Analysis of incomplete multivariate data. New York: Chapman & Hall.

Su, Y.S., Gelman, A., Hill, J., & Yajima, M. (2011). Multiple Imputation with Diagnostics (mi) in R:Opening Windows into the Black Box. *Journal of Statistical Software*, 45(2), 1{31. URL: http://www.jstatsoft.org/v45/i02/.

Stekhoven*, D*. J., & Bühlmann, P. (2012). Missforest-non-parametric missing value imputation for mixed-type data.*Bioinformatics*, 28(1), 112-118.

Si, Y., & Reiter, J. P. (2013). Nonparametric bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of educational and behavioral statistics*, 38(5), 499-521.

Stata Corporation, Stata statistical software, Release 13, College Station, Texas, TX, USA. 2013.

SAS Institute, Base SAS 9. 4 Procedures Guide: Statistical Procedures. Cary: SAS Institute; 2014.

Schafer, J. L., & Zhao, J. H. (2014). pan: Multiple imputation for multivariate panel or clustered data (Version 0.9) [Computer software]. Retrieved from http://CRAN.R-project.org/package=pan

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. *American Journal of Epidemiology*, 179, 764–774.

Templ, M., Andreas, A., Alexander, K., & Bernd, P. (2012). VIM: Visualization and Imputation of Missing Values. http://cran.r-project.org/web/packages/VIM/VIM.pdf.

van Buuren, S., & Oudshoorn, C.G.M. (1999). Flexible multivariate imputation by MICE. Technical report, TNO Prevention and Health, Leiden.

van Buuren, S., & Brand, J.P., Groothuis-Oudshoorn, C., & Rubin, D.B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12), 1049–64.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3), 219–42.

Vermunt, J. K., Van Ginkel, J. R., Van der Ark, L. A., Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology*, 38, 369-397.

van Buuren, S., & Groothuis-Oudshoon, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), 1–67.

van Buuren, S. 2012. Flexible imputation of missing data. Florida: CRC press.

White, I.R., Royston, P., & Wood, A.M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, 30(4), 377–99.

Zhu, J. & Raghunathan, T.E. (2015). Convergence properties of a sequential regression multiple imputation algorithm. *Journal of the American Statistical Association*, 110(511),1112–1124.

Supplementary file

TableS1. ESEs and RMSEs for all coefficients for various MI methods and hybrid architecture 1

	ESEs				RMSEs			
Coef.	**$MICE_{DEF}$**	**$MICE_{CART}$**	**$H.DEF_1$**	**$H.CART_1$**	**$MICE_{DEF}$**	**$MICE_{CART}$**	**$H.DEF_1$**	**$H.CART_1$**
X_{b_1}	0.51	0.51	0.53	0.52	2.04	2.04	1.99	2.03
X_{b_2}	0.41	0.41	0.40	0.41	1.38	1.38	1.31	1.39
X_{b_3}	0.40	0.40	0.41	0.40	1.57	1.57	1.53	1.55
X_{b_4}	0.40	0.40	0.42	0.40	1.29	1.29	1.23	1.32
X_{b_5}	0.44	0.44	0.42	0.44	1.42	1.42	1.48	1.43
X_{b_6}	0.40	0.40	0.41	0.41	1.65	1.65	1.64	1.64
X_{b_7}	0.41	0.41	0.40	0.40	1.24	1.24	1.21	1.24
X_{b_8}	0.41	0.41	0.42	0.42	1.46	1.46	1.39	1.45
X_{b_9}	0.48	0.48	0.48	0.49	1.46	1.46	1.44	1.49
$X_{b_{10}}$	0.39	0.39	0.40	0.41	1.32	1.32	1.28	1.32
$X_{b_{11}}$	0.40	0.40	0.39	0.39	0.88	0.88	0.87	0.85
$X_{b_{12}}$	0.68	0.68	0.67	0.65	1.03	1.03	0.92	0.84
$X_{b_{13}}$	0.49	0.49	0.49	0.48	0.98	0.98	0.98	1.00
$X_{b_{14}}$	0.40	0.40	0.42	0.40	1.36	1.36	1.37	1.38
$X_{b_{15}}$	0.51	0.51	0.50	0.50	1.93	1.93	1.94	1.95
$X_{b_{16}}$	0.41	0.41	0.41	0.41	1.18	1.18	1.18	1.19
$X_{b_{17}}$	0.58	0.58	0.60	0.56	1.46	1.46	1.42	1.43
$X_{b_{18}}$	0.39	0.39	0.39	0.39	1.49	1.49	1.47	1.47
$X_{b_{19}}$	0.43	0.43	0.43	0.43	0.98	0.98	0.98	0.99
$X_{b_{20}}$	0.39	0.39	0.40	0.38	1.98	1.98	1.94	1.95
$X_{b_{21}}$	0.36	0.36	0.39	0.37	1.52	1.52	1.47	1.49
$X_{b_{22}}$	0.40	0.40	0.41	0.38	1.61	1.61	1.57	1.57
$X_{b_{23}}$	0.42	0.42	0.41	0.42	1.55	1.55	1.53	1.54
$X_{b_{24}}$	0.42	0.42	0.43	0.39	1.43	1.43	1.38	1.40
$X_{b_{25}}$	0.44	0.44	0.42	0.41	1.37	1.37	1.26	1.35
$X_{b_{26}}$	0.41	0.41	0.42	0.41	1.76	1.76	1.66	1.74
$X_{b_{27}}$	0.42	0.42	0.42	0.41	1.64	1.64	1.61	1.64
$X_{b_{28}}$	0.39	0.39	0.41	0.40	1.48	1.48	1.45	1.48
$X_{b_{29}}$	0.42	0.42	0.44	0.42	1.38	1.38	1.30	1.32
$X_{b_{30}}$	0.47	0.47	0.47	0.47	1.58	1.58	1.56	1.54
$X_{b_{31}}$	0.42	0.42	0.42	0.41	1.69	1.69	1.59	1.63
$X_{m_{1\_2}}$	0.48	0.48	0.46	0.45	1.30	1.30	1.36	1.36
$X_{m_{1\_3}}$	0.51	0.51	0.51	0.48	1.17	1.17	1.27	1.24
$X_{m_{1\_4}}$	0.67	0.67	0.63	0.64	0.71	0.71	0.76	0.72
$X_{m_{1\_5}}$	0.59	0.59	0.57	0.55	0.60	0.60	0.61	0.58
$X_{m_{1\_6}}$	0.75	0.75	0.74	0.71	0.83	0.83	0.88	0.77
$X_{m_{2\_2}}$	0.52	0.52	0.51	0.50	1.64	1.64	1.59	1.61
$X_{m_{2\_3}}$	0.80	0.80	0.78	0.79	2.27	2.27	2.19	2.23
$X_{m_{2\_4}}$	1.10	1.10	1.06	1.06	2.61	2.61	2.55	2.60
X_{n_1}	0.35	0.35	0.34	0.33	1.51	1.51	1.60	1.61

X_{n_2}		0.21	0.20	0.21	1.32	1.32	1.28	1.31
$X_{b_{32}}$	0.21	0.11	0.12	0.11	0.36	0.36	0.37	0.39
$X_{b_9}X_{b_{15}}$	0.11	0.69	0.70	0.71	1.75	1.75	1.77	1.79
$X_{b_1}X_{b_{17}}$	0.69	0.71	0.76	0.72	1.61	1.61	1.60	1.58
$X_{b_{13}}X_{b_{30}}$	0.71	0.75	0.72	0.71	1.34	1.34	1.31	1.35
	0.75							

TableS2. Point estimates and Standard errors for all coefficients under various MI methods and hybrid architecture 1.

	Point estimates				Standard errors			
Coef.	**$MICE_{DEF}$**	**$MICE_{CART}$**	**$H.DEF_1$**	**$H.CART_1$**	**$MICE_{DEF}$**	**$MICE_{CART}$**	**$H.DEF_1$**	**$H.CART_1$**
X_{b_1}	-1.329	-1.029	-1.084	-1.037	0.935	0.760	0.773	0.759
X_{b_2}	2.183	1.681	1.754	1.674	0.754	0.596	0.608	0.598
X_{b_3}	1.887	1.481	1.530	1.502	0.744	0.588	0.604	0.595
X_{b_4}	2.230	1.776	1.848	1.745	0.767	0.604	0.613	0.601
X_{b_5}	-1.981	-1.654	-1.581	-1.634	0.756	0.610	0.612	0.608
X_{b_6}	1.816	1.404	1.408	1.416	0.731	0.580	0.586	0.581
X_{b_7}	-2.245	-1.831	-1.858	-1.827	0.737	0.581	0.591	0.583
X_{b_8}	2.017	1.600	1.679	1.612	0.757	0.604	0.615	0.602
X_{b_9}	1.961	1.616	1.640	1.592	0.821	0.674	0.682	0.673
$X_{b_{10}}$	2.242	1.743	1.789	1.741	0.750	0.593	0.604	0.594
$X_{b_{11}}$	1.474	1.219	1.221	1.244	0.697	0.570	0.574	0.568
$X_{b_{12}}$	3.055	2.229	2.372	2.470	1.239	0.985	0.987	0.979
$X_{b_{13}}$	-1.418	-1.154	-1.156	-1.121	0.867	0.708	0.714	0.702
$X_{b_{14}}$	2.127	1.696	1.692	1.676	0.744	0.586	0.604	0.587
$X_{b_{15}}$	1.417	1.136	1.123	1.114	0.890	0.730	0.734	0.722
$X_{b_{16}}$	2.346	1.889	1.896	1.884	0.724	0.575	0.579	0.574
$X_{b_{17}}$	-3.259	-2.666	-2.711	-2.682	1.024	0.822	0.832	0.826
$X_{b_{18}}$	-1.947	-1.558	-1.579	-1.582	0.728	0.579	0.588	0.578
$X_{b_{19}}$	-2.636	-2.116	-2.115	-2.102	0.763	0.602	0.612	0.606
$X_{b_{20}}$	-1.369	-1.062	-1.103	-1.090	0.698	0.565	0.575	0.565
$X_{b_{21}}$	-1.964	-1.526	-1.583	-1.557	0.707	0.561	0.571	0.561
$X_{b_{22}}$	-1.850	-1.436	-1.485	-1.481	0.717	0.571	0.582	0.574
$X_{b_{23}}$	-1.869	-1.504	-1.521	-1.514	0.730	0.590	0.597	0.589
$X_{b_{24}}$	-2.090	-1.634	-1.688	-1.653	0.738	0.586	0.597	0.586
$X_{b_{25}}$	-2.117	-1.703	-1.815	-1.713	0.745	0.588	0.602	0.594
$X_{b_{26}}$	-1.738	-1.291	-1.391	-1.307	0.721	0.576	0.591	0.577
$X_{b_{27}}$	-1.760	-1.417	-1.445	-1.409	0.765	0.612	0.624	0.616
$X_{b_{28}}$	-1.924	-1.575	-1.614	-1.575	0.733	0.590	0.595	0.590
$X_{b_{29}}$	-2.181	-1.688	-1.774	-1.745	0.773	0.606	0.618	0.611
$X_{b_{30}}$	1.981	1.490	1.506	1.534	0.911	0.733	0.742	0.732
$X_{b_{31}}$	1.812	1.368	1.470	1.423	0.776	0.621	0.637	0.621
	2.204	1.794	1.717	1.718	0.818	0.669	0.672	0.658

$X_{m_{1\_2}}$	2.246	1.946	1.840	1.851	0.827	0.686	0.686	0.677
$X_{m_{1\_3}}$	0.806	0.764	0.568	0.678	1.074	0.897	0.907	0.894
$X_{m_{1\_4}}$	1.000	0.876	0.772	0.835	0.976	0.810	0.825	0.814
$X_{m_{1\_5}}$	0.892	0.635	0.527	0.693	1.404	1.132	1.153	1.140
$X_{m_{1\_6}}$	1.797	1.440	1.492	1.464	0.989	0.801	0.808	0.795
$X_{m_{2\_2}}$	1.129	0.881	0.958	0.913	1.563	1.271	1.283	1.265
$X_{m_{2\_3}}$	0.674	0.630	0.680	0.623	2.224	1.806	1.818	1.803
$X_{m_{2\_4}}$	-1.832	-1.531	-1.433	-1.421	0.628	0.501	0.504	0.496
X_{n_1}	1.996	1.695	1.735	1.707	0.429	0.326	0.332	0.321
X_{n_2}	0.774	0.663	0.645	0.624	0.215	0.169	0.171	0.166
$X_{b_{32}}$	-1.592	-1.394	-1.373	-1.351	1.194	0.996	1.010	1.003
$X_{b_9}X_{b_{15}}$	-1.973	-1.557	-1.592	-1.587	1.320	1.092	1.119	1.109
$X_{b_1}X_{b_{17}}$	2.258	1.893	1.904	1.849	1.260	1.040	1.061	1.043
$X_{b_{13}}X_{b_{30}}$								

TableS3. ESEs and RMSEs for all coefficients for various MI methods and hybrid architecture 2

