Multiple imputation of large scale complex surveys

Humera Razzak

München 2020

Multiple imputation of large scale complex surveys

Humera Razzak

Dissertation
an der Fakultät fur Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Humera Razzak
aus Faisalabad

München, den 9 December 2019

Erstgutachter: Prof. Dr. Christian Heumann

Zweitgutachter: Prof. Dr. Martin Spieß

Drittgutachter: Prof. Dr. Thomas Augustin

Tag der Disputation: 27.02.2020

For Zarish and Saim

Acknowledgments

First, I thank ALLAH for granting me health, courage, strength, and knowledge to accomplish this task.

I have no words to express my sincere gratitude to my respectable and worthy supervisor Prof. Dr. Christian Heumann for the highly dedicated and patient supervision of my dissertation throughout all the years, the countless consultation meetings and his skillful guidance in statistical research questions. Thank you for your motivation and your trust.

Prof. Dr. Martin Spieß and Prof. Dr. Thomas Augustin who kindly agreed to examine this dissertation.

Dr. Fabian Scheipl and Dr. Bernd Bischl for being members of the examination committee.

I gratefully acknowledge the financial support of Higher Education Commission (HEC) Pakistan and Deutscher Akademischer Austauschdienst (DAAD) Germany throughout the study period.

My heartiest thanks goes to my husband Mr. Mehboob Ali for his dexterous guidance, valuable suggestions, and moral support.

I would like to thank my late parents, my sisters, my brother, and parent’s in-law for their prayers, encouragement, love, and patience. I really want to thank all of those who have raised their hands in front of ALLAH for my success and who have kept thumbs up to say “Good Luck”.

Final thanks goes to my daughter Zarish Mehboob and Saim Mehboob, whose cute smiles gave me the strength to complete this work.

Humera Razzak

Zusammenfassung

Im großen Maßstab angelegte, komplexe Umfragen, wie der in dieser Arbeit verwendete Multiple Indicator Cluster Survey (MICS, http://mics.unicef.org/, http://mics.unicef.org/surveys), hier speziell für Punjab (Pakistan), enthalten in der Regel eine große Anzahl von Variablen, die an einer noch größeren Anzahl von Befragten gemessen werden. Variablen mit verschiedenen Skalenniveaus (zum Beispiel nominale, ordinale oder metrische Variablen) können die Aufgabe der multiplen Imputation (Multiple Imputation, MI) für Umfragedaten erschweren, insbesondere wenn die Anzahl der kategorialen Variablen hoch ist. Diese Arbeit stellt eine allgemeine Methode für die Behandlung des Problems fehlender Daten vor, bei dem vorhandene MI-Methoden kombiniert werden. Die Entwicklung dieses Konzepts wurde durch die Forschungsfrage motiviert, wie Variablen gemischten Skalentyps ersetzt werden können, wenn vorhandene MI-Methoden nur für einen Skalentyp, zum Beispiel metrische Variablen, gut funktionieren und für andere, zum Beispiel hochdimensionale kategoriale Daten, fehlschlagen oder auch von der Rechenzeit her zu aufwendig werden. Wir konzentrieren uns dabei auf die Suche nach flexiblen und rechentechnisch effizienten Imputationsalgorithmen.

In **Beitrag 1** wird eine sogenannte hybride multiple Imputationsmethode (HMI) zur Behandlung fehlender Werte vorgestellt und deren prädiktive Güte untersucht.

Beitrag 2 untersucht die HMI Methode in einer komplexen Simulationsstudie, wobei die Zielvariable binär ist (logistisches generalisiertes lineares Modell (logistisches GLM)). Die Kovariablen können dabei fehlende Werte aufweisen, die Zielvariable wird als vollständig beobachtet angenommen. Umfangreiche Vergleiche der vorgeschlagenen und vorhandenen Methoden für MI werden mit verschiedenen statistischen Kennzahlen (zum Beispiel dem sogenannten root mean squared error, RMSE) durchgeführt. Anschließend wird die Methode auf einen großen Datensatz, MICS 2014 Punjab, angewendet.

Beitrag 3 konzentriert sich auf die statistischen Eigenschaften der HMI-Methode für eine metrische Zielvariable und Kovariablen mit verschiedenen Skalentypen unter Verwendung linearer Modelle (LMs). Umfangreiche Simulationsstudien werden durchgeführt. Die HMI-Methode wird zusätzlich auf einen Teildatensatz von MICS 2014 Punjab angewendet.

Beitrag 4 beschreibt und bewertet die in *R* allgemein verfügbaren Softwarepakete und vergleicht deren Ergebnisse mit denen der vorgeschlagenen HMI-Methode am Beispiel eines künstlichen, simulierten Datensatzes.

Beitrag 5 implementiert eine Erweiterung der vorgeschlagenen HMI-Methode. In diesem Beitrag wird das Konzept der Imputation des Datensatzes basierend auf hilfsweise kategorisierten Versionen der metrischen Variablen angewendet. Die erweiterte HMI-Methode nutzt dabei für die Imputation der kategorialen Variablen auch die Information der metrischen,

kategorisierten Variablen aus. Ein spezielles, sequentielles Verfahren hierzu wird vorgestellt, implementiert und getestet.

Summary

Large scale complex surveys e.g. multiple indicator surveys or MICS typically contain a large number of variables measured on an even larger number of respondents. Mixed type variables (i.e. categorical and continuous) can complicate the task of Multiple Imputation (MI) for survey data, especially if the number of categorical variables is high. This thesis introduces a general framework for the missing data problem by combining existing MI methods. The development of this framework was motivated by the research question how to impute variables of mixed type when existing MI methods perform well for only one type of variables and fail for the other. We focus our attention on seeking several flexible and computational efficient imputation algorithms.

In **contribution 1**, the use of a hybrid multiple imputation (HMI) method to handle missing values in large scale surveys is introduced, highlighting its predictive performance by measuring the accuracy of predictive models.

Contribution 2 evaluates the performance of HMI with repeated sampling simulation studies for generalized linear models (GLM's) with binary response and mixed type missing covariates. Extensive comparisons of the proposed and existing methods for MI are made based on various statistical measures. Complex structured simulation studies are conducted in order to assess the ability of HMI and applied to a large dataset from MICS 2014 Punjab.

Contribution 3 focuses on the statistical properties of the HMI method for continuous response and mixed type covariates using linear models (LM's). The HMI method is applied to a child dataset from MICS 2014 Punjab. Further, extensive simulation studies are performed.

Contribution 4 describes and evaluates software packages commonly available in *R* and compares the results with the proposed HMI method by using an artificial data set as an example.

Contribution 5 implements an extension to the proposed HMI method. In this contribution the concept of imputing the dataset based on the categorized versions of the continuous variables is applied. The extended HMI method utilizes the information on continuous and categorical variables to impute each other through a sequential procedure.

Contents

Contributions

The present cumulative dissertation is composed of the following five contributions. They are arranged according to the order they appear in this thesis.

Contribution 1: Razzak, H. and Heumann, C. (2019a). Predictive performance of a hybrid technique for the multiple imputation of survey data. Proceedings of the *New Techniques and Technologies for Statistics (NTTS) conference*. Brussels: Belgium.

Available under: https://coms.events/ntts2019/data/abstracts/en/abstract_0108.html.

The following technical report that is an extended version of the conference proceedings is included in Chapter 3 of this thesis instead:

Razzak, H. and Heumann, C. (2019b). Predictive performance of a hybrid technique for the multiple imputation of survey data. Department of Statistics (LMU München): *Technical Reports*, Nr. 228, last updated 3. December 2019.
Available under: url: https://doi.org/10.5282/ubm/epub.69897

The whole extended version was drafted and written by the author. Additionally, several rounds of discussing preliminary versions of the paper with co-author lead to improvements of the presentation.

Contribution 2: Razzak, H. and Heumann, C. (2019c): A Hybrid Technique for the Multiple Imputation of Survey Data. Revision under review at the *Journal of Official Statistics*.

As the paper is still under review, the following technical report that is identical to the submitted revision is included in Chapter 3 of this thesis instead:

Razzak, H. and Heumann, C. (2020a): A Hybrid Technique for the Multiple Imputation of Survey Data. Department of Statistics (LMU München): Technical Reports, Nr. 229, last updated 7. January 2020.
Available under: url: https://doi.org/10.5282/ubm/epub.70064

The extended version is revised according to remarks of the two anonymous referees. Additionally, several rounds of discussing preliminary versions of the paper with co-author lead to improvements of the presentation.

Contribution 3: Razzak, H. and Heumann, C. (2019d). Hybrid multiple imputation in a large scale complex survey. *Statistics in transition new series*, 20(4): 33-58.

The original publication is available under: url: https://doi.org/10.21307/stattrans-2019-033

Contribution 3 builds up on results that the author developed for a presentation for "Deutsche Arbeitsgemeinschaft Statistik (DAGStat)" 2019 conference with a different title. This contribution is drafted by the author and revised according to remarks of the co-author as well as review comments of the two anonymous referees.

Contribution 4: Razzak, H. and Heumann, C. (2019e). The Ability of Different Imputation Methods to Capture Complex Dependencies in High Dimensions. Article under review at *Romanian Statistical R*eview, since 29$_{th}$ March 2019.

As the paper is still under review, the following technical report that is identical to the submitted revision is included in Chapter 3 of this thesis instead:

Razzak, H. and Heumann, C. (2020b):The Ability of Different Imputation Methods to Capture Complex Dependencies in High Dimensions (LMU München): *Technical Reports*, Nr. 230, last updated 7. January 2020.
Available under: url: https://doi.org/10.5282/ubm/epub.70080

The whole paper was mainly drafted and written by the author. The co-author contributed to proofreading and improving the paper in several discussion rounds.

Contribution 5:

Razzak, H. and Heumann, C. (2019f): Application of Iterative Hybrid MI Approach to Household Survey Data with Complex Dependency Structures. Article under review at *Communications in Statistics - Simulation and Computation*, since 10$_{th}$ October 2019.

As the paper is still under review, the following technical report that is identical to the submitted revision is included in Chapter 3 of this thesis instead:

Razzak, H. and Heumann, C. (2020c): Application of Iterative Hybrid MI Approach to Household Survey Data with Complex Dependency Structures (LMU München): *Technical Reports*, Nr. 231, last updated 7. January 2020.
Available under: url: https://doi.org/10.5282/ubm/epub.70081

The whole paper was mainly drafted and written by the author. The idea for the concept of categorization of continues variables is developed by Christian Heumann.

List of Figures

List of Tabels

Chapter 1

1 Introduction

1.1 Multiple Indicator Cluster Survey (MICS)

In recent days, all countries around the globe are committed to the advancement of the statistical systems both at national and district levels. These statistical systems or bureaus not only compile and disseminate data but also develop new methodologies to standardize statistical methods, classifications of geographical regions and definitions. Many large scale complex surveys such as the Multiple Indicator Cluster Survey or MICS are conducted to recognize forces that contribute to the public's health factors that interact at individual, family, community, population, and policy levels.

MICS is an international household survey. United Nations Children's Fund (UNICEF) started MICS and assists countries in collecting and analyzing data in order to fill data gaps for monitoring the situation of children and women. MICS is a main source of information on the background characteristics of women, children and households. It contains items on health, demographic, and socioeconomic characteristics. The data is collected at both, family and person levels, and it allows the study of relationships between health and other characteristics. MICS contains information on more than 100 key indicators of the health and well-being of women and children. The MICS program was started in 1995 and since then nearly 300 surveys have been implemented in more than 100 low and middle income countries. Face-to-face interviews with household members are conducted to collect the data. MICS questionnaires have a modular structure and can be adapted for national and sub-national monitoring priorities. Questionnaires are widely used in surveys but other methods like structured, and in-depth interviews, observation and content analysis can also be used in surveys. The questionnaire consists of several items and these items may measure qualitative variables like health, fertility, attitudes toward domestic violence, etc. The data can be analyzed by cross tabulating several variables and results obtained can further be used to understand obstacles to implement new public health programs.

The Bureau of Statistics Punjab has conducted the MICS Punjab 2014 in Pakistan in collaboration with UNICEF. The Government of Punjab has provided the major funding through the Annual Development Program 2014-15. The documents related to MICS Punjab consisting final report, key findings, survey plan, list of indicators and questionnaires can be found on the MICS website (www. http://bos.gop.pk).

1.2 Description of the Datasets

We use the women's, child's and household's datasets from MICS Punjab 2014 household survey to illustrate the techniques developed and compare them with other frequently used methods throughout the thesis as key examples. Brief description of these datasets is provided as:

1.2.1 MICS Women Data

The MICS questionnaire for women of age 15-49 contains information of a woman's background, access to mass media, use of information and communication technology, fertility/birth history, desire for last birth, maternal and newborn health, illness symptoms, contraception, attitudes toward domestic violence, marriage/union, sexual behavior, HIV/AIDS, tobacco and alcohol use, life satisfaction and victimization etc. Most of the background variables are categorical with lots of categories whereas few variables like age are continuous. The MICS Punjab 2014 women data comprises more than 200 background variables on 61286 observations from 36 districts of Punjab. An analysis based on these characteristics can prove to be very helpful in decision making policies regarding women and child health. For example, women data can be used to determine the effect of various factors on feeding practices in the district Punjab (WHO, 2003).

1.2.2 MICS Child Data

Questionnaires for child labor and child discipline in MICS household survey contains information of children of age $1-7$ years living in the household. The questionnaire for children under age five is administered to the mother of the child. It includes information of following dimensions of children's life: family characteristics, care and protection, child health, child development, education, early labor engagement, birth registration, breastfeeding and dietary intake, immunization and attitudes toward children with disabilities. The MICS Punjab 2014 children dataset contains more than 200 child related health background variables on 31083 children. For example: indicators on child mental development (e.g. child follows simple directions, child is able to pick up small object with 2 fingers, child identifies at least ten letters of the alphabet, attends early childhood education program and child is able to do something independently etc.), child's nutrition intake in diet (e.g. child drank or ate vitamin or mineral supplements, child still being breastfed, child ate white potatoes, green leafy vegetables, eggs, cheese, meat etc.), vaccinations (e.g. ever had vaccination card, times child given Polio and BCG vaccinations etc.). MICS data also contains indicators related to child's discipline (e.g. took away privileges, spanked, hit or slapped child on bottom with bare hand, hit child on the bottom or elsewhere with belt, brush, stick, called child dumb, lazy or another name etc.) and participation in household chores (e.g. shopping, repairing equipment, cooking or cleaning, washing clothes and caring for children etc.) Information based indicators described above can prove to be very useful in policy making in order to improve children's health conditions in the district Punjab. For example, stunting, also known as "insufficient longitudinal growth" is one of the forms of chronic malnutrition. Many factors, like socio economic status, imbalanced intake of the nutrients, inequitable distribution of food within the household, vaccination, and infectious diseases are attributed to the main cause of persistent underweight and stunted growth among children in developing countries (Black et al., 2013, McDonald et al., 2013).

1.2.3 MICS Household Data

The MICS household questionnaire is responded by a household head. Household head could be any knowledgeable adult member living in the household. It contains information of the following dimensions of a household head's life: education, household characteristics, water and sanitation, salt iodization, hand washing facilities, water quality testing and results etc. The MICS Punjab 2014 household survey covers 38,405 households to provide estimates of around 125 indicators for the province, 9 divisions and 36 districts. 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.). The household dataset has mixed type variables. 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 further be used for prediction. Indicators described above can prove to be very useful in policy making in order to improve quality of drinking water and sanitation in the district Punjab (Tabassum, 2017, Daud et al., 2017).

1.3 Complications with MICS

Like other cross sectional studies, MICS data is often prone to a lot of missing values and various complications. In the following we give a description of complications with survey data.

1.3.1 Missing Values

The MICS data has a significant amount of missing values. Missing data problems arise when a sampled unit does not respond to the entire survey (also called unit non response (UNR)) or to a particular question (also called item non response (INR)). Variables in the MICS Punjab 2014 women's dataset have between 14 to 95 per cent missing values. Only few variables are completely observed. Respondents feeling shy to answer sexual activity related questions can result in INR. MICS child's dataset also has a significant amount of INR. The missing data rates in the MICS Punjab 2014 child's data range from 10% to 95% and most of the variables have more than 50% missing values. Questions related to 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 the MICS 2014 household data file, only 26819 out of 41413 observations have complete data. The missing data rates range from 7% to 83%. Respondents can be reluctant to answer questions related to income, wealth and marriages which can be the reason for missing values in household surveys.

1.3.2 Inconsistency

Inconsistency in MICS data may happen when one or more sets of exclusive iterative questions are answered in the survey. Rates of inconsistency may vary by respondent's race, education and cognitive ability. For example, in MICS child's dataset, DPT2 and polio vaccination dates differ, the polio 3 vaccination date must not come before the polio 1 vaccination date. Poor recall for the exact number of doses and vaccination of children by illiterate mothers is also subject to inconsistency problems (Gareaballah, Loevinsohn, 1989, Angelillo et al., 1999) Sensitivity or stigma related with the activity being reported is also a main cause of inconsistent responses. Incorrect information on child labor can hinder child education or trends in estimated poverty (Siddiqi and Patrinos, 1995). Inconsistencies in surveys also happen due to edit restrictions i.e. disaggregation totals not adding up; reported numbers far lower or higher than for previous reporting period. Complex consistency errors, however, must be resolved by carefully examining the questionnaire.

1.3.3 Complex Dependencies

Apart from missing values and inconsistency problems, MICS datasets may contain the categorical variables having complex dependencies and various distributional features (i.e. continuous variables having different variances and skewness at different combinations of the categorical variables). The precision of survey estimates is affected if complex dependency structures in the items are not taken into account (Chromy and Abeyasekera, 2005). Complex dependency structures among units may also occur due to the clustered nature of the data. For example, a child feeding index is linked to responses to items on breastfeeding, use of baby bottles, dietary diversity, the number of days the child receives selected food groups in past seven days, and feeding frequency (Ruel and Menon, 2002). In depth qualitative interviews with complex logical structured questionnaires are a main reason for complex dependencies where answers to questions depend on whether previous questions were answered or not. Analysis of data becomes more complex in such situations. Complex relationships linked to large households, multiple generation households or two or more households increase the number of potential predictors for each imputation model. More information on complex household survey designs can be found in Binder (1983). However, we mainly focus on the missing data problem and inconsistency and complex dependency problems are not addressed in this thesis.

1.4 Statement of the Problem

Presence of complex dependency structures in large scale complex surveys can make estimates biased (Bishop et al., 1975). Traditional methods to deal missing values can fail to detect complex dependencies structure among categorical variables. For example, implementation of conditional specification models become challenging when incompatibility issue arises due to high dimensions in large scale complex data (White et al., 2011). Other techniques to handle missing values are limited to categorical variables or require transformations (or other tricks) for continuous variables or require knowledge of complicated models to create dependence between the continuous and the high dimensional categorical variables (Murray and Reiter, 2016). Multiple imputation (MI) was originally introduced to handle nonresponse in public use data files or shared databases (Rubin, 1987). Despite of its popularity and acceptance it is applied to a handful of variables in many studies (e.g. NNS, 2011). Applications of MI in large scale studies with complicated datasets such as MICS are very few. These limitations create serious problems for researchers to obtain complete datasets with mixed type variables. The objectives of this study are:

1. To develop methods for imputing mixed type data from large scale complex surveys.
2. To avoid difficulties of complicated models in high dimensions.
3. To combine existing techniques to handle incomplete large scale complex datasets.
4. To gain computational efficiency.

Chapter 2

2 Fundamentals of Missing Data and Multiple Imputation (MI)

2.1 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). In MCAR, the probability of missing data on a variable is not correlated to its self and to the other measured variables. In MAR probability of missing depends on other, observed, variables. Finally, data are MNAR if the probability of missing depends on the variable value itself. MCAR and MAR are 'ignorable' because we don't have to include any information about the missing data itself when we deal with the missing data. MAR or MNAR results in the loss of power due to missing information and the possibility of a biased estimate. Practically all methods implemented in software assume MAR. MNAR is called "non-ignorable" because the missing data mechanism itself has to be modeled as you deal with the missing data. Exact missing data mechanisms are often unknown when dealing with large scale datasets therefore, most of the time certain assumptions are made accordingly. Li et al. (2013) address some problems with the MI in large data. Little's MCAR test is used commonly for testing missing cases being MCAR. Notations and assumptions for the missing mechanisms are given as:

Let Y denote the $n \times p$ 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 dataset $Y = \{Y^{miss}, Y^{obs}\}$ where the first component denotes the observed part of the data and the second component is the missing data. Let H be a response indicator matrix with the same dimensions as Y indicating, if an element of Y is observed or missing:

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

Data is MCAR when $Pr(H|Y^{miss},Y^{obs})$=$Pr(H)$, MAR when $Pr(H|Y^{miss},Y^{obs})$=$Pr(H|Y^{obs})$ and MNAR when $Pr(H|Y^{miss},Y^{obs}) \neq Pr(H|Y^{obs})$ (Little and Rubin, 2002). We treat item non response as MAR throughout the thesis.

