STATISTICS IN TRANSITION new series, December 2019
Vol. 20, No. 4, pp. 33–58, DOI 10.21307/stattrans-2019-033
Submitted – 25.02. 2019; Paper ready for publication – 19.10.2019

HYBRID MULTIPLE IMPUTATION IN A LARGE SCALE COMPLEX SURVEY

Humera Razzak[1], **Christian Heumann**[2]

ABSTRACT

Large-scale complex surveys typically contain a large number of variables measured on an even larger number of respondents. Missing data is a common problem in such surveys. Since usually most of the variables in a survey are categorical, multiple imputation requires robust methods for modelling high-dimensional categorical data distributions. This paper introduces the 3-stage Hybrid Multiple Imputation (HMI) approach, computationally efficient and easy to implement, to impute complex survey data sets that contain both continuous and categorical variables. The proposed HMI approach involves the application of sequential regression MI techniques to impute the continuous variables by using information from the categorical variables, already imputed by a non-parametric Bayesian MI approach. The proposed approach seems to be a good alternative to the existing approaches, frequently yielding lower root mean square errors, empirical standard errors and standard errors than the others. The HMI method has proven to be markedly superior to the existing MI methods in terms of computational efficiency. The authors illustrate repeated sampling properties of the hybrid approach using simulated data. The results are also illustrated by child data from the multiple indicator survey (MICS) in Punjab 2014.

Key words: complex surveys, high-dimensional data, missing data, multiple imputation.

1. Introduction

Large scale complex surveys contain high-dimensional data with a large number of variables measured on an even larger number of respondents. The Multiple Indicator Cluster Surveys (MICS) is such a popular large scale international household survey. Like other cross-sectional surveys, the data sets from MICS contain complex survey features (e.g. many categorical variables). Missing values are also a problem in MICS surveys. Missing data problems arise when a sampled unit does not respond to the entire survey (also called unit non response) or to a particular question (also called item non response). For example, the MICS Punjab 2014 child data set contains more than 200 child health background variables on 31083 children under the age of 5. Among all

[1] humera.razzak@stat.uni-muenchen.de.
[2] christian.heumann@stat.uni-muenchen.de.

these variables, the missing data rates per variable range from 10% to 95% and 26 variables have more than 50% missing values. Questions related to a child cleaning utensils or washing clothes and physical punishment, etc. may make participants reluctant to provide full information, which results in incomplete data (Akmatov (2011)) (Cappa and Khan (2011)).

In recent decades, considerable efforts have been made into the development of statistical methods to treat the problem of missing data. Complete-case or available-case analysis, or single imputation methods such as mean and regression imputation, often result in potentially biased estimates when applied to incomplete data (Anderson et al. (1983)). Rubin (1987) proposed multiple imputation (MI) as an appropriate alternative under certain assumptions. Predictive distributions are used to draw repeated samples in order to simulate values for missing data. *M>1* complete data sets are generated and point and variance estimates of interest are estimated and combined using the formulas developed by Rubin (1987). One advantage of MI is the decoupling of the imputation task and the analysis task although one has to be careful in choosing the imputation and the analysis model (Xie et al. (2017)).

In this paper, we propose a computationally efficient and an easy to implement 3-stage Hybrid Multiple Imputation (HMI) approach to impute complex survey data sets that contain both continuous and categorical variables. The HMI approach applies sequential regression MI techniques to impute continuous variables by using information of categorical variables already imputed by a non-parametric Bayesian MI approach. This blended version of joint and sequential modelling MI techniques makes it possible to obtain complete datasets with both types of variables. This approach is motivated by missing values in background variables related to children's life and health in MICS. In order to get valid and accurate results, it becomes important to impute all types of variables in MICS. As we noted earlier, handling mixed continuous and categorical data in high dimensions presents unique challenges to MI. Existing MI methods can be difficult to implement in the presence of complex dependence structures among categorical variables, whereas some recently developed methods focus on missing values of few variables (Zhao and Long (2016)). Moreover, various MI techniques are limited to categorical variables or require transformations (or other tricks) for continuous variables (Si and Reiter (2013)).

The reminder of the paper is organized as follows. We begin in Section 2 by describing missing data mechanisms. In Section 3, we review imputation methods dedicated to categorical, continuous and mixed data in high dimensions. In section 4 we illustrate Rubin's inference and various estimates used for comparing the performance of the imputation algorithms. Section 5 presents the proposed hybrid architecture. In Section 6 we present the simulation studies and relevant results to evaluate our proposed approach. Section 7 presents the imputation of the MICS Child Data. We conclude with a discussion at the end.

2. Missing data mechanisms

There are three missing data mechanisms. Missing values in any data can be missing completely at random (MCAR), or missing at random (MAR), or missing not at random (MNAR) (Rubin (1987)), (Little and Rubin (2002)). Let Y denote the

n × p data matrix with n rows (cases) and p variables. Let y_{ij} refer to the value in row i and column j of Y, where i=1,…,n and j=1,…,p. Further, suppose that 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 U be a response indictor matrix with the same dimensions as Y indicating whether an element of Y is observed or missing:

$$U_{ij} = \begin{cases} 0 \ if \ y_{ij} \ is \ missing, \\ 1 \ if \ y_{ij} \ is \ observed. \end{cases}$$

Data is MCAR when $Pr(U|Y^{miss}, Y^{obs})=Pr(U)$, MAR when $Pr(U|Y^{miss}, Y^{obs})=Pr(U|Y^{obs})$ and MNAR when $Pr(U|Y^{miss}, Y^{obs}) \neq Pr(U|Y^{obs})$ (Little and Rubin (2002)). MNAR is also called "non-ignorable" (NI).

3. Imputation methods for large scale complex surveys

A complete overview of the state of the art MI methods for accommodating nonlinear relationships and best ways to impute categorical and non-normal continuous variables is given in Vermunt et al. (2008), Yucel et al. (2011), Lee et al. (2012), Seaman et al. (2012) and Lee and Carlin (2017). Information on missing categorical data can be obtained by log-linear models (Schafer (1997)).

Imputation of large scale survey data can become challenging due to the presence of irregular missing patterns, interdependent logical constraints and data inconsistencies. There exist several approaches for MI for high-dimensional data. For example, in hot-deck imputation, which replaces missing values with observed values of pre-defined "donor" cells (Marker et al. (2002)), the probabilities of donor selection can be modified by respondent sampling weights (Andridge (2009)), or a k nearest neighbours (KNN) MI approach based on the distance metric for high-dimensional data (Holder (2015)) may be used or a principal component method to impute missing values (Audigier et al. (2016)). But most of the existing methods are not designed to handle mixed data (quantitative and categorical), become difficult to implement in the situation of large dimensions and are extremely time-consuming (Erosheva et al. (2002)). Moreover, the presence of complex dependence structures can also lead to biased estimates (Wirth and Tchetgen (2014)).

Sequential regression models (Raghunathan et al. (2001)) or fully conditional specification (FCS) (Su et al. (2011)), (van Buuren and Oudshoorn (1999)) is another general approach for MI. It is an iterative process. It specifies univariate conditional distributions on a variable-by-variable basis, and it draws missing values iteratively from the specified conditional distributions. FCS is also known as MI by chained equations (MICE) (Raghunathan et al. (2001)), (van Buuren and Groothuis-Oudshoorn (2011)), (White et al. (2011)), (Su et al. (2011)). The researcher can choose a suitable regression model for each incomplete variable where all the other variables are included as predictor variables, and a suitable imputation method, e.g. predictive mean matching (PMM) (Morris et al. (2014)). Examples are a linear regression model for a continuous variable or a logistic

regression model for a binary variable. Also, classification and regression trees (CART; Breiman (2001)) can be used. Vermunt et al. (2008) and van Buuren (2007) applied FCS to impute a small number of categorical and continuous variables. The theoretical implementation of this approach may become challenging when specified conditional densities become incompatible due to high dimensions (White et al. (2011)). Chained equations, when implemented by default settings (i.e. ignoring interaction effects in the conditional models) can also result in biased estimates. Moreover, standard MICE methods cannot handle high-dimensional data (Deng et al. (2016)). Sometimes problems of convergence and incompatibility arise when MICE is used to specify univariate conditional distributions (Arnold and Press (1989)), (Gelman and Speed (1993)) and due to the presence of complex dependencies, implementation of MICE may fail. Similar to log-linear models, conditional models in MICE suffer from model selection and estimation problems in high dimensions, which makes the regression imputations very time-consuming.