	ESEs				RMSEs			
Coef.	**$MICE_{DEF}$**	**$MICE_{CART}$**	**$H.DEF_2$**	**$H.CART_2$**	**$MICE_{DEF}$**	**$MICE_{CART}$**	**$H.DEF_2$**	**$H.CART_2$**
X_{b_1}	0.65	0.51	0.51	0.54	1.79	2.04	1.96	2.01
X_{b_2}	0.53	0.41	0.42	0.41	0.97	1.38	1.32	1.36
X_{b_3}	0.52	0.40	0.40	0.40	1.23	1.57	1.54	1.56
X_{b_4}	0.55	0.40	0.41	0.40	0.95	1.29	1.20	1.31
X_{b_5}	0.54	0.44	0.43	0.42	1.15	1.42	1.49	1.44
X_{b_6}	0.51	0.40	0.42	0.41	1.29	1.65	1.68	1.65
X_{b_7}	0.53	0.41	0.41	0.39	0.93	1.24	1.21	1.26
X_{b_8}	0.54	0.41	0.42	0.41	1.12	1.46	1.38	1.46
X_{b_9}	0.60	0.48	0.48	0.51	1.20	1.46	1.44	1.50
$X_{b_{10}}$	0.55	0.39	0.41	0.41	0.94	1.32	1.30	1.35
$X_{b_{11}}$	0.50	0.40	0.38	0.40	0.73	0.88	0.87	0.86
$X_{b_{12}}$	0.94	0.68	0.65	0.69	0.94	1.03	0.94	0.89
$X_{b_{13}}$	0.61	0.49	0.49	0.49	0.84	0.98	0.96	0.97
$X_{b_{14}}$	0.56	0.40	0.39	0.40	1.04	1.36	1.36	1.39
$X_{b_{15}}$	0.59	0.51	0.49	0.49	1.69	1.93	1.94	1.94
$X_{b_{16}}$	0.54	0.41	0.40	0.40	0.85	1.18	1.20	1.18
$X_{b_{17}}$	0.78	0.58	0.58	0.58	1.07	1.46	1.45	1.49
$X_{b_{18}}$	0.55	0.39	0.39	0.39	1.19	1.49	1.49	1.48
$X_{b_{19}}$	0.59	0.43	0.42	0.43	0.69	0.98	0.99	1.01
$X_{b_{20}}$	0.50	0.39	0.39	0.40	1.71	1.98	1.94	1.96
$X_{b_{21}}$	0.49	0.36	0.38	0.36	1.14	1.52	1.50	1.49
$X_{b_{22}}$	0.52	0.40	0.41	0.41	1.26	1.61	1.56	1.58
$X_{b_{23}}$	0.54	0.42	0.42	0.41	1.25	1.55	1.52	1.53
$X_{b_{24}}$	0.54	0.42	0.42	0.42	1.06	1.43	1.37	1.39
	0.55	0.44	0.42	0.43	1.04	1.37	1.27	1.35

$X_{b_{25}}$	0.53	0.41	0.41	0.41	1.37	1.76	1.65	1.74
$X_{b_{26}}$	0.54	0.42	0.41	0.42	1.35	1.64	1.63	1.64
$X_{b_{27}}$	0.54	0.39	0.42	0.41	1.20	1.48	1.48	1.45
$X_{b_{28}}$	0.57	0.42	0.43	0.43	1.00	1.38	1.29	1.33
$X_{b_{29}}$	0.61	0.47	0.48	0.47	1.19	1.58	1.59	1.57
$X_{b_{30}}$	0.54	0.42	0.42	0.41	1.30	1.69	1.59	1.64
$X_{b_{31}}$	0.62	0.48	0.46	0.47	1.01	1.30	1.35	1.36
$X_{m_{1\_2}}$	0.64	0.51	0.49	0.49	0.99	1.17	1.25	1.24
$X_{m_{1\_3}}$	0.81	0.67	0.65	0.63	0.83	0.71	0.77	0.69
$X_{m_{1\_4}}$	0.72	0.59	0.57	0.57	0.72	0.60	0.61	0.60
$X_{m_{1\_5}}$	0.97	0.75	0.71	0.71	0.98	0.83	0.86	0.79
$X_{m_{1\_6}}$	0.70	0.52	0.51	0.50	1.39	1.64	1.59	1.61
$X_{m_{2\_2}}$	1.16	0.80	0.79	0.77	2.20	2.27	2.20	2.19
$X_{m_{2\_3}}$	1.61	1.10	1.09	1.06	2.83	2.61	2.56	2.55
$X_{m_{2\_4}}$	0.43	0.35	0.32	0.34	1.25	1.51	1.61	1.62
X_{n_1}	0.27	0.21	0.20	0.22	1.04	1.32	1.27	1.31
X_{n_2}	0.15	0.11	0.11	0.11	0.27	0.36	0.38	0.40
$X_{b_{32}}$	0.75	0.69	0.67	0.67	1.60	1.75	1.77	1.78
$X_{b_9}X_{b_{15}}$	0.89	0.71	0.71	0.75	1.36	1.61	1.57	1.60
$X_{b_1}X_{b_{17}}$	0.86	0.75	0.68	0.70	1.14	1.34	1.27	1.29
$X_{b_{13}}X_{b_{30}}$								

TableS4. Point estimates and Standard errors for all coefficients under various MI methods and hybrid architecture 2

Point estimates					Standard errors			
Coef.	**$MICE_{DEF}$**	**$MICE_{CART}$**	**$H.DEF_2$**	**$H.CART_2$**	**$MICE_{DEF}$**	**$MICE_{CART}$**	**$H.DEF_2$**	**$H.CART_2$**
X_{b_1}	-1.329	-1.029	-1.106	-1.061	0.935	0.760	0.768	0.758
X_{b_2}	2.183	1.681	1.754	1.701	0.754	0.596	0.605	0.596
X_{b_3}	1.887	1.481	1.517	1.489	0.744	0.588	0.602	0.590
X_{b_4}	2.230	1.776	1.869	1.754	0.767	0.604	0.615	0.598
X_{b_5}	-1.981	-1.654	-1.574	-1.622	0.756	0.610	0.609	0.604
X_{b_6}	1.816	1.404	1.371	1.402	0.731	0.580	0.584	0.579
X_{b_7}	-2.245	-1.831	-1.861	-1.799	0.737	0.581	0.590	0.580
X_{b_8}	2.017	1.600	1.685	1.595	0.757	0.604	0.612	0.598
X_{b_9}	1.961	1.616	1.640	1.591	0.821	0.674	0.676	0.669
$X_{b_{10}}$	2.242	1.743	1.770	1.710	0.750	0.593	0.602	0.595
$X_{b_{11}}$	1.474	1.219	1.221	1.242	0.697	0.570	0.576	0.565
$X_{b_{12}}$	3.055	2.229	2.321	2.431	1.239	0.985	0.986	0.972
$X_{b_{13}}$	-1.418	-1.154	-1.175	-1.165	0.867	0.708	0.713	0.700
$X_{b_{14}}$	2.127	1.696	1.701	1.670	0.744	0.586	0.598	0.584
$X_{b_{15}}$	1.417	1.136	1.119	1.124	0.890	0.730	0.730	0.720
$X_{b_{16}}$	2.346	1.889	1.874	1.890	0.724	0.575	0.578	0.574
$X_{b_{17}}$	-3.259	-2.666	-2.669	-2.628	1.024	0.822	0.829	0.820

$X_{b_{18}}$	-1.947	-1.558	-1.562	-1.577	0.728	0.579	0.585	0.580
$X_{b_{19}}$	-2.636	-2.116	-2.098	-2.086	0.763	0.602	0.612	0.598
$X_{b_{20}}$	-1.369	-1.062	-1.100	-1.077	0.698	0.565	0.572	0.564
$X_{b_{21}}$	-1.964	-1.526	-1.548	-1.551	0.707	0.561	0.568	0.563
$X_{b_{22}}$	-1.850	-1.436	-1.492	-1.478	0.717	0.571	0.582	0.570
$X_{b_{23}}$	-1.869	-1.504	-1.539	-1.521	0.730	0.590	0.596	0.582
$X_{b_{24}}$	-2.090	-1.634	-1.693	-1.679	0.738	0.586	0.592	0.584
$X_{b_{25}}$	-2.117	-1.703	-1.797	-1.716	0.745	0.588	0.602	0.592
$X_{b_{26}}$	-1.738	-1.291	-1.406	-1.308	0.721	0.576	0.592	0.577
$X_{b_{27}}$	-1.760	-1.417	-1.426	-1.417	0.765	0.612	0.620	0.612
$X_{b_{28}}$	-1.924	-1.575	-1.586	-1.609	0.733	0.590	0.593	0.588
$X_{b_{29}}$	-2.181	-1.688	-1.786	-1.744	0.773	0.606	0.619	0.608
$X_{b_{30}}$	1.981	1.490	1.489	1.504	0.911	0.733	0.740	0.727
$X_{b_{31}}$	1.812	1.368	1.469	1.412	0.776	0.621	0.635	0.617
$X_{m_{1\_2}}$	2.204	1.794	1.728	1.729	0.818	0.669	0.671	0.660
$X_{m_{1\_3}}$	2.246	1.946	1.852	1.864	0.827	0.686	0.688	0.679
$X_{m_{1\_4}}$	0.806	0.764	0.596	0.720	1.074	0.897	0.903	0.895
$X_{m_{1\_5}}$	1.000	0.876	0.785	0.833	0.976	0.810	0.820	0.813
$X_{m_{1\_6}}$	0.892	0.635	0.515	0.658	1.404	1.132	1.149	1.130
$X_{m_{2\_2}}$	1.797	1.440	1.495	1.469	0.989	0.801	0.801	0.793
$X_{m_{2\_3}}$	1.129	0.881	0.949	0.953	1.563	1.271	1.274	1.259
$X_{m_{2\_4}}$	0.674	0.630	0.687	0.685	2.224	1.806	1.811	1.791
X_{n_1}	-1.832	-1.531	-1.424	-1.413	0.628	0.501	0.500	0.493
X_{n_2}	1.996	1.695	1.742	1.708	0.429	0.326	0.331	0.319
$X_{b_{32}}$	0.774	0.663	0.636	0.620	0.215	0.169	0.169	0.166
$X_{b_9}X_{b_{15}}$	-1.592	-1.394	-1.364	-1.350	1.194	0.996	1.006	0.995
$X_{b_1}X_{b_{17}}$	-1.973	-1.557	-1.593	-1.587	1.320	1.092	1.122	1.110
$X_{b_{13}}X_{b_{30}}$	2.258	1.893	1.927	1.920	1.260	1.040	1.058	1.041

TableS5. Real data: Summary of all categorical variables

No.	Variable	Description	Levels	%miss
1	T.fuel	Energy use for cooking	3	68
2	Cooking_loc	Cooking location	3	43
3	physically_punished	Child needs to be physically punished to be brought up properly	2	37
4	Mother_tongue	Mother tongue of household head	4	7
5	Elec	Electricity	2	7
6	material_floor	Main material of flooring	3	7
7	material_exterior	Main material of exterior walls	15	7
8	area	Area of Residence	2	5
9	refrigrator	Refrigerator	2	7
10	wash_machine.dryer	Washing machine/ Dryer	2	7
11	A.C	Air conditioner	2	7

12	Air_cooler.fan	Air cooler/ Fan	2	7
13	copmuter	Computer	2	7
14	Radio	Radio	2	7
15	no _mobile	Non-mobile phone	2	7
16	gas	Gas	2	7
17	water_filter	Water filter	2	7
18	Microwave	Cooking range/ Micro wave	2	7
19	sew.nitt_machine	Sewing/ Knitting Machine	2	7
20	iron	Iron	2	7
21	Dunkey_pump.turbine	Dunky pump/ Turbine	2	7
22	watch	Watch	2	7
23	Trac_troly	Tractor trolley	2	7
24	Bicycle	Bicycle	2	7
25	Animal_drawn_cart	Animal-drawn cart	2	7
26	motercycle	Motorcycle or scooter	2	7
27	boat_w_moter	Boat with motor	2	7
28	car_or_van	Car or Van	2	7
29	Bus.truck	Bus or truck	2	7
30	mobile	Mobile telephone	2	7
31	soap_avilb_wash_hand	Soap or detergent present at place of handwashing	2	9
32	water_place_hand_wash	Water available at the place for handwashing	2	9
33	gov_init_lowincome	Government initiatives are benifiting the low income groups	2	7
34	HH_rec_remmitence	HH recieved any remittances during last year	2	7
35	HH_rec_pension	Any HH member recieved any pension benefits during last year	2	7
36	HH_bought_utility_store	HH purchased consumable items from utility store	2	7
37	HH_rec_benif_gov	HH received any benifit from Government	2	7
38	memb_outside.V.C.	Family member working outside village/city/country	2	7
39	sex_head_HH	Sex of household head	2	7
40	fam_memb_work_outside	Number of HH member working outside		7
41	person_coll_water	Person collecting water	7	83
42	loc_water_source	Location of the water source	2	19
43	bank_acc_saving_sertif	Any household member have account in Bank, PO or National Saving Centre		7
44	HH_own_animal	Household own any animals	2	7
45	HH_own_dwelling	Household owns the dwelling	3	7
46	treat_water_make_safe	Treat water to make safer for drinking	2	7

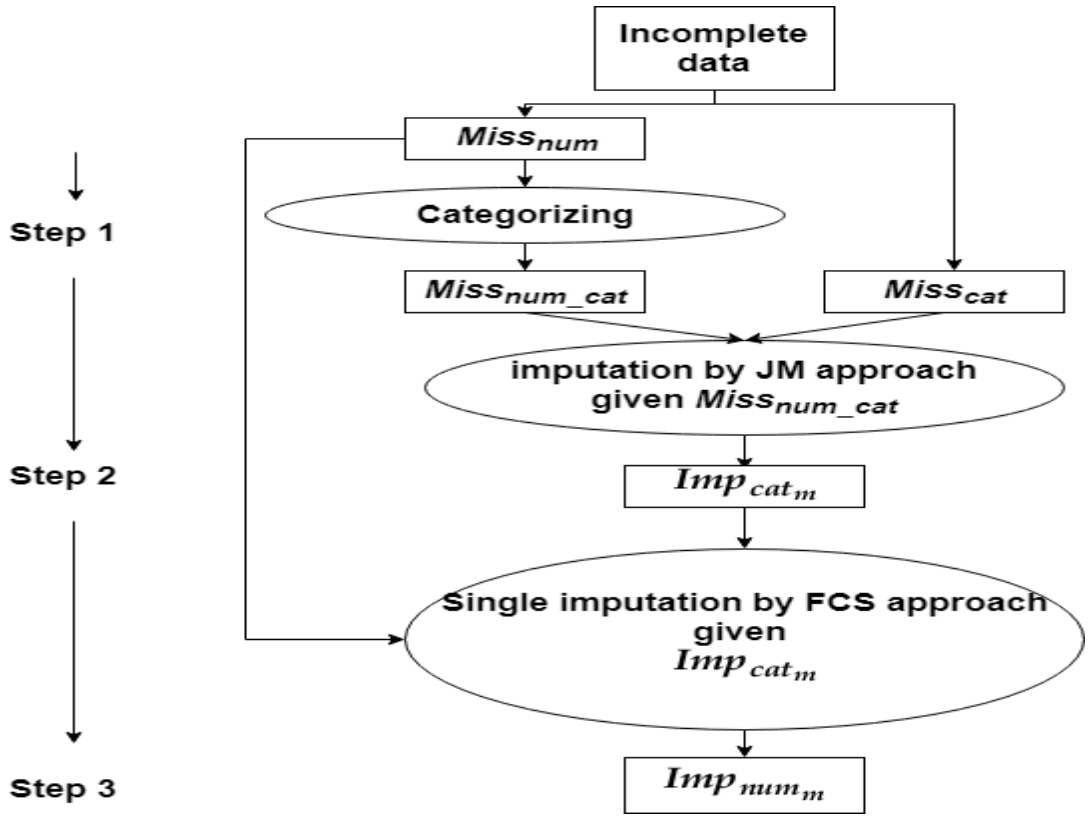
47	HH_own_land_agri	Any household member own land that can be used for agriculture	2	7
48	Type.of.toilet.facility	Type of toilet facility	13	7
49	T.V.	Television	2	7

"Levels" indicates number categories of categorical variables and "% mis" indicates percentage of missing observations in all variables.