2.2 Problems Associated with Missing Data

In population surveys, respondents 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. Problems associated with missing data are:

1. Systematic nonresponse may make sample non representative. For instance, evidence exists that in sample surveys, the sample results may over represent the middle incomes due to non-response behavior of individuals with low-income or high-income when they are asked to fill in their incomes. Questions related to income or wealth are often related to high rate of INR (e.g. Riphahn and Serfling, 2005, Hawkes and Plewis, 2006). High rate of INR occurs for simple demographic variables such as age, sex or marital status. According to various studies (Colsher and Wallace, 1989, Dillman, 1978, Herzog and Rodgers, 1992), high INR is commonly observed for responded with less education and elder age. Missing values may also occur due to interview errors i.e. a variable that should have a response, but the question was not asked.

2. Information about the parameters of interest is less in incomplete dataset as compared to the hypothetical complete dataset. Analysis of incomplete dataset may result in larger standard errors, wider confidence intervals and less significant p-values consequently, resulting in loss in statistical power and making conclusions less powerful.

3. Missing values can make analysis more complicated and can reduce the efficiency of statistical analysis, for instance it is complicated to directly apply logistic regression when covariates are incompletely observed.

2.3 Simple Methods for Missing Data

Most popular approaches for INR include list wise deletion (LD), available case (AC) analysis, maximum likelihood of the incomplete data (MLID) approach and imputation. In LD method, complete response patterns provided by a responded are included in statistical analysis by discarding the cases which are not completely observed. Loss of power is one of the main downside of LD. Moreover, LD can make results biased unless strong assumption about the mechanism that caused the missingness are met. In order to get a satisfactory solution for loss of power problem, AC analysis can be used. AC also requires same strong assumption about mechanism working behind missingness. AC analysis selects all respondents who provided complete information on the variables that are used in analysis. The amount of data used by AC analysis is more as compared to the LD method. MLID is yet another approach to handle missing data. Maximum likelihood estimates of the parameters of the desired statistical model are estimated in presence of missing data (e.g. Rasch, 1960, Birnbaum, 1968, Masters, 1982, Samejima, 1969). Imputation is a gold standard for fill in the blanks in incomplete data. According to the definition of Cambridge dictionary *"imputation is a way of calculating something when you do not have the full or correct data"*. In imputation method, a statistical model is used to estimate missing values. There are several imputation techniques describe by Little and Rubin (2002). In mean imputation values of mean, median or mode are imputed for metrical, ordinal or nominal variables respectively. Hot deck imputation is a nonparametric technique to imputation. A similar but observed unit, whose value serves as a donor for the record of the similar but incompletely observed unit is specified in hot deck methods in order to impute missing values. Various techniques are used to measure similarity e.g. distance measures etc. The predictive mean matching (PMM) (Little, 1988) is one of the popular known methods of k-nearest neighbor (kNN) algorithms for generating hot-deck imputations. The nearest neighbor donor distance based on expected values of the missing variables which are conditional on the observed covariates is employed to impute missing values. The main advantages of kNN imputation are that it is a simple method without strong parametric assumptions and various types of variables can be imputed by applying it easily (e.g. Andridge and Little, 2010, Little, 1988, Schenker and Taylor, 1996). Harrell (2015) proposes “aregImpute” algorithm which combines PMM with the various aspects of model-based imputation methods in the form of flexible nonparametric models. Another imputation approach is called cold deck imputation. It uses external data to generate substitute values. Regression imputation predicts the missing values by fitting a regression model of the variables with missing values on the other variables where these variables are observed. Stochastic regression imputation is one step higher than regression imputation. It adds a random noise to the predictions from the regression imputation. Amelia (Honaker et al., 2011) is a modern method which uses explicit or implicit linear imputation models. The dependent variable used in a homoscedastic linear model with incompletely observed metric predictors has the property of conditional normality. However, the argument to apply linear imputation model on the incompletely observed variable requires more justification. Thus, assumed linear imputation models would be incompatible with the true data generating process in general. Moreover, the transformation of variables to assume multivariate normality (e.g. Honaker et al., 2011, Schafer, 1997) does not seems to work well in general and can led to biases in the estimators (Hippel, 2013). Approaches for missing values described above are simple, reasonable in case of a small fraction of incomplete cases and are commonly implemented in existing statistical softwares for complete data. However, these methods have serious disadvantages when fraction of missing cases gets larger. Software packages used for imputations are “mi” (Su et al., 2011) in *R (R* Core Team, 2018), “ICE” (Royston, 2004) in STATA (Stata Corporation, 2013) and “IVEware” (Raghunathan et al., 2002) in SAS (SAS Institute, 2014) implement imputations. Salfran and Spiess (2015) provide details of these imputation techniques. We take R under consideration throughout the thesis due to its open source character and popularity.

2.4 Multiple Imputation

Multiple imputation (MI) (Rubin, 1987) replaces missing values in a dataset by drawing random values from the predictive posterior distribution of the missing data given the observed data. MI creates *M* complete datasets. Inference of interest (e.g. mean, regression) can be run on each newly created imputed dataset. Estimates can be combined by using "Rubin's rules". Final estimates obtained are unbiased on the average.

2.4.1 Fully Conditional Specification Model

Algorithm 1: MICE (FCS)

```
1: Fill in missing data $Y^{mis}$ bootstrapping the observed data $Y^{obs}$
2: For $j = 1, . . . , p$
    a. Draw $\theta_j^*$ , from the posterior distribution of the imputation parameters.
    b. Impute $Y_j^*$ from the conditional model $f_j(Y_j | Y_{-j}, \theta_j^*)$
3: Repeat step 2 until convergence
```

Fully conditional specification (FCS) model is a general approach to MI. FCS specifies univariate conditional distributions on a variable-by-variable basis, and draws missing values iteratively from the specified conditional distributions. MI by chained equations (MICE) (Raghunathan et al., 2001, van Buuren and Groothuis-Oudshoorn, 2011, Royston and White, 2011, Su et al., 2011) is a fully conditional specification (FCS) approach to MI. The FCS uses following iterative algorithm. For each incomplete variable a density, $f_j(Y_j | Y_{-j}, \theta)$, conditional on all other variables is specified, where θ is the unknown parameters of the imputation model. A conditionally specified imputation model known as MICE, visits sequentially each incomplete variable and draws alternately the imputation parameters and the imputed values. The FCS method is summarized in algorithm 1. The researcher can choose a suitable regression model for each variable for example classification and regression trees (CART) (Breiman et al., 1984) for categorical variables, PMM for continuous variables or default method which uses logistic regression models for categorical and PMM for continuous variables. Sometimes problems of convergence and incompatibility arises when MICE is used for specifying univariate conditional distributions (Arnold and Press, 1989, Gelman and

Speed, 1993). Moreover, “regression imputations” is very time consuming. *R* package “mice” (van Buuren and Groothuis-Oudshoorn, 2011) implements MICE.

2.4.2 Joint Model

Algorithm 2: Joint Modeling Gibbs Sampler[1]

```
1: Fill in missing data $Y^{mis}$ bootstrapping the observed data $Y^{obs}$
2: Estimate $\overline{Y}$ and $S$
3: Draw $\mu$ and $\Sigma^{-1}$ from equations (2) and (3)
4: Draw $Y_*^{mis} \sim N(\mu_*, \Sigma_*)$
5: Update the estimation of $\overline{Y}$ and $S$
6: Repeat steps 3 to 5 a large number of times to allow the sampler to reach its
   stationary distribution.
```

The joint modeling (JM) specification is another approach used for MI. Joint modeling (JM) draws missing values simultaneously for all incomplete variables. JM involves specifying a multivariate distribution for the missing data and draws imputations from their conditional distributions by Markov Chain Monte Carlo (MCMC) methods (Schafer, 1997). For simplicity, let’s assume that

$$Y \sim N(\mu, \Sigma), \tag{1}$$

where $\mu = (\mu_1, \ldots, \ \mu_p)$ and Σ a $p \times p$ covariance matrix. The posterior distribution of (μ, Σ) given Y (where Y is fully observed) with a prior distribution for μ and a prior $W_p(v, S_p)$ for Σ^{-1}could be written as the product of

$$\mu | Y, \Sigma \sim N(\overline{Y}, n^{-1}\Sigma) \tag{2}$$

and

[1] The “*” symbol in algorithm 2 denotes that the variable or parameter randomly drawn from a posterior conditional distribution

$$\Sigma^{-1}|Y \sim W_{p(S_p^{-1}+S)^{-1}}(n+v,(S_p^{-1}+S)^{-1}) \tag{3}$$

where $\overline{Y}$ and $(n-1)^{-1}S$ are the sample mean and covariance matrix respectively (Carpenter and Kenward, 2013). Above method can be generalized for incomplete Y method. See algorithm 2 for summary. 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, 2013). Modeling mixed type variables can make the specification of a joint distribution very difficult. *R* packages "pan" (Schafer and Zhao, 2014) and "jomo" (Quartagno and Carpenter, 2014) implement JM approach.

2.4.3 Dirichlet Process Mixture of Products of Multinomial Distributions Model (DPMPM)

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 the DPMPM for the first time. This approach uses nonparametric Bayesian versions of latent class models (LCM) to multiply impute high-dimensional categorical data (Vermunt et al., 2008).This approach automatically models complex dependencies whereas other MI methods (log linear model or logistic regressions) can fail to detect complex structures in high dimensional categorical variables. Before describing the DPMPM, few notations related to LCM in the context of categorical data are as follow. Let Y represent the data of *n* individuals with *p* categorical variables that is, $Y = (Y_1, \dots, Y_p)$ subject to INR. An associated response vector for each individual *i* can be defined as $Y_i = (Y_{i1}, \dots, Y_{ip})$. Elements of vector Y_i can take on a set of L_j levels such as each $Y_{ij} \epsilon \{1, \dots, L_j\}$, thus $Y_i \epsilon C = \Pi_{j=1}^{p}\{1, \dots, L_j\}$. Missing and observed parts of Y_j can be presented as Y_j^{mis} and Y_j^{obs} respectively, so that $Y^{obs} = \{Y_1^{obs}, \dots, Y_p^{obs}\}$ and $Y^{mis} = \{Y_1^{mis}, \dots, Y_p^{mis}\}$ be the observed and missing parts in Y, respectively. LCM without any missing data is a finite mixture of product-multinomial distributions,

$$\mathrm{p}(\mathrm{Y}|\lambda,\pi) = f^{LCM}(\mathrm{Y}|\lambda,\pi) = \Sigma_{k=1}^{K}\pi_k \; \Pi_{j=1}^{p}\lambda_{jk}[Y_j], \tag{4}$$

$$\text{where } \lambda = \lambda_{jk}[l], \text{ all } \lambda_{jk}[l] > 0 \text{ and } \Sigma_{l=1}^{L_j}\lambda_{jk}[l] = 1. \tag{5}$$

Here, $\pi \ = \{\pi_1, \dots, \pi_k\}$, where $\Sigma_{k=1}^{K}\pi_k = 1$. This model can be use to generate data using

$$Y_{ij}|z_i \overset{ind}{\sim} Discrete_1: L_j(\lambda_{jz_i}[1], \dots, \lambda_{jz_i}[L_j]) \text{ for all } i \text{ and } j, \tag{6}$$

$$z_i|\theta \overset{iid}{\sim} Discrete_{1:K} \{\theta_1, \dots, \theta_K\} \text{ for all } i. \tag{7}$$

For prior distributions[2], Si and Reiter (2013) and Manrique-Vallier and Reiter (2012) use

$$\lambda_{kj}[.] \sim Dirichlet \ (1L_j)\ , \tag{8}$$

$$\pi_k = V_k \ (\textstyle\prod_{h<k} 1 - V_h)\ , \tag{9}$$

$$V_k \overset{iid}{\sim} Beta\ (1,\ \alpha) \text{ for } k=1,\dots,K\text{-}1,\ V_k=1, \tag{10}$$

$$\alpha \sim Gamma \ (a_\alpha,\ b_\alpha). \tag{11}$$

In order to get complete data sets, first the latent class indicator for each individual is drawn from the full conditional and then, second, each missing Y_i is drawn from class-specific, independent categorical distributions. Like many complex models, the effectiveness of DPMPM still lag in capturing the many features of empirical data. The DPMPM imputation routines are implemented in the *R* software package "NPBayesImputeCat" (Quanli et al., 2018).

2.4.4 Combining rules

In order to incorporate the uncertainty introduced by missing data and imputation into the inferences, the estimates for quantities of interest obtained by analyzing each completed dataset are combined by utilizing rules proposed by Rubin (1987). For instance, let Q be any quantity of interest (a population proportion or probability). For $m = 1,\dots,M,$ let $q^{(m)}$ and $u^{(m)}$ be respectively the point estimates of, Q and variance estimates of $q^{(M)}$. Valid inferences for scalar Q by combining the $q^{(m)}$ and $u^{(m)}$, by Rubin (1987) are as follows:

$$\overline{q} \ = \textstyle\sum_{m=1}^{M} \frac{q^{(m)}}{M}\ , \tag{12}$$

$$b = \textstyle\sum_{m=1}^{M} \frac{(q^{(m)} - \overline{q})^2}{M-1}\ , \tag{13}$$

[2] We examined different vague prior specifications for $a_\alpha,\ b_\alpha$ in contributions.

$$\overline{u} = \sum_{m=1}^{M} \frac{u^{(m)}}{M}, \tag{14}$$

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

$$T = \left(1 + \frac{1}{M}\right) b + \overline{u} \tag{15}$$

with degrees of freedom $v = (M - 1)(1 + r^{-1})^2$, (16)

where

$$r = \frac{(1+M^{-1})b}{\overline{u}} \tag{17}$$

represents the relative increase in the conditional variance due to the missing data (see Rubin, 1987). Confidence intervals can be constructed using standard multiple imputation confidence interval construction rules, possibly based on a t-distribution. For more details see Rubin (1996), Barnard and Meng (1999), Reiter and Raghunathan (2006), Harel and Zhou (2006).

2.4.5 Number of Imputations

It has been shown that 2 to 5 multiple imputations are usually sufficient to yield excellent results (Carpenter and Kenward, 2013, van Buuren, 2012). However, there is no general consensus about for which situation this is an appropriate number for imputations, because over time, more and more examples occurred where that proved to be problematic. Various factors, i.e. the number of observations and missing values, the exact patterns in the missingness, the extent of complications in the imputation and the substantive model, play an important role in the determination of the exact number for imputations. Previous guidelines suggest the use of relative efficiency for determining a sufficient *M*. The relative efficiency of the *M* imputations given a fraction of missing data (Rubin, 1987) is computed in standard error units as

$$RE=\left(1+\frac{\lambda}{M}\right)^{-1/2} \tag{18}$$

where

$$\lambda = \left(\frac{r+2/(v+3)}{r+1}\right) \tag{19}$$

is the estimated fraction of missing information, with v and r_m given by equations (16) and (17) (Rubin, 1987). For instance, a relative efficiency of 0.92 can be obtained if λ = 0.9 and *M* is set to 5 imputations. Consequently, more than five imputations are rarely required, and 10 imputations are more than suitable in almost any realistic application. According to van Buuren (2012), a small number of imputations may be created in the beginning when building the imputation model with an exploratory analysis, and increase *M* gradually for the final analysis.

2.5 Imputation Methods for Large Scale Complex Survey

A complete overview of state of the art MI methods for accommodating non-linear relationships and best ways to impute categorical and non-normal continuous variables is available in Vermunt et al. (2008), Yucel et al. (2011), Seaman et al. (2012) and Lee and Carlin (2010). 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 multiple imputation of high dimensional data (Marker et al., 2002, Andridge and Little, 2010, Zhu and Eisele, 2013, Audigier et al., 2018) but most of the existing methods are not designed to handle mixed data (quantitative and categorical) and become difficult to implement with large dimensions and are extremely time consuming (Erosheva et al., 2002). Moreover, the presence of complex dependency structures can also make estimates biased (Bishop et al., 1975). Random forest imputation is yet another method for handling missing data (Stekhoven and Bühlmann, 2012). Random forest imputation is a machine learning technique for nonlinearity and interaction problems and does not require a particular regression model to be specified. Shah and Anoop (2014) use random forest imputation for imputing complex epidemiologic datasets. They find 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 focus on the setting where few variables have missing values. One disadvantage of the 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. (2015) implement CART and forests to overcome incomplete data problems when the auxiliary variables are numerous. A study shows that CART and forest methods are more reliable than likelihood methods for MI but CART can be biased towards selecting variables that allow more splits (Loh and Shih, 1997, Kim and Loh, 2001). A 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 behaviors 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 the CART algorithms computationally inefficient for standard computers. Joint distributions of the missing covariates are also specified by parametric, non-parametric and semi parametric models. Akande et al. (2017) compare the performance of various default MI methods for categorical data. According to their study, the Bayesian mixture model approach dominates the application of the chained equations approach based on Generalized Linear Models (GLM's) (Nelder and Wedderburn, 1972) and is as reliable as imputation engines based on CART. They also found that the Bayesian joint modeling approach is substantially computationally expedient as compared to the FCS methods for MI. However, in the presence of a large number of categorical and continuous variables, the sequential behavior 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 modeling approach as suggested by Akande et al. (2017), one has to either discard all continuous variables or to categorize them. Murray and Reiter (2016) extend the Bayesian, joint modeling approach for multivariate continuous and categorical variables. 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 model for a mixture of continuous and categorical variables. Despite of being superior to CART in many ways, He (2009) suggests that the JM approaches can lack the flexibility needed to represent complex data structures arising in various 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) are proposed to overcome the problem of ignoring interactions in chained equations but most of the proposed methods combines recursive partitioning with single imputation instead of multiple imputation. A multilevel singular value decomposition (SVD) approach to missing values 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) address main effects and interactions challenges in mixed and incomplete data frames. MI by multiple correspondence analysis (MIMCA) (Audigier et.al, 2017) utilizes the dimensionality reduction property of multiple correspondence analysis to impute categorical data with a high number of categories. Estimates obtained by MIMCA are reliable as MI methods using log linear models or conditional logistic regressions. MIMCA is less time consuming on datasets of 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). Iterative singular value decomposition (SVD) algorithms for MI can be a good choice for quantitative (Hastie et al., 2015), qualitative (Audigier et al., 2017) and mixed data (Audigier et al., 2016) because of better performance than their counter parts. De Jong, van Buuren, and Spiess (2016) propose a new method based on generalized additive models for location scale, and shape (GAMLSS) which uses spline functions to fit a nonparametric regression model. The individual conditional distribution of the variables with missing values is specified using these functions. The conditional distribution can be further used in the framework of chained equations. However, simulation studies for the GAMLSS imputation method were limited to missing values in only one covariate and the variables were all independent and normally distributed. Salfran (2018) extends the GAMLSS imputation method to the multivariate case by relaxing the distributional assumption of the error and requiring weak distributional assumptions. Extensive empirical comparisons of the GAMLSS approaches with available modern techniques in the context of complex datasets show that the extended method allows valid inference. However, applications of GAMLSS-based methods are limited and they did not perform well with non-monotone missing patterns. Moreover, GAMLSS-based methods are very time consuming as compared to the available standard methods. Further research is required in high dimensions.

2.6 Hybrid Techniques for Imputations

Recently, hybrid techniques for imputations have gained a lot of attention (Ankaiah and Ravi, 2011, Tang et.al, 2015, Shukur and Lee, 2015). For example, Ankaiah and Ravi (2011) propose a hybrid two stage imputation method involving K-means algorithm and multi-layer perceptron (MLP) in stage 1 and stage 2, respectively. Also, Nishanth et al. (2012) propose a hybrid clustering and model based method where k-means are combined with an artificial neural network (ANN). Nishanth and Ravi (2013) propose the online data imputation framework incorporating data mining techniques. Considering the local similarity of data, Li et al. (2013) borrow 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 propose a novel missing value imputation method using stacked auto-encoder and incremental clustering (SAIC). However, obtaining 100% correct clustering results may become challenging due to the expansion of the data volume with existing clustering algorithms. MI 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 variables 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 of nested imputation. These methods explicitly manage two MI procedures in a dependent structure (Rubin, 2003). Weirich (2014) extends the nested imputation methods in both continuous and categorical background variables for large-scale assessment. However, these procedures are computationally more extensive and implemented in limited ways and require further research. Zhao and Long (2016) have done 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) propose a new hybrid imputation method that deals with the missing data issue of 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.