Random forest imputation is a method for handling missing data (Stekhoven et al. (2012)). Random forest imputation is a machine learning technique for nonlinearity and interaction problems and does not require a particular model to be specified. Shah et al. (2014) used random forest imputation for imputing complex epidemiological data sets. They found that MI based on random forest techniques tends to be more efficient and produced narrower confidence intervals as compared to standard MI methods. However, they focused on the setting where few variables have missing values. One disadvantage of algorithms based on random forests is that they are computationally expensive to implement in high dimensions and do not account for the uncertainty of estimating parameters in the imputation models (Rubin (1987)).

Loh et al. (2016) implement CART and forests to overcome incomplete data problems when the auxiliary variables are numerous. The study shows that the CART and forests methods are more reliable than likelihood methods for MI but CART can be biased toward selecting variables that allow more splits (Loh and Shih (1997)), (Kim and Loh (2001)). The study by Burgette and Reiter (2010) suggests that inferences based on the CART imputation engine can be more reliable than default applications of MICE based on main-effects generalized linear models. However, despite of various merits, CART methods and other fully conditional specifications are subject to odd behaviours, e.g. CART can be biased toward selecting variables that allow more splits in high dimensions (Raghunathan et al. (2001)), (van Buuren and Oudshoorn (1999)). Categorical predictors with many levels can be a major hurdle for CART algorithms. For example, over two billion potential partitions are formed for a categorical predictor with 32 levels, which makes CART algorithms computationally inefficient for standard computers.

The joint modelling (JM) specification is an alternative to the FCS approach. JM involves specifying a multivariate distribution for the data and draws imputations from their conditional distributions by Markov Chain Monte Carlo (MCMC) methods. Joint distributions of the variables with missing values are also specified by parametric, non-parametric and semi parametric models. A non-parametric Bayesian joint modelling approach for MI for multivariate categorical

data presented by Si and Reiter (2013) uses the Dirichlet process mixtures of multinomial distributions (DPMPM) (Dunson and Xing (2009)). This approach automatically models complex dependencies whereas other MI methods (log linear model or conditional logistic regressions) can fail to detect complex structures in high-dimensional categorical variables. Akande et al. (2017) compared the performance of various MI methods for categorical data. According to their study, the Bayesian mixture model approach dominates the approach based on chained equations (which uses generalized linear models) and is as reliable as imputations based on CART in MICE. They also found that the Bayesian joint modelling approach is substantially faster than the FCS methods for MI. However, in the presence of a large number of categorical and continuous variables, the sequential behaviour of CART can form suboptimal and unstable trees (Hastie et al. (2001)), (Marshall and Kitsantas (2012)), (Strobl et al. (2009)). Moreover, to implement a fully Bayesian, joint modelling approach as suggested by Akande et al. (2017), one has to either discard all continuous variables or to categorize them. Murray and Reiter (2016) extended the Bayesian, joint modelling approach for multivariate continuous and categorical variables. However, this approach involves knowledge of complicated models to create the dependence structure between the continuous and the categorical variables. Schafer (1997) uses a JM approach called general location models for a mixture of continuous and categorical variables. Despite of being superior to FCS and CART in many ways, He (2010) suggests that the JM approaches can lack the flexibility needed to represent complex data structures arising in many studies (van Buuren (2007)).

Various recursive partitioning (RP) techniques (Iacus and Porro (2007, 2008)), (Nonyane and Foulkes (2007)), (Burgette and Reiter (2010)), (Stekhoven and Bühlmann (2012)), (Doove et al. (2014)) were proposed to overcome the problem of ignoring interactions in chained equations but most of the proposed methods combine recursive partitioning with single imputation instead of multiple imputation.

An approach called multilevel singular value decomposition (SVD) is used by Husson et al. (2018) for mixed data. SVD uses the between and within groups variability to impute values. One major drawback of SVD is that it cannot be implemented with MI. Geneviève et al. (2018) addressed main effects and interaction challenges in mixed and incomplete data frames.

MI by multiple correspondence analysis (MIMCA) (Audigier et al. (2017b)) utilizes the dimensionality reduction property of multiple correspondence analysis to impute categorical data with a high number of categories. Estimates obtained by MIMCA are as reliable as methods using MI with log linear models or conditional logistic regressions. MIMCA is less time-consuming on data sets with high dimensions than the other multiple imputation methods. However, MIMCA is limited to only categorical variables. Imputation methods that treat the categorical data as continuous, for example, as multivariate normal, can work well for some problems but are known to fail in others, even in low dimensions (Ake (2005)), (Allison (2000)), (Bernaards et al. (2007)), (Finch (2010)), (Graham and Schafer (1999)), (Horton et al. (2003)), (Yucel et al. (2011)).

An iterative singular value decomposition (SVD) algorithm for MI can be a good choice for quantitative (Hastie et al. (2015)), qualitative (Audigier et al. (2017a)) and mixed data (Audigier et al. (2016)) because of better performance

than their counterparts. However, these methods cannot be suitable for the complex data we address in this paper.

Recently, hybrid techniques for imputations have gained a lot of attention (Ankaiah et al. (2011)), (Tang et al. (2015)), (Liyong et al. (2016)), (Shukur and Lee (2015)). For example, Ankaiah and Ravi (2011) propose a hybrid two stage imputation method involving the K-means algorithm and a multi-layer perceptron (MLP) in stage 1 and stage 2, respectively. Also, Nishanth et al. (2012) proposed a hybrid clustering and model based method, where they combine the K-means with an artificial neural network (ANN). Nishanth and Ravi (2013) propose an online data imputation framework incorporating data mining techniques. Considering the local similarity of data, Li et al. (2013) borrowed the idea from clustering and applied it to the problem of missing data imputation. Azim et al. (2014) present a hybrid model that uses a multi-layer perceptron and a fuzzy c-means clustering working in sequence for data imputation. Liang et al. (2015) also proposed a novel missing value imputation method using the stacked auto-encoder and incremental clustering (SAIC). However, obtaining good clustering results may become challenging due to the expansion of the data volume with existing clustering algorithms. Multiple Imputation using grey theory and entropy based on clustering (MIGEC) is another hybrid missing data method proposed by Ting et al. (2014). The MIGEC method divides the complete data into clusters and selects the nearest cluster based on grey theory for each incomplete instance and imputes values using a weighted average based on the information entropy.

Various other MI approaches are suggested in nested imputation (Rubin (2003)), where a set of a variable is imputed based on the former set. Two-stage multiple imputation by Harel (2007), Harel and Schafer (2003), Reiter and Drechsler (2007), Reiter and Raghunathan (2007) are examples for nested imputations. These methods explicitly manage two multiple imputation procedures in a dependent structure (Rubin (2003)). Weirich et al. (2014) extended nested imputation methods in both continuous and categorical background variables for a large-scale assessment. However, we think these procedures are computationally more extensive, implemented in limited ways and require further research. Zhao and Long (2016) did some recent work for imputation methods in the presence of high-dimensional data. However, they focused on the setting where only one variable has missing values. Most recently, Nikfalazar et al. (2019) proposed a new hybrid imputation method that deals with the missing data issue in the Mobility in Cities Database (MCD). Their hybrid method combines features of decision trees and fuzzy clustering into an iterative algorithm for missing data imputation.

When dealing with large scale complex data with missing values in high-dimensional situations, we desire a hybrid multiple imputation approach (HMI) that (i) avoids odd behaviours of FCS techniques in high dimensions (ii) avoids difficulties of creating complicated models for the dependence between the continuous and the categorical variables as in JM approaches (iii) avoids the problem of a specification of clusters (iv) offers efficient computation. HMI is a flexible and practical technique, which combines properties of existing approaches to handle missing values in large scale complex surveys. We propose a HMI technique which applies FCS MI techniques to impute continuous

variables based on information obtained by categorical variables that are already imputed by a JM MI approach.

4. Materials and methods

Before introducing the proposed hybrid architecture, a brief description of FCS and JM MI methods, Rubin's inference and various estimates used for comparing the performance of the imputation algorithms is given below.

4.1. Fully Conditional Specification (FCS): Chained Equations

The FCS method specifies an imputation model for each variable with missing values conditional on the other variables in the data set. Missing values are sequentially imputed in each iteration. Imputation starts from the first variable with missing values.

In the first step, initial values for missing values in all variables are specified, i.e. Y_1^0 , ... ,Y_1^0.