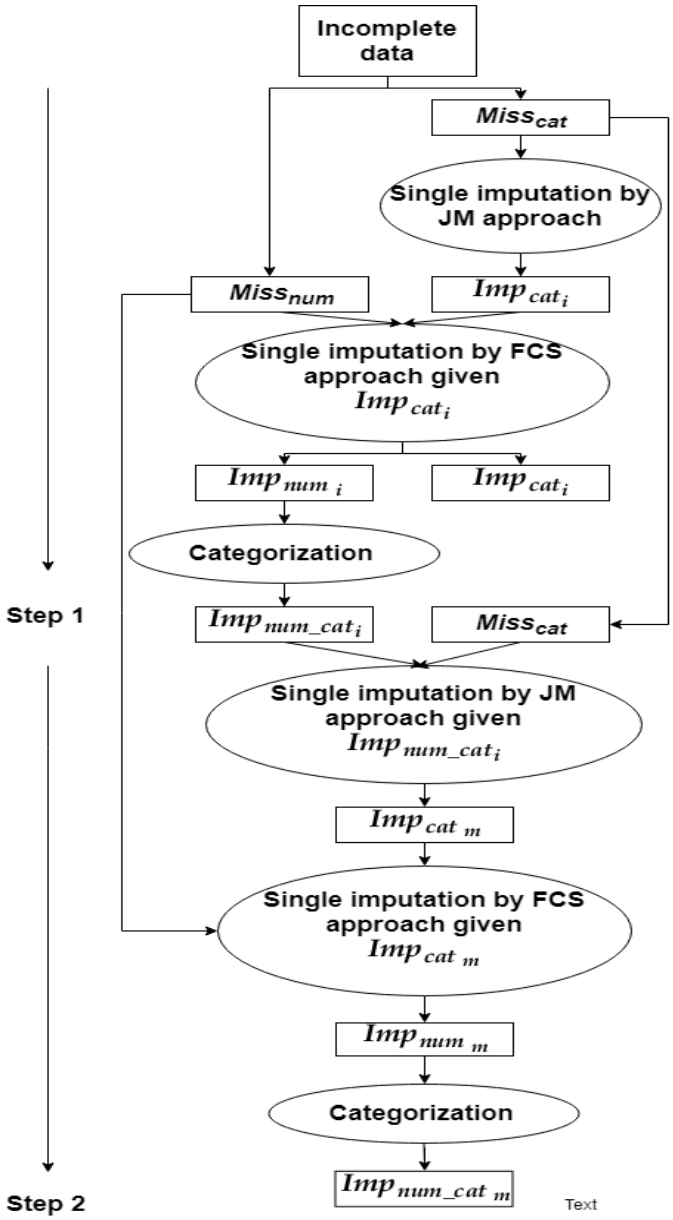
TableS6. Real data: Summary of all continuous variables

No.	Variabels	Discription	%miss
1	time_inmin_get_water	Time (in minutes) to get water and come back	83
2	no.HHmem	Number of HH members	13
3	T.C.age_1_17	Total children aged 1-17 years	7
4	no.W._15_19	Number of women 15 - 49 years	7
5	No_rooms_use_sleeping	Number of rooms used for sleeping	7
6	no.C._und5	Number of children under age 5	7
7	hhweight	Household sample weight	0
8	stweight	Salt testing's sample weight	0

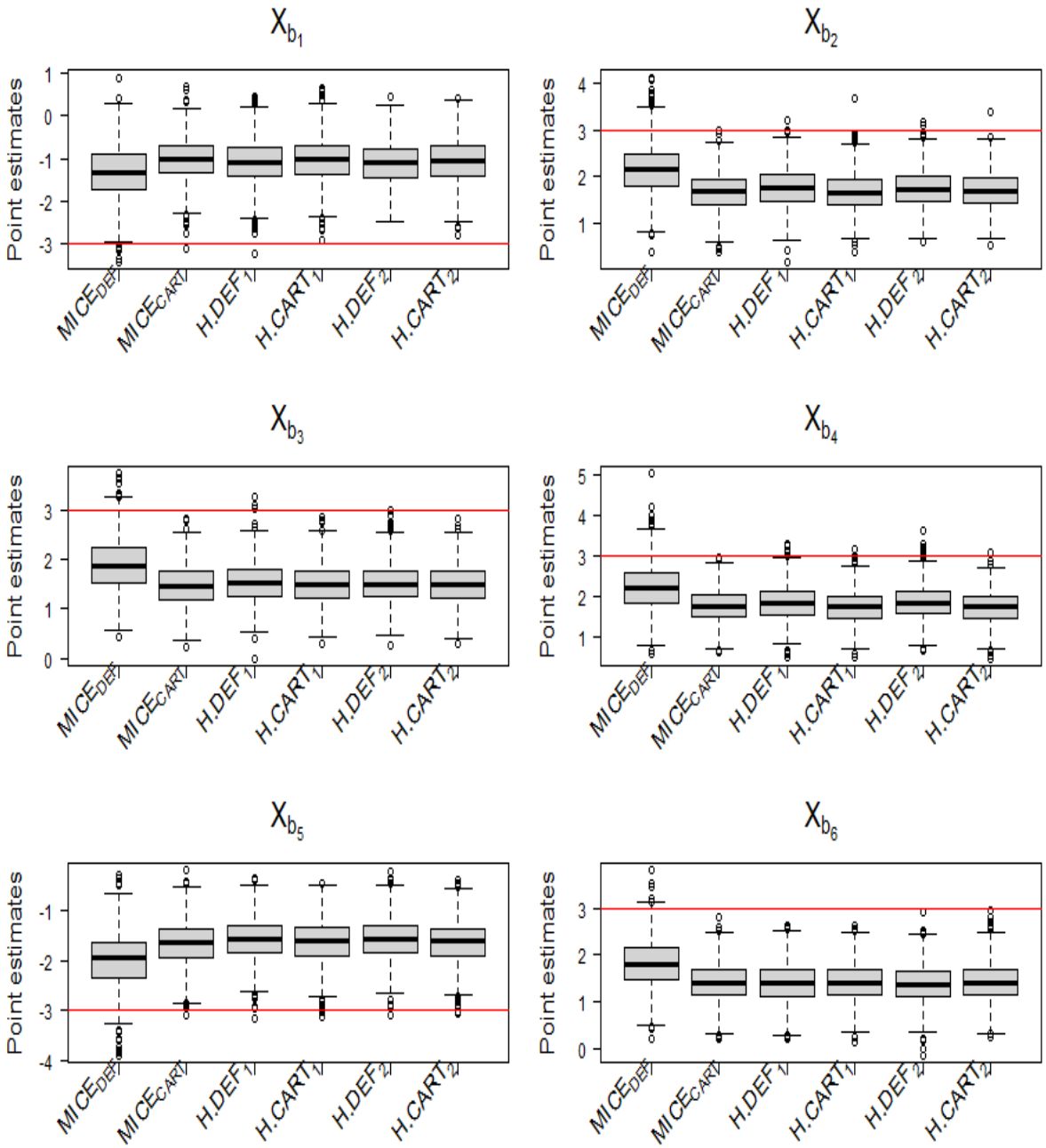
"% mis" indicates percentage of missing observations in all variables



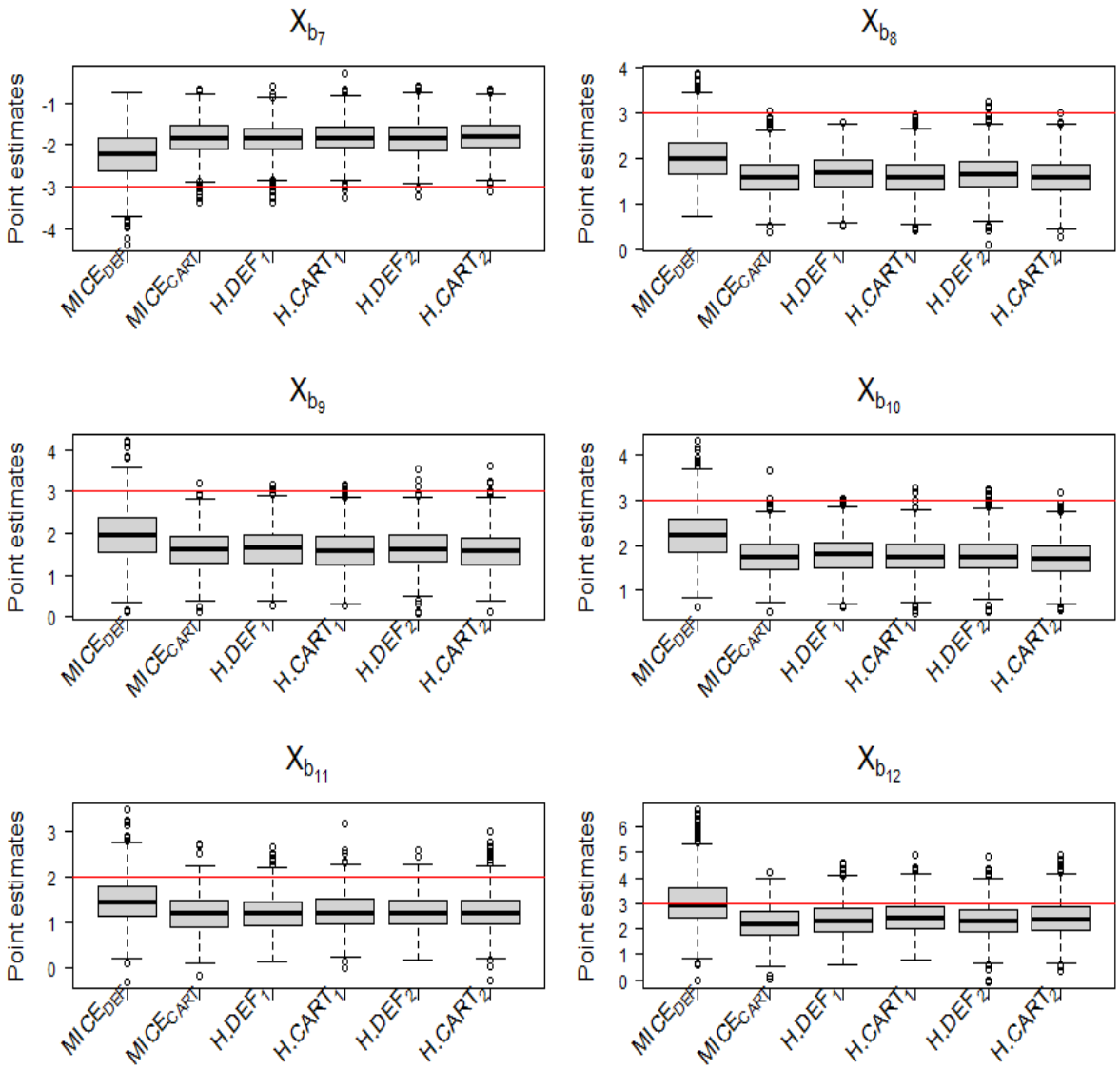
FigureS1. Schematic diagram illustrating the proposed hybrid architecture 1



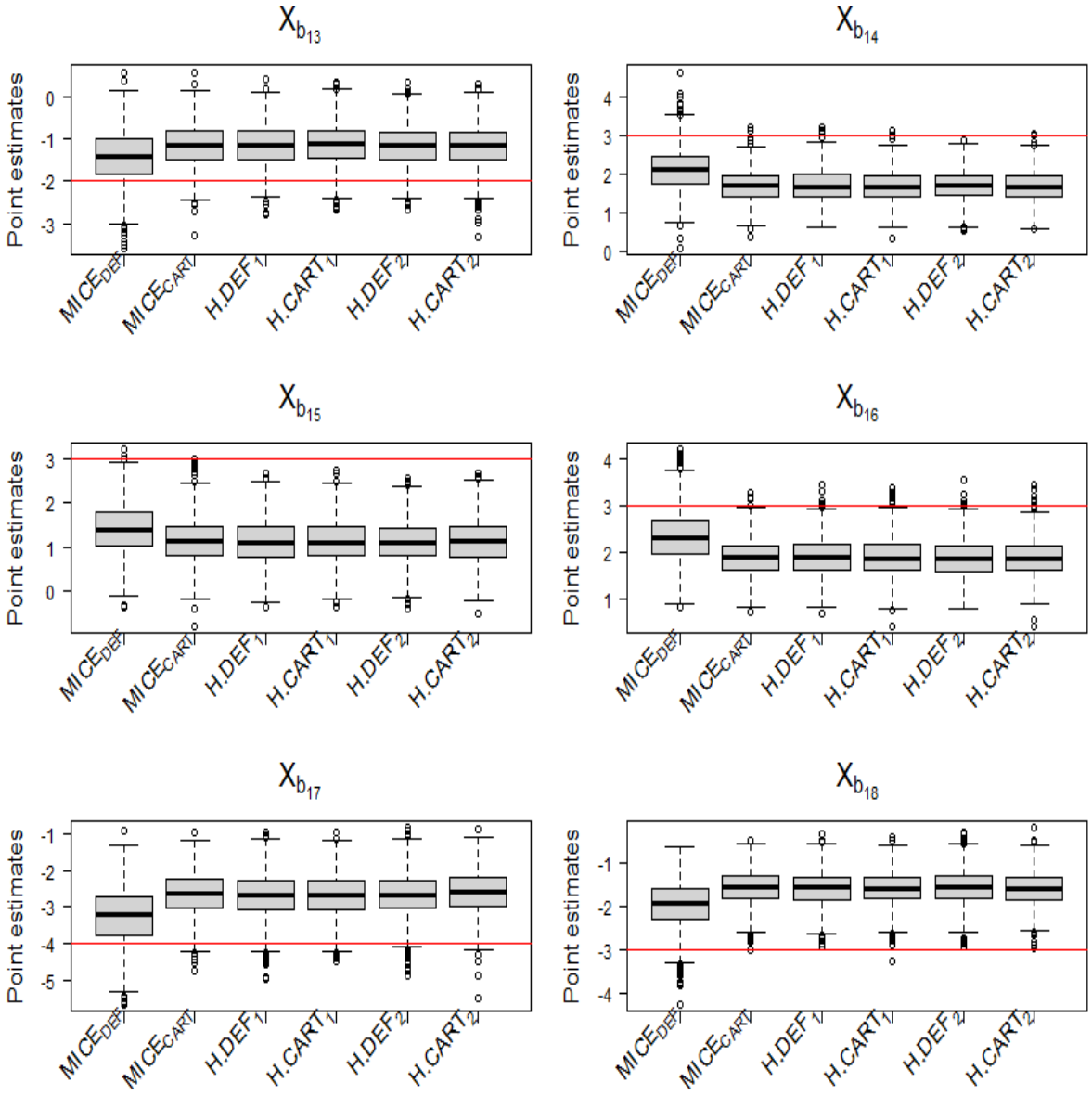
FigureS2. Schematic diagram illustrating the proposed hybrid architecture 2