2.7 New Approach to the Imputation of Large Scale Survey Data

Algorithm 3: Hybrid MI

```
Require: $P$ $nxp$ matrix with incomplete data
Miss.cat , Miss.num ← Initial division of $p$ variables into factor and numeric subsets
    for z= $1, \ldots, Z$ do
        for $m= 1, \ldots, M$ do
            Imp.$P^{z}_{cat_m}$← Imputation using fully Bayesian joint modelling MI
            Imp.$P^{z}_{cat_m}$ $Miss.^{z}_{num_m}$← Combining Imp.$\mathrm{P}^{z}_{cat_m}$and Miss.$^{z}_{num_m}$ to generate partially imputed dataset
            $Imp^{z}_{m}$← Imputing Imp.$P^{z}_{cat_m}$ $Miss.^{z}_{num_m}$using MICE i.e. $f(\,Miss.^{z}_{num_m} \mid Imp.P^{z}_{cat_m})$
            $Imp^{z}_{m}$ ← Final imputed dataset
        end for
    end for
```

This thesis deals exclusively with the development of new methods for the imputation of mixed data type in complex surveys. All contributions are based on the assumption of fully conditional specification models for incomplete continuous variables dependent on complete categorical variables obtained by fully Bayesian, non-parametric joint models. Although it can be argued that this assumption for imputation is unjustified, there certainly exist situations in high dimensions which are enough complex that existing methods for imputation are difficult to implement seperately. As already mentioned by van Buuren and Groothuis-Oudshoorn (2011), "fitting a series of conditional distributions, as is done using a series of regression models, may not be consistent with a proper joint distribution". Also motioned by Speidel et al. (2018), that "the specification of a joint distribution can be difficult, if different variable types need to be modeled". Hybrid MI (HMI) methods are able to combine conditional and joint models to impute mixed type variables. In addition, already present knowledge about complete categorical variables can be directly included to impute continuous variables. The same holds for categorical variables when categorized continuous variables are used in dependence models with or without initial values. Another advantage of hybrid models is that they are computational efficient. Different dependence models using a categorization approach are presented in the last contribution whereas, in the first

four contributions, dependence models that only use information of categorical variables are implemented with a variety of settings. The proposed hybrid MI (HMI) approach is a 3-stage approach. In step 1, imputations for a large number of categorical variables are created under the JM MI techniques. Incomplete continuous variables are combined with complete categorical variables in step 2, resulting in a dataset where values in the continuous variables may be missing and values in the categorical variables are imputed. The continuous variables in each dataset are then imputed using FCS MI techniques, such that the means of the draws from the posterior predictive distribution of the unobserved data depend on the data already imputed by the JM MI in step 3. Steps 1-3 are repeated *M* times to generate multiple completed versions of the data. Therefore we provide a flexible and practical hybrid MI approach to obtain complete data, which sometimes cannot be possible to be obtained when both MI approaches are applied separately. Step 1 is implemented by using the DPMPM MI technique due to its computational efficiency, its ability to automatically model complex dependencies and its successful implementation for the case of high dimensional categorical variables (Chib and Hamilton, 2002, Hirano, 2002, Kyung and Gill, and Casella, 2010). Step 3 is implemented by using MICE due to its open source character and popularity. Algorithm 3[3] provides an outline of how the simulations are run for the HMI method. HMI is denoted as H.CART, H.PMM and H.DEF when the MICE algorithms CART, PMM and default (which uses logistic regression for categorical and PMM for continuous variables) are used to impute the continuous variables in step 3, respectively.

[3] The various experimental conditions can be controlled according to steps 1 to 7 of the algorithm. These distinct settings will be discussed in the next sections.

3 Contributions

3.1 Contribution 1

In this Section an overview of the theoretical background of those concepts are presented, which are essential for the understanding of the contributions[1]. At first, in Section 3.1.1 an introduction into generalized linear regression (GLM's) is given. To check the capacity of a hybrid and common MI approaches for predictive performance in GLM's, the experimental conditions after step 7 of algorithm 3 are further defined in Section 3.1.2. We take a close look at various dimensions of proposed approach and highlight its predictive performance in simulation study and a real data example in Sections 3.1.3 and 3.1.4 respectively. The Section ends with an outlook.

3.1.1 Generalized Linear Models

The theory of generalized linear models (GLM's) was first introduced by Nelder and Wedderburn in 1972. They proposed that interdependencies and causalities between the dependent (response) variable $Y = (y_1, . . . , y_n)$ and $p \geq 1$ independent variables (covariates) $X = (X_1, . . . , X_p)$ can be analysed by an entire class of regression models where the response variable of the model is hypothesized to follow exponential family of distributions e.g. (Gaussian, binomial, poisson, gamma, inverse Gaussian, geometric, and negative binomial). Many types of response variables e.g. count, binary, proportions and positive valued continuous distributions can be accommodated by GLM's (Nelder and Wedderburn (1972) and Hoffmann (2004). Nelder and Wedderburn (1972) discovered that the assumptions of linear models can be relaxed to develop general models in order to the specify relationship between response variable and some number of covariates and relationship which initially seems to be nonlinear can be linearized by restructuring the relationship between the linear predictor and the fit. This flexibility makes GLM's a valuable statistical tool and is widely implemented in softwares since past twenty years (Hoffmann, 2004).

[1] The following sections are written in a way to aid the reader in understanding the contributions, but are by no means exhaustive or even complete with respect to the theory described.

3.1 Contribution 1

GLM's use a random sample of *n* observations to make inference on the whole population under investigation. The observations on Y are normally distributed with constant variance σ^2. Parameters of interest and the variance are estimated from sample by solving a linear equation system using ordinary least square (OLS) method. Properties of OLS include independence and constant variance whereas maximum likelihood (ML) linear regression has a more restrictive distributional assumption of normality. The response is linked to a linear combination of covariates in classical setting of ordinary linear regression and a random error term $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ is added to the model to observe the deviation of the each observation from global sum. The main assumptions are independent, identical and normally distributed error terms, expectation $E(\varepsilon_i) = 0$ and variance $Var(\varepsilon_i) = \sigma^2$, and error terms are uncorrelated to the covariates. Intercept β_0 and weights $\beta_1, \ldots, \beta_p$ are the parameters of interest in the model. A GLM consists of:

1. A linear predictor

$$\gamma_i = \beta 0 + \sum_{j=1}^{p} x_{ij} \beta_j , i = \{1, \ldots, n\}.$$

2. A smooth linearizing response function[2] (e.g. w). The link function linearizes the GLM's by transforming the expectation of the response variable, $\mu_i = E(y_i)$ to the linear predictor:

$$w(\mu_i) = \gamma_i .$$

3. A variance function that describes how the variance, $var(y_i)$ depends on the mean

$$var(y_i) = \varphi \mathrm{V}(\mu),$$

where the dispersion parameter φ is a constant. More flexible modelling of the linear predictor can be made using an approach called generalized estimation equations (GEE) and several other models branched off from the stems of basic ideas of GLM's. A sum of unknown functions (to be estimated) of the covariates can be used as a linear predictor in nonparametric regression. A (penalised) spline-based approach is used to estimate these functions. This approach transforms the linear predictor back to the estimation of a generalized linear regression. Some covariates can be modelled in the predictor using aspects of functions whereas others may use aspects of linear combinations. Generalized additive regression (e.g. Hastie and Tibshirani, 1999) is another approach which combines these aspects. More flexible modelling of the linear predictor can be

[2] The inverse function w^{-1} is called link function

formed using these approaches. These approaches are beyond the scope of this thesis since only standard generalized linear regression models are discussed in contributions.

GLM's approach has some key assumptions which need to be met when computing a p-value. Violation of these assumptions can produce biased standard errors and can make p-values unreliable. However, key assumptions for linear modeling are not properly specified. For instance, the chi-square distribution assumes homogeneity, normality, and independent deviations centred on zero to calculate the type I error (the p-value) on the improvement in fit with the GLM's (Dobson, 2002). In this scenario, it follows that these properties are considered as key assumptions for GLM's. A general consensus is developed on the assumptions of homogeneity and independence of residuals (see e.g. Nelder, Wedderburn, 1972, Hoffmann, 2004, Dobson, 2002, Breslow, 1996, McCullagh, Nelder, 1989). On the other hand the importance of normality of residuals in GLM's is not clearly specified. According to Hoffmann (2004) and Dobson (2002) the assumption of normality of the residuals is important in order to correctly interpret the results while Gill (2001) noted that normally distributed errors are simply a description of model behaviour not a condition of GLM's quality.

3.1.2 Predictive performance for HMI

Algorithm 4: Holdout cross validation and estimation of AUROC for HMI method

Require: Imp_m^z i.e. M complete datasets over Z simulations runs

```
for $z = 1, \dots, Z$ do
    for $m = 1, \dots, M$ do
        $Imp_{testing_m}^z, Imp_{training_m}^z$ ← Divide matrix $Imp_m^z$ into testing and training subsets[3]
        $P(y = 1|\ x_{1,\dots,}x_p)_m^z = 1/(1 + e^{-(a+\sum_{j=1}^{p}(b\,{}^{z}_{jm}x_{jm}^{z})}$ ) ← Train a GLM model on $Imp_{training_m}^z$
        $P_m^z$ ← Make prediction on $Imp_{testing_m}^z$
        $AUROC_m^z$ ← AUROC curve based on $P_m^z$
    end for
    $\overline{AUROC}^z = \frac{\sum_{m=1}^{M}(AUC_m^z)}{M}$ ←Pooled[4] AUROC curve
end for
```

[3] The test and train datasets are generated randomly using unequal split.

[4] The arithmetic mean is taken of M AUROC estimates obtained by M fitted GLM's.

3.1 Contribution 1

Algorithm 4 explains the experimental conditions for obtaining the predictive performance for HMI proposed in Chapter 2. Holdout cross validation is used to access the predictive performance of GLM's for binary response. We focus on the area under the Receiver Operating Characteristic (ROC) curve or AUROC as diagnostic check to evaluate performance of the new method.

3.1.3 Simulation Study

A small simulation study is conducted with five covariates generated from a Multivariate Normal (MVN) distribution. Further, three out of five covariates are binariazed using some thresh hold criterias. Lastly, covariates dependent binary response is generated with probabilities governed by logistic model. Artificial data is generated under two scenarios i.e. variables having moderate to somewhat high correlations. Logistic probabilities satisfying Rubin's definition for MAR missngness are defined to yield missing observations in covariates and binary response. *M* complete datasets are made for proposed and existing MI methods. A total of 200 simulations are made for each method. We use GLM's[5] because effect of various factors on a binary response (breastfeeding) is analysed later in a real data example. Various numbers of imputations are generated using three MICE methods (i.e. CART, PMM, default) and two HMI methods (H.CART, H.PMM) in both types of scenarios. The AUROC values for Cross validated complete datasets are used as benchmark ("theoretical" AUROC) for comparison. Computational times are also recorded as a measure of performance to compare different MI methods. We noticed, for highly correlated data, H.PMM and H.CART showed better predictive performance in terms of larger median values of pooled AUROC over all simulations as compared to the default and PMM methods in simulation studies.

3.1.4 Real Data Application

In order to confirm potential of proposed approach we apply hybrid approach to create multiple-imputed background characteristics for individual women in 2014 Punjab MICS data. A subset of background characteristics of women is selected because implication of existing MI approaches may become problematic for large number of categorical variables due too high missing rates, inconsistencies and complex dependency structures. Women's background characteristics like demographics, age, education, motherhood and recent births are included in this dataset. The

[5] We used the GLM's with link "logit" throughout all contributions.

number of categorical variables is high as compared to continuous variables. Fifty sampling simulations are run and *M=5* completed datasets are generated for each MI method. The binary response is modeled using GLM's depending on various categorical and continuous covariates. The AUROC is pooled for each MI method after cross validation over all simulation runs. Results suggest overall better predictive performance of the GLM's for the two hybrid methods as compared to CART and PMM methods. Surprisingly, the computational time taken by MICE methods is reduced from days to hours when the proposed methods are applied.

3.1.5 Outlook

The source of low rates of AUC for hybrid methods as compared to CART in simulation studies is still unknown. Further research for complex simulation studies, large-sample results or large number of imputations could be needed to find an answer.

3.2 Contribution 2

This contribution links with the previous one in the way that it is a review of inference in GLM’s with binary response and mixed type missing covariates for the presented and existing methods. Repeated sampling properties for the imputation techniques are highlighted in this Section. More complex data structures are generated in simulation studies for the evaluation of various statistical properties for GLM’s. A brief description of the simulation study is provided and the major findings of contribution 3 are discussed. The section ends with an outlook.

3.2.1 Inference on GLM’s for HMI method

Algorithm 5 explains the experimental conditions for the hybrid approach described in Chapter 2 for inference in GLM’s.

```
Algorithm 5:  Inference on GLM’s for HMI method
Require: $Imp_m^z$ i.e. $M$ complete datasets over Z simulations runs
    for $z= 1, \dots ,Z$ do
        for $m= 1, \dots,M$ do
    $\bar{q}^{(z)} \leftarrow \sum_{m=1}^{M} \frac{q^{(m)}}{M}$        Pooled point estimates[1].
    $b^{(z)} \leftarrow \sum_{m=1}^{M} \frac{(q^{(m)} - \bar{q}^{(z)})^2}{M-1}$
    $\bar{u}^{(z)} \leftarrow \sum_{m=1}^{M} \frac{u^{(m)}}{M}$
    $T^{(z)} \leftarrow \left(1 + \frac{1}{M}\right) b^{(z)} + \bar{u}^{(z)}$      Pooled variances[2].
        end for
    $\bar{q} \leftarrow \sum_{z=1}^{Z} \frac{\bar{q}^{(z)}}{Z}$        Average of pooled point estimate[3].
    $\bar{T} \leftarrow \sum_{z=1}^{Z} \frac{T^{(z)}}{Z}$      Average of pooled variance[4].
  end for
```

[1] $\bar{q}^{(z)}$ are pooled point estimates over *M* imputed datasets across *Z* simulations.

[2] $T^{(z)}$ are pooled variances over *M* imputed datasets across *Z* simulations.

[3] $\bar{q}$ is an average of pooled variances *(*$\bar{q}^{(z)}$*)* across *Z* simulations.

[4] $\bar{T}$ is an average of pooled variances *(*$T^{(z)}$*)* across *Z* simulations.

3.2.2 Simulation Study

Flexibility and ability of new MI method to detect complex dependencies structures in categorical variables motivated us to conduct more complex simulations. A simulation study is conducted to check the theoretical findings and to quantify the bias. Artificial data is generated under the missing at random mechanism with various percentages of item nonresponse in all covariates. The number of categorical variables is kept more than the number of continuous variables, aiming to compare strategies in a realistic data situation. Bernoulli distributions with probabilities governed by the logistic regression are used to generate binary variables. Statistical properties of different MICE based MI methods (i.e. CART, PMM, default) are compared with two HMI methods (i.e. H.CART, H.DEF) based on the root mean square errors (RMSEs), empirical standard errors (ESEs) and coverage rates of 95% confidence intervals for GLM's with binary response and mixed covariates. We also check the performance of imputation models with graphical diagnostics (i.e. boxplots for the point estimates and standard errors across all simulations) for various regression coefficients under all MI methods. Results from simulations showed that HMI methods tend to produce minimum bias as compared to MICE methods for most of the co-variants. The average point estimates based on proposed methods are closer to the corresponding true values in most of the cases. Average standard errors, RMSEs and ESEs are also smaller for most of the cases hence, suggesting reasonable performance.

3.2.3 Real Data Application

MICS Punjab 2014 women data (used in contribution 1) is also used for real data application in this contribution. The binary response is modeled using a GLM's depending on four categorical variables and continuous. Twenty simulations are run and *M=10* completed datasets are generated for each MI method. We noted that, HMI MI methods tend to have comparatively smaller pooled standard errors are for all coefficients as compared to MICE default and PMM MI methods and similar pooled standard errors as compared to MICE CART which suggests a reasonable performance. Computational time is also reduced significantly for most of the settings of proposed HMI methods.

3.2.4 Outlook

Statistical properties of the proposed approach can be further studied for continuous response with mixed type co-variants. The proposed method is practical and computationally efficient. New method for MI allows the user to choose a set of incomplete categorical that the regular MICE can sometimes fails to impute due to various restrictions i.e. large dataset, complex dependencies, high percentage of missing data, specification of higher order interactions, multicolinearity and other instability problems. To implement this method no knowledge of complicated models is required. Further experiments with strong relationship between continuous and categorical covariates can be made to improve the estimated values for coverage rates and point estimates. However, of note, one limitation of proposed method is that, the information available in the continuous variables is not used for imputing the categorical variables.

3.3 Contribution 3

The features of HMI method for GLM's in previous contributions suggest its potential to handle large scale survey data with complexities and dependencies. Motivated by the performance, we further made a comprehensive comparison of the MI methods in contribution 3 for child data from the multiple indicator survey (MICS) in Punjab 2014. We evaluated estimators of regression coefficients for a linear regression model in the presence of incomplete binary and continuous predictors. Results are also illustrated with more complex and somewhat large simulated data.

3.3.2 Simulation study

In this simulation, some part of the artificial data is generated from a MVN distribution and all variables are continuous. Further, all continuous variables are discretized to be binary variables by taking small steps. Some part of the data is generated from normal distributions (ND) and random variables are split into various homogeneous groups between 4 and 6 nominal categories. To encode complex dependence relationships with higher order interactions in simulation, another binary covariate is generated from Bernoulli distributions with probabilities governed by the logistic regression. Two highly correlated continuous covariates having strong relationship with categorical covariates are generated from NDs. Finally, a covariate dependent continuous response with a random error component is generated. Hence, simulated data is generated with complex dependence structure in order to insure complications which are difficult to capture with log linear models or chained equation methods for MI. Missing values for independent covariates are generated using a novel approach in each simulation that conforms to Rubin's (1987) definition of missing at random (MAR). We begin with complete data by imposing MAR missingness onto the data. A point to note, default version of chained equations using "mice" was unable to impute missing values in the child data which gives an indication that may be complex dependence structures in the data make it complicated to identify them by the default application of MICE. Rubin's combining rule are used to estimate the parameters of interest for linear models (LM's) with continuous response and mixed covariates. In order to make model as rich as possible, we included all of the variables from the generated data in the imputation model ensuring that the imputation model preserves the relationships between the variables of interest. RMSEs, empirical standard errors ESEs, coverage rates of 95% confidence intervals and bias are compared for the evaluation of performance. Means for CI coverage and RMSEs over all beta coefficients are also calculated. For the purpose of graphical diagnostic check, comparisons are also made based on boxplots of standard errors and point estimates for various regression coefficients for the 1000 simulation runs. There seem to be similarities in structure among all MI methods for a binary covariate generated with higher order interactions. We noticed that H.CART tends to be less biased as compared to CART for all types of covariates and interaction terms. The H.DEF method led to more overall accuracy with smaller means for RMSEs over all beta coefficients as compared to CART. For the most part, coverage rates for the H.CART are in line to 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. H.DEF method tends to have smaller standard errors, ESEs and slightly higher RMSEs as compared to CART for all covariates.

3.3.3 Real Data Application

In order to demonstrate the validity of proposed MI method on a real dataset, we use the MICS Punjab 2014 child's data. For description of child data used in this contribution see Chapter 1. We impute a subset of variables that includes background variables which are continuous and categorical with multiple categories. Graphics of incomplete predictors are used to explore the missing data patterns. Multiple categories for categorical variables were reduced for proper comparisons because it can be tedious for MICE to specify imputation models and interaction terms in the presence of large data. But to keep the analysis comparable and challenging a demographical variable with 36 levels was retained. To create multiple imputations, we included all covariates (especially variables that will be used in subsequent analyses) as predictors in the imputation model. We formulate linear model for a continuous outcome with two continuous and two categorical variables predictors. Since there are no true values to compare for in the real data example, we calculated complete case (CC) estimates for comparison purposes. Similar to simulation study ESEs, average point estimates, average standard across the 200 simulations and computational time are calculated for real data. The results showed smaller standard errors for

H.DEF as compared to CART (see Figure 1). ESEs for HMI variants are also smaller as compared to CART for most of the cases, suggesting better performance over CART. There is a noticeable difference in computational time. Hybrid methods require 3 times less computational time to run all stimulations as compared to MICE methods.