In the second step, at iteration t: for j = 1 to p, most recently imputed values, i.e. X, Y_1^t, ... ,Y_{j-1}^t,Y_{j+1}^{t-1} , ... ,Y_p^{t-1} of all other variables, X, Y_2^{t-1}, ... , Y_p^{t-1} for j=1 and Y_1^{t-1}, ... ,Y_p^{t-1} use a univariate method to impute all missing values in the jth variable Y_j^t. Here, X denotes a set of variables that have no missing values. Repeat the second step until the maximum number of iterations is reached. The above steps (including the first one) are repeated M times to get M imputations. The starting values for each chain are generated with a different seed for random numbers to generate different initial values.

4.2. Fully Bayesian joint modelling (JM) using Dirichlet process infinite mixtures of products of multinomials (DPMPM)

The fully Bayesian, joint modelling (JM) approach known as "Dirichlet process mixtures of products of multinomial distributions model" (DPMPM) (Dunson and Xing, (2009)) is described as:

1. Assume that each individual *i* belongs to exactly one of $K < \infty$ latent classes.
2. For *i = 1,…, n*, let $g_i \in \{1, \dots, k\}$ indicate the class of individual *i*, and let $\pi_k = Pr(g_i = k)$. Assume further that $\pi = \{\pi_1, \dots, \pi_k\}$ is the same for all individuals.
3. Within any class, we suppose that each of the *j* variables independently follows a class-specific multinomial distribution, i.e. for any value

$$y_j \in \{1, \dots, d_j\}, \text{let } \phi_{k_c j}^{(j)} = Pr(y_{ij} = y_j \mid g_i = k).$$

Note that d_j denotes the number of categories of the *j*-th variable.

Mathematically, the finite mixture model can be expressed as follows:

$$y_{ij} | g_i , \phi \overset{ind}{\sim} \text{Multinomial } (\phi_{g_i 1}^{(j)}, \dots, \phi_{g_i d_j}^{(j)}) \text{ for all } i \text{ and } j \quad (4.1)$$

$$g_i | \pi \sim \text{Multinomial } (\pi_1, \dots, \pi_K) \text{ for all } i \quad (4.2)$$

For prior distributions on ϕ and π , we have

$$\pi_k = V_k \left(\prod_{l<k} 1 - V_g\right) \text{ For } k=1,\ldots,K$$

$$V_k \overset{iid}{\sim} Beta\ (1,\ \alpha)$$

$$\alpha \sim Gamma\ (a_\alpha,\ b_\alpha\ )$$

$$\phi_{kj} \sim Dirichlet \quad (\ a_{j1}\ ,\ldots,\ a_{jd_j})$$

We set a_{j1}=…= a_{jd_j} = 1 for all *j*, and (a_α = 0.25; b_α = 0.25). In order to get complete data sets, first the latent class indicator for each individual is drawn from the full conditional and then each missing y_{ij} is drawn from the class specific, independent multinomial distributions.

4.3. Rubin's inference:

For *m = 1,…,M,* let $q^{(m)}$ and $u^{(m)}$ be respectively the point estimates of *Q* (e.g. the estimated regression coefficient in an analysis model) and the variance estimates of $q^{(m)}$ of the interesting analysis model, e.g. a parametric regression model. Valid inferences for a scalar *Q* are obtained by combining the $q^{(m)}$ and $u^{(m)}$, using Rubin's (1987) rules as follows:

$$\overline{q}_M = \sum_{m=1}^{M} \frac{q^{(m)}}{M}, \tag{4.3}$$

$$b_M = \sum_{m=1}^{M} \frac{(q^{(m)} - \overline{q}_M)^2}{M-1}, \tag{4.4}$$

$$\overline{u}_M = \sum_{m=1}^{M} \frac{u^{(m)}}{M}, \tag{4.5}$$

$\overline{q}_M$ can be used to estimate *Q* and the variance of $\overline{q}_M$ can be estimated by

$$T_M = \left(1 + \frac{1}{M}\right) b_M + \overline{u}_M, \tag{4.6}$$

with degrees of freedom $v_M = (M-1)\left(1 + \frac{\overline{u}_M}{\left(\left(1+\frac{1}{M}\right) b_M\right)^2}\right)$. (4.7)

5. Proposed hybrid architecture

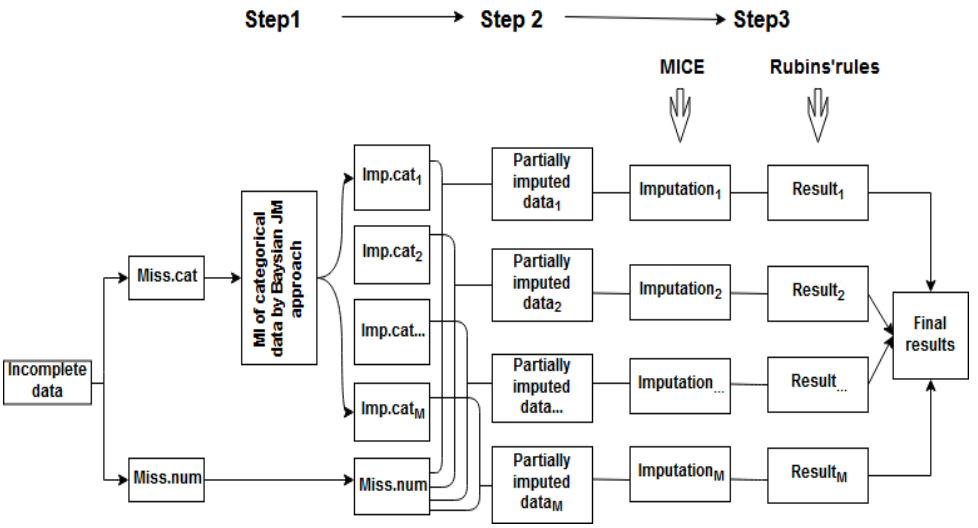

Figure 1. Schematic diagram illustrating the proposed hybrid architecture

A schematic diagram illustrating the proposed hybrid architecture is provided in Figure 1. The proposed missing data imputation approach is a 3-stage approach. **Step 1**: We begin by partitioning incomplete data into two different groups, i.e. categorical data → Miss.cat and incomplete continuous data → Miss.num, where Miss.cat and Miss.num may contain missing values. After partitioning, multiple complete versions → Imp.cat are created for Miss.cat by applying a fully Bayesian joint modelling approach to MI. In this step, Miss.num still contains missing values. **Step 2**: All variables in the data set Miss.num are added to each of the Imp.cat data sets, resulting in *M* partially imputed datasets where values in the continuous variables may be missing and values in the categorical variables have already been imputed in step 1. **Step 3:** Incomplete continuous variables in the *M* partially imputed datasets are imputed using MICE such that the draws from the posterior predictive distribution of the unobserved continuous data depend on the given categorical variables, which have been already imputed by the fully Bayesian joint modelling MI.

To implement the HMI approach, we combine a JM approach “DPMPM” with the FCS approach MICE. We select “DPMPM” due to its computational efficiency, its ability to automatically model complex dependencies and its successful implementation for the case of high-dimensional categorical variables in various fields, i.e. econometrics (Chib and Hamilton (2002), Hirano (2002)), social science (Kyung, Gill, and Casella (2010)), and finance (Rodrı´guez and Dunson (2011)). MICE is selected due to its open source character and popularity. R (R Core Team (2018)) software, version 3.0.1 is used to perform all calculations. The two R packages “mice” (van Buuren and Groothuis-Oudshoorn, (2011)), version 2.17

and “NPBayesImputeCat” (Quanli et al. (2018)), version 0.1 are used to implement the HMI approach. The “default” function of “mice” uses predictive mean matching (PMM) for continuous variables, logistic regression for factor variables with two levels and multinomial logit model for more than two categories. We also use the package 'mitools' (Thomas (2019)) to combine the results from MI. Default versions of chained equations using “mice” fail to impute missing values in the child data. The neural net function, called by “mice” for categorical variables with more than two categories, stops the default version because of exceeded "maximum allowable number of weights". The function “nnet” is used to prevent running code that will take a very long time to complete when there are factor variables with many levels. This gives an indication that complex dependence structures in the data make it complicated to identify them by the default application of MICE. Therefore, we did not implement the default version and compare two HMI approaches, i.e. “H.CART” and “H.DEF” with the MICE based method “$\text{Mice}_{\text{CART}}$” (classification and regression trees (CART)). “H.CART” and “H.DEF” combine a fully Bayesian joint modelling approach with the MICE algorithms “CART” and “Default”, respectively. To implement the hybrid approach, we examine a small prior specification for a_α and b_α (i.e. a_α= 0.25, b_α = 0.25) with a moderate number of mixture components (i.e. *k*=80).