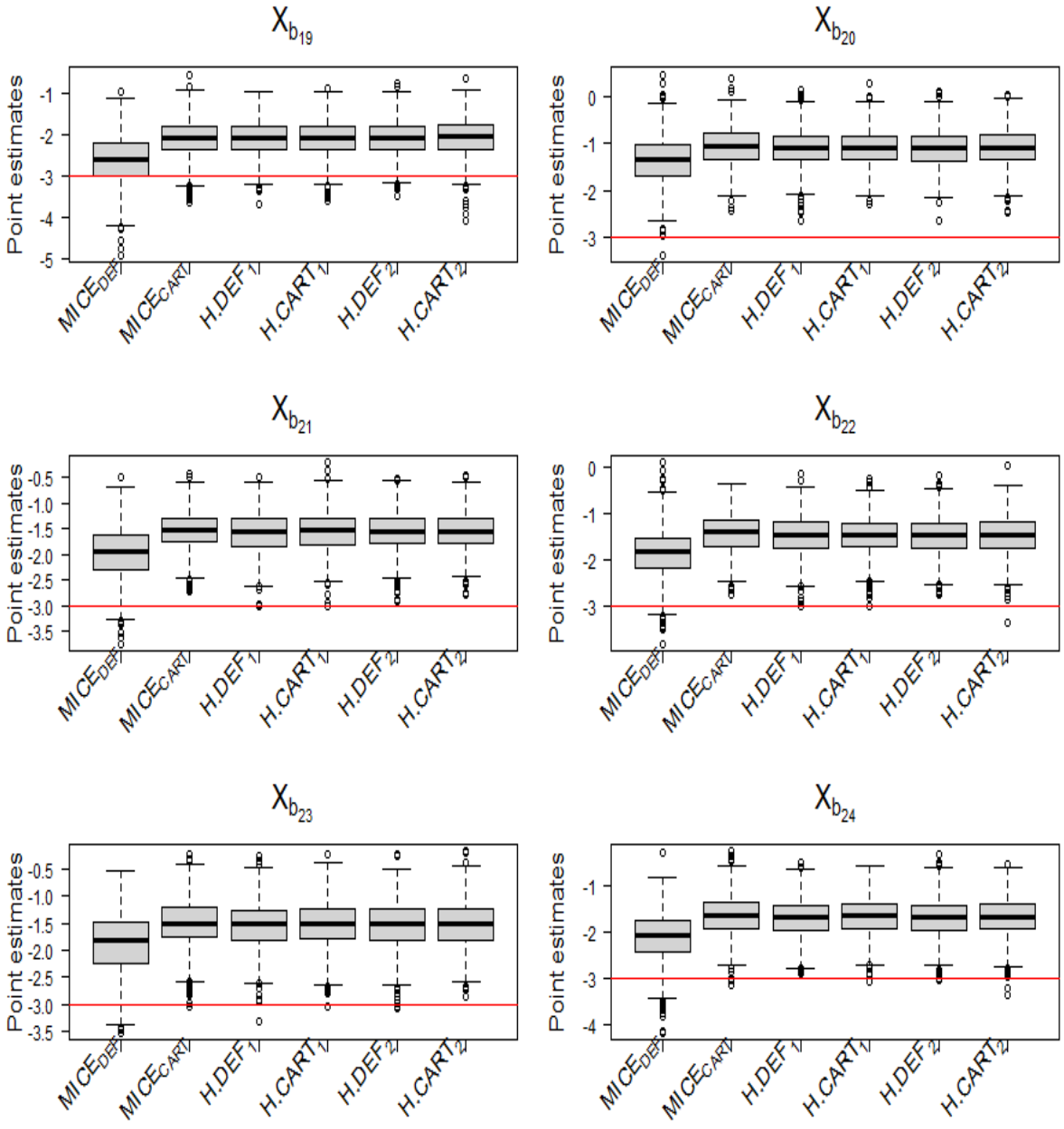
FigureS3. Simulated data: Boxplots of point estimates for coefficients X_{b_1}, X_{b_2} , $X_{b_3}, X_{b_4}, X_{b_5}, X_{b_6}$ under various MI methods over 1000 simulations



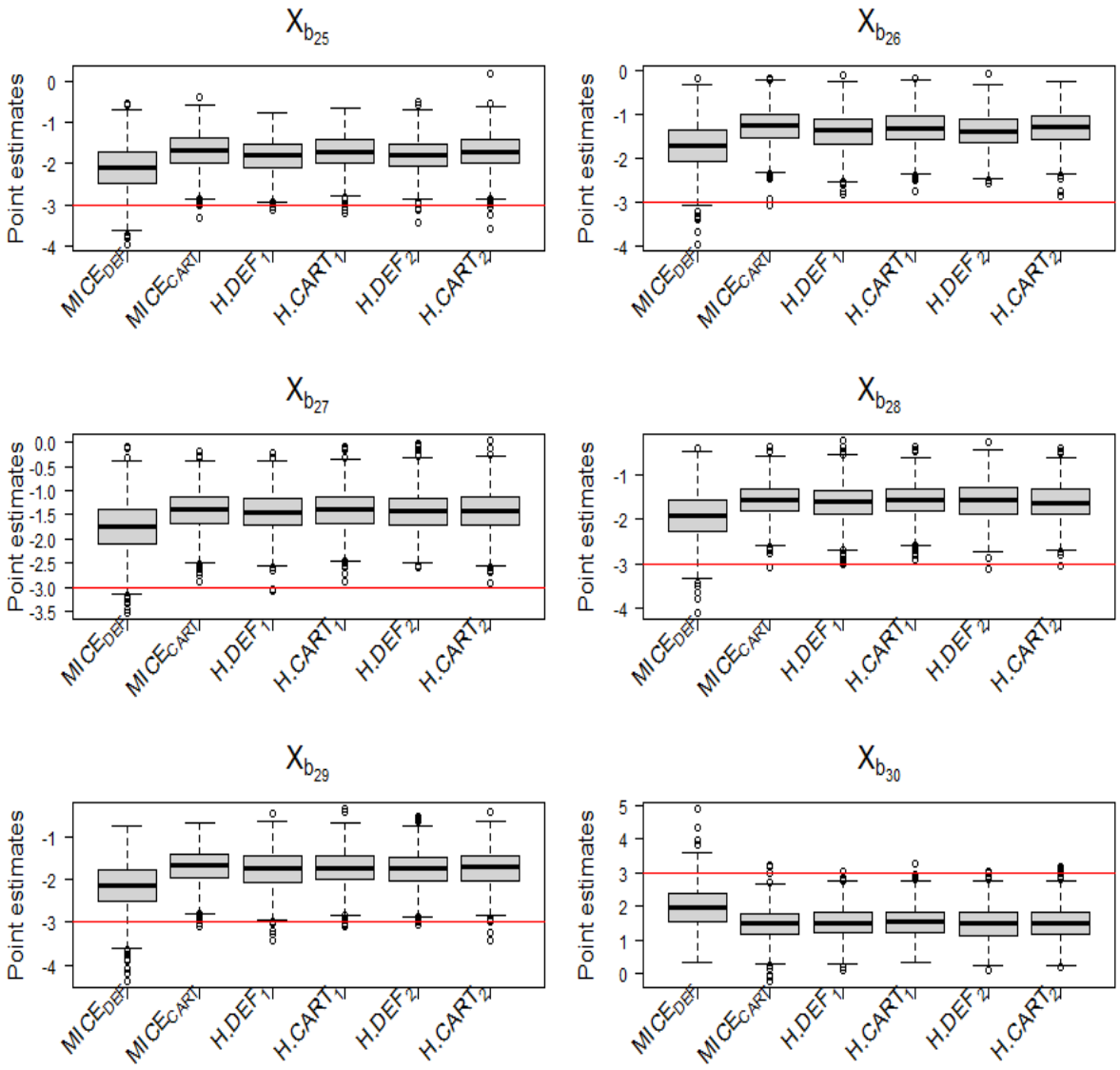
FigureS4. Simulated data: Boxplots of point estimates for coefficients $X_{b_7}, X_{b_8}, X_{b_9}, X_{b_{10}}, X_{b_{11}}, X_{b_{12}}$ under various MI methods over 1000 simulations



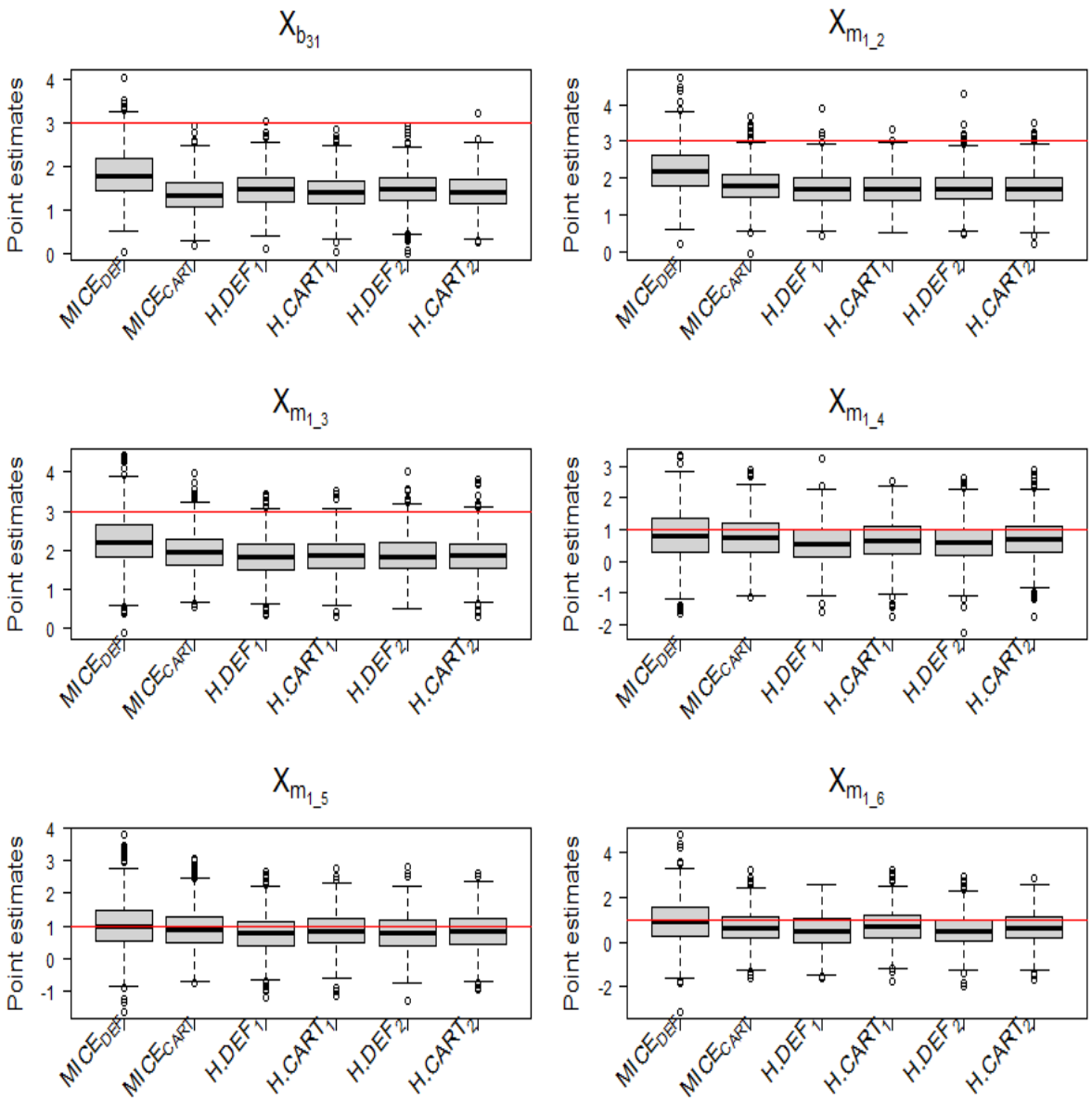
FigureS5. Simulated data: Boxplots of point estimates for coefficients $X_{b_{13}}, X_{b_{14}}, X_{b_{15}}, X_{b_{16}}, X_{b_{17}}, X_{b_{18}}$ under various MI methods over 1000 simulations