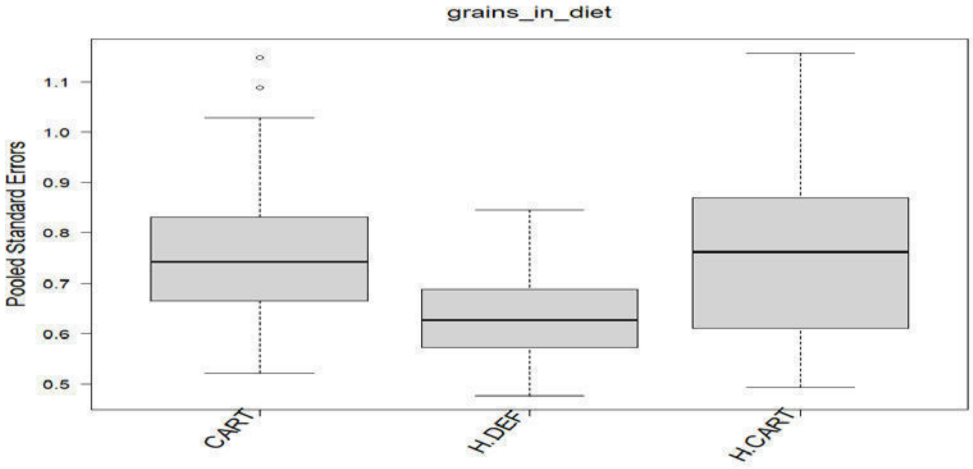


Figure 1. Real data: Boxplots for standard errors across 200 simulations by various imputation methods under Missing at Random (MAR) and ten imputations.

3.3.4 Outlook

A drawback 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.

3.4 Contribution 4

This contribution is an extension to the contribution 2. The statistical properties of GLM's are compared for three dependence models for incomplete variables in proposed HMI method. Therefore three related hybrid imputation strategies are proposed to generate a complete complex data. The objective of this study is to perform an extensive empirical study that evaluates the performance of three software packages commonly available in R. The results are compared for the proposed hybrid MI method and available modern techniques by using an artificial dataset as an example.

3.4.1 HMI for dependence models

In this contribution we investigate the ability of various approaches to detect complex dependency structures in high dimensions using the HMI approach. HMI method is implemented by combining DPMPM MI technique with MICE, expectation-maximization with bootstrapping (EMB) and additive regressions/bootstrapping technique. These techniques are denoted as MICE, Amelia and Hmisc respectively. We denote HMI algorithm as H.MICE, H.Amelia and H.Hmisc when incomplete continuous variables are imputed by MICE, expectation-maximization with bootstrapping (EMB) and additive regressions/bootstrapping technique respectively. Basic information of MI used in this contribution is provided in Table 1. Algorithm 6 explains that how simulations are carried out for different dependence models.

3.4.2 Summary

To access the efficiency, we applied existing MI methods to both incomplete continuous and categorical data and contrast the results with HMI methods. A simple artificial data is generated with a covariate dependent binary response in this study. First, five predictors are generated from a MVN distribution which are further binarized using thresh hold criteria. Further two continuous predictors are generated from normal distributions with probabilities governed by regression models. A binary response is generated from Bernoulli distributions with probabilities governed by the logistic regression. Moderate missing rates are induced using a particular (MAR) dropout model in each item except the response variable which is completely observed. The missing values are then multiply imputed somewhat large times using three existing and three HMI methods. Finally, GLM's with binary response are fitted to the completed data.

Estimates obtained from prescribed models are compared with the parameters used for data generation. The impact of the various procedures is evaluated in terms of the RMSEs and ESEs indices and the coverage rates of the 95% confidence intervals. Five thousand sampling simulations are run for each MI procedure. The coverage rates of the 95% confidence intervals for Hmisc tend to be larger than the coverage rates for H.Hmisc. H.Hmisc tends to result in smaller RMSEs as well for most of the cases. H.Amelia tends to have high coverage rates for most of the estimands with slight bias. H.MICE tends to have lower bias for most of the cases as compared to the MICE. Standard errors are also often lower for the three HMI methods. We considered only binary response with binary and continuous covariates in this study. Challenging issues which need further research include consideration of continuous response with mixed type covariates in HMI approach. Additionally, data with ordinal nature and more categories can be included for further comparisons.

Table 1. Basic information of MI methods

#Method	Acronym	Description
1	Amelia [1]	Uses a bootstrap +EM algorithm
2	Hmisc[2]	Uses Additive Regression, Bootstrapping and PMM algorithms
3	NPBayesImpute[3]	Uses a fully Bayesian, joint modeling approach
4	MICE[4]	MI using FCS
5	H.Amelia	Amelia+NPBayesImpute
6	H.Hmisc	Hmisc+NPBayesImpute
7	H.MICE	Mice+NPBayesImpute

Source: Based on Manuals available on http://www.r-project.org/ in R and Hybrid Multiple Imputation (HMI).

[1] We use the *R* package “Amelia II” (version 1.6.1, Honaker, King, and Blackwell, 2011) with defaults as basic command.

[2] We implement bootstrap and PMM MI methods using 13 (for convenience) iterations with the “aregImpute” function in the “Hmisc”.

[3] The main *R* tool “NPBayesImpute” seems not anymore available at CRAN. The new version called “NPBayesImputeCat” implements the same routines.

[4] We implement a default version of chained equations in “mice” software package in *R*.

3.4 Contribution 4

Algorithm 6: HMI for different dependence models

```
Require: Imp.$P_{cat_m}^{z}$ and $Miss._{num_m}^{z}$
**for** $z = 1, \ldots, Z$ **do**
    **for** $m = 1, \ldots, M$ **do**
        $Imp_m^z \leftarrow$ Imputing Imp.$P_{cat_m}^{z}$ $Miss._{num_m}^{z}$using MICE | Amelia | Hmisc i.e. $f(\,Miss._{num_m}^{z} \mid Imp.\,P_{cat_m}^{z})$
        $Imp_m^z \leftarrow$ Final imputed dataset
        $\bar{q}^{(z)} \leftarrow \sum_{m=1}^{M} \frac{q^{(m)}}{M}$
        $b^{(z)} \leftarrow \sum_{m=1}^{M} \frac{(q^{(m)} - \bar{q}^{(z)})^2}{M-1}$
        $\bar{u}^{(z)} \leftarrow \sum_{m=1}^{M} \frac{u^{(m)}}{M}$
        $T^{(z)} \leftarrow \left(1 + \frac{1}{M}\right) b^{(z)} + \bar{u}^{(z)}$
    **end for**
    $\bar{q} \leftarrow \sum_{z=1}^{Z} \frac{\bar{q}^{(z)}}{Z}$
    $\bar{T} \leftarrow \sum_{z=1}^{Z} \frac{T^{(z)}}{Z}$
**end for**
```

3.5 Contribution 5

Performance of previously developed hybrid approach to handle missing values in large scale complex surveys was limited and was not equipped to use information of continuous variables to impute categorical variables. Therefore, we developed two conditional scenarios of hybrid architectures which use the concept of categorizing continuous covariates. Unlike existing approaches, where categorizing results in loss of power, proposed approaches restore the continuous variables in their original form. These variants are computationally fast and can be applied to both categorical and continuous data in high dimensions.

3.5.1 Iterative hybrid architecture 1

In first conditional scenario of hybrid architecture, we use the concept of categorizing continuous variables before the imputation of categorical data by using three steps (see Figure 2). Incomplete data is divided in to two sub groups i.e. one containing incomplete continuous data and other having incomplete categorical data. In Step 1 incomplete continuous variables are categorized. Further, in Step 2 the JM technique is applied on these categorized variables given additional covariates i.e. (incomplete categorical data) to generate complete categorical data. Complete categorical data generated in this step contains complete categorical variables. In Step3, the FCS technique is applied to impute missing values in original continuous variables given additional completed categorical variables. Points worth noticing, in first step, categorization allows the information on continuations variables to impute categorical variables and 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 datasets. 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'.

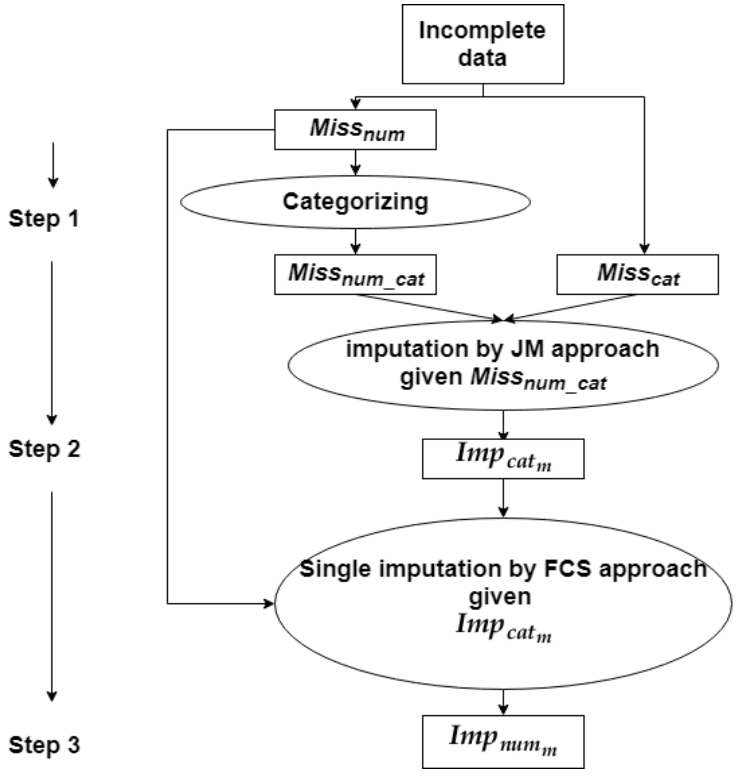


Figure 2. Schematic diagram illustrating the proposed hybrid architecture 1.

3.5.2 Iterative hybrid architecture 2

The second variant uses initial imputed values. These values are obtained by categorization of continuous data before the imputation of categorical data (see Figure 3). Second variant of proposed hybrid architecture is a two steps approach. In first step the initial values for categorical variables are obtain by applying JM approach to missing categorical data. Given the initial values for categorical variables in dependence model, initial values for continuous data are generated. Further, in step 2 these initial values for continuous are categorized and used in conditional models with incomplete categorical variables to obtain complete categorical data with updated values. Given the updated values of categorical covariates, complete continuous variables are generated by applying single iteration of FCS approach to incomplete continuous data with updated values**.** These updated values are further categorized for another initial values for categorical variables. These steps are repeated *M* times with new updated values and *M* complete datasets are obtained.

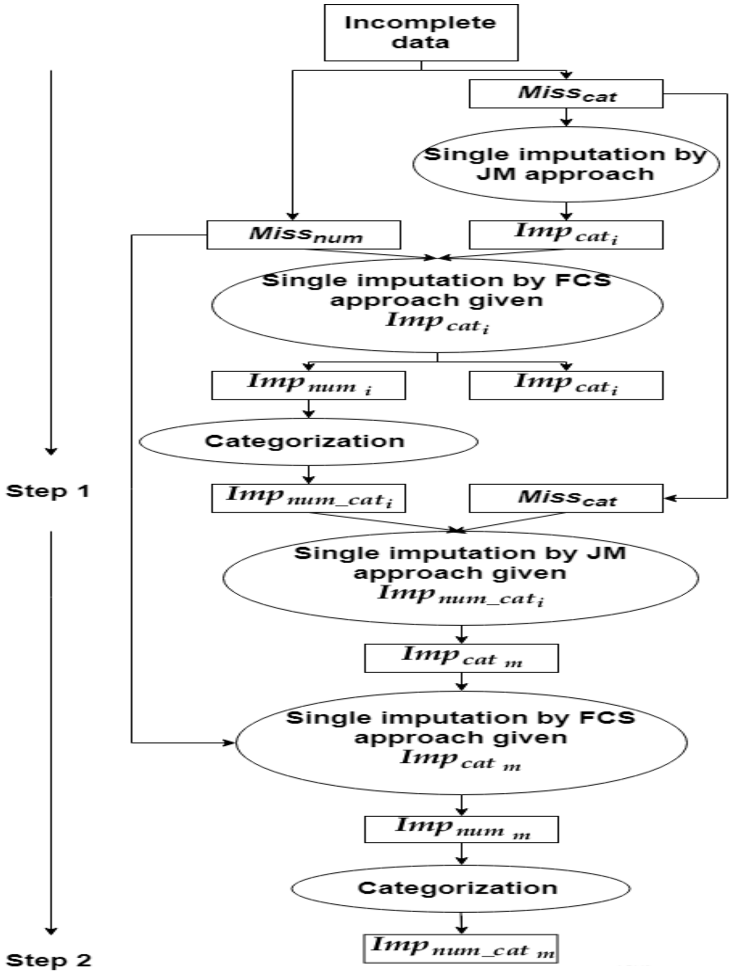

Incomplete data
$Miss_{cat}$
Single imputation by JM approach
$Miss_{num}$
Imp_{cat_i}
Single imputation by FCS approach given Imp_{cat_i}
$Imp_{num\ i}$
Imp_{cat_i}
Categorization
Step 1
$Imp_{num\_cat_i}$
$Miss_{cat}$
Single imputation by JM approach given $Imp_{num\_cat_i}$
$Imp_{cat\ m}$
Single imputation by FCS approach given $Imp_{cat\ m}$
$Imp_{num\ m}$
Categorization
$Imp_{num\_cat\ m}$
Step 2

Figure 3. Schematic diagram illustrating the proposed hybrid architecture 2.

3.5.1 Simulation Study

In this contribution, the complex simulation experiment design similar to the contribution 3 is generated with a covariate dependent binary response. The GLM's are used as a analysis models of interest to explore the performance of the imputation methods. HMI methods which use CART and default algorithms to impute continuous variables under the first and second conditional scenarios are denoted as $H.DEF_1$, $H.CART_1$ and $H.DEF_2$ and $H.CART_2$ respectively. For comparison, two MICE based MI methods i.e. CART and default are used. Ten imputed datasets for each of the proposed and the MICE MI methods are generated. The parameters of interest are estimated using Rubin's aforementioned method for $Z =1000$ simulation runs. Hybrid and CART methods tend to have smaller standard errors as compared to default method for covariates, whereas the hybrid methods tend to have similar standard errors as compared to CART for most of the cases. All hybrid methods tend to have smaller RMSEs for most of the cases where $H.DEF_2$ shows smallest RMSE among the remaining methods. Hence, suggesting overall better performance.

3.5.2 Real Data

The data for this example were taken from a secondary household data from the Punjab Multiple Indicator Cluster Survey in 2014. The substantive goal of this study is to determine the association between access to water and sanitation, and geographic, demographic, and socio-economic factors. A geographical variable "district" is given a special importance in imputation model because it has 36 levels. Most of the background variables related to geographic, demographic, and socio-economic characteristics in MICS household data are categorical with many categories having complex data structures and missing values. 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 proper comparisons, multiple categories for categorical variables were reduced by merging them and a sub-sample 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, a variable which has fifteen levels is included in the sub-sample. The variables are in a sub sample are mostly categorical with multiple categories. The missing data rates in most items are moderate. Only two variables are fully observed. We assume items are MAR in data under consideration. To identify key determinants of water quality we use various explanatory variables associated with the binary response using GLM's. Since there are no true values to compare for real data example, we calculated complete case (CC) estimates for comparison purpose. Point estimates and standards for ten completed datasets across 50 simulations are calculated for real data. Computational time, ESEs and means of point estimates (standard errors) for completed datasets across all simulation runs under various MI methods are estimated. The empirical example with real data indicated that the both hybrid variants yielded smaller standard errors as compared to remaining methods. ESEs and means of standard errors for hybrid variants are also smaller as compared to other methods, suggesting better performance. Moreover there exist significant differences in terms of the computational efficiency among the MI methods.

3.5.3 Outlook

Since there is a variety of MI methods for different types of variables, using these methods in conditional models for multiply imputing missing values in the presence of high-dimensional mixed data seems logical. Our numerical results show that the hybrid imputation achieves, in most cases, better performance than the other existing imputation methods. In addition, the iterative hybrid architectures 1 and 2 have an other advantage. It is straightforward to include information of both types of variables in conditional models, where hybrid approach used in previous contributions do not use the information on continuous variables to impute categorical data. Our current work is limited to MAR mechanism, however, 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. Issues like convergence and appropriate selection of predictors is are not addressed in this contribution therefore further evaluations with diversity of experimental settings will undoubtedly be needed to account for this.

4 Concluding Remarks and Conjectures for future research

Investigation of optimal strategies for fitting MI on the classical regression techniques in the presence of a large number of variables is questionable. There is no general agreement on especially to the how many variables should the imputation model have. According to van Buuren et al. (2012) the number of predictors should be as large as possible for the generally accepted principle for imputation. On the other hand a Hardt et al. (1999) recommends that the small number of variables will be sufficient to successfully implement MI in the R package “mice”. It is worth noting that the performance of the regression techniques is known to deteriorate as number of variables increases and it is generally not feasible to include all variables in imputation models. Little (2018) focuses on the flexibility of MICE by referring a large list of references to the application of chained equation MI in real applications. As opposed to Little (2018), we claim that high-dimensional real applications in these references are limited. Many of the references applied MICE to epidemiological real data in context of large sample sizes rather than a large number of mixed type variables. Consequently the application of classic regression models to the high-dimensional setting as investigated in this thesis may be questionable. The new imputation techniques open the door for us to conduct imputation in the high-dimensional setting by combining various properties of existing MI approaches, the main advantages of the proposed methodology are as follows: (1) it is flexible and can be implemented to mixed type high-dimensional data, (2) it does not rely on heuristic rules of thumb for predictor selection and (3) it is fast. Despite of favorable features of HMI methods for missing data imputation in large-scale studies, in the following some global remarks are given.

Various issues concerning the implementation of the hybrid imputation models need further research. For example, CART method resulted similar or improved performance over hybrid models in most our all applications in simulation studies where we have considered moderate rates of missingness. Whereas, for real world applications where we have high missing rates, hybrid models performed relatively better than CART which gives an indication that we may need even higher rates of missingness than we used in our simulations to get improved performance over CART. Moreover for better performance it may be that we need an even larger number of imputations than we used in our application. Moreover, the specification of the priors in the context of HMI needs further study. The performance of the hybrid models is compared with three existing algorithms for MI in contribution 4 already. In parallel with existing classical imputation approaches, the performance of the hybrid models can be improved by adapting other existing algorithms such as GAMLSS imputation method by using *R* package "ImputeRobust" to multiply impute missing values in the presence of high-dimensional data, which can relax some modeling assumptions. It may be interesting to evaluate the hybrid approach in comparison to a 2-stage and nested multiple imputation approaches. Hybrid models can also be compared to the Bayesian nonparametric hierarchical model developed by Murray (2019) which imputes missing multivariate continuous and categorical data using *R* package "MixedDataImpute". Iterative hybrid models proposed in the last contribution can easily be extended with categorization of continuous variables up to multiple levels. It may be interesting to investigate whether inclusion of more continuous variables may improve the obtained imputations. Future work will considering the diagnostics or sensitivity analysis in the real data applications can be helpful to provide justification of the method recommendation.

Hybrid MI methods are applied to various datasets from MICS Punjab. However, applications can be made to MICS datasets from other provinces of Pakistan (e.g. Sindh, Khyber Pakhtunkhwan and Baluchistan). Lastly, as previously mentioned, datasets with more complex structures between the variables will be considered.

Bibliography

Arnold, B. C. and Press, S. J. (1989). 'Compatible conditional distributions'. *Journal of the American Statistical Association 84*(405), 152–156.

Angelillo, I.F., G. Ricciardi, P. Rossi, P. Pantisano and E. Langiano, et al. (1999). 'Mothers and vaccination: knowledge, attitudes, and behaviour in Italy'. *Bulletin of the World Health Organization 77*(3), 224-229.

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

Allison, D.P. (2009). *Missing data*. Thousand Oaks, CA, Sage Publications Ltd., 72-89.

Andridge, R.R. and R.J. Little (2010). 'A Review of Hot Deck Imputation for Survey Non response'. *International Statistical Review 78*(1), 40–64. doi:10.1111/j.1751-5823.2010.00103.x

Akmatov, M.K. (2011). 'Child abuse in 28 developing and transitional countries--results from the Multiple Indicator Cluster Surveys'. *International Journal of Epidemiology 40*(1), 219–27.

Ankaiah, N. and V. Ravi (2011). 'A novel soft computing hybrid for data imputation'. *Proceedings of the 7th international conference on data mining (DMIN)*, Las Vegas:USA.

Azim, S. and S. Aggawal (2014). 'Hybrid model for data imputation: using fuzzy c means and multi layer perceptron'. *Proceeding of International Conference on Advance Computing*, 1281–1285.

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

Akande, O., F. Li and J.P. Reiter (2017). 'An empirical comparison of multiple imputation methods for categorical data'. *American Statistician 71*(2), 162–170.