6. Simulation studies

To investigate the performance of the HMI method via simulation, we generate a large number *(X=39)* of mixed type variables. First, we generate 31 binary *(X_b)* variables. A multivariate normal (MVN) distribution is used to generate correlated random covariates C_i comprising 1000 observations. 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 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:

$$\mu_1 = 0.1 + 0.1 \sum_{i=1}^{31} X_{b_i} + 0.1 X_{b_2} X_{b_3} + 0.1 X_{b_5} X_{b_8} + 0.1 X_{b_2} X_{b_{29}}. \quad (6.1)$$

$$\mu_2 = 0.1 + 0.1 \sum_{i=1}^{31} X_{b_i} + 0.1 C_{32} + 0.1 X_{b_2} X_{b_3} + 0.1 X_{b_5} X_{b_8} + 1.1 X_{b_2} X_{b_{29}}. \quad (6.2)$$

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 distributions with probabilities governed by the logistic regression with

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

(6.3

We then generate two continuous covariates, i.e. X_{n_1} and X_{n_2} from normal distributions (ND) as follows:

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

Where, $\mu_3 = -2 - 1.5X_{b_1} + 2.15X_{b_2} + 2.25\ X_{b_3} - 3.6\ X_{b_4} - 1.88X_{b_5} + 1.11\ X_{b_6} + 2X_{b_7} - 5X_{b_8} + X_{b_9} - 2X_{b_{10}} + 2X_{b_{11}} + 5X_{b_{12}} - 2X_{b_{13}} + 3X_{b_{14}} + 4X_{b_{15}} + X_{b_{16}} + X_{b_{17}} - X_{b_{18}} - X_{b_{19}} - X_{b_{20}} - X_{b_{21}} - X_{b_{22}} + 2X_{b_{23}} - X_{b_{24}} + X_{b_{25}} + X_{b_{26}} + X_{b_{27}} + X_{b_{28}} + X_{b_{29}} + X_{b_{30}} + X_{b_{31}} + 2C_{32} - C_{33} + X_{b_{32}} + 2X_{b_{11}}\ X_{b_{12}}\ X_{b_{13}} - 2\ X_{b_{15}}X_{b_{18}} + 2X_{b_{12}}\ X_{b_{29}}.$

(6.4)

And

$$X_{n_2} \sim N\ (\mu_4; \sqrt{0.5}). \tag{6.5}$$

Where, $\mu_4 = \mu_3 + X_{n_1}$.

(6.6)

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

We then define a covariate dependent continuous response with expectation

$$\mu_y = 1 + \sum_{i=1}^{32} X_{b_i} + \sum_{i=1}^{4} X_{n_i} + \sum_{i=2}^{4} X_{m_{1\_i}} + \sum_{i=2}^{6} X_{m_{2\_i}} + X_{b_9}\ X_{b_{15}} + X_{b_1}\ X_{b_{17}} + X_{b_{14}}\ X_{b_{20}} + \epsilon.$$

(6.7)

Additionally, a random component $\epsilon \sim N\ (0;\ 0.5)$ is added. The regression coefficients for categorical variables with multiple levels are expressed as dummy variables, e.g. $\sum_{i=2}^{4} X_{m_{1\_i}}$ and $\sum_{i=2}^{6} X_{m_{2\_i}}$ in the predictor (all coefficients are 1.0).

Equations 6.1–6.7 include higher-order interactions to represent 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 linear model. Observations in all covariates can be missing (at random) with probabilities based on a logistic probability distribution model. Probabilities for missing for a random covariate *X* are given as:

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

Here, *i*={*1,…,39*} and *j* ≠ *i.* Missingness in X_i is attributed solely to other observed variable X_j. This yields 10% of the observations to be MAR. Based on recommendations in the MI literature (White et al. (2011)), (van Buuren (2012)), we decided to include all of the variables from the generated data in the imputation model to ensure that the imputation model preserves the relationships between the variables of interest (Schafer (1997)), (Moons et al. (2006)). Based on *Z =1000* simulation runs, the parameters of interest are estimated using the aforementioned Rubin's method. According to Rubin (1987), the number of suitable imputations for useful statistical inferences can be determined by a fraction of missing data. A surprisingly high relative efficiency can be obtained with no more than five imputations. Fichman and Cummings (2003) suggest, that *M=10* imputations are more than suitable in almost any realistic application. Therefore, ten imputed datasets are generated for each of the proposed and the MICE MI methods. Two hundred iterations (for each imputation step) are run to insure convergence and to obtain results of the simulations in a reasonable time. To compare the performance of the imputation algorithms, two error-based measurements were chosen to evaluate the quality of MI: Root mean square error (RMSE) and empirical standard errors (ESE) (Akande et al. (2017)), (Armina et al. (2017)). Smaller values for RMSEs and ESEs indicate better performance (Oba et al. (2003)). RMSE and ESE are calculated using the following formulas:

$$\text{Root mean square error } (\text{RMSE}_{\bar{q}_m}) = \sqrt{\frac{\sum_{z=1}^{Z}(\bar{q}_M^z - \beta)^2}{Z}}, \tag{6.8}$$

$$\text{Empirical standard errors } (\text{ESE}_{\bar{q}_m}) = \sqrt{\frac{\sum_{z=1}^{Z}(\bar{q}_M^z - \bar{q})^2}{Z}}, \tag{6.9}$$

where $\bar{q}_M^z$ denotes the estimated parameter pooled over *M* imputed data sets in simulation run number *z* and *β* denotes the original parameter. The arithmetic mean of $\bar{q}_M^z$ and ($\sqrt{T_M}$) across all ***z = {1,…,Z}*** simulations are denoted as $\bar{q}$ and $\overline{\sqrt{T}}$. The amount of bias can be calculated by a simple difference, i.e.

$$Bias = RMSE - ESE \tag{6.10}$$

The coverage rates of at least 95% are calculated as:

$$\text{Coverage rate}_{\bar{q}_m} = \frac{\sum_{z=1}^{Z} 1\,[\beta \in CI\,(\bar{q}_M^z, T_M^z)]}{Z}, \tag{6.11}$$

where $1\,[\beta \in CI\,(\bar{q}_M^z,\ T_M^z)]$ is an indicator function. The indicator function is equal to one when the confidence interval based on $\bar{q}_M^z$ and T_M^z contains β and equal to zero otherwise.

Table 1 gives the performance of the MI methods. Means for CI coverage and RMSEs over all beta coefficients are presented in Table 2. Various researchers (White et al. (2011)), (van Buuren, 2012)) recommend graphical comparisons of the imputation methods. For that purpose, boxplots of standard errors ($\sqrt{T_M}$) and point estimates ($\overline{q}_M$) for the regression coefficients for the 1000 simulation runs are presented in Figures 2 and 3 respectively.

6.1. Results

Table 1. Performance of methods for MI

Estimates	Parameter	MICE$_{CART}$	H.DEF	H.CART
RMSEs (ESEs)	$X_{b_{23}}$	0.158(0.114)	0.148(**0.089**)	**0.122**(0.110)
	$X_{m_{1\_2}}$	0.158(0.155)	0.228(**0.122**)	0.173(0.158)
	$X_{m_{2\_3}}$	0.187(0.148)	0.167(**0.114**)	**0.164**(0.145)
	$X_{b_{32}}$	0.045(0.032)	0.071(**0**)	**0.032**(0.032)
	X_{n_2}	0.063(0.063)	0.071(**0.032**)	**0.055**(0.055)
	$X_{b_1}X_{b_{17}}$	**0.190**(0.182)	0.239(**0.130**)	0.195(0.190)
$\bar{q}(\overline{\sqrt{T}})$	$X_{b_{23}}$	0.891(0.192)	1.119(**0.137**)	0.947(**0.137**)
	$X_{m_{1\_2}}$	1.038(0.266)	0.808(**0.193**)	0.928(0.272)
	$X_{m_{2\_3}}$	0.887(0.245)	1.122(**0.176**)	0.920(0.249)
	$X_{b_{32}}$	0.969(0.049)	1.065(**0.027**)	1.006(0.047)
	X_{n_2}	1.014(0.088)	0.935(**0.049**)	0.995(0.086)
	$X_{b_1}X_{b_{17}}$	0.951(0.319)	0.800(**0.255**)	0.958(**0.225**)
Bias	$X_{b_{23}}$	0.044	0.059	**0.012**
	$X_{m_{1\_2}}$	0.772	**0.615**	0.656
	$X_{m_{2\_3}}$	**0.039**	0.053	0.671
	$X_{b_{32}}$	0.013	0.071	**0**
	X_{n_2}	0.956	**0.886**	0.909
	$X_{b_1}X_{b_{17}}$	0.008	0.109	**0.005**
Coverage(%)	$X_{b_{23}}$	99	95	100
	$X_{m_{1\_2}}$	100	94	100
	$X_{m_{2\_3}}$	100	97	100
	$X_{b_{32}}$	97	29	99
	X_{n_2}	99	83	100
	$X_{b_1}X_{b_{17}}$	100	96	100