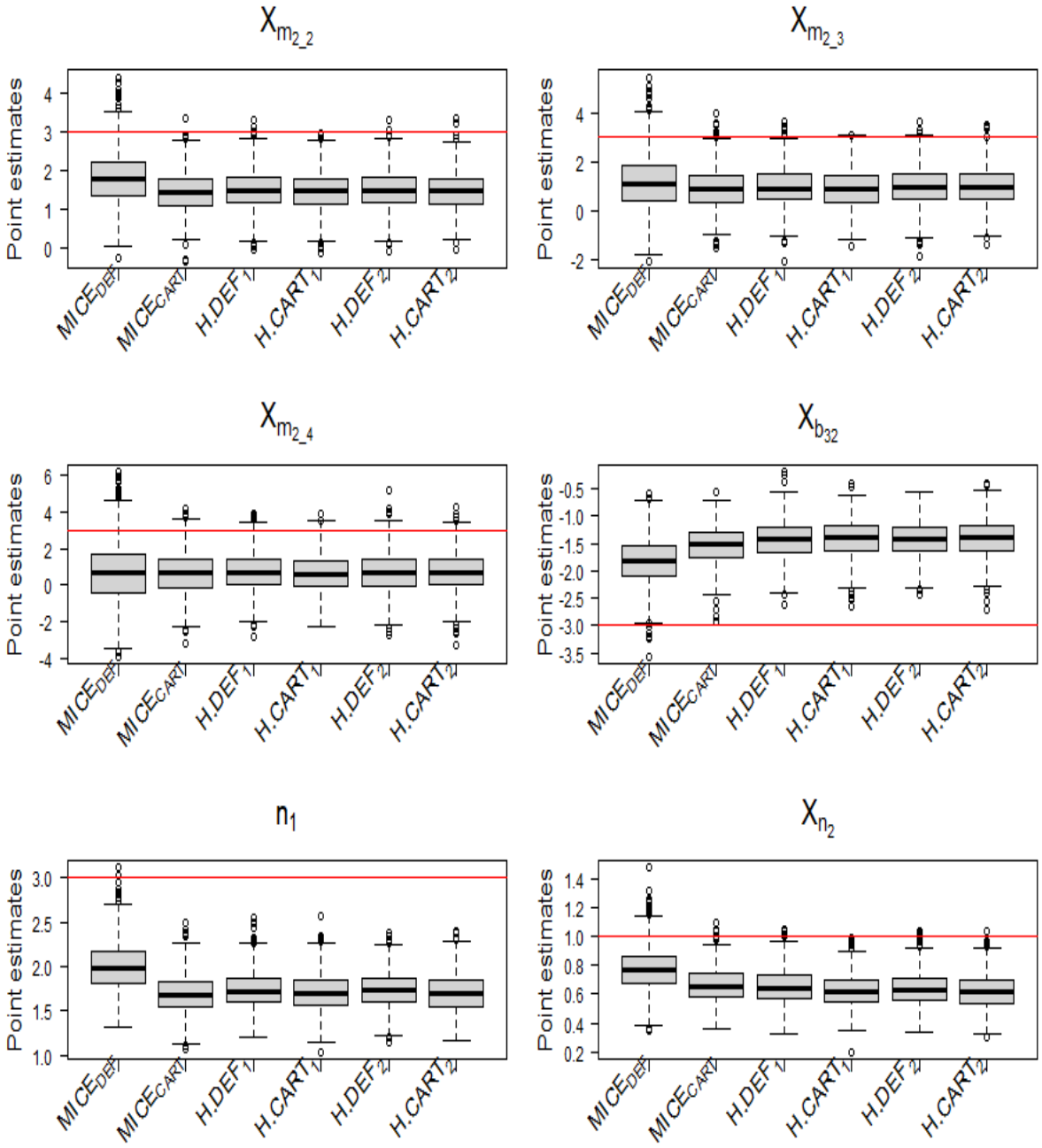
FigureS6. Simulated data: Boxplots of point estimates for coefficients $X_{b_{19}}, X_{b_{20}}, X_{b_{21}}, X_{b_{22}}, X_{b_{23}}, X_{b_{24}}$ under various MI methods over 1000 simulations



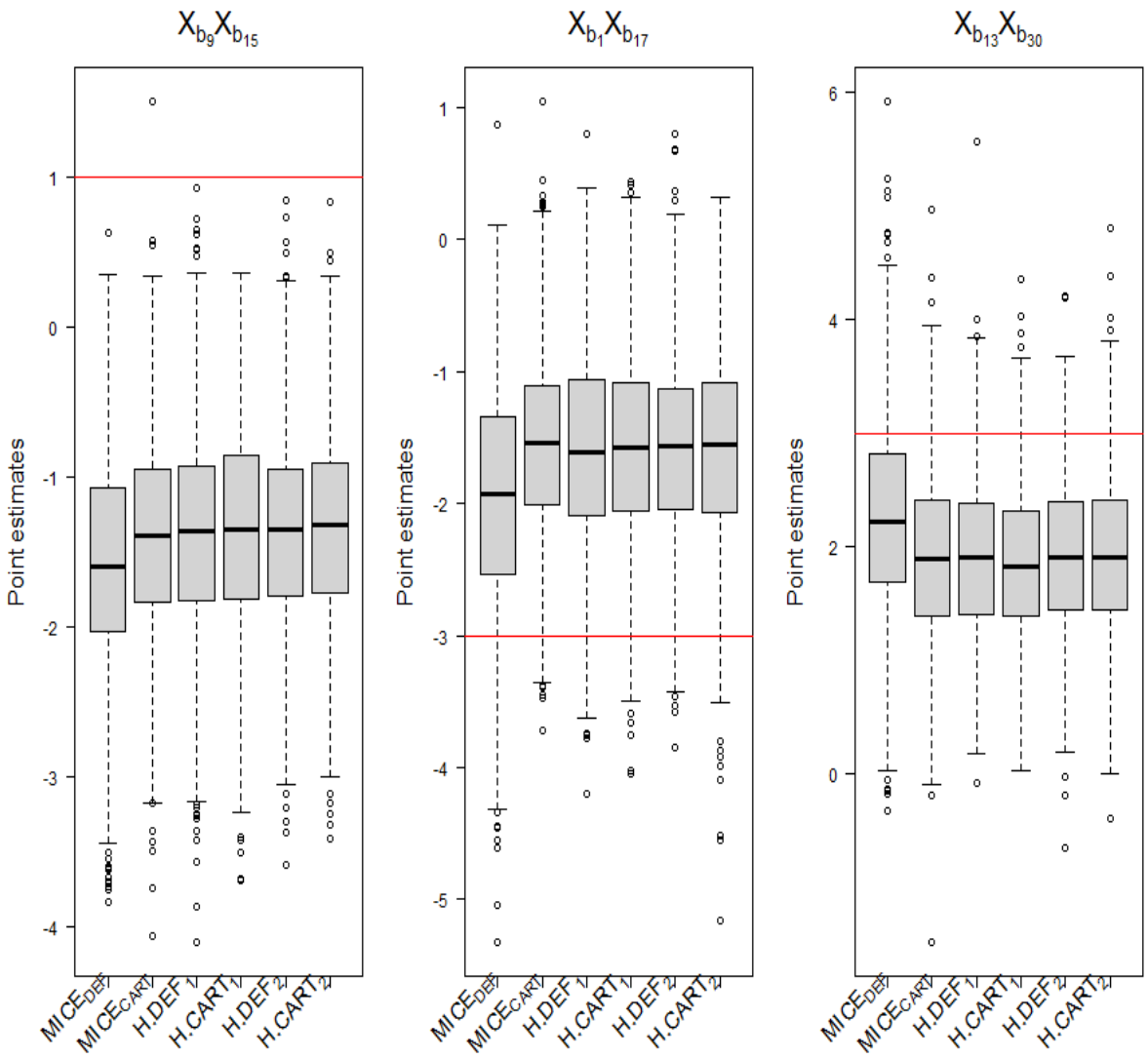
FigureS7. Simulated data: Boxplots of point estimates for coefficients $X_{b_{25}}, X_{b_{26}}, X_{b_{27}}, X_{b_{28}}, X_{b_{29}}, X_{b_{30}}$ under various MI methods over 1000 simulations



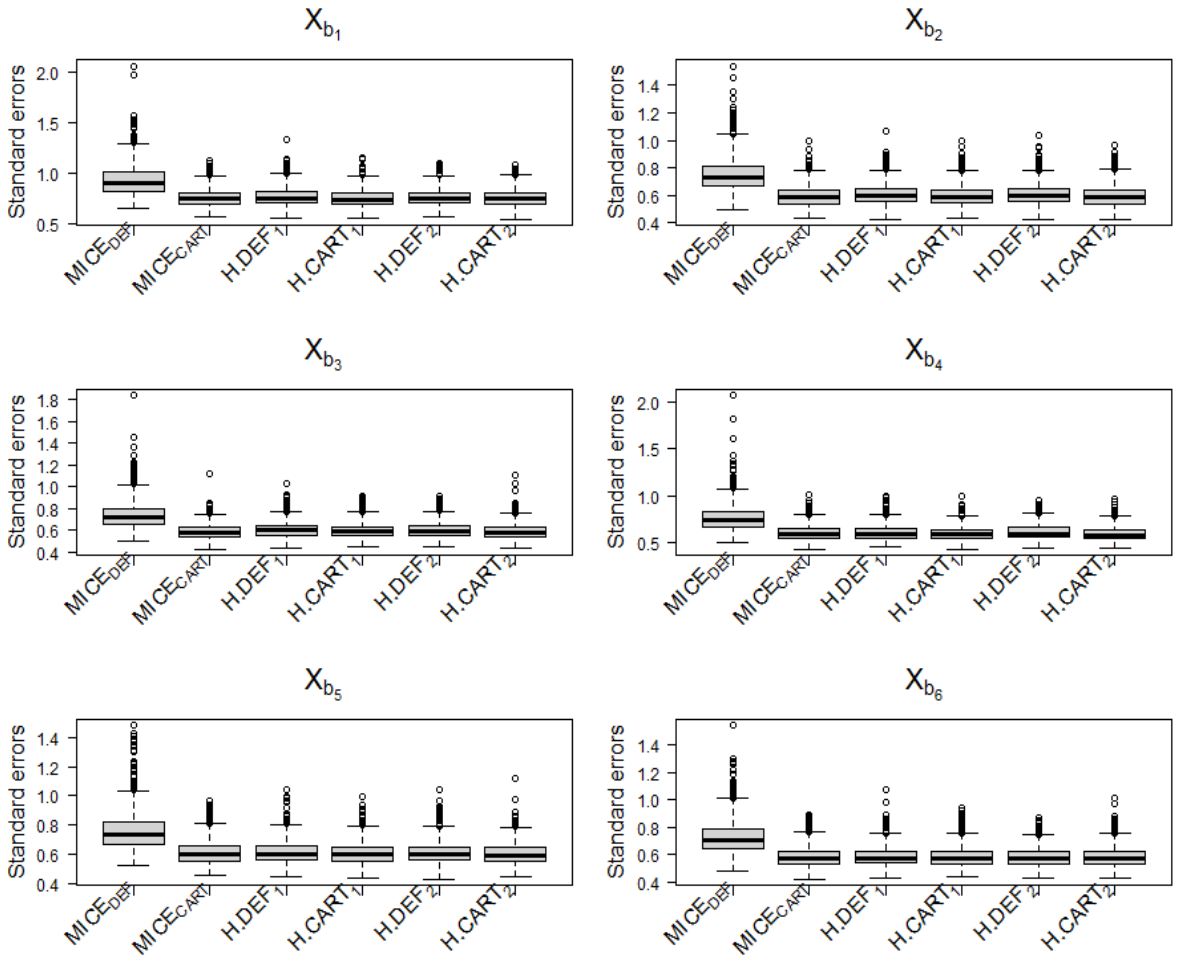
FigureS8. Simulated data: Boxplots of point estimates for coefficients $X_{b_{31}}$, $X_{m_{1\_2}}$, $X_{m_{1\_3}}$, $X_{m_{1\_4}}$, $X_{m_{1\_5}}$, $X_{m_{1\_6}}$ under various MI methods over 1000 simulations



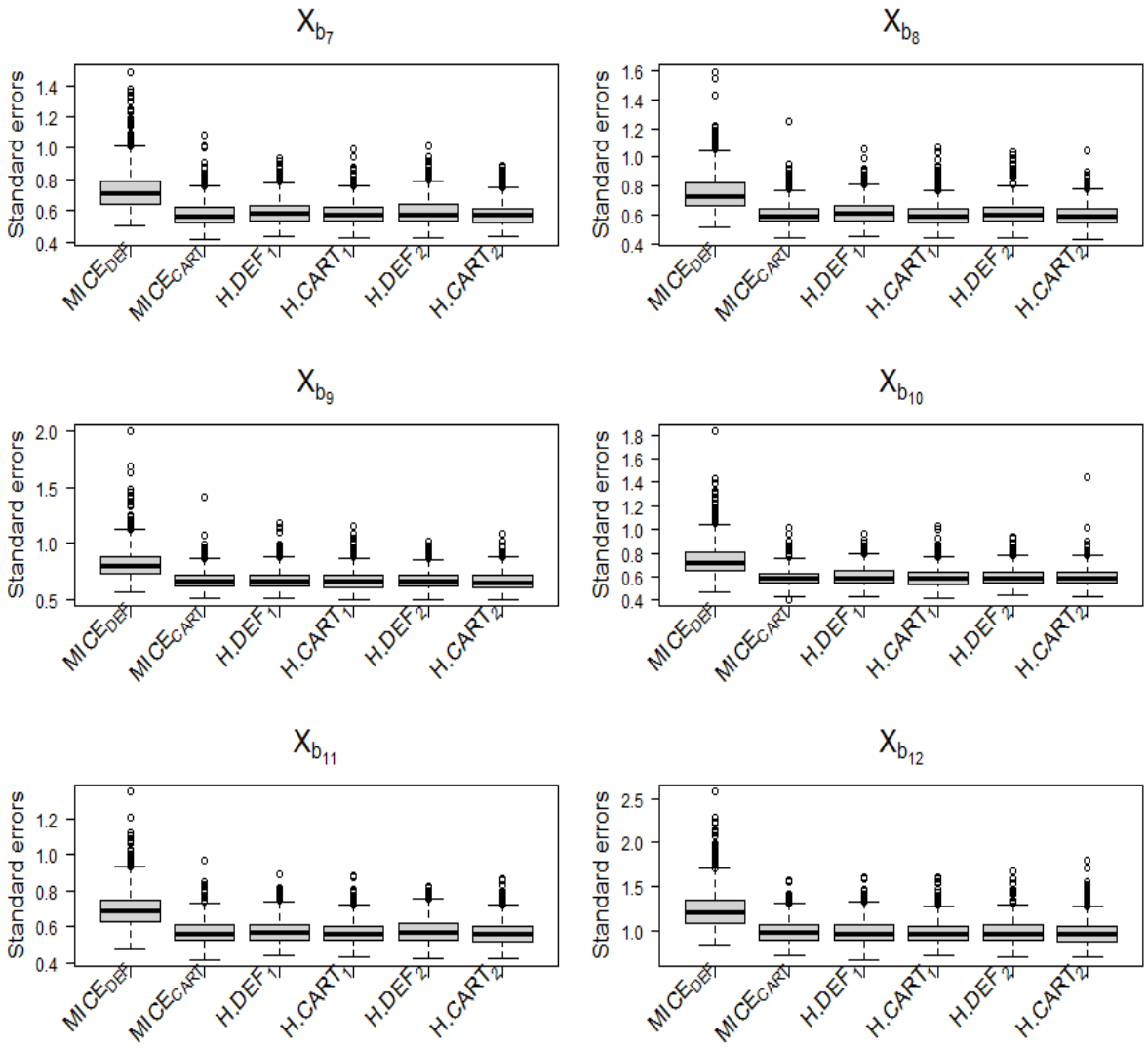
FigureS9. Simulated data: Boxplots of point estimates for coefficients $X_{m_{2\_2}}$,$X_{m_{2\_3}}$,$X_{m_{2\_4}}$,X_{b32}, X_{n_1} , X_{n_2} under various MI methods over 1000 simulations