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

Audigier,V., I.R. White, S. Jolani, T. Debray, M. Quartagno, J. Carpenter, S. van Buuren, and M. Resche-Rigon (2018). 'Multiple imputation for multilevel data with continuous and binary variables'. *Statistical Science 33*(2), 160-183.

Birnbaum (1968). 'Some latent trait models and their use in inferring an examinee's ability'. In F. In Lord & M. Novick (Eds.), Statistical theories of mental test scores. Mass: Addison-Wesley, 453-479.

Bishop, Y.M.M., S. E. Feinberg, and P. W. Holland (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press. Springer-Verlag: New York.

Binder, D. (1983). 'On the variance of asymptotically normal estimators from complex surveys'. *International Statistical Review 51*(3), 279-292.

Breiman, L., J.H. Friedman, R.A. Olshen and C.J. Stone (1984). *Classification and Regression Trees*. Wadsworth: Belmont.

Breslow, N. (1996). 'Generalized linear models: Checking assumptions and strengthening conclusions'. *Statistica applicata* 8, 23-41.

Barnard, J. and X. Meng (1999). 'Applications of multiple imputation in medical studies: From aids to nhanes'. *Statistical Methods in Medical Research 8*(1), 17–36.

Bernaards, C.A., T.R. Belin and J.L. Schafer (2007). 'Robustness of a multivariate normal approximation for imputation of binary incomplete data'. *Statistics in Medicine 26*(6*)*, 1368–1382.

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

Black, R.E., C.G. Victora, S.P. Walker, Z.A. Bhutta, P.S. Christian, M. De Onis and R. Uauy (2013). 'Maternal and child undernutrition and overweight in low-income and middle-income countries'. *The Lancet 382*(9890), 427-451.

Colsher, P.L. and R.B. Wallace (1989). 'Data quality and age: Health and psycho behavioral correlates of item nonresponse and inconsistent responses'. *Journal of Gerontology 44*(2), 45-52.

Chib, S. and B.H. Hamilton (2002). 'Semiparametric Bayes analysis of longitudinal data treatment models'. *Journal of Econometrics 110*(1), 67–89.

Chromy, J.R. and S. Abeyasekera (2005). 'Statistical analysis of survey data. In: Department of Economic and Social Affairs Statistics Division, United Nations'. *Household Sample Surveys in Developing and Transition Countries*, New York: United Nations, 389- 417.

Cappa, C. and S.M. Khan (2011). 'Understanding caregivers' attitudes towards physical punishment of children: evidence from 34 low- and middle-income countries'. *Child Abuse and Neglect 35*(12), 1009–1021.

Carpenter, J.R. and M.G. Kenward (2013). *Multiple Imputation and Its Applications*. Chichester, UK: John Wiley & Sons.

Dillman, D.A. (1978). *Mail and Telephone Surveys: The Total Design Method*. New York: Wiley & Sons.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*. Chapman and Hall, London, UK.

Dunson, D.B. and C. Xing (2009). 'Nonparametric Bayes modeling of multivariate categorical data'. *Journal of the American Statistical Association 104*(487), 1042-1051.

Doove, L.L., S. van Buuren, and D. Elise (2014). 'Recursive Partitioning for Missing Data Imputation in the Presence of Interaction Effects'. *Computational Statistics and Data Analysis 72*, 92-104.

de Jong, R., S. van Buuren and M. Spiess (2014). 'Multiple imputation of predictor variables using generalized additive models'. Communications in Statistics – Computation and Simulation. doi: 10.1080/03610918.2014.911894.

Daud, M., M. Nafees, S. Ali, M. Rizwan, R.A. Bajwa, M.B. Shakoor, M.U. Arshad, S.A.S. Chatha, F. Deeba and W. Murad (2017). 'Drinking Water Quality Status and Contamination in Pakistan'. *BioMed Research International.*

Erosheva, E.A., S.E. Fienberg and B.W. Junker (2002). 'Alternative statistical models and representations for large sparse multi-dimensional contingency tables'. *Annales de la Faculte´ des Sciences de Toulouse 11*, 485-505.

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

Gareaballah, E.T. and B.P. Loevinsohn (1989). 'The accuracy of mother's reports about their children's vaccination status'. *Bull World Health Organ 67*(6), 669–874.

Gelman, A. and T. P. Speed (1993). 'Characterizing a Joint Probability Distribution by Conditionals'. *Journal of the Royal Statistical Society B 55*(1), 185–188.

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

Gill, J. (2001). *Generalized linear models: a unified approach.* Sage University Paper: London.

Geneviève, R., K. Olga, J. Julie, M. Éric and T. Robert (2018). 'Main effects and interactions in mixed and incomplete data frames'. arXiv preprint arXiv:1806.09734.

Herzog, A.R. and W.L. Rodgers (1992). 'The use of survey methods in research on older Americans'. In: R. B. Wallace & R. F. Woolson (eds). *The Epidemiological Study of the Elderly*. Oxford: Oxford University Press.

Hastie, T., R. Tibshirani and J. Friedman (2001). *The Elements of Statistical Learning; Data Mining, Inference, and Prediction* (2 ed.). Springer Verlag, New York.

Hirano, K. (2002). 'Semiparametric Bayesian inference in autoregressive panel data models'. *Econometrica 70*(2), 781–799.

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

Hotron, N.J., S.P. Lipsitz and M. Parzeen (2003). 'A potential for bias when rounding in multiple imputation'. *The American Statistician 57*(4), 229–232.

Hoffmann, J.P. (2004). *Generalized linear models: An applied approach.* Pearson: Boston.

Harel, O. and X. H. Zhou (2006). ‘Multiple Imputation for Correcting Verification Bias’. *Statistics in Medicine 25*(22), 3769–3786.

Hawkes, D. and I. Plewis (2006). ‘Modelling non-response in the National Child Development Study’. *Journal of the Royal Statistical Society Series A 169*(3), 479–491.

Harel, O. (2007). ‘Inferences on missing information under multiple imputation and two-stage multiple imputation’. *Statistical Methodology 4*(1), 75-89.

He, Y. and T.E. Raghunathan (2009). ‘On the Performance of Sequential Regression Multiple Imputation Methods with Non Normal Error Distributions’. *Communications in Statistics - Simulation and Computation 38*(4), 856–883. ISSN: 0361-0918. DOI: 10.1080/03610910802677191. URL: http: //dx.doi.org/10.1080/03610910802677191

Honaker, J., K. Gary, and B. Matthew (2011). ‘Amelia II: A Program for Missing Data’. *Journal Of Statistical Software 45*(7), 1–47.

Hardt, J., M. Herke and R. Leonhart (2012). ‘Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research’. *BMC Medical Research Methodology 12*(1).

Hippel, P.T. (2013). ‘Should a Normal Imputation Model be Modified to Impute Skewed Variables?’ *Sociological Methods & Research 42*(1),105– 138. ISSN: 0049-1241, 1552-8294. DOI: 10.1177/0049124112464866.

Harrell, F.E. (2015). *Regression Modeling Strategies* (2 ed.). Springer Series in Statistics. New York, NY: Springer New York. ISBN: 9783319194257. DOI: 10. 1007/978-3-319-19425-7. URL: http://dx.doi.org/10.1007/978-3-319- 19425-7

Hussion, 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. and G. Porro (2007). ‘Missing data imputation, matching and other applications of random recursive partitioni*ng’. Computational Statistics and Data Analysis* 52(2), 773–789.

Iacus, S.M. and G. Porro (2008). ‘Invariant and metric free proximities for data matching: an R package’. *Journal of Statistical Software 25*(11), 1–22.

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

Kyung, M., J. Gill and G. Casella (2010). ‘Estimation in Dirichlet random effects models’. *Annals of Statistics 38*(2), 979–1009.

Little, R.J.A. (1988). ‘Missing-Data Adjustments in Large Surveys’. *Journal of Business & Economic Statistics 6*(3), 287–296. ISSN: 07350015. DOI: 10.2307/ 1391881.

Loh, W.Y. and Y.S. Shih (1997). ‘Split selection methods for classification trees’. *Statistica Sinica 7*, 815–840.

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

Lee, K.J. and J. B. Carlin (2010). ‘Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation’. *American journal of epidemiology 171*(5), 624–632.

Li, D., H. Gu and L.Y. Zhang (2013). ‘A hybrid genetic algorithm-fuzzy c-means approach for incomplete data clustering based on nearest-neighbor intervals’. *Journal of Soft Computing 17*(10), 1787–1796.

Loh, W.-Y., X. He and M. Man (2015). ‘A regression tree approach to identifying subgroups with differential treatment effects’. *Statistics in Medicine 34*(11), 1818–1833.

Liang, Z., C. Zhikui, Y. Zhennan and Hu. Yueming (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, NY, 1725-1730.

Little, R. J. (2018). ‘On Algorithmic and Modeling Approaches to Imputation in Large Data Sets’. *Statistica Sinica* (to appear)

Masters, G.N. (1982). ‘A Rasch model for partial credit scoring’. *Psychometrika 47*(2), 149-174.

McCullagh, P. and J.A. Nelder (1989). *Generalized Linear Models* (2 ed.). London: Chapman & Hall.

Marker, D.A., D. R. Judkins, and M. Winglee (2002). ‘Large-Scale Imputation for Complex Surveys’. *Survey Nonresponse*, Wiley: New York, 329–341.

Manrique-Vallier, D. and J.P. Reiter (2012). ‘Estimating Identification Disclosure Risk Using Mixed Membership Models’. *Journal of the American Statistical Association 107*(500), 1385-1394.

McDonald, C.M.M., I. Olofin, S. Flaxman, W.W. Fawzi, D. Spiegelman., L.E. Caulfield, R.E. Black, M. Ezzati and G. Danaei (2013). ‘The effect of multiple anthropometric deficits on child mortality: meta-analysis of individual data in 10 prospective studies from developing countries’. *The American Journal of Clinical Nutrition 97*(4), 896-901.

Murray, J.S. and J.P. Reiter (2016). ‘Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence’. *Journal of the American Statistical Association 111*(516), 1466–1479.

Murray, J.S. (2019). *MixedDataImpute: Missing Data Imputation for Continuous and Categorical Data using Nonparametric Bayesian Joint Models*. *R* package version 0.1.

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

Nonyane, B.A.S. and A.S. Foulkes (2007). 'Multiple imputation and random forests (MIRF) for unobservable, high-dimensional data'. *The international journal of biostatistics 3*(1), 1-18.

National Nutrition Survey (NNS) of Pakistan, Islamabad (2011). Government of Pakistan. Aga Khan University &UNICEF.

Nishanth, K.J., V. Ravi, N. Ankaiah and L. Bose (2012). 'Soft computing based imputation and hybrid data and text mining: The case of predicting the severity of phishing alerts'. *Expert Systems with Applications 39*(12), 10583–10589.

Nishanth, K.J. and V. Ravi (2013). 'A computational intelligence based online data imputation method: An application for banking'. *Journal of information Processing Systems 9*(4), 633–650.

Nikfalazar, S., C.H. Yeh, S. Bedingfield and H.A. Khorshidi (2019). 'A Hybrid Missing Data Imputation Method for Constructing City Mobility Indices'. *Communications in Computer and Information Science*, Volume 996, pp. 135-148. Springer: Singapore.

Quartagno, M. and J. Carpenter (2019). *jomo: A package for multilevel joint modelling multiple imputation.* R package version 2.6.

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

Rasch, G. (1960). 'Probabilistic Models for some Intelligence and Attainment Tests'. *Danish Institute for Educational Research.*

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

Rubin, D.B. (1996). 'Multiple Imputation After 18+ Years'. *Journal of the American Statistical Association 91*(434), 473–489.

Raghunathan, T. E., J. M. Lepkowski, J. van Hoewyk, and P. Solenberger (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* 27 (1), 85–95.

Raghunathan, T.E., P.W. Solenberger, J. van Hoewyk (2002). *IVEware: Imputation and Variance Estimation Software user guide*. Michigan: University of Michigan. [accessed May 19, 2008]. Available at http://www.isr.umich.edu/src/smp/ive/ [Google Scholar]

Ruel, M.T. and P. Menon (2002). *Creating a Child Feeding Index Using the Demographic and Health Surveys: An Example from Latin America Washington*. DC: International Food Policy Research Institute.

Rubin, D.B. (2003). 'Nested multiple imputation of NMES via partially incompatible MCMC'. *Statistica Neerlandica 57*(1), 3-18.

Royston, P. (2004). 'Multiple Imputation of Missing Values'. *The Stata Journal 4*(3), 227-241.

Riphahn, R.T. and O. Serfling (2005). ‘Item Non-response on Income and Wealth Questions’. *Empirical Economics 30*(2), 521-538.

Reiter, J.P., T.E. Raghunathan and S.K. Kinney (2006). ‘The importance of modeling the sampling design in multiple imputation for missing data’. *Survey methodology* 32(2), 143 -150.

Reiter, J.P. and J. Drechsler (2007). ‘Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality’. *IAB Discussion Paper 20*, 1-18.

Reiter, J.P. and T.E. Raghunathan (2007). ‘The multiple adaptions of multiple imputation’. *Journal of the American Statistical Association 102*(480), 1462-1471.

Royston, P. and I.R. White (2011). ‘Multiple imputation by chained equations (MICE): implementation in Stata’. *Journal of Statistical Software 45*(4), 1-20.

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

Samejima, F. (1969). ‘Estimation of latent ability using a response pattern of graded scores’. *Psychometrika Monograph* 17.

Siddiqi, F. and H. Patrinos (1995). ‘Child labor: Issues, causes and interventions’. *Human Resources and Operations Policy Working Paper 56*, World Bank, Washington, DC.

Schenker, N. and J. M. G. Taylor (1996). ‘Partially parametric techniques for multiple imputation’. *Computational Statistics & Data Analysis 22*(4), 425– 446. ISSN: 01679473. DOI: 10.1016/0167-9473(95)00057-7.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data.* CRC Press. ISBN: 978-1-4398-2186-2.

Strobel, C., J. Malley and A. Zeileis (2009). ‘An introduction to recursive partitioning: rationale, application and characteristics of classification and regression trees, bagging and random forests’. *Psychol Methods 14*(4), 323–348.

Su, Y.S., A. Gelman, J. Hill and M. Yajima (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/.

Seaman, S.R., J.W. Bartlett and I.R. White (2012). ‘Multiple Imputation of Missing Covariates with Non-Linear Effects and Interactions: An Evaluation of Statistical Methods’. *BMC Medical Research Methodology 12* (1), 46.

Stekhoven, D.J. P. Buhlmann (2012). ‘MissForest–non-parametric missing value imputation for mixed-type data’. *Bioinformatics 28*(1), 112–118.

Si, Y. and J.P. Reiter (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 (2013). *Stata statistical software,* Release 13, College Station, Texas, TX, USA. 2013.

SAS Institute (2014). *Base SAS 9. 4 Procedures Guide: Statistical Procedures*. Cary: SAS Institute; 2014.

Schafer, J.L. and J.H. Zhao (2014). *pan: Multiple Imputation for Multivariate Panel or Clustered Data.* R package version 0.9.

Shah (2014). *CALIBERrfimpute: Imputation in MICE using Random Forest. R* package version 0.1.

Salfran, D. and S. Martin (2015). *A Comparison of Multiple Imputation Techniques.* Tech. rep. Discussion Paper.

Shukur, O.B. and M.H. Lee (2015). 'Imputation of missing values in daily wind speed data using hybrid AR-ANN method'. *Modern Applied Science 9*(11).

Salfran., D. (2018). *ImputeRobust: Robust Multiple Imputation with Generalized Additive Models for Location Scale and Shape.* R package version 1.3.

Speidel, M., J. Drechsler and S. Jolani (2018). *R package hmi: a convenient tool for hierarchical multiple imputation and beyond (No. 16/2018).* IAB-Discussion Paper.

Tang, J., G. Zhang, Y. Wang, H. Wang and F. Liu (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*, 29-40.

Ting, J., B. Yu, D. Yu, S. Ma (2014). 'Missing data analyses: a hybrid multiple imputation algorithm using gray system theory and entropy based on clustering'. *Applied intelligence 40*(2), 376-388.

Tabassum, R.A. (2017). 'Arsenic assessment and removal from drinking water of tehsil Hasilpur 1162 using agricultural byproducts'. Department of Environmental Sciences, Vol. MS. COMSATS Institute of Information Technology, Vehari, 1-81.

van Buuren, S., H.C. Boshuizen and D.L. Knook et al. (1999) 'Multiple imputation of missing blood pressure covariates in survival analysis'. *Statistics in Medicine 18*(6), 681-694.

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

van Buuren, S. (2007). 'Multiple imputation of discrete and continuous data by fully conditional specification'. *Statistical Methods in Medical Research 16*(3), 219-242. ISSN: 0962-2802. DOI: 10.1177/0962280206074463.

Vermunt, J.K., J.R. van Ginkel, L.A. van der Ark and K. Sijtsma (2008). 'Multiple imputation of incomplete categorical data using latent class analysis'. *Sociological Methodology 38*(1), 369-397.

van Buuren, S. and K. Groothuis-Oudshoorn (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.

World Health Organization (WHO) (2003). *Community-based Strategies for Breastfeeding Promotion and Support in Developing Countries,* Department of child and adolescent health and development, Geneva, World Health Organization.

White, I.R., P. Royston and A.M. Wood (2011). ‘Multiple imputation using chained equations: issues and guidance for practice’. *Statistics in Medicine 30*(4), 377–399.

Weirich, S., N. Haag, M. Hecht, K. Bohme, T. Siegle and O. Ludtke (2014). ‘Nested multiple imputation in large-scale assessments’. *Large-scale Assessments in Education 2*(9), 1–18.

Yucel, R.M., Y. He and A.M. Zaslavsky (2011). ‘Gaussian-based routines to impute categorical variables in health surveys’. *Statistics in Medicine 30*(29), 3447–3460.

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

Zhao, Y. Q. Long (2016). ‘Multiple imputation in the presence of high-dimensional data’. *Statistical Methods in Medical Research 25*(5), 2021-2035.

Attached Contributions

Contribution 1:

This contribution is presented as a conference paper in the NTTS 2019 conference:

Razzak, H. and Heumann, C. (2019a). Predictive performance of a hybrid technique for the multiple imputation of survey data. Proceedings of the *New Techniques and Technologies for Statistics (NTTS) conference*. Brussels: Belgium.

Available under: url: https://coms.events/ntts2019/data/abstracts/en/abstract_0108.html.

The following technical report that is an extended version of the conference proceedings is included in Chapter 3 of this thesis instead:

Razzak, H. and Heumann, C. (2019b). Predictive performance of a hybrid technique for the multiple imputation of survey data. Department of Statistics (LMU Munchen): *Technical Reports*, Nr. 228, last updated 3. December 2019.

Available under: url: https://doi.org/10.5282/ubm/epub.69897

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Humera Razzak, Christian Heumann

Predictive performance of a hybrid technique for the multiple imputation of survey data

Technical Report Number 228, 2019
Department of Statistics
University of Munich

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

Predictive performance of a hybrid technique for the multiple imputation of survey data

Humera Razzak*
Christian Heumann†
Department of Statistics, University of Munich

January 6, 2020

Abstract

We discuss the development of a multiple imputation (MI) method for analysing data from the Multiple Indicator Cluster Survey (MICS). A popular chained equations approach to MI called MICE fails to perform sometimes because of computational inefficiency, a complex dependency structure among categorical variables and high percentage of missing information in large scale survey data. On the other hand, a MI approach based on fully Bayesian joint modeling seems to perform very well for categorical variables having complex dependencies but requires transformation and other techniques to impute continuous variables. A hybrid approach is presented here where imputations for a large number of categorical variables are created under a fully Bayesian joint modeling MI technique and regular MICE is used to create imputations for continuous variables. This provides a flexible and practical hybrid MI approach to obtain complete data, which sometimes cannot be obtained when both MI approaches are applied separately. The method proposed is used to analyse data from the MICS 2014 survey women's data investigating the association between various factors and breastfeeding practices among women in Punjab. The relationship between the binary response (breastfeeding) and explanatory variables is modelled using generalized linear models (GLM's). The accuracy of a predictive model is assessed by the area under the receiver operating characteristic (ROC) curve, known as AUROC, and the results obtained under the proposed and existing MI methods are compared. The proposed method outperforms the MICE algorithms CART and PMM in most of the cases requiring less computational time and only minimal tuning by the analyst. The results obtained by the simulation study are supported by a real data example.

Key Words: Complex dependencies; Hybrid multiple imputation

1 Introduction

Many large scale complex surveys such as the Multiple Indicator Cluster Survey or MICS are conducted to recognize forces that contribute to the public health factors that interact at individual, family, community, population, and policy levels. Generally, MICS contains a large number of categorical variables with lots of categories, a complex dependency structure and missing values. For example, the data set of individual women from MICS 2014 used in the real data example has more than 60 per cent data missing on 44 background variables.