Root mean square errors and empirical standard errors (top), point estimates, standard errors and bias for different methods (middle) and estimated coverage probability (bottom) for MI methods under the Missing at Random (MAR) assumption. The middle panel lists the mean estimated standard errors and point estimates across the simulated data sets. All results are based on 10 imputations. Estimates are shown for only six regression coefficients, i.e. for variables $X_{b_{23}}$, $X_{m_{1\_2}}$, $X_{m_{2\_3}}$, $X_{b_{32}}$, X_{n_2}, $X_{b_1} X_{b_{17}}$. Bold figures indicate the smallest mean root mean square errors, mean empirical standard errors and amount of bias among the three imputation variants.

Table 2. Results over all beta coefficients

Estimates	**MICE$_{CART}$**	**H.DEF**	**H.CART**
CI coverage	98.66	91.91	99.89
RMSEs	0.184	0.170	**0.146**

Means for CI coverages and RMSEs are estimated over all regression coefficients for all MI methods. Bold values indicate the smallest mean for RMSEs over all regression coefficients among the three imputation variants.

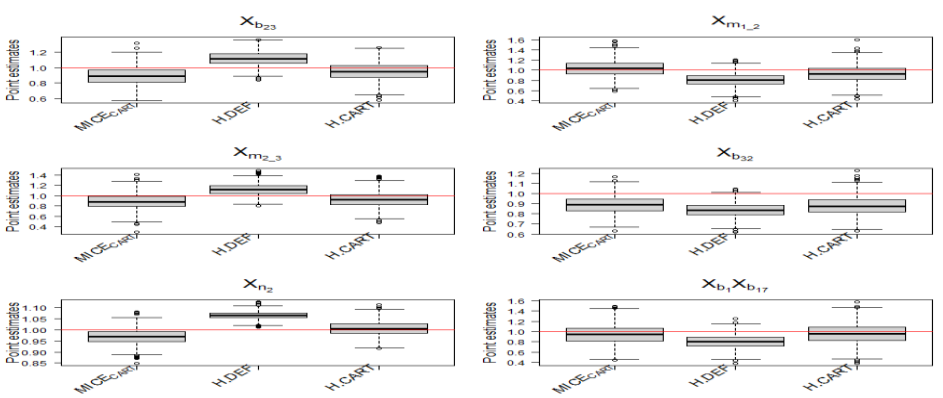

Figure 2. Simulated data: Boxplots for the point estimates ($\overline{q}_M$) across 1000 simulations by imputation methods under Missing at Random (MAR) and ten imputations. Point estimates are shown for only six regression coefficients, i.e. for variables $X_{b_{23}}$, $X_{m_{1\_2}}$, $X_{m_{2\_3}}$, $X_{b_{32}}$, X_{n_2}, $X_{b_1} X_{b_{17}}$.The horizontal red lines indicate the respective “true” values

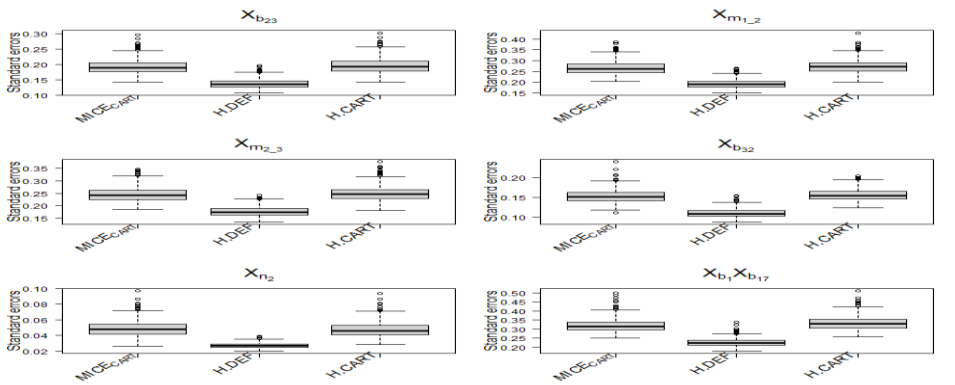

Figure 3. Simulated data: Boxplots for standard errors ($\sqrt{T_M}$) across 1000 simulations by imputation methods under Missing at Random (MAR) and ten imputations. Standard errors are shown for only six regression, i.e. $X_{b_{23}}$, $X_{m_{1\_2}}$, $X_{m_{2\_3}}$, $X_{b_{32}}$, X_{n_2} , $X_{b_1}X_{b_{17}}$ coefficients

The average point estimates based on H.CART are closer to the corresponding true values than those based on CART. H.CART tends to be less biased as compared to the CART method for all types of covariates and interaction terms, whereas H.DEF tends to be upward biased for binary and the multilevel covariate with four levels and slightly downward biased for the multilevel covariate with six levels, for the continuous covariates and the interaction terms as compared to the CART method (Figure 2). There seem to be similarities in the structure among all MI methods (i.e. all methods are downward biased) for binary covariate $X_{b_{32}}$, which was generated with higher order interactions. The H.DEF method tends to have smaller standard errors as compared to two relevant methods for all covariates, whereas the H.CART method tends to have similar standard errors as compared to CART for most of the cases (Figure 3). The estimated $RMSE_S$, ESEs and averages of standard errors for the H.CART method are smaller for all types of covariates except the multilevel covariate with many categories. H.CART shows similar ESEs and averages of standard errors and slightly higher $RMSE_S$ for the multilevel covariate with more categories as compared to CART. The H.DEF method shows smaller ESEs and averages of standard errors for all types of covariates and slightly higher RMSEs for most of the covariates as compared to the other methods (Table 1). The H.DEF method led to more overall accuracy with smaller means for RMSEs over all beta coefficients as compared to CART (Table 2). A possible explanation for the efficiency gain with H.DEF is that it was able to make better use of the available information by accommodating nonlinearities among the predictors. For the most part, coverage rates for H.CART are in line with those from CART and produce almost identical results. In most cases, coverage probabilities for H.CART were 100%, which suggests that these confidence intervals may be too conservative. The simulated coverage rates of the 95% confidence intervals based on H.DEF are near to nominal 95% for most cases. Few of the incidences in H.DEF led to under-coverage. All but one of the

incidences, i.e. $X_{b_{32}}$in which coverages dip below 30% occur. This severe under-coverage suggests that H.DEF (which uses the Bayesian approach for categorical and PMM as default for continuous covariates) might performing not well for continuous covariates but works well for categorical covariates. This might be one of the reasons that H.DEF gets biased results. Increasing *M* can lead to obtain coverage rates that are close to nominal in the case of under-coverages. Nevertheless, the H.DEF method led to coverage rates that are close to nominal over all beta coefficients as compared to CART (Table 2).