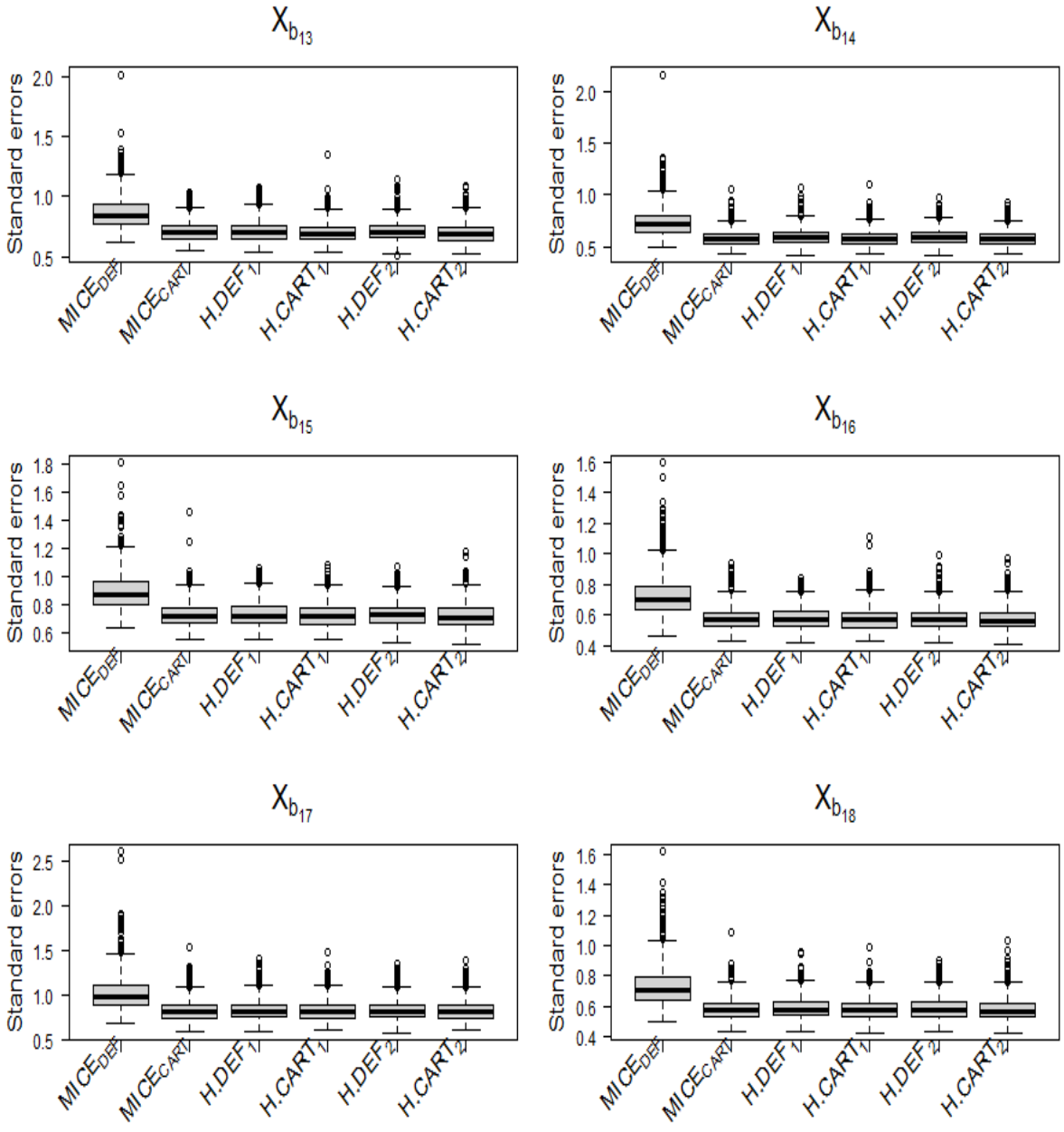
FigureS10. Simulated data: Boxplots of point estimates for coefficients $X_{b_9}\,X_{b_{15}}$, $X_{b_1}\,X_{b_{17}}$ $X_{b_{13}}\,X_{b_{30}}$ under various MI methods over 1000 simulations



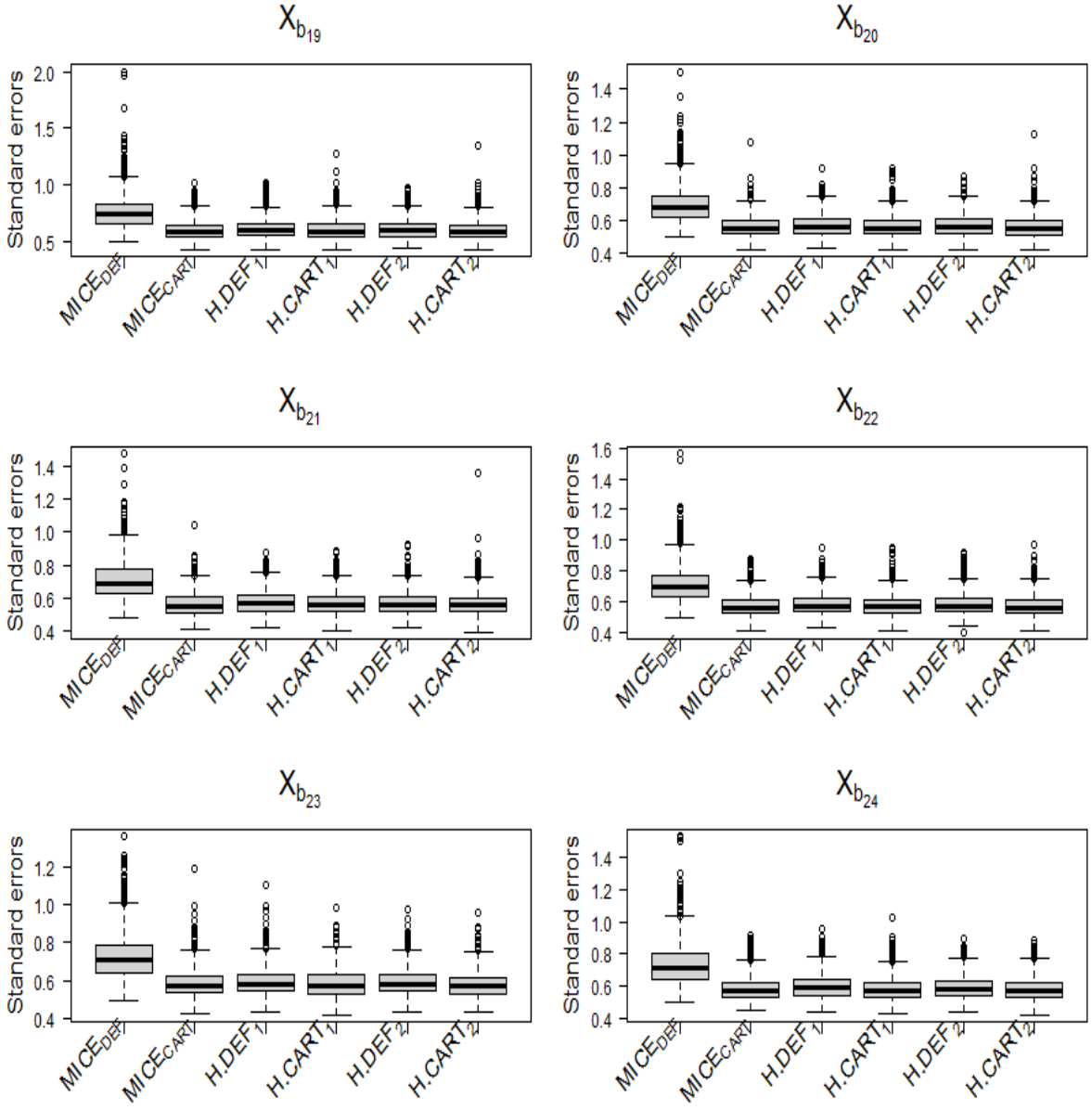
FigureS11. Simulated data: Boxplots of standard errors for coefficients $X_{b_1}, X_{b_2}, X_{b_3}, X_{b_4}, X_{b_5}, X_{b_6}$ under various MI methods over 1000 simulations



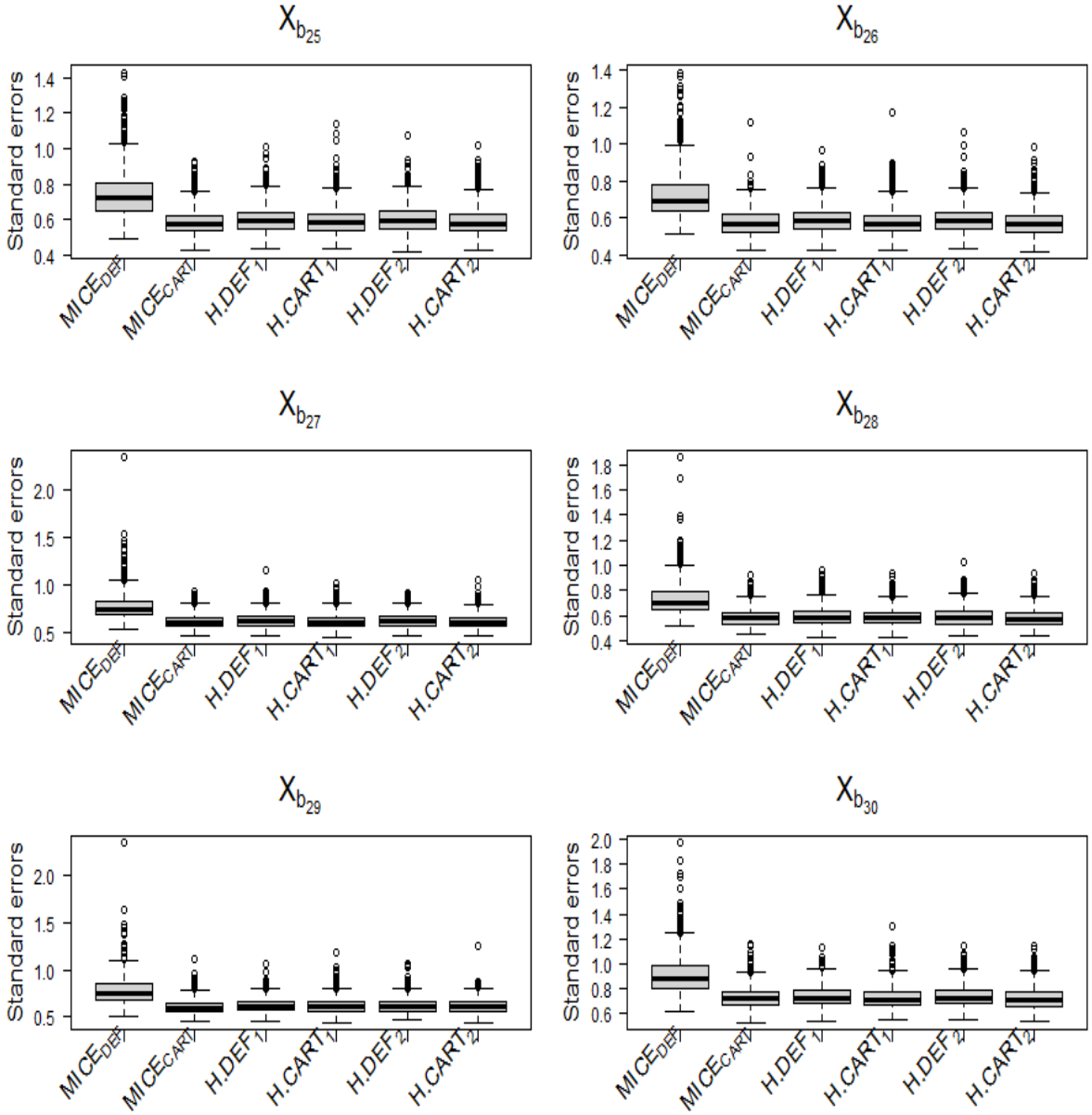
FigureS12. Simulated data: Boxplots of standard errors for coefficients $X_{b_7}, X_{b_8}, X_{b_9}, X_{b_{10}}, X_{b_{11}}, X_{b_{12}}$ under various MI methods over 1000 simulations