Missing data often implicates a biased or an inefficient analysis. Missing mechanisms are: (i) missing completely at random (MCAR), (ii) missing at random (MAR), (iii) missing not at random (MNAR)[1] (Little and Rubin, 2002). MCAR occurs if the probability of missing variable X does not depend on the values of any other variable in the data set (Bennett, 2001). This means that the value of the missing variable is unrelated to any other variable. For example, if the probability that the gender of the child is missing in a household database does not depend on any other variable of the database then MCAR holds. Although it is difficult to detect whether data are MCAR, however, Little (1988) provides a statistical test of MCAR. Schafer and Graham (2002) describe MCAR to be a special case of MAR. With MAR, the probability of having a missing data point in a certain variable is related

*Razzak@stat.uni-muenchen.de
†chris@stat.uni-muenchen.de

[1] MNAR is also called non-ignorable (Ankaia and Ravi, 2011) and not further used in the paper.

to atleast one other variable in the data set but is not related to the variable itself (Allison, 2002). MAR occurs if the probability that a variable X is missing depends on observed data set but not on the variable X itself. For example, if the probability that income of a person is missing depends on profession and age, then the missing data process is MAR. MNAR occurs if the probability that a variable X is missing depends on the variable X itself. For example, if the probability that income is missing dependes on the income itself (often the probability that income is missing is higher for low incomes than for higher incomes) then MNAR occurs.

It is critical to impute the data since multiply imputed data usually provides more accurate inference as compared to complete case analysis or single imputation (Abdella and Marwala, 2005, Little and Rubin, 2002), if the missing data is missing at random (MAR). In recent decades, lots of efforts have been made in the development of statistical methods to treat the problem of missing data. According to studies (Vach and Blettner, 1991 and Kleinbaum et al., 1981), the estimation of regression coefficients can be biased when ad hoc methods and complete case analysis for handling missing data are used. Various approaches based on the Expectation-Maximization (EM) algorithm (Little and Schluchter, 1985), a fully Bayesian analysis (Dellaportas and Smith, 1993), maximum likelihood (Vach and Schumacher, 1993), a mixture of independent multinomial distributions (Dunson and Xing, 2009) and weighted estimating equations (Robins et al., 1994) have been proposed. Multiple-imputation (MI) introduced by Rubin (1987) is nowadays considered as a gold standard to handle the missing data problem. MI replaces missing values in a data set by drawing random values from the predictive posterior distribution of the missing data given the observed data. MI creates M complete data sets. Inference of interest (e.g. mean, regression) can be run on each newly created imputed data set and estimates can be combined by using "Rubin's rules" (Rubin, 1987). One approach for MI is the so-called Fully conditional specification (FCS) model. FCS specifies univariate conditional distributions on a variable-by-variable basis, and draws sequentially missing values iteratively from the estimated conditional distributions. MI by chained equations (MICE) (Raghunathan et al., 2001, van Buuren and Groothuis-Oudshoorn, 2011) is such a fully conditional specification (FCS) approach to MI. The researcher can choose a suitable regression model for each variable, for example classification and regression trees (CART) (Breiman et al., 1984) for categorical variables, predictive mean matching (PMM) (Little, 1988) for continuous variables or just rely on the default method which e.g. uses logistic regression models for binary and PMM for continuous variables. Sometimes, problems of convergence and incompatibility arise when MICE is used for specifying univariate conditional distributions (Gelman and Speed, 1993). MICE fails to perform sometimes due to a complex dependency structure among the categorical variables and a high percentage of missing information which is typical for large scale survey data. Moreover, regression imputations are very time consuming. The R (R Core Team, 2018) package "mice" (van Buuren and Groothuis-Oudshoorn, 2011) implements MICE. The joint modeling (JM) specification is another approach used for MI. JM draws missing values simultaneously for all incomplete variables. JM involves specifying a multivariate distribution for the variables and draws imputations from their conditional distributions by the Markov Chain Monte Carlo (MCMC) methods (Schafer, 1997). Modeling variables of different types can make the specification of a joint distribution very difficult. The Dirichlet Process Infinite Mixtures of Products of Multinomials (DPMPM) is a full Bayesian JM approach (Dunson and Xing, 2009). Si and Reiter (2013) implement DPMPM to impute missing values for categorical variables. The R package "NPBayesImputeCat" by Quanli et al. (2018) implements the DPMPM approach for MI. The implemented DPMPM JM technique to handle missing values is therefore limited to categorical variables and requires transformations (or other tricks) for continuous variables.

The complex dependencies in the MICS data sets containing mixed type covariates (i.e. both categorical and continuous) can be difficult to be identified by the mentioned MI approaches. It has been shown that the MI approach based on DPMPM performs very well for categorical variables having complex dependencies but requires knowledge of complicated models to create the dependence structure between the continuous and the (possibly high) dimensional categorical variables (Murray and Reiter, 2016). These limitations sometimes create serious problems for researchers to obtain complete data sets with mixed type variables. Therefore, we need to develop methods for imputing mixed type data from large scale complex surveys which avoid difficulties of complicated models in high dimensions, combine existing well studied techniques to handle incomplete large scale complex data sets and which are computationally efficient.

We develop a Hybrid Multiple-Imputation (HMI) approach for handling data for the problem described above. We propose to apply the DPMPM MI approach to impute categorical variables having potentially complex dependencies and to use MICE to create imputations for the continuous variables after the categorical variables have been imputed beforehand. The HMI method enables us to utilize the good properties of the DPMPM MI approach and the simplicity of MICE to obtain complete data sets in the mixed data type situation in a flexible and practical manner.

The method proposed is used to analyse data from the MICS 2014 survey women's data. The association between various factors and breastfeeding practices among women in Punjab is investigated. The relationship between the binary response (breastfeeding) and explanatory variables is modelled using generalized linear models (GLM's). The accuracy of the predictive model is assessed by the area under the receiver operating characteristic (ROC) curve, known as AUROC. The predictive performance of the proposed and existing MI methods is compared under a large spectrum of data characteristics. The hybrid mechanism is described in section 2. In Section 3 and 4, cross validation and the measure of performance used for comparison are described. Through simulation studies, we evaluate two software packages used for implementing the hybrid procedure in section 5. Section 6 shows an applications of the proposed method for a real data set. Finally, we give concluding remarks.

2 Proposed hybrid architecture

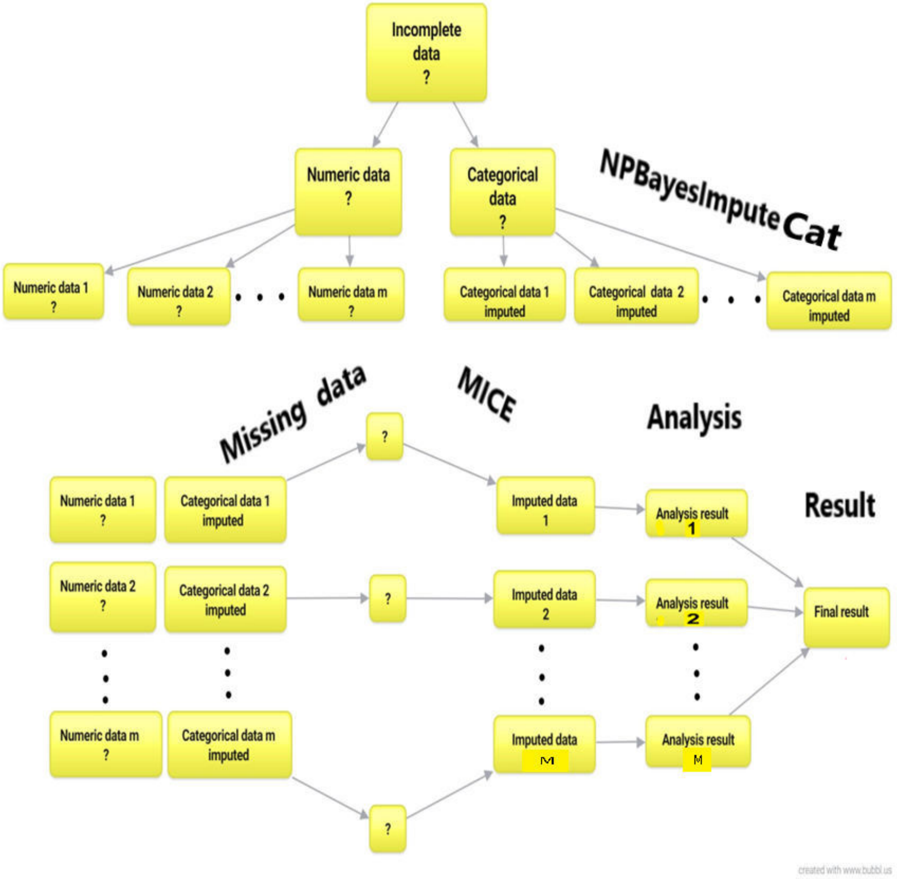


Figure 1: The schema of the hybrid imputation method

The proposed missing data imputation approach is a 3-stage approach. The dataflow diagram (Figure 1) presents the schema of the hybrid imputation method. Step 1: Only the categorical variables ($Imp._{cat}$) are imputed utilizing the R package NPBayesImputeCat (Quanli et al., 2018) which uses a fully Bayesian joint modeling approach. Step 2: The incomplete continuous variables ($Miss._{num}$) are combined with the already imputed categorical variables, $Imp._{cat}$, resulting in M incomplete data sets where values in the continuous variables may be missing and values in the categorical variables have been imputed. M incomplete data sets are made such that the rows of each $Miss._{num}$ data set correspond to the same rows of each $Imp._{cat}$ data set. Hence, one ensures that MI using chained equations for continuous variables uses the information of the imputed categorical covariates for the same unit. Step 3: MICE with various algorithms is used to yield M complete datasets. The R package mice (van Buuren and Groothuis-Oudshoorn 2011) is used for this purpose. The draws from the posterior predictive distribution of the incomplete continuous variables therefore depend on the (in the first step) imputed categorical variables. This process is repeated M times to generate multiple complete data sets. Two Hybrid MI based methods are H.CART and H.PMM. H.CART combines DPMPM with the CART and H.PMM combines DPMPM with PMM. For comparisons, CART, PMM and the Default method in MICE are used.

3 Cross validation

Holdout cross validation is used to assess the predictive performance of a logistic regression model used for the binary response. The logistic regression model [2]is used because the effect of various factors on a binary response (breastfeeding) is analysed later in the real data example. Train and test data sets are generated randomly using a 70% / 30% split. The basic reason to select this method is its simplicity.

4 Evaluation of Performance

Algorithm 1: Holdout cross validation and estimation of AUROC for HMI method

```
Require: $P$ $nxp$ matrix with incomplete data
$Miss._{cat}$, $Miss._{num}$ ← Initial division of $p$ variables into factor and numeric subsets
for $z = 1, \dots, Z$ do
    for $m = 1, \dots, M$ do
        $Imp.P^{z}_{cat_m}$ ← Imputation using "NPBayesImputeCat" for $Miss._{cat}$
        $Imp.P^{z}_{cat_m}$ $Miss.^{z}_{num_m}$ ← Combining $Imp.P^{z}_{cat_m}$ and $Miss.^{z}_{num_m}$ to generate partially imputed dataset
        $Imp^{z}_{m}$ ← Imputing $Imp.P^{z}_{cat_m}$ $Miss.^{z}_{num_m}$ using MICE i.e. $f(Miss.^{z}_{num_m} \mid Imp.P^{z}_{cat_m})$
        $Imp^{z}_{m}$ ← Final imputed data set
        $Imp^{z}_{testing_m}$, $Imp^{z}_{training_m}$ ← Divide matrix $Imp^{z}_{m}$ into testing and training subsets
        $P(y = 1 \mid x_{1,\dots,}x_p)^{z}_{m} = 1/(1 + e^{-(a+\sum_{j=1}^{p}(b^{z}_{jm}x^{z}_{jm})})$ ← Train a GLM model on $Imp^{z}_{training_m}$
        $P_m{}^{z}$ ← Make prediction on $Imp^{z}_{testing_m}$
        $AUROC^{z}_{m}$ ← AUROC curve based on $P_m{}^{z}$
    end for
    $\overline{AUROC} = \frac{\sum_{m=1}^{M}(AUC^{z}_{m})}{M}$ ←Pooled AUROC curve
end for
```

The area under the receiver operating characteristic (ROC) curve, known as AUROC, is used to compare different MI methods. For more detail see McNeil and Hanley (1984), Metz (1986), Swets (1979) and Wieand et al., (1989). Algorithm 1 describes how the AUROC curve is pooled[3].

[2] A special generalized linear model with link logit.

[3] The arithmetic mean is taken of M AUROC values obtained by M fitted GLM's

5 A small scale study

A small scale study is conducted to examine the impact of MI by our proposed method. The incomplete data is generated MAR to compare the methods in a realistic data situation. The number of categorical variables is kept higher than the number of continuous variables due to the fact that the simulation is aimed to be similar to the survey data. Table 1 represents a large spectrum of practially occurring data characteristics used for generating data according to a variety of settings. Series of simulations are run varying the correlation among covariates, the number of imputations, different hybrid methods and the algorithms used in MICE.

Simulation study: Five (X_1, X_2, X_3, X_4 and X_5) dimensional correlated normal data is generated using the R package Binorm (Demirtas et al., 2014). The marginal distribution of $X_1, X_2, X_3 \sim Bernoulli(0.5), X_4 \sim N\ (80, 250)$ and $X_5 \sim N\ (80, 250)$. The correlation structure is given as:

$$\text{H}= \begin{bmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{bmatrix}$$

Here, $\rho = 0.5$ and 0.7 stand for moderate and high correlations, respectively. The dichotomization of X_1, X_2 and X_3 is based on the following criteria

$$P(X_i = 1) = P(X_i \leq \mu_i) = 0.5.$$

Where $i = 1, 2, 3$ and 0.5 is the mean value of X_i. A population consists of $N = 1000$ observations is generated. By defining and standardizing $\mu_y = \beta_1 X_{i1} + ... + \beta_p X_{ip}$, $\theta = \beta_{true} = (2, 2, 2, 2, 2), p = 5, i = 1...N$. We generate the covariate dependent binary response y using the probability

$$\pi = \frac{1}{[1+\exp(a-b\mu_y)]}.$$

Where $a = -1$ and $b = -8$. By using the following probability, it is ensured that the missing mechanism is MAR in each variable:

$$\text{p}= 1 - \frac{\epsilon^{(-0.5-\mu_y)}}{(1+\epsilon^{(-0.5-\mu_y)})}.$$

The probability defined above yields about 20% of the observations in X_i and y to be missing (at random). R version 3.0.1 is used to perform all calculations. The packages mice, version 2.17 and NPBayesImputeCat, version 0.1 are used to perform MICE for continuous data and Non-Parametric Bayesian Multiple Imputation for categorical variables, respectively.

Table 1: Simulation settings

Perameters	Notations	Values
Population size	N	1000
No. of covariates	p	5
No.imputations	Imp.	$2, 5, 10$
Correlation	ρ	$0.5, 0.7$
Prior specifications	a_α, b_α	$0.25, 0.25$
Missing mechanism		MAR
Algorithms		CART, PMM, Default, DPMPM
No. of mixture componenta	k	80
No.simulations	Z	$50, 200$

Various numbers of imputations ($M = 2, 5, 10$) are generated using five MI methods for moderately and highly correlated simulated data. Numbers of imputations are small to facilitate beginners because manuals and descriptions for statistical software often use small number of imputations in examples whereas, large number of imputations is made for better estimates. A total of 200 simulations were made for each method. The binary response is modeled using GLM's depending on various categorical and continuous covariates. Predictive performance of the GLM's for binary response is compared using pooled AUROC curves after cross validation. The actual times taken for MI using all methods for high and moderate correlated data sets are displayed in Tables 2 and 3 respectively. Median values of pooled AUROC curves for all MI methods and different correlations are shown in Table 4. Since no noticeable differences in the posterior distributions of y are observed for different prior specifications in the similar study by Si and Reiter (2013), we limited the examination of different vague prior specifications for a_α and b_α to $(a_{\alpha=0.25}, b_{\alpha=0.25})$. The maximum number of mixture components k is set to 80 in all simulation runs. The AUROC values for moderate and high correlated, cross validated complete data sets are 98 per cents. These values can be used as benchmark (theoretical AUROC) for comparison. For moderate correlation, the predictive performance of the Hybrid MI methods is low but at least comparable to the MICE MI methods (see Figure 2). Figure 3 shows that for the highly correlated data, the Hybrid MI methods perform better than PMM and Default. The performance of H.CART is slightly less than CART. The number of multiple imputations has no significant effect on the results in the simulation study. It is noticeable, that although there is no significant difference among computational time taken for two Hybrid MI methods and Default MI method, but this difference increases when comparison is made with PMM and CART.

Table 2: Similated data $\rho = 0.7$: The time to complete M multiple imputation by variants of MI across 200 simulations

Imp.	Default	CART	PMM	H.CART	H.PMM
2	15.12m	15.74m	25.52m	13.51m	15.50m
5	36.47m	38.08m	59.48m	30.99m	36.45m
10	1.13h	1.21h	1.83h	1.04h	1.17h

Note: m = minutes and h = hours to complete multiple imputation on this subset.

Table 3: Similated data $\rho = 0.5$: The time to complete M multiple imputation by variants of MI across 200 simulationss

Imp.	Default	CART	PMM	H.CART	H.PMM
2	12.19m	16.92m	25.46m	13.79m	15.65m
5	28.84m	41.09m	59.80m	32.63m	35.40m
10	53.35m	1.34h	1.85h	1.02h	1.15h

Note: m = minutes and h = hours to complete multiple imputation on this subset.

Table 4: Simulated data: Median values of the pooled AUROC curve for various MI methods across 200 simulations

	$\rho = 0.5$					$\rho = 0.7$				
Imp.	Default	CART	PMM	H.CART	H.PMM	Default	CART	PMM	H.CART	H.PMM
2	0.9511	0.9593	0.9507	0.9282	0.9270	0.9658	0.9720	0.9662	0.9715	0.9700
5	0.9533	0.9593	0.9505	0.9286	0.9272	0.9658	0.9718	0.9660	0.9714	0.9703
10	0.9509	0.9592	0.9504	0.9288	0.9273	0.9657	0.9718	0.9662	0.9713	0.9703

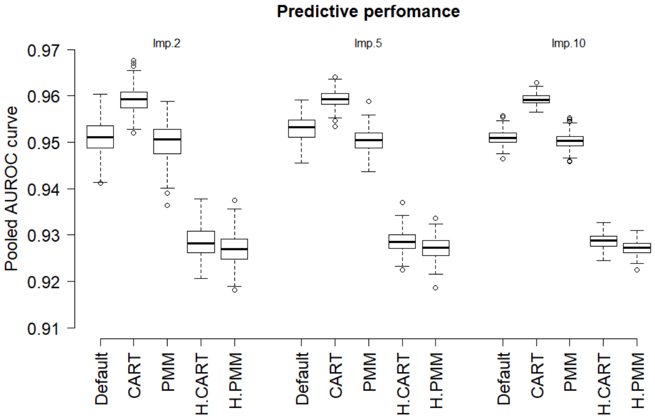


Figure 2: Simulation study: Boxplots of pooled AUROC under various MI methods for$\rho = 0.5$

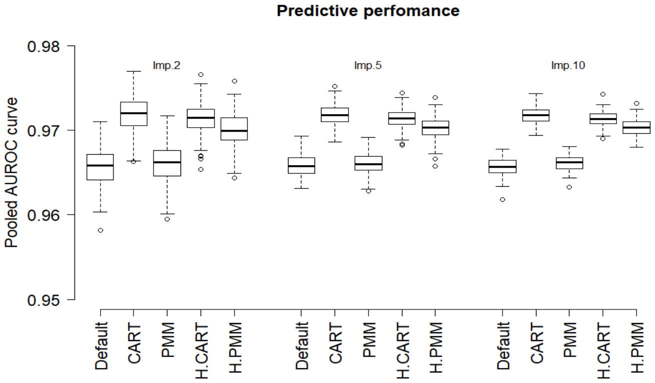


Figure 3: Simulation study: Boxplots of pooled AUROC under various MI methods for$\rho = 0.7$

6 Real data-based example: Imputation of MICS Background Variables

We use the MICS 2014 women's data as real data based example. This data contains more than 200 variables with 61286 observations around all districts of Punjab. Due to compatibility problems and for demonstration purposes, we include only forty four background variables in the analysis. Women's background characteristics like demographics, age, education, motherhood and recent births are included in this data set. The number of categorical variables is high as compared to continuous variables. According to WHO (2003), breastfeeding is important for the well-being of both child and a mother. MICS 2014 women's data can be used to determine the effect of various factors affecting feeding practices in Punjab. We treat item non response as MAR. Information on the global MICS may be obtained from mics.unicef.org and information about Bureau of Statistics, Punjab is available at bos.gop.pk. Fifty sampling simulations are run and $M = 5$ completed data sets are generated for each MI method. The binary response (Ever Breastfeed) compromising two categories (Yes / No) is modeled using the GLM's depending on various categorical and continuous covariates. The AUROC is pooled for each MI method after cross validation. Predictive performance of the GLM's for two hybrid methods is slightly less than the Default MI method and better than the remaining two, see Figure 4. Surprisingly, there is a great difference between the computational time required by the proposed and the MICE MI methods. It can be seen in Table 5 that the time taken by MICE methods is reduced from days to hours when the proposed methods are applied. The median values of the pooled AUROC curves for all methods can be seen in Table 6.