7. Imputation of MICS child data

The data for MICS is collected at both family and person level and it allows the study of relationships between health indicators and other characteristics. In this study, we use the child data set from the MICS Punjab 2014 household survey. The MICS Punjab data for children contains more than two hundred indicators on a variety of a child's conditions. For example, indicators on a child's mental development (e.g. a child is able to pick up small object with 2 fingers, etc.), a child's nutrition intake in diet (e.g. a child drank or ate vitamin or mineral supplements, etc.) and vaccinations (e.g. ever had vaccination card, etc.). The MICS data for children contains a complex data structure for categorical variables with multiple levels and large amounts of missingness, which can be problematic for MICE. It can be tedious for MICE to specify imputation models and interaction terms in the presence of large databases with hundreds of variables and multicollinearity (Van Buuren and Oudshoorn 1999). It was not possible to have a proper comparison of the proposed and existing MI approaches in such case. Therefore, multiple categories for categorical variables were reduced by merging them, and a sub-sample of 52 variables, which contains information on child health, nutrition and development, is selected from MICS Punjab 2014 children data. Among these variables, 43 background variables are categorical with multiple categories and the remaining are continuous. Demographical variables like "district" and "area" are also included in the sub-sample. In this sub-sample, 5 variables have between 6 and 21% of missing values, 17 variables have 48% of missing values, 27 variables have between 50% and 86% of missing values, and 1 variable has more than 90% of missing values. Of all variables, only 3, i.e. "sex", "wealth" and "area", have complete records (see additional file). The variable "district" has 36 levels, hence keeping the analysis comparable and challenging at the same time. There are various reasons listed for item non response in the methodology of MICS i.e. nonresponse, don't know and not reached, etc. Without distinguishing reasons for item non response, we assume that the items are MAR in the data under consideration. Similar to the simulation study, all of the variables from the sub-sample are included in the imputation model.

After imputations, parameters of interest for the child health are estimated using linear models for continuous response (height for age percentiles NCHS). The response variable, "height for age percentiles NCHS", is obtained by using a table of Z-scores (percentile = the area from infinity to Z). Based on the evidence from demographical and behavioural risk factors associated to height, two continuous covariates i.e. "age", "polio_vacc." and two categorical variables, i.e.

"grains_in_diet" (Yes/ No) and "eggs_in_diet" (Yes/ No) are selected as potential determinants in the analysis model. Since there are no true values to compare for in the real data example, we calculated complete case (CC) estimates for comparison purposes (Table 5). The R package "VIM" (Templ et al. (2012)) is utilized for exploring data and the pattern of missing values. Figure 4 shows graphics of the incomplete predictors. Graphics for the remaining variables in the sub-sample are provided in an additional file. Similar to the simulation study ESEs, average point estimates and average standard across the 200 simulations are calculated for real data. Computational time and ESEs for MI methods are shown in Tables 3 and 4 respectively. Figures 5 and 6 display the average point estimates and average standard errors for the MI methods across the 200 simulations.

7.1. Results

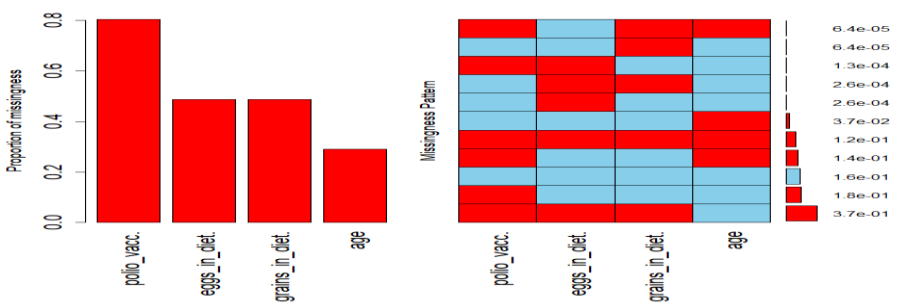

Figure 4. Real data: Aggregateplot in R, graphics of incomplete predictors. For purposes of displaying the graphical depiction, only four variables with proportions of missing values ranges between 18-28 were selected

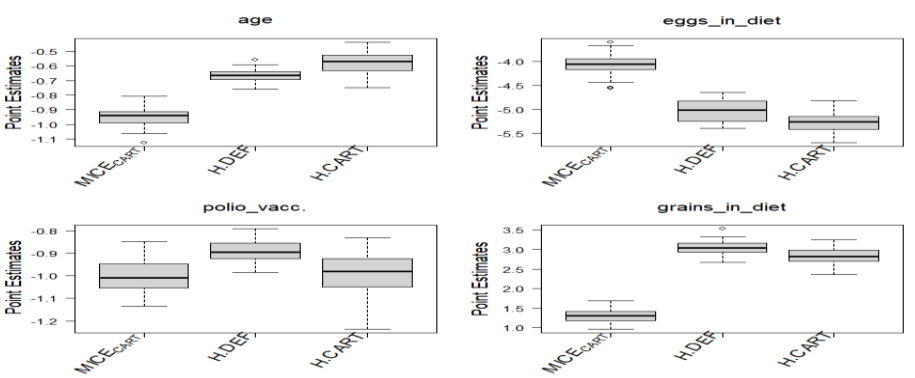

Figure 5. Real data: Boxplots for point estimates ($\overline{q}_M$) across 200 simulations by imputation methods under Missing at Random (MAR) and ten imputations

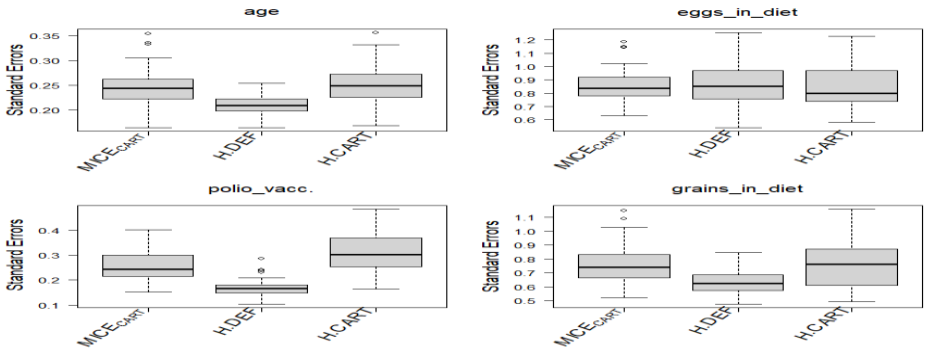

Figure 6. Real data: Boxplots for standard errors ($\sqrt{T_M}$) across 200 simulations by imputation methods under Missing at Random (MAR) and ten imputations.

Table 3. Real data: Time taken for various MI methods

Method	Default	CART	H..DEF	H.CART
Time	No run	3.25$_d$	22.78$_{h.}$	21.21$_h$

Note: time = the time to complete 10 multiple imputation by variants of MI across 1000 simulations, h = hours, d = days, and Not Run = the program not able to complete multiple imputation on this subset. The maximum number of iterations is set to 200.

Table 4. Real data: ESEs for various MI methods

Variables	CART	H.DEF	H.CART
age	0.06	**0.04**	**0.06**
eggs_in_diet	0.21	0.22	**0.20**
polio_vacc.	0.07	**0.04**	0.09
grains_in_diet	0.17	**0.16**	0.21

Empirical standard errors by imputation methods under Missing at Random (MAR) and ten imputations. Cases where both HMI methods result in minimum between imputation variances (ESEs) as compared to CART are highlighted in bold.

Table 5. Real data: complete case (CC) estimates

Variables	est	se
age	3.542	0.899
eggs_in_diet	-9.866	1.305
polio_vacc.	-0.808	0.242
grains_in_diet	0211	1,342

The CC analysis uses only the complete cases (n = 4264), “est” and “se” denote the point estimates and standard errors of the coefficients of the linear model, respectively.

Figure 4 displays graphics of incomplete predictors. The bar plot on the left side shows the proportions of missing values in the predictors. The continuous predictor “polio_vacc.” has the highest amount of missing values (i.e. about 80%) while the amount is rather small in the other three variables (i.e. less than 60% for two binary predictors and less than 40% for predictor “age”). An aggregation plot on the right side shows all existing combinations of missing (red) and imputed observed (blue) values. Additionally, the frequencies of different combinations are visualized by a small bar plot and by the number of their occurrences on the right side (Templ et al. (2012)). The aggregation plot reveals that missing values in the variable “polio_vacc.” are also missing in the two binary variables. We note that the standard errors for all of the coefficients are smaller compared to the (absolute) point estimates under all MI methods (see Figures 5-6). This happens most likely due to sampling variability in the multiple imputation inferences. The empirical example with real data indicated that the CART and HMI variants yielded differing point estimates. We noticed that point estimates in CART are nearer to the estimates in complete case analysis for most of the cases with larger standard errors as compared to hybrid methods (see Table 5, Figures 5-6). Figure 6 displays smaller standard errors for H.DEF as compared to CART. ESEs for HMI variants are also smaller as compared to CART for most of the cases (see Table 5), suggesting better performance over 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. Almost 4 days were taken by CART to run on standard computers, whereas, surprisingly, this time was reduced to almost one day when HMI methods were applied. We also applied the proposed methods to the full MICS data set with hundreds of variables and categories with multiple levels. We found that the proposed methods have a good capacity to perform for the MICS data where the MICE methods simply fail.