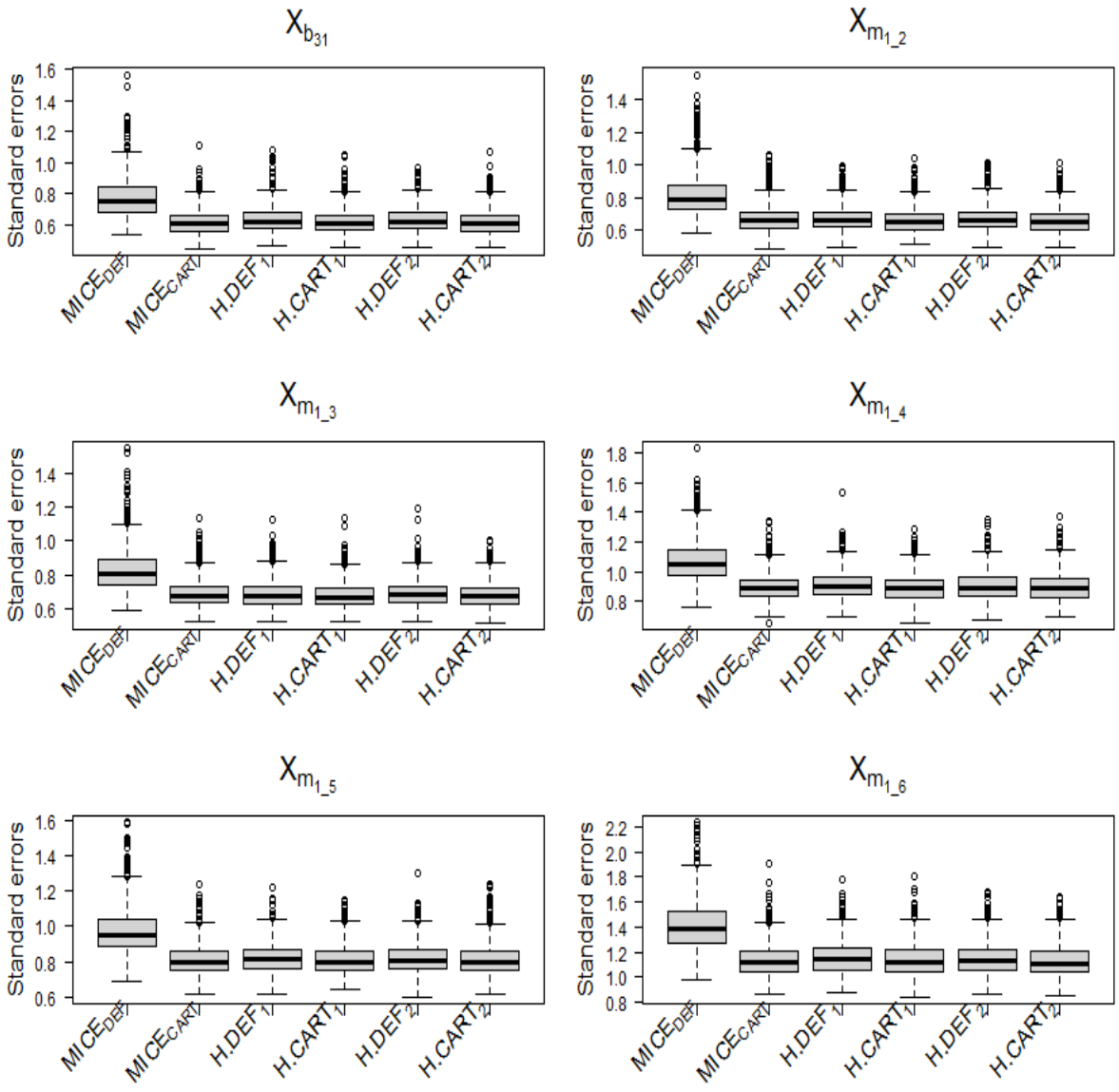
FigureS13. Simulated data: Boxplots of standard errors for coefficients $X_{b_{13}}, X_{b_{14}}, X_{b_{15}}, X_{b_{16}}, X_{b_{17}}, X_{b_{18}}$ under various MI methods over 1000 simulations



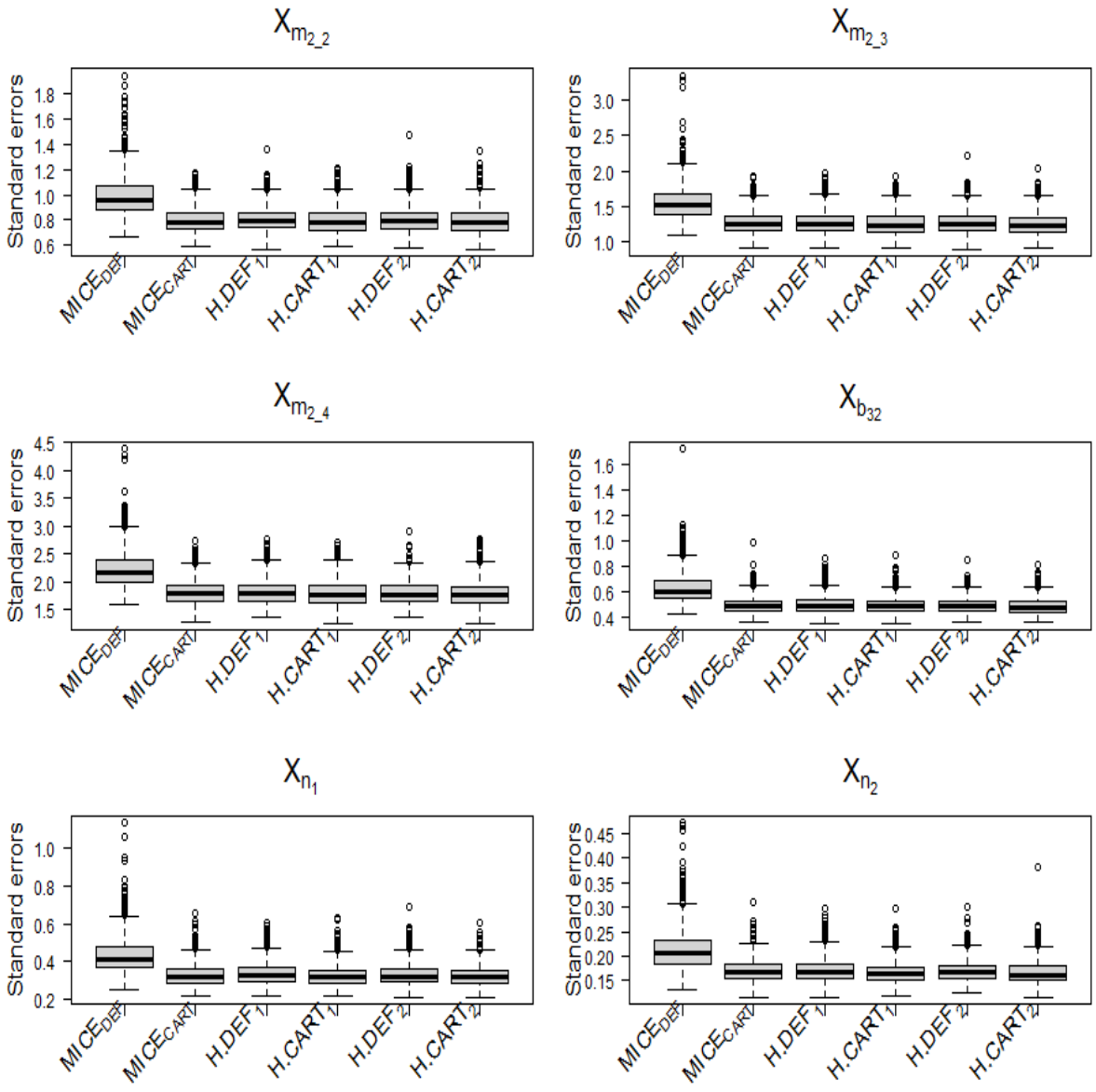
FigureS14. Simulated data: Boxplots of standard errors for coefficients $X_{b_{19}}, X_{b_{20}}, X_{b_{21}}, X_{b_{22}}, X_{b_{23}}, X_{b_{24}}$ under various MI methods over 1000 simulations



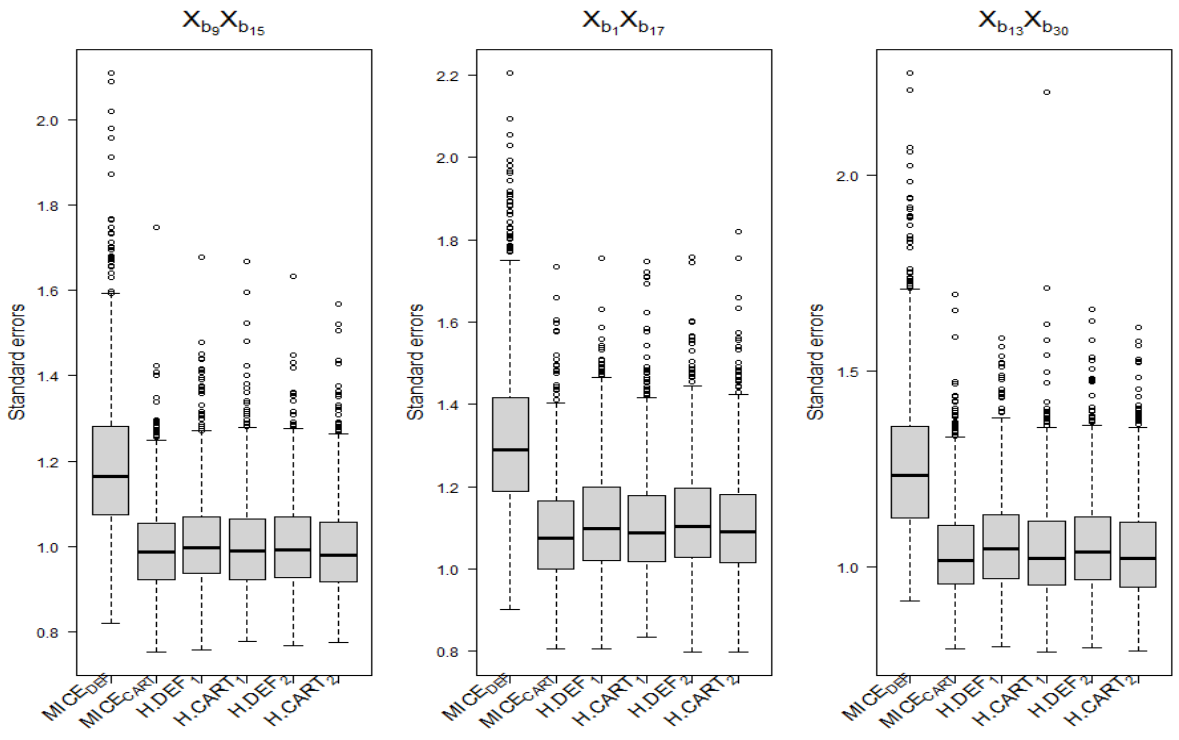
FigureS15. Simulated data: Boxplots of standard errors for coefficients $X_{b_{25}}, X_{b_{26}}, X_{b_{27}}, X_{b_{28}}, X_{b_{29}}, X_{b_{30}}$ under various MI methods over 1000 simulations



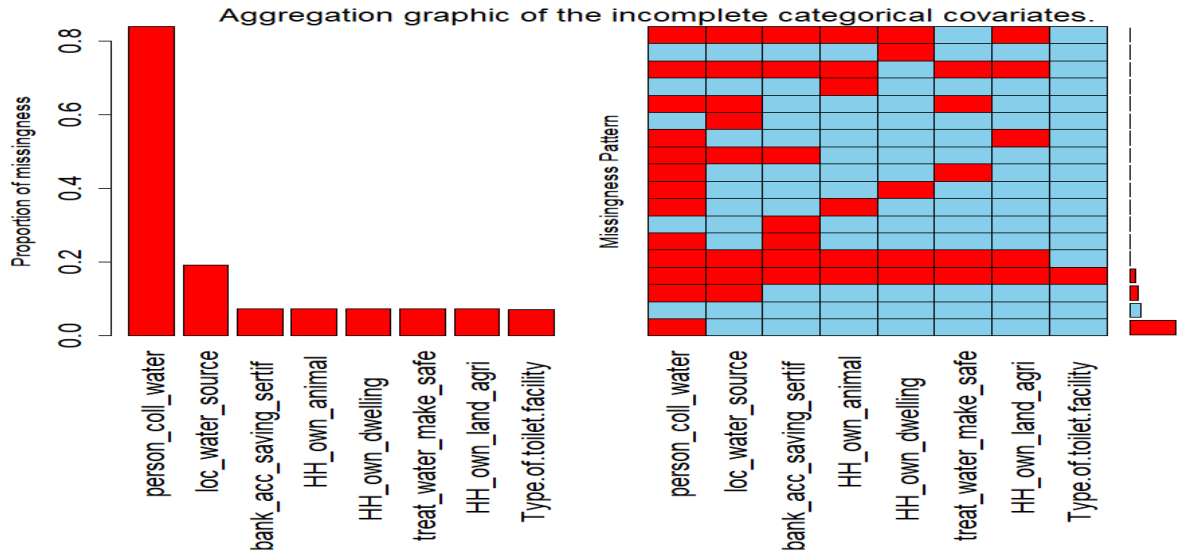
FigureS16. Simulated data: Boxplots of standard errors for coefficients $X_{b_{31}}$, $X_{m_{1\_2}}$, $X_{m_{1\_3}}$, $X_{m_{1\_4}}$, $X_{m_{1\_5}}$, $X_{m_{1\_6}}$ under various MI methods over 1000 simulations



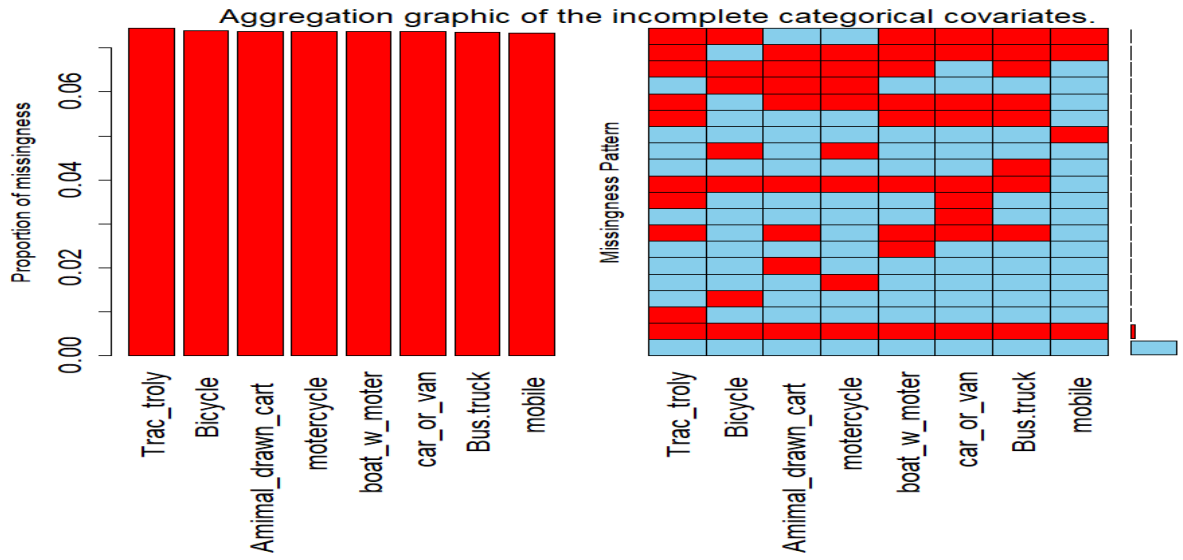
FigureS17. Simulated data: Boxplots of standard errors for coefficients $X_{m_{2\_2}}$,$X_{m_{2\_3}}$,$X_{m_{2\_4}}$,X_{b32}, X_{n_1} , X_{n_2} under various MI methods over 1000 simulations