Table 5: Real data: The time to complete 5 multiple imputations by variants of MI across 50 simulationss

Imp.	Default	CART	PMM	H.CART	H.PMM
5	1.93d	1.88d	1.80d	10.78h	11.59h

Note: d = days and h = hours to complete multiple imputation on this subset.

Table 6: Real data: Median values of the pooled AUROC curve for various MI methods across 50 simulations

Imp.	Default	CART	PMM	H.CART	H.PMM
5	0.94	0.63	0.82	0.88	0.88

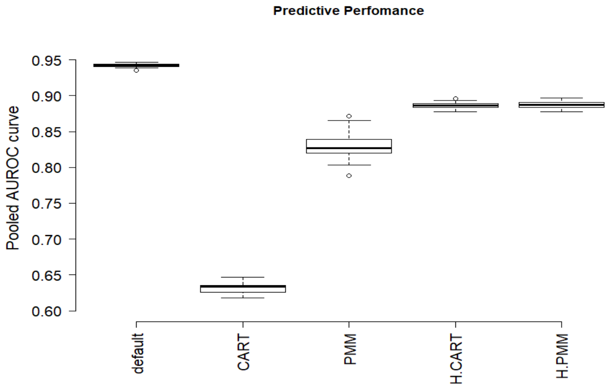


Figure 4: Real data : Boxplots of pooled AUROC obtained for 5 imputations under various MI methods

7 Concluding remarks

We proposed a computational efficient hybrid MI method. Our proposed method makes it possible to MI both types of variables (categorical with large numbers of outcomes and continuous) in survey data in the presence of complex dependencies. This method combines MI by chained equations and mixtures of multinomial. In this method, chained equations of MI continuous variables are made dependent on categorical variables MI by DPMPMs. This approach can prove to be very appropriate for a large number of variables with complex association structures especially coming from sample surveys. To implement this method no knowledge of complicated models is required. The dependence among continuous and categorical variables can be made through an easy engine. Better predictive performance with minimum computational time as compared to the existing methods is partly achieved in simulation studies. However, of note, one limitation of the proposed method is that the information available in the continuous variables is not used for imputing the categorical variables. The source of low rates of AUC for hybrid methods as compared to CART in simulation studies is still unknown. Further research for complex simulation studies, large-sample results or large number of imputations could be needed to find an answer.

8 References

M. Abdella and T. Marwala. The use of genetic algorithms and neural networks to approximate missing data in database. *IEEE 3rd International Conference on Computational Cybernetics*, 24: 207-212, 2005.

P.D. Allison. *Missing Data.* Thousand Oaks, CA: Sage Publications, 2002.

N. Ankaiah and V. Ravi. A Novel Soft Computing Hybrid for Data Imputation. In Proceedings of the 7th International Conference on Data Mining (DMIN), Las Vegas, USA, 2011.

L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone. *Classification and Regression Trees.* Wadsworth: Belmont, 1984.

D.A. Bennett. How can I Deal with Missing Data in my Study. *Australian and New Zealand Journal of Public Health*, 25:464 469, 2001.

P. Dellaportas and A.F.M. Smith. Bayesian Inference for Generalized Linear and Proportional Hazards Models via Gibbs Sampling. *Applied Statistics*, 42:443-459, 1993.

D.B. Dunson and C. Xing. Nonparametric Bayes Modeling of Multivariate Categorical Data. *Journal of the American Statistical Association*, 104:1042-1051, 2009.

H. Demirtas, A. Amatya and B. Doganay . BinNor: An R Package for Concurrent Generation of Binary and Normal Data. *Communications in Statistics - Simulation and Computation*, 43(3):569 579, 2014. A. Gelman and T. P. Speed. Characterizing a Joint Probability Distribution by Conditionals. *Journal of the Royal Statistical Society*, 55(1), 185-88, 1993.

D.G. Kleinbaum, H. Morgernstern and L.L. Kupper. Selection Bias in Epidemiological Studies. *Am. J. Epidem.*, 113:452-463, 1981.

R.J.A. Little and M.D. Schluchter. Maximum Likelihood Estimation for Mixed Continuous and Categorical Data with Missing Values. *Biometrika*, 72: 497-512, 1985.

R.J.A. Little. Missing-Data Adjustments in Large Surveys. *Journal of Business Economic Statistics*, 6: 287-296, 1988.

R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data.* Hoboken, NJ: John Wiley Sons, 2002.

B.J. McNeil and J.A. Hanley. Statistical Approaches to the Analysis of Receiver Operating Characteristic (ROC) Curves. *Med Decis Making*, 4: 137-50, 1984.

C.E. Metz. ROC Methodology in Radiological Imaging. *Invest Radiol*, 21: 720-33, 1986.

J.S. Murray and J.P. Reiter. Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. *Journal of the American Statistical Association*, 111(516), 1466-1479, 2016.

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

J. M. Robins, L.P. Zhao, A. Rotnitzky and S. Lipsitz. Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Ass.*, 89: 846-866, 1994.

D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys.* Wiley series in probability and mathematical statistics. John Wiley Sons, New York, USA, 1987.

T.E. Raghunathan, J. M. Lepkowski, J. van Hoewyk, and P. Solenberger. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27 (1), 85-95, 2001.

R Core Team. R: *A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2018. Url: https://
www.R-project.org.

J.A. Swets. ROC analysis applied to the evaluation of medical imaging techniques. *Invest Radiol*, 14:109-21, 1979.

J.L. Schafer. *Analysis of Incomplete Multivariate Data.* CRC Press, 1997. ISBN: 978-1-4398-2186-2.

J.L. Schafer and J.W. Graham. Missing data: Our view of the state of the art. *Psychological methods*, 7:147-177, 2002.

Y. Si and J.P. Reiter. Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, 38:499-521, 2013.

M. Vach and M. Schumacher. Logistic regression with incompletely observed categorical covariates: a comparison of three approaches. *Biometrika*, 80:353-362, 1993.

W. Vach and M. Blettner. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values of confounding variables. *Am. J. Epidem.*, 134: 895-907, 1991.

S.van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45: 1-67, 2011. Doi: http://dx.doi. org/10.18637/jss.v045.i03.

S. Wieand, M.H. Gail, B.R. James and K.L. James. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*, 76: 585-92, 1989.

WHO World Health Organization. *Community-based Strategies for Breastfeeding Promotion and Support in Developing Countries.* Dept. of child and adolescent health and development, Geneva, World Health Organization, 2003.

Contribution 2:
Razzak, H. and Heumann, C. (2019c): A Hybrid Technique for the Multiple Imputation of Survey Data. Revision under review at *Journal of Official Statistics*.

As the paper is still under review, the following technical report that is identical to the submitted revision is included in Chapter 3 of this thesis instead:

Razzak, H. and Heumann, C. (2020a): A Hybrid Technique for the Multiple Imputation of Survey Data. Department of Statistics (LMU Munchen): *Technical Reports*, Nr. 229, last updated 7. January 2020.

Available under: https://doi.org/10.5282/ubm/epub.70064

LMU
LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Humera Razzak, Christian Heumann

A hybrid technique for the multiple imputation of survey data

Technical Report Number 229, 2019
Department of Statistics
University of Munich

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

A hybrid technique for the multiple imputation of survey data

HUMERA RAZZAK[1] CHRISTIAN HEUMANN[2]

Abstract

Most of the background variables in MICS (multiple indicator cluster surveys) are categorical with many categories. Like many other survey data, the MICS 2014 women's data suffers from a large number of missing values. Additionally, complex dependencies may be existent among a large number of categorical variables in such surveys. The most commonly used parametric multiple imputation (MI) approaches based on log linear models or chained equations (MICE) become problematic in these situations and often the implemented algorithms fail. On the other hand, nonparametric MI techniques based on Bayesian latent class models have worked very well if only categorical variables are considered. This paper describes how chained equations MI for continuous variables can be made dependent on categorical variables which have been imputed beforehand by using latent class models. Root mean square errors (RMSEs) and coverage rates of 95% confidence intervals (CI) for generalized linear models (GLM's) with binary response are estimated in a simulation study and a comparison is made among proposed and various existing MI methods. The proposed method outperforms the MICE algorithms in most of the cases with less computational time. The results obtained by the simulation study are supported by a real data example.

Keywords: Complex dependencies; MICE; Multiple Indicator Cluster Surveys

1 Introduction

Information on many variables is collected in different large-scale surveys like Multiple Indicator Cluster Surveys (MICS). The MICS provides opportunities to fill data gaps for monitoring the health situation of children and women in under developed countries. MICS collects data on various indicators like mortality, nutrition, child and reproductive health, etc. Face to face interviews with household members are conducted to collect data. Information based on background variables of the indicators mentioned

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

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

above is very important for data analysis, and for policy making (Corsi, Perkins and Subramanian, 2017). However, the problem of missing data is inevitable in such studies. For example, the data set of individual women from MICS 2014, which has been used in the real data example latter, has a high percentage of data missing on 200 background variables. This problem arises, for example, due to item non response (INR) or entry errors etc. Beside INR, general reasons for the missing datasets include data entry errors, system failures etc. 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). In MCAR, the probability of missing data on a variable is not correlated to itself and or other measured variables. In MAR, the probability of missing depends on other, observed, variables. Finally, data are MNAR if the probability of missing depends on the variable value itself. Practically all methods implemented in software assume MAR. MNAR is called "non-ignorable", if the parameters driving the missing data process and the parameters driving the data generating process are distinct (or independent in a Bayesian analysis), but this is not further considered in the paper. Exact missing data mechanisms are often unknown when dealing with large scale data sets. Therefore, most of the time, certain assumptions are made accordingly. Li et al. (2012) addresses some problems with missing large data. Little's MCAR test proposed by Little (1988) is used commonly for testing missing data being MCAR.

The representativeness of the sample can be reduced and inferences about the population can be distorted due to missing values. Moreover, ignoring missing data can lead to a bias of unknown direction and magnitude in the estimated parameters. Therefore, it is critical to impute the data, which usually provides more accurate inference compared to ad-hoc methods (e.g. complete case (CC) analysis or single imputation) in case of missing at random (MAR) (Abdella and Marwala, 2005; Little and Rubin, 2002). The CC analysis sacrifices all units where at least the value of one variable is missing. Such methods are still very popular in psychological research (Schlomer et al., 2010). However, the CC analysis (listwise deletion) can lead to biased estimates (Little and Rubin, 2002). The CC method also results in a loss of power, which can make the analysis inefficient (Little and Rubin, 2002). Despite of being the worst available method (Wilkinson and Task Force on Statistical Inference, 1999), CC is still the most applied technique due to the simplicity and availability as default option in statistical software packages (van Ginkel, 2007). The hot-deck method is another approach and belongs to the family of single-imputation approaches. This method replaces missing values with values from a "similar" responding unit (Andridge and Little, 2010) and the empirical distribution obtained is used to draw the imputed values. In the case that the entire sample of respondents is being used as a single donor pool, this method produces consistent and unbiased estimates for missing completely at random (MCAR) data (Rubin, 1976; Little and Rubin, 2002). This method uses covariate information, avoids strong parametric assumptions and requires no careful modelling to develop selection criteria for imputing a value because it does not have any parametric model (Schafer and Graham, 2002). However, the problem with this method is that it lacks the clear criteria to guide the selection of the donor set of complete cases (Pérez et al., 2002). Bayesian bootstrap (Rubin, 1987) is a useful alternative when standard hot-deck becomes unsuitable to impute in the presence of a large number of variables (Andridge and Little, 2017). Other proposed methods for missing data use various statistical methods including self-organizing maps (SOM) (Kohonen, 1995; Oja and Kaski, 1999), k-nearest neighbour (kNN) (Batista and Monard, 2003), multi-layer perceptron (Sharpe and Solly, 1995), recurrent neural networks (Bengio and Gingras, 1995). Auto-associative neural network imputations with genetic algorithms are proposed by Pyle (1999), Narayanan et al. (2002), Chung and Merat (1996). Marseguerra and Zoia (2005) and Marwala and Chakraverty (2006) also implement some of the well-known methods used for handling missing data. Multi-task learning approaches are some other techniques based on machine learning methods (Ankaiah and Ravi, 2011). According to the studies of Horton and Kleinman (2007), Honaker et al. (2011), Royston and White (2011) and van Buuren and Groothuis-Oudshoorn (2011), over the last three decades, a wide range of variety and settings of multiple imputation (MI) techniques has been introduced for catering missing data problems in different research areas (Abdella and Marwala, 2005; Honaker et al., 2011; Little and Rubin, 2002; Schafer and Graham, 2002). MI, likelihood based analysis, and weighting approaches are alternatives to listwise and pairwise deletion methods. These methods usually make the assumption that the missing data is missing at random (MAR), hence making the estimates unbiased, consistent, and asymptotically normal (Allison, 2002; Barnard and Meng, 1999; Roth, 1994; Schafer and Graham, 2002) if that assumption holds. Model-based MI is currently considered the most popular method of addressing missing data problems. The true complete-data distribution and the missing-data mechanism form the basis of the imputation model which can be explicit or implicit by nature (Rubin, 1987). Draws from the posterior predictive distribution of the unobserved data given the observed data can be used to impute missing values. This process is repeated and *M* imputed data sets are created. By conducting the analysis on each of these data sets, the resulting *M* point and *M* variance estimates are then combined by a set of rules (Rubin, 1987). Missing values in continuous variables are often treated using a multivariate normal MI. These models are often robust to departure from normality by nature (Graham and Schafer, 1999; Schafer, 1997). Indicators in survey datasets are mostly categorical. Schafer (1997) describes that MI with log-linear models can be used to generate imputed values for such indicators by capturing the associations in the joint distribution. A severe restriction is that the number of variables must in general be small (Vermunt et al., 2008). The fully conditional specification (FCS) (van Buuren, 2007), also known as MI by chained equations (MICE) (Raghunathan et al., 2001; van Buuren, 2007) is another important tool. Missing values are sequentially imputed by estimating a series of univariate conditional models. Normal regressions and logistic or multinomial logistic regressions are used for continuous and categorical dependent variables, respectively. Alternatively, a method called predictive mean matching (PMM) can be used. Newer implementations also allow classification and regression trees (CART). MICE is an iterative method and imputes missing values variable by variable. It uses the current regression estimates for the response variable, where the response variable in this context is the actual target variable in the iterative process for which missing values are imputed. MICE assumes that equivalent, or at least nearly as good, draws for the joint distribution of the variables can be approximated by the sequential draws from the univariate conditional models. There are three main limitations or difficulties in the implementation of MICE. First, there is a possible lack of compatibility among the set of univariate conditional regression models and the joint distribution of the variables being imputed (Arnold and Press, 1989; Gelman and Speed, 1993). Although an algorithm is proposed which selects the sequence of regression models such that they are assumed to be a good fit for the data, it is very complicated to establish exact conditions for convergence (Zhu and Raghunathan, 2016). Second, the risk of overlooking higher order interactions arises when MICE includes only the main effects in the univariate conditional regression models, although using CART may resolve this problem. Third, the procedure is very time consuming when higher-order interactions are included parametrically in the model (Vermunt et al., 2008). To resolve such complications, a fully Bayesian Joint Modelling (JM) approach, called Dirichlet process mixture of products of multinomial distributions (DPMPM), is proposed by Si and Reiter (2013). This approach uses nonparametric Bayesian versions of latent class models to multiply impute high-dimensional categorical data (Vermunt et al., 2008). This approach has two stages. In stage one, a mixture of independent multinomial distributions is modelled for a contingency table of the categorical variables. In the second stage, the mixture distributions are estimated non-parametrically with Dirichlet process prior distributions. Arbitrarily complex dependencies can be described by such mixtures of multinomials. Since the computation of these dependencies is practical and generally easy, they can serve as an effective general purpose MI engine. These models have been successfully used to impute missing values in up to 80 categorical variables (Si and Reiter, 2013).Murray and Reiter (2016) have also worked on combining Dirichlet process mixtures of multinomial and multivariate normal distributions for categorical and continuous variables, but this approach involves complicated models to create the dependence structure between the continuous and the categorical variables. The R (R Core Team, 2018) package "NPBayesImputeCat" by Quanli et al. (2018) is a tool for non-parametric Bayesian JM MI, but the implementation of this package is restricted to categorical variables. Since categorical variables are internally represented as dummy variables which could easily double the actual number of predictors, the implementation of the FCS MI by chained equations algorithm becomes extremely slow or difficult in the presence of categorical variables with missing values. The R package "mice" by van Buuren and Groothuis-Oudshoorn (2011) implements MI by chained equations. Usually, household surveys based on health studies include data on a range of risk factors and health outcomes, including categorical variables with many categories mainly, and often the number of numeric variables is less as compared to categorical variables in such studies (Chandra et al., 2005; Gulliford et al., 1999). Therefore, one is limited in the choice of both MI methods, i.e. for using the former (JM), one has to sacrifice continuous variables in the analysis (or categorize them) and the latter (FCS) becomes problematic if many categorical variables are involved. Due to certain limitations, both approaches cannot be used together without correct modifications. An easy to implement hybrid technique is proposed in this paper which describes how FCS MI by chained equations for continuous variables can be blended with JM MI by latent class models for categorical variables.

The paper is organized as follows: A detailed description of a fully Bayesian, JM approach for multiple imputations of large categorical datasets is given in Section 2. In Section 3, the measures of performance used for the comparisons are described. The hybrid algorithm is described in Section 4. Section 5 compares the performance of different imputation methods in simulation studies. In Section 6, the proposed method is applied to a real data set and results are discussed. Concluding remarks are given at the end.

2 Latent class models and multiple imputation

2.1 Bayesian latent class imputation model and MI

To understand a fully Bayesian, JM approach to multiply impute large categorical datasets, it is important to understand a few details regarding how mixture models are used for density estimation and MI. The distribution of categorical data can be described by a mixture model known as latent class model (Lazarsfeld, 1950). Mixture models are considered as flexible tools which model the association structure of a set of variables (their joint density) by utilizing a finite mixture of simpler densities (McLachlan and Peel, 2000). The probability of having a specific response pattern is defined by each mixture component in a Latent Class Analysis (LCA). A weighted average of the class-specific densities generates the estimated overall density. As described by Lazarsfeld (1950), the scores of different items are independent of each other within latent classes due to local independence assumptions in LCA. A brief introduction to the mathematical form of an LC model as a tool for density estimation is given in the following: Let y_{ij} be the score of the i_{th} person on the j_{th} categorical item belonging to an $n \times J$ data-matrix Y $(i = 1, ..., n, j = 1, ..., J)$, y_i the J-dimensional vector with all scores of person i, and x_i a discrete (unobserved) latent variable with K categories. In the LC model, the joint density P (y_i; $\boldsymbol{\pi}$) has the following form:

$$\mathrm{P}\,(\boldsymbol{y_i};\boldsymbol{\pi}) = \textstyle\sum_{k=1}^{\mathrm{K}} P(x_i = k;\pi_x)\,\mathrm{P}\left(y_i|x_i = k;\pi_y\right)$$

$$= \textstyle\sum_{k=1}^{\mathrm{K}} P(x_i = k;\pi_x)\prod_{j=1}^{J}\mathrm{P}\left(y_{ij}|x_i = k;\pi_{yj}\right) \qquad (1)$$

where $\boldsymbol{\pi}=(\pi_{x,}\ \pi_y)$ is a set of LC model parameters which can be partitioned into two parts. The first part contains the latent class proportions (π_x) and the second contains class-specific item response probabilities (π_y). A separate set of parameters for each of the J items ($\boldsymbol{\pi}_{yj}$) is assigned to the second part. Due to the fact that a mixture distribution is used, a weighted sum of the K class-specific multinomial densities P$(y_i|x_i = k;\ \pi_y)$ generates the overall density. In this generation, the latent proportions are used as weights. From (1) it can be seen that the product over the J independent multinomial distributions (conditional on the *k-th* latent class) makes use of the local independence assumption. The first, second, and higher-order moments of the J response variables can be captured in LC models by setting the number of latent classes large enough (McLachlan and Peel 2000). The generated higher-order moments are actually the univariate margins, bivariate associations, and higher-order interactions when dealing with categorical variables (Vermunt et al., 2008). The unit's posterior class membership probabilities, i.e. the probability that a unit belongs to the *k*-th class given the observed data pattern $\mathrm{y_i}$, is the quantity of interest when using LC models. According to the theorem of Bayes, we can define this quantity as follows:

$$\mathrm{P}(x_i = k|y_i\,;\pi) = \frac{P(x_i = k;\pi_x)P(y_i|x_i = k;\pi_y)}{P(y_i;\pi)} \qquad (2)$$

2.2 Dirichlet process infinite mixtures of products of multinomials

The fully Bayesian, joint modeling (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 $i = 1,\ldots, n$, let $\mathrm{x}_i \epsilon \; \{1,\ldots, k\}$ indicate the class of individual *i*, and let $\pi_k{=}P\left(x_i = k\right)$. Assume further, that $\pi \; = \{\, \pi_1,\ldots, \pi_\infty\}$ is the same for all individuals. Within any class, we suppose that each of the *j* variables independently follows a class-specific multinomial distribution i.e. for any value $y_j \epsilon \; \{1,\ldots, d_j\}$ let $\text{¥}^{(j)}_{kjy} = P\left(y_{ij} = y_j \;|x_i = k\right)$. Here, d_j is the the total number of categories for the variable *j*.