8. Conclusion and remarks

We acknowledge that results of MI can be biased even when complex multivariate data is MAR (White and Carlin, 2010). However, in this paper, we

assumed that the missing data mechanism is MAR. We applied our hybrid strategy to handle missing data in large scale survey data with complex dependence structures among categorical variables and a high percentage of missing information. Identification of complex dependence structures among mixed type covariates will be difficult for JM and FCS MI methods in high dimensions. We obtain promising results by performing an illustrative analysis. The results obtained from the simulation studies and a real data example confirm the potential of our proposed approach to handle missing data under MAR. Superiority of H.DEF was its efficiency relative to the other imputation inference methods. The H.DEF method outperformed the other methods with respect to RMSEs, ESEs and standard errors but its point estimates were downwardly biased for a few regression coefficients, which led to under-coverage of the confidence intervals. H.CART gives estimates with less bias but over-coverage of confidence intervals. There was no noticeable difference in coverage and standard errors between H.CAT and CART. H.CART produces smaller RMSEs and ESEs for most parts and 3 times less computational cost as compared to MICE. A problem of the HMI approach is that it does not use the information available on the continuous variables for imputing the categorical variables. Further work is needed to use iterative procedures to develop strong relationships between the categorical and continuous variables. Currently, we are implementing solutions for this problem and we use the concept of categorizing continuous variables. We are working on the development of a new R package that will implement the proposed HMI approach with the hope that it will contribute in MI of large scale survey data.

Acknowledgments

The authors thank the editor and two referees for their suggestions, which greatly helped improve the article.

REFERENCES

ANDERSON, A. B., BASILEVSKY, A., HUM, D. P., (1983). Missing data: A review of the literature. In J. D. W. P. H. Rossi and A. B. Anderson (Eds.), Handbook of survey research, New York: Academic Press.

ARNOLD, B. C., PRESS, S. J., (1989). Compatible Conditional Distributions. Journal of the American Statistical Association, 84, pp. 152–156.

ALLISON, P. D., (2000). Multiple imputation for missing data: A cautionary tale. Sociological Methods and Research, 28, pp. 301–309.

AKE, C. F., (2005). Rounding after multiple imputation with non-binary categorical covariates (paper 112-30). In Proceedings of the Thirteenth Annual SAS Users Group International Conference, SAS Institute Inc., Cary, NC, pp. 1–11.

ANDRIDGE, R. R. (2009). Statistical methods for missing data in complex sample surveys. PhD thesis, The University of Michigan.

AKMATOV, M. K., (2011). Child abuse in 28 developing and transitional countries--results from the Multiple Indicator Cluster Surveys, Int J Epidemiol, 40(1), pp. 219–27.

ANKAIAH, N., RAVI, V., (2011). A novel soft computing hybrid for data imputation, Proceedings of the 7th international conference on data mining (DMIN), Las Vegas, USA.

AZIM, S., AGGARWAL, S. (2014). Hybrid model for data imputation: using fuzzy c means and multi layer perceptron. Advance Computing Conference (IACC), 2014 IEEE International. IEEE, pp. 1281–1285.

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), pp. 5–26.

AKANDE, O., LI, F., REITER, J., (2017). An empirical comparison of multiple imputation methods for categorical data, Amer. Statist, 71, pp. 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, pp. 012004.

AUDIGIER, V., WHITE, I. R., JOLANI, S., DEBRAY, T., QUARTAGNO, M., CARPENTER, J., ESCHE-RIGON, M., (2017a), Multiple imputation for multilevel data with continuous and binary variables, arXiv preprint, arXiv:1702.00971.

AUDIGIER, V., HUSSON, F., JOSSE, J., (2017b). MIMCA: Multiple imputation for categorical variables with multiple correspondence analysis. Statistics and Computing, 27, pp. 501–518.

BREIMAN, L., (2001). Random Forests. Machine Learning, 45(1), pp. 5–32.

BERNAARDS, C. A., BELIN, T. R., SCHAFER, J. L., (2007). Robustness of a multivariate normal approximation for imputation of binary incomplete data, Statistics in Medicine, 26, pp. 1368–1382.

BURGETTE, L. F., REITER, J. P., (2010). Multiple Imputation for Missing Data via Sequential Regression Trees. American Journal of Epidemiology, Oxford University Press, 172(9), pp. 1070–6.

CHIB, S., HAMILTON, B. H., (2002). Semiparametric Bayes analysis of longitudinal data treatment models, Journal of Econometrics, 110, pp. 67–89.

CAPPA, C., KHAN, S.M., (2011). Understanding caregivers' attitudes towards physical punishment of children: evidence from 34 low- and middle-income countries, Child Abuse Negl, 35(12), pp. 1009–21.

DUNSON, D. B., XING, C., (2009). Nonparametric Bayes modeling of multivariate categorical data, Journal of the American Statistical Association, 104, pp. 1042-1051.

DENG, Y., CHANG, C., IDO, M.S., LONG, Q., (2016). Multiple imputation for general missing data patterns in the presence of high-dimensional data. Scientific Reports, 6.

DOOVE, LISA, L., VAN BUUREN, S., ELISE, D., (2014). Recursive Partitioning for Missing Data Imputation in the Presence of Interaction Effects, Computational Statistics and Data Analysis, Elsevier, 72, pp. 92–104.

EROSHEVA E. A., FIENBERG S. E., JUNKER B. W. (2002). Alternative statistical models and representations for large sparse multi-dimensional contingency tables, Annales de la Faculté des Sciences de Toulouse, 11, pp. 485–505.

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

FINCH, W. H., (2010). Imputation methods for missing categorical questionnaire data: A comparison of approaches. Journal of Data Science, 8, pp. 361–378.

GELMAN, A., SPEED, T. P., (1993). Characterizing a joint probability distribution by conditionals, Journal of the Royal Statistical Society Series B: Statistical Methodology, 55, pp. 185–188.

GRAHAM, J. W., SCHAFER, J. L., (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. H. Hoyle (Ed.), Statistical strategies for small sample research, Thousand Oaks, CA: Sage, pp.1–29.

GENEVIÈVE, R., OLGA, K., JULIE, J., ÉRIC M., ROBERT, T., (2018). Main effects and interactions in mixed and incomplete data frames. arXiv preprint, arXiv:1806.09734.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J., (2001). The Elements of Statistical Learning; Data Mining, Inference, and Prediction, second ed. Springer Verlag, New York.

HIRANO, K., (2002). Semiparametric Bayesian inference in autoregressive panel data models. Econometrica, 70, pp. 781–799.

HAREL, O., SCHAFER, J. L., (2003). Multiple Imputation in two Stages. Proceedings of the Federal Committee on Statistical Methodology Research Conference, Washington D. C.

HORTON, N. J., LIPSITZ, S. P., PARZEN, M., (2003). A potential for bias when rounding in multiple imputation. The American Statistician, 57, pp. 229–232.

HAREL, O., (2007). Inferences on missing information under multiple imputation and two-stage multiple imputation. Statistical Methodology, 4, pp. 75–89.

HE, Y., (2010). Missing data analysis using multiple imputation: getting to the heart of the matter. Circ Cardiovasc Qual Outcomes, 3, pp. 98–105.

HASTIE, T., MAZUMDER, R., LEE, J. D., ZADEH,R., (2015). Matrix completion and low-rank svd via fast alternating least squares, J. Mach. Learn. Res., 16(1), pp. 3367–3402.

HOLDER, L., (2015). Multiple Imputation in Complex Survey Settings: A Comparison of Methods within the Health Behaviour in School-aged Children Study, Queen's University

HUSSON, F., J. JOSSE, B. NARASIMHAN, G. ROBIN., (2018). Imputation of mixed data with multilevel singular value decomposition, arXiv e-prints, arXiv:1804.11087.

IACUS, S. M., PORRO, G., (2007). Missing data imputation, matching and other applications of random recursive partitioning. Comput. Statist. Data Anal, 52, pp. 773–789.

IACUS, S. M., PORRO, G., (2008). Invariant and metric free proximities for data matching: an R package. J. Stat. Softw, 25, pp. 1–22.

KIM, H., LOH, W.Y., (2001). Classification trees with unbiased multiway splits. Journal of the American Statistical Association, 96, pp. 589–604.