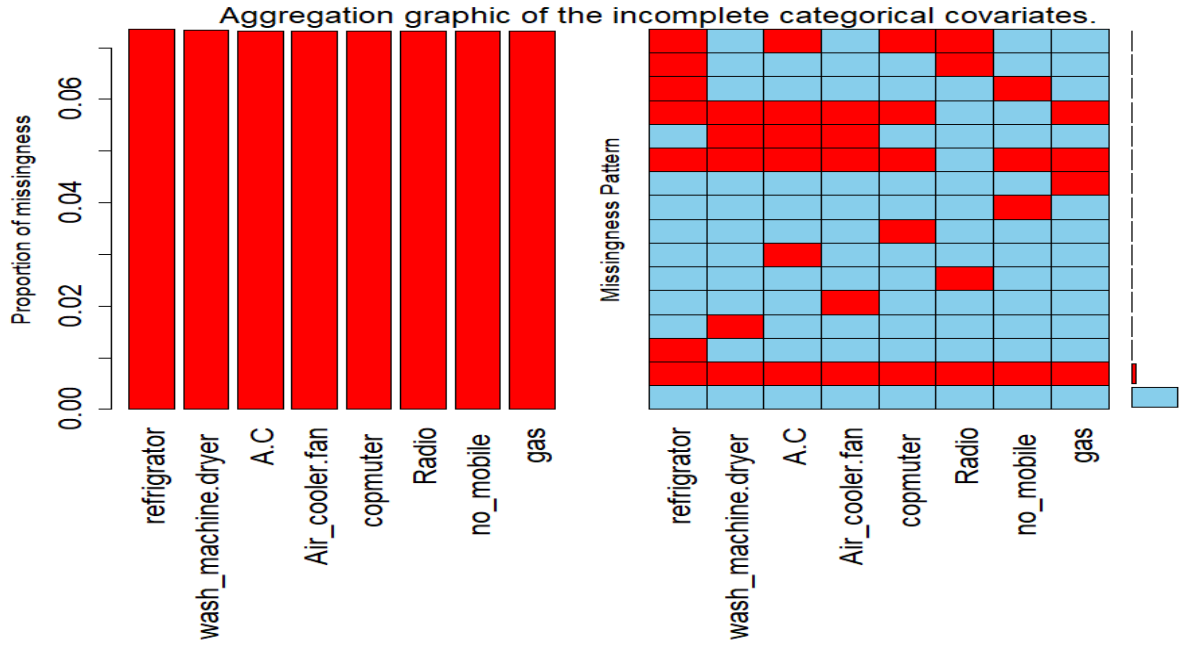
FigureS18. Simulated data: Boxplots of standard errors for coefficients $X_{b_9}X_{b_{15}}$, $X_{b_1}X_{b_{17}}$, $X_{b_{13}}X_{b_{30}}$ under various MI methods over 1000 simulations



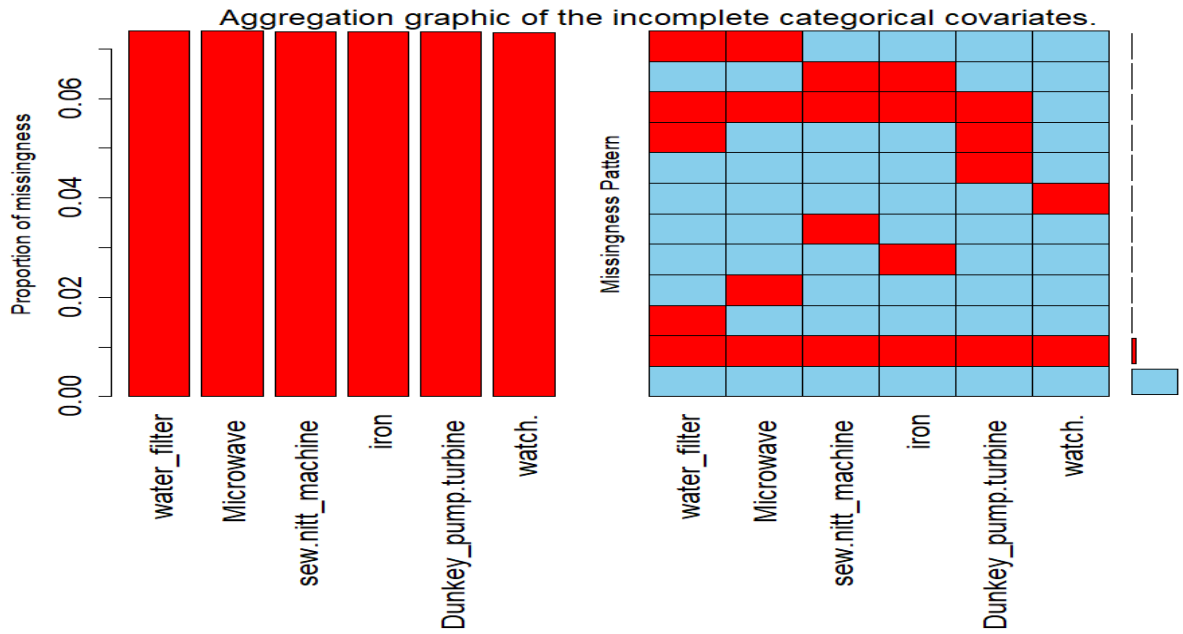
FigureS19. Real data: Aggregate plots in R, graphics of incomplete variables i.e. "HH_own_dwelling,"HH_own_land_agri","Type.of.toilet.facility","HH_own_animal","treat_water_make_safe", "bank_acc_saving_sertif","loc_water_source","person_coll_water"



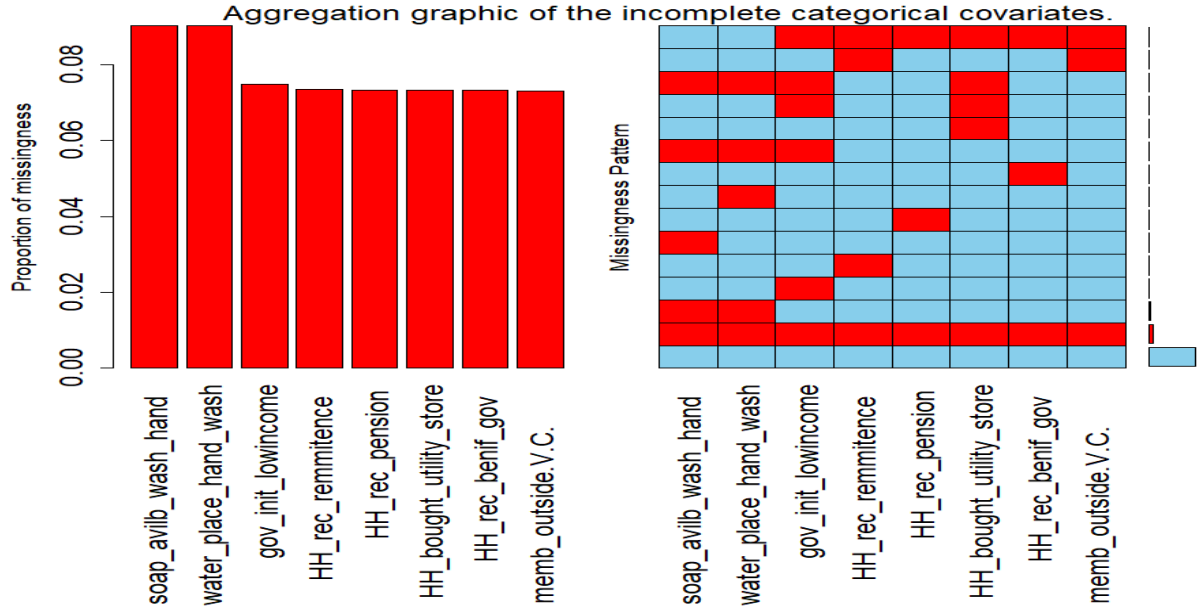
FigureS20. Real data: Aggregate plots in R, graphics of incomplete variables i.e. "mobile ","Bicycle","motercycle","Amimal_drawn_cart","Bus.truck","boat_w_moter","car_or_van","Trac_troly"



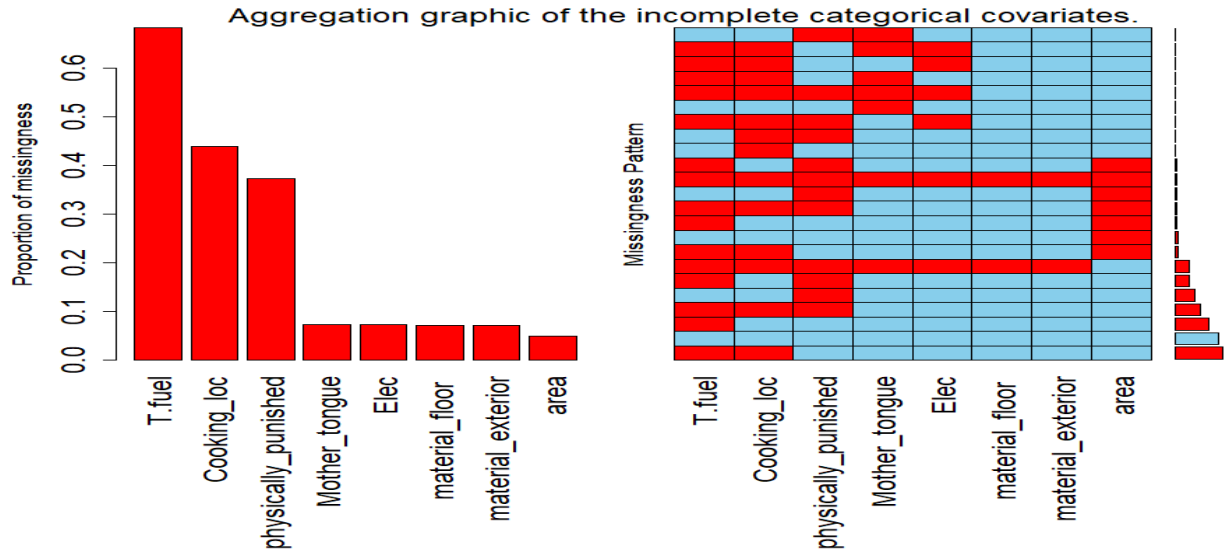
FigureS21. Real data: Aggregate plots in R, graphics of incomplete variables i.e. "Radio", "no_mobile", refrigrator","gas"," copmuter ", "A.C", "wash_machine.dryer ", "Air_cooler.fan"



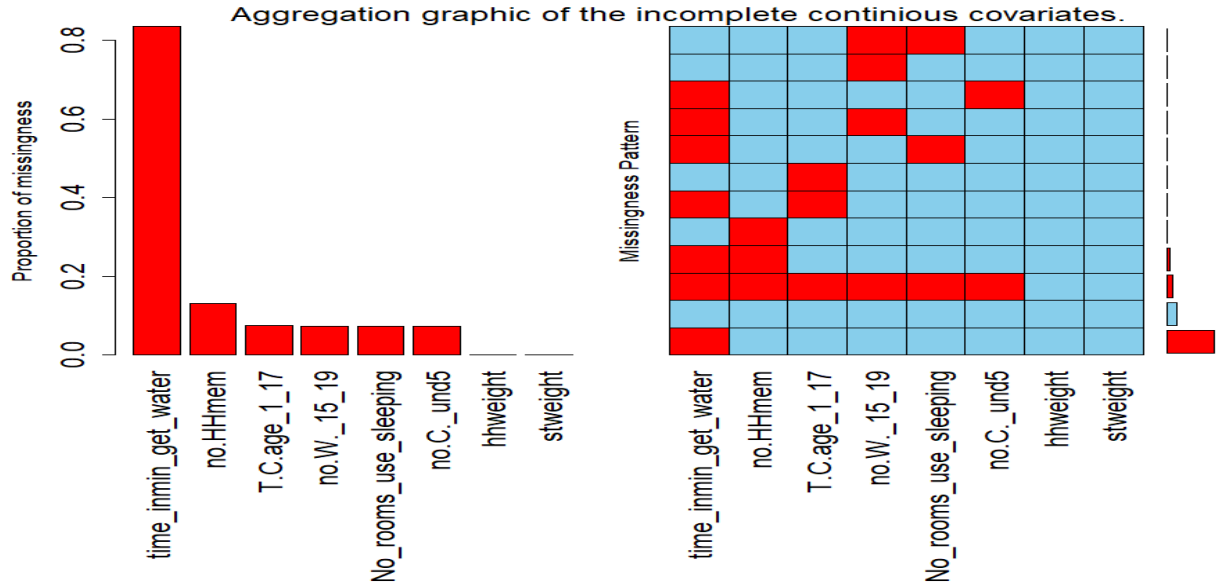
FigureS22. Real data: Aggregate plots in R, graphics of incomplete variables i.e. “Microwave", "sew.nitt_machine ","iron", "water_filter", "Dunkey_pump.turbine ", "watch”



FigureS23. Real data: Aggregate plots in R, graphics of incomplete variables i.e. "memb_outside.V.C.","HH_rec_remmitenc","HH_rec_pension","HH_rec_benif_gov","HH_bought_utility_store","gov_init_lowincome","water_place_hand_wash","soap_avilb_wash_hand"



FigureS24. Real data: Aggregate plots in R, graphics of incomplete variables i.e. "area", "physically_punished","Mother_tongue","material_floor","material_exterior","T.fuel","Cooking _loc","Elec"



FigureS25. Real data: Aggregate plots in R, graphics of incomplete variables i.e."no.HHmem", "no.W._15_19","no.C._und5","T.C.age_1_17","No_rooms_use_sleeping","time_inmin_get_wat er", "hhweight", "stweight"