KYUNG, M., GILL, J., CASELLA, G., (2010). Estimation in Dirichlet random effects models. Annals of Statistics, 38, pp.979–1009.

WIRTH, K. E., TCHETGEN TCHETGEN, E. J., (2014). Accounting for selection bias in association studies with complex survey data. Epidemiology (Cambridge, Mass.), 25(3), pp. 444–453.

LOH, W., SHIH, Y., (1997). Split selection methods for classification trees. Statistica Sinica, 7, pp. 815–840.

LITTLE, R. J. A., RUBIN, D. B., (2002). Statistical analysis with missing data (2nd ed.). New York: Wiley.

LEE, K.J., GALATI, J. C., SIMPSON, J. A., CARLIN, J. B., (2012). Comparison of methods for imputing ordinal data using multivariate normal imputation: a case study of non-linear effects in a large cohort study. Stat Med, 31(30), pp. 4164–74.

LI, D., GU, H., ZHANG, L.Y., (2013). A hybrid genetic algorithm-fuzzy c-means approach for incomplete data clustering based on nearest-neighbor intervals. J. Soft Computing, 17, pp. 1787–1796.

LIANG, Z., ZHIKUI, C., ZHENNAN, Y., YUEMING, HU., (2015). A Hybrid Method for Incomplete Data Imputation. 2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems, New York, pp. 1725–1730.

LIYONG, Z., WEI, L., XIAODONG, L., WITOLD, P., CHONGQUAN, Z., LU, W., (2016). A Global Clustering Approach Using Hybrid Optimization for Incomplete Data Based on Interval Reconstruction of Missing Value, International Journal of Intelligent Systems, 31(4), pp. 297–313.

LOH, W. Y., ELTINGE, J., CHO, M., LI, Y., (2016). Classification and Regression Tree Methods for Incomplete Data from Sample Surveys, arXiv preprint arXiv:1603.01631.

LEE, K. J., CARLIN, J. B., (2017). Multiple imputation in the presence of non-normal data. Stat Med, 36(4), pp. 606–17.

MARKER, D. A., JUDKINS, D. R., WINGLEE, M. (2002), Large-Scale Imputation for Complex Surveys. Survey Nonresponse, Wiley: New York, pp. 329–341.

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. J Clin Epidemiol, 59(10), pp. 1092–101.

MORRIS, T. P., IAN, R. W., PATRICK, R., (2014). Tuning Multiple Imputation by Predictive Mean Matching and Local Residual Draws. BMC Medical Research Methodology, BioMed Central, 14(1), 75.

MARSHALL, R. J., KITSANTAS, P., (2012). Stability and structure of cart and span search generated data partitions for the analysis of low birth weight. J. Data Sci, 10, pp. 61–73.

MURRAY, J. S., 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, pp. 1466–1479.

NONYANE, B. A. S., FOULKES, A. S., (2007). Multiple imputation and random forests (MIRF) for unobservable, high-dimensional data. Int J Biostat, 3, pp. 1–18.

NISHANTH, K. J., RAVI, V., ANKAIAH, N., BOSE, I., (2012). Soft computing based imputation and hybrid data and text mining: The case of predicting the severity of phishing alerts. Expert Sys Appl, 39(12), pp. 10583–10589.

NISHANTH, K. J., RAVI, V., (2013). A computational intelligence based online data imputation method: An application for banking. J. Inf. Process. Syst. 9, pp. 633–650.

NIKFALAZAR, S., YEH C. H., BEDINGFIELD, S., KHORSHIDI, H. A., (2019). A Hybrid Missing Data Imputation Method for Constructing City Mobility Indices. In: Islam R. et al. (eds.) Data Mining. AusDM 2018. Communications in Computer and Information Science, Vol. 996. Springer, Singapore.

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, pp. 2088–2096.

QUANLI, W., DANIEL, 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., (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley.

RAGHUNATHAN, T. W., LEPKOWKSI, J. M., VAN HOEWYK, J., SOLENBEGER, P. A., (2001). Multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology, 27, pp. 85–95.

RUBIN, D. B., (2003). Nested multiple imputation of NMES via partially incompatible MCMC. Statistica Neerlandica, 57(1), pp. 3–18.

REITER, J. P., DRECHSLER, J., (2007). Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality. IAB Discussion Paper, 20, pp. 1–18.

REITER, J. P., RAGHUNATHAN, T. E., (2007). The multiple adaptions of multiple imputation, Journal of the American Statistical Association, 102, pp. 1462–1471.

RODRI´GUEZ, A., DUNSON, D. B., (2011). Nonparametric Bayesian models through probit stick-breaking processes. Bayesian Analysis, 6, pp. 145–178.

R Core Team (2018). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria,

https://www.Rproject.org/.

SCHAFER, J. L., (1997). Analysis of Incomplete Multivariate Data. London: Chapman and Hall.

STROBL, C., MALLEY, J., ZEILEIS, A., (2009). An introduction to recursive partitioning: rationale, application and characteristics of classification and regression trees, bagging and random forests. Psychol. Methods, 14, pp. 323–348.

SU, Y.S., GELMAN, A., HILL, J., YAJIMA, M., (2011). Multiplebimputation with diagnostics (mi) in R: Opening windows into the black box. Journal of Statistical Software, 45(2), pp. 1–31.

SEAMAN, S., BARTLETT, J., WHITE, I., (2012). Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. BMC Med Res Methodol, 12(1), pp. 1–13.

STEKHOVEN, D. J., BÜHLMANN, P., (2012). MissForest–non-parametric missing value imputation for mixed-type data. Bioinformatics, 28, pp.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, pp. 499–521.

SHAH, A.D., JONATHAN, W. B., JAMES, C., OWEN, N., HARRY, H., (2014). Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using Mice: A Caliber Study. American Journal of Epidemiology, 179 (6). Oxford University Press, pp. 764–74.

SHUKUR, O. B., LEE, M. H., (2015). Imputation of missing values in daily wind speed data using hybrid AR-ANN method. Modern Applied Science.

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.

TING, J., YU, B., YU, D., MA, S., (2014). Missing data analyses: a hybrid multiple imputation algorithm using gray system theory and entropy based on clustering, Applied intelligence, 40(2), pp. 376–388.

TANG, J., ZHANG, G., WANG, Y., WANG, H., LIU, F., (2015). A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation. Transportation Research Part C: Emerging Technologies, 51, pp. 29–40.

THOMAS, L., (2019). mitools: Tools for Multiple Imputation of Missing Data. R package version 2.4, https://CRAN.R-project.org/package=mitools.

VAN BUUREN, S., OUDSHOORN, C. G. M., (1999). Flexible multivariate imputation by MICE. Tech. rep., TNO Prevention and Health, Leiden.

VAN BUUREN, S., GROOTHUIS-OUDSHOON, K., (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), pp. 1–67.

VAN BUUREN, S., (2007). Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification. Statistical Methods in Medical Research, Sage Publications Sage UK: London, England, 16(3), pp. 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, pp. 369–397.

VAN BUUREN, S., (2012). Flexible imputation of missing data. Boca Raton: CRC Press.

WHITE, I. R., ROYSTON, P., WOOD, A. M., (2011). Multiple imputation using chained equations: issues and guidance for practice. Stat Med, 30(4), pp. 377–99.

WHITE, I.R., CARLIN, J. B., (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Stat Med, 29(28), pp. 2920–31.

WEIRICH, S., HAAG, N., HECHT, M., BÖHME, K., SIEGLE, T., LÜDTKE, O., (2014). Nested multiple imputation in large-scale assessments. Large Scale Assess. Educ., 2, pp. 1–18.

XIE, X., MENG, X.-L., (2017). Dissecting multiple imputation from a multi-phase inference perspective: what happens when God's, imputer's and analyst's models are uncongenial? Statistica Sinica 27, pp. 1485–1594 (including discussion).

YUCEL, R.M., HE, Y., ZASLAVSKY, A. M., (2011). Gaussian-based routines to impute categorical variables in health surveys. Stat Med, 30(29), pp. 3447–60.

ZHU, J., M., EISELE, M., (2013). Multiple Imputation in a Complex Household Survey, The German Panel on Household Finances (PHF): Challenges and Solutions. PHF User Guide.

ZHAO, Y., LONG, Q., (2016). Multiple imputation in the presence of high-dimensional data. Statistical Methods in Medical Research, 25, pp. 2021–2035.