Mathematically expressing the finite mixture model as

$$y_{ij}|\mathrm{x}_i\, , \text{¥} \;\; \overset{\sim}{ind} \;\; \text{Multinomial } (\text{¥}^{(j)}_{x_i 1},\ldots,\text{¥}^{(j)}_{x_i d_j}) \text{ for all } i \text{ and } j \qquad (3)$$

$$\mathrm{x}_i|\; \pi \sim \text{Multinomial } (\pi_1,\, \ldots,\, \pi_\infty) \text{ for all } i \qquad (4)$$

For prior distributions on ¥ and π , we have

$$\pi_k = \mathrm{V_k} \; (\textstyle\prod_{l<k} 1 - \mathrm{V_x}) \;\; \text{For } k{=}1,\ldots,\, \infty$$

$$\mathrm{V_w} \;\; \overset{\sim}{iid} \;\; \text{Beta } (1,\, \alpha)$$

$$\alpha \;\sim \text{Gamma } (\mathrm{a}_\alpha,\, \mathrm{b}_\alpha\,)$$

$$\text{¥}_{\mathrm{kj}} \;\sim \text{Dirichlet } (\, a_{j1}\, ,\, \ldots,\, a_{jd_{\mathrm{j}}})$$

Here $(\mathrm{a}_\alpha, \mathrm{b}_\alpha)$ and $(a_{j1},\, \ldots,\, a_{jd_{\mathrm{j}}})$ are analyst-supplied constants. Each element of $(a_{j1}\, ,\, \ldots,\, a_{jd_{\mathrm{j}}})$ is set to one in order to correspond to the uniform prior distribution. Following Dunson and Xing (2009), we set $a_\alpha = 0.25$ and $b_\alpha = 0.25$ and $k{=}80,\ 150$ and 400 as numbers for the mixture components.

3 Evaluation of performance

In order to incorporate the uncertainty introduced by missing data and the imputations into the inferences, the estimates for quantities of interest obtained by analyzing each completed dataset are combined by utilizing rules proposed by Rubin (1987). Let Q be any quantity of interest (e.g. a population proportion or a probability or a regression coefficient). For $m = 1,\ldots,M,$ let $\mathrm{q}^{(\mathrm{m})}$ and $\mathrm{u}^{(\mathrm{m})}$ be respectively the point estimate of Q in the *m*-th imputed data set with variance estimate $\mathrm{q}^{(\mathrm{m})}$. Valid inferences for a scalar Q by combining the $\mathrm{q}^{(m)}$ and $\mathrm{u}^{(m)}$ according to Rubin (1987) are obtained as follows:

$$\overline{q}_M = \textstyle\sum_{m=1}^{M} \frac{\mathrm{q}^{(m)}}{M} \;\; , \qquad (5)$$

$$b_M = \sum_{m=1}^{M} \frac{(q^{(m)} - \overline{q}_M)^2}{M-1} \,, \qquad (6)$$

$$\overline{u}_M = \sum_{m=1}^{M} \frac{u^{(m)}}{M} \,, \qquad (7)$$

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

$$T_M = \left(1 + \frac{1}{M}\right) b_M + \overline{u}_M \,, \qquad (8)$$

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

Confidence intervals can be constructed using standard multiple imputation confidence interval construction rules, which approximately follow a t-distribution. For more detail see Rubin (1996), Barnard and Meng (1999), Reiter et al. (2006), Harel and Zhou (2007).

4 Proposed hybrid architecture

Since the application of the package “NPBayesImputeCat” (Quanli et al., 2018) is limited to only categorical variables, the incomplete dataset is proposed to be partitioned into two sets, one consisting of categorical variables (*Miss.*$_{cat}$), (which MICE may not be able to impute due to reasons described in the introduction) and the other consisting of continuous variables (*Miss.*$_{num}$), where variables may be missing in both sets. A fully Bayesian JM (DPMPM) approach is used to fill in missing values by utilizing the package "NPBayesImputeCat" in *Miss.*$_{cat}$. This results in a complete version (*Imp.*$_{cat}$) of categorical variables independent of information available in the continuous variables. This complete version (*Imp.*$_{cat}$) of categorical variables can be used by MICE to construct chained equations based on categorical variables which have already been imputed by the fully Bayesian joint models to now impute the continuous variables. To achieve this, the dataset (*Miss.*$_{num}$) is added to the dataset (*Imp.*$_{cat}$) and MICE is run. This provides one completely imputed dataset where the imputations of the continuous variables obtained by FCS using chained equations depend on the information available in the imputed categorical variables. This process is repeated *M* times to obtain multiple imputed datasets using different algorithms offered by the R package “mice” (van Buuren and Groothuis-Oudshoorn, 2011) along with some prior specifications and a number of mixture components used in the R package "NPBayesImputeCat" (Quanli et al., 2018). Algorithm 1 explains the proposed hybrid architecture in detail.

Algorithm 1: Proposed hybrid architecture

```
Require: $P$ $nxp$ matrix with incomplete data
$Miss._{cat}$, $Miss._{num}$ ← Initial division of $p$ variables into factor and numeric subsets.
for $z = 1, \dots, Z$ do
    for $m = 1, \dots, M$ do
        $Imp.^{z}_{cat_m}$ ← Imputing $Miss._{cat}$ using R package “NPBayesImputeCat”.
        $Imp.^{z}_{cat_m}$ $Miss.^{z}_{num_m}$ ← Combining $Imp.^{z}_{cat_m}$ and $Miss.^{z}_{num_m}$ to generate partially imputed dataset.
        $Imp^{z}_{m}$ ← Imputing $Imp.^{z}_{cat_m}$ $Miss.^{z}_{num_m}$ using R package “mice” i.e. $f( Miss.^{z}_{num_m} \mid Imp.^{z}_{cat_m})$
        $Imp^{z}_{m}$ ← Final imputed data set.
        $\bar{q}^{(z)} \leftarrow \sum_{m=1}^{M} \frac{q^{(m)}}{M}$    Pooled point estimates[1].
        $b^{(z)} \leftarrow \sum_{m=1}^{M} \frac{(q^{(m)} - \bar{q}^{(z)})^2}{M-1}$
        $\bar{u}^{(z)} \leftarrow \sum_{m=1}^{M} \frac{u^{(m)}}{M}$
        $T^{(z)} \leftarrow \left(1 + \frac{1}{M}\right) b^{(z)} + \bar{u}^{(z)}$    Pooled variances[2].
    end for
    $\bar{q} \leftarrow \sum_{z=1}^{Z} \frac{\bar{q}^{(z)}}{Z}$    Average of pooled point estimate[3].
    $\bar{T} \leftarrow \sum_{z=1}^{Z} \frac{T^{(z)}}{Z}$    Average of pooled variance[4].
end for
```

1: $\bar{q}^{(z)}$ are pooled point estimates over M imputed datasets across z simulations.

2: $T^{(z)}$ are pooled variances over M imputed datasets across z simulations.

3: $\bar{q}$ is an average of pooled point estimates *(* $\bar{q}^{(z)}$*)* across z simulations.

4: $\bar{T}$ is an average of pooled variances *(*$T^{(z)}$*)* across z simulations.

5 Simulation studies

Simulation studies are conducted to examine the impact of MI by our proposed method. The incomplete data is generated as MAR with (known) effects and the number of categorical variables is kept more than the number of continuous variables, aiming to compare strategies in a realistic data situation.

We generate a sample of size $n=\{1000\}$ for five $(X_1, X_2, X_3, X_4, X_5)$ dimensional correlated random covariates from a multivariate normal distribution MVN. The marginal distributions of X_1, X_2, X_3, X_4, X_5 are normal and we set the mean and variance of each variable to 0 and 0.5 respectively. The correlation structure is given as:

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

where $\rho = 0.5$. The following component-wise threshold is used to transform random covariates into binary values.

$$X_i = \begin{cases} 0 & if \quad X_i \leq 0.5 \;, \\ 1 & if \quad X_i > 0.5 \;, \end{cases}$$

where $i=1, 2, 3, 4, 5$.

We than define $\mu_6 = -0.2X_1-0.3X_2+0.5X_3 -0.2X_4+ 0.22X_5$ and $\mu_7 = -2+\mu_6$. Outcomes for two continuous covariates are generated from normal distributions (ND) described as below:

$$X_6 \sim N(\mu_6; \sqrt{2}),$$

$$X_7 \sim N(\mu_7; \sqrt{2}).$$

We generate X_8 from Bernoulli distributions with probabilities governed by the logistic regression with $logit\ Pr\ (X_8) =-3+1.5X_1 -2.15X_2+2.25\ X_3+1.6X_4-1.88X_5 +1.11X_6 -0.96\ X_2X_3 +2.3\ X_1X_3+0.5\ X_2X_6 -2\ X_5\ X_6+1.21\ X_1\ X_5-2.7\ X_1X_2+1.2\ X_1X_2X_3+3\ X_6X_7.$

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

$logit\ Pr\ (y) = 0.2-0.1X_1 -0.1\ X_2-0.1\ X_3+0.3X_4 -0.5X_5 +0.2\ X_6-0.1\ X_7-0.1\ X_8$ and $\beta true = (0.2;-0.1\ ;-0.1\ ;-0.1\ ;0.3\ ;-0.5;0.2\ ;-0.1\ ;-0.1)$. We suppose that values in all covariates are missing at random with the following probabilities

$$p = 1 - \frac{e^{(-\tau-X_7)}}{(1 + e^{(-\tau-X_7)})} \;,$$

where τ is a constant. The probabilities defined above yield about 10% to 15 % of the observations in X_j to be missing (at random) for τ=-1.5 and τ= -0.5 respectively. We repeat the process 1000 times, each time generating new binary response variables and new missing patterns. We use three purely MICE based MI methods, namely classification and regression trees (CART) (Breiman, 2001), ppredictive mean matching (PMM) (Morris et al., 2014) and the Default (which uses logistic models for categorical and PMM for continuous variables). We use two Hybrid Multiple Imputation (HMI) methods e.g. H.CART and H.DEF depending on various combinations with MICE algorithms (Default and CART) and different tuning parameters (a_α, b_α; k). We further define H.CART$_1$ which is a combination of MICE.$_{\text{CART}}$ and (a_α= 0.25, b_α= 0.25, k = 80), H.CART$_2$ which is a combination of MICE.$_{\text{CART}}$ and (a_α= 0.25, b_α= 0.25, k =150) and H.CART$_3$ which is a combination of MICE.$_{\text{CART}}$ and (a_α=0.25, b_α=0.25, k =

400). Also we define H.DEF$_1$ which is a combination of MICE.$_{\text{DEF}}$ and (a_α= *0.25,* b_α= *0.25,k = 80),* H.DEF$_2$ which is a combination of MICE.$_{\text{DEF}}$ and (a_α= *0.25,* b_α= *0.25,k = 150)* and H.DEF$_3$ which is a combination of MICE.$_{\text{DEF}}$ and (a_α= *0.25,* b_α= *0.25,k = 400).* In order to achieve convergence and estimates from simulations in a reasonable time, a Gibbs sampler with 100 Markov-Chain-Monte-Carlo (MCMC) iterates is used. Two hundred iterations are run to insure convergence and to have the results of the simulations in a reasonable time when using the HMI methods. The R (R Core Team, 2018) version 3.0.1 is used to perform all calculations. The packages "mice" (van Buuren and Groothuis-Oudshoorn, 2011), version 2.17 and "NPBayesImputeCat" (Quanli et al., 2018), version 0.1 are used to perform MICE for continuous data and non-parametric Bayesian MI for categorical variables, respectively. Three sets of *M=10* imputed datasets are generated using MICE methods, i.e. MICE.$_{\text{PMM}}$, MICE.$_{\text{DEF}}$ and MICE.$_{\text{CART}}$, three sets of (*M=10)* imputed datasets are generated using H.CART$_1$*,* H.CART$_2$ and H.CART$_3$ and three sets of *(M=10)* imputed datasets are generated using H.DEF$_1$*,* H.DEF$_2$ and H.DEF$_3$. The number of multiple imputations *(M=10)* is large in order to get better estimates of standard errors. Even a higher number of *M* would have been desirable but would have led to further increased computing times. Simulated root mean square errors (RMSEs), empirical standard errors (ESEs) and coverage rates of 95% confidence intervals for generalized linear models (GLM's) with binary response and mixed covariates are estimated via combining rules described above and a comparison is made among the proposed and various existing MI methods. Tables 1-2 and Tables 3-4 display the coverage rates of 95% confidence intervals (CI) and RMSEs (ESEs) for the 10% and 15% MAR datasets, respectively, across *1000* simulations. Figures 1-2 and Figures 3-4 show boxplots of the pooled point estimates and standard errors for 10% and 15% MAR datasets, across *1000* simulations respectively.

5.1 Results

As discussed, we used two HMI methods i.e. ("H.CART" and "H.DEF") for comparison with three MICE based MI methods, i.e. ("MICE.$_{\text{DEF}}$", "MICE.$_{\text{CART}}$" and "MICE.$_{\text{PMM}}$"). In the simulation study in section 5, we generated datasets with two missing rates, i.e. 10% and 15%, using a MAR process. The HMI method "H.DEF$_1$" provides almost equal 95% CI coverage rates for the most parts and the remaining two "H.DEF" methods, i.e. ("H.DEF$_2$" and "H.DEF$_3$") provide better results for the most parts as compared to the "MICE.$_{\text{DEF}}$" and "MICE.$_{\text{PMM}}$" MI methods. This may imply that larger values for *k* have an effect on the overall performance of the "H.DEF" MI methods. All three MI methods based on "H.CART" provide better 95% CI coverage rates for the most parts as compared to "MICE.$_{\text{DEF}}$" and "MICE.$_{\text{PMM}}$", but slightly worse coverage than "MICE.$_{\text{CART}}$" for some of the simulations. Surprisingly, the coverage rates for the regression coefficient β_8 of all three “H.CART” based MI methods are higher for the 10% MAR datasets, indicating a better ability to detect complex dependency structure as compared to “MICE.$_{\text{CART}}$”. See Tables 1-2. However, we observe no such real differences in the monte carlo errors (Koehler et al., 2009). This can be due to the limited number of simulation runs used. We observe for the most parts that the between imputation variations (i.e. ESEs) for all HMI MI methods are smaller compared to “MICE.$_{\text{DEF}}$” and “MICE.$_{\text{PMM}}$” and almost equal compared to “MICE.$_{\text{CART}}$”. The amount of bias is also relatively low for the proposed HMI methods, see Tables 3-4. The average point estimates based on the proposed HMI methods are close to the corresponding true values in most of the cases, see Figures 1-2. Average standard errors based on the proposed HMI methods are also smaller for all cases as compared to the three MICE based MI methods, see Figures 3-4.

Table 1. Simulated data: 95% confidence intervals (CI) coverage rates for 10% MAR.

Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
MICE.$_{\text{PMM}}$	95	95	96	95	95	94	95	96
MICE.$_{\text{CART}}$	97	96	97	96	96	96	95	96
MICE.$_{\text{DEF}}$	95	95	96	96	95	96	95	95
H.DEF$_1$	96	96	96	94	95	95	96	96
H.CART$_1$	95	96	97	94	96	96	97	97
H.DEF$_2$	96	96	96	95	95	95	95	97
H.CART$_2$	95	96	96	94	97	95	96	96
H.DEF$_3$	96	96	96	94	95	94	95	97
H.CART$_3$	96	96	96	95	97	96	96	97

Table 2. Simulated data: 95% confidence intervals (CI) coverage rates for 15% MAR.

Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
MICE.$_{\text{PMM}}$	97	95	95	95	95	96	95	97
MICE.$_{\text{CART}}$	98	96	97	95	94	95	95	97
MICE.$_{\text{DEF}}$	94	95	95	96	96	96	96	96
H.DEF$_1$	97	97	97	96	96	96	95	98
H.CART$_1$	96	97	97	95	96	96	96	96
H.DEF$_2$	98	96	97	96	95	96	95	97
H.CART$_2$	96	96	96	95	96	96	96	97
H.DEF$_3$	98	96	96	96	95	96	96	97
H.CART$_3$	96	96	97	96	96	96	96	97

Table 3. Simulated data: RMSEs (ESEs) for 10% MAR**.**

Variables	Mice.$_{PMM}$	MICE.$_{DEF}$	MICE.$_{CART}$	H.DEF$_1$	H.CART$_1$	H.DEF$_2$	H.CART$_2$	H.DEF$_3$	H.CART$_3$
β_1	0.16(0.16)	0.16(0.16)	0.14(0.14)	0.15(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)
β_2	0.16(0.16)	0.16(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)
β_3	0.16(0.16)	0.16(0.16)	0.15(0.15)	0.16(0.16)	0.15(0.15)	0.16(0.16)	0.15(0.15)	0.16(0.16)	0.15(0.15)
β_4	0.16(0.16)	0.16(0.16)	0.15(0.15)	0.16(0.16)	0.16(0.16)	0.16(0.16)	0.16(0.16)	0.16(0.16)	0.16(0.16)
β_5	0.16(0.16)	0.16(0.16)	0.16(0.16)	0.16(0.15)	0.15(0.15)	0.16(0.15)	0.15(0.15)	0.16(0.15)	0.15(0.15)
β_6	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)
β_7	0.05(0.05)	0.04(0.04)	0.04(0.04)	0.04(0.04)	0.04(0.04)	0.05(0.05)	0.04(0.04)	0.05(0.04)	0.04(0.04)
β_8	0.19(0.19)	0.19(0.19)	0.17(0.17)	0.17(0.17)	0.17(0.16)	0.17(0.17)	0.17(0.16)	0.17(0.17)	0.17(0.16)

Table 4. Simulated data: RMSEs (ESEs) for 15% MAR**.**

Variables	Mice.$_{PMM}$	MICE.$_{DEF}$	MICE.$_{CART}$	H.DEF$_1$	H.CART$_1$	H.DEF$_2$	H.CART$_2$	H.DEF$_3$	H.CART$_3$
β_1	0.15(0.15)	0.17(0.17)	0.14(0.14)	0.14(0.14)	0.15(0.15)	0.14(0.14)	0.15(0.15)	0.14(0.14)	0.15(0.15)
β_2	0.16(0.16)	0.17(0.17)	0.15(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)
β_3	0.17(0.17)	0.17(0.17)	0.15(0.15)	0.15(0.15)	0.16(0.16)	0.15(0.15)	0.16(0.16)	0.16(0.15)	0.16(0.16)
β_4	0.16(0.16)	0.16(0.16)	0.16(0.16)	0.16(0.15)	0.16(0.16)	0.16(0.15)	0.16(0.16)	0.16(0.15)	0.16(0.16)
β_5	0.17(0.17)	0.17(0.17)	0.17(0.16)	0.16(0.16)	0.16(0.16)	0.17(0.16)	0.16(0.16)	0.17(0.16)	0.16(0.16)
β_6	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)
β_7	0.05(0.05)	0.05(0.05)	0.05(0.05)	0.05(0.05)	0.05(0.05)	0.05(0.05)	0.05(0.05)	0.05(0.05)	0.05(0.05)
β_8	0.20(0.20)	0.21(0.21)	0.18(0.17)	0.17(0.17)	0.18(0.17)	0.18(0.17)	0.18(0.17)	0.18(0.17)	0.18(0.17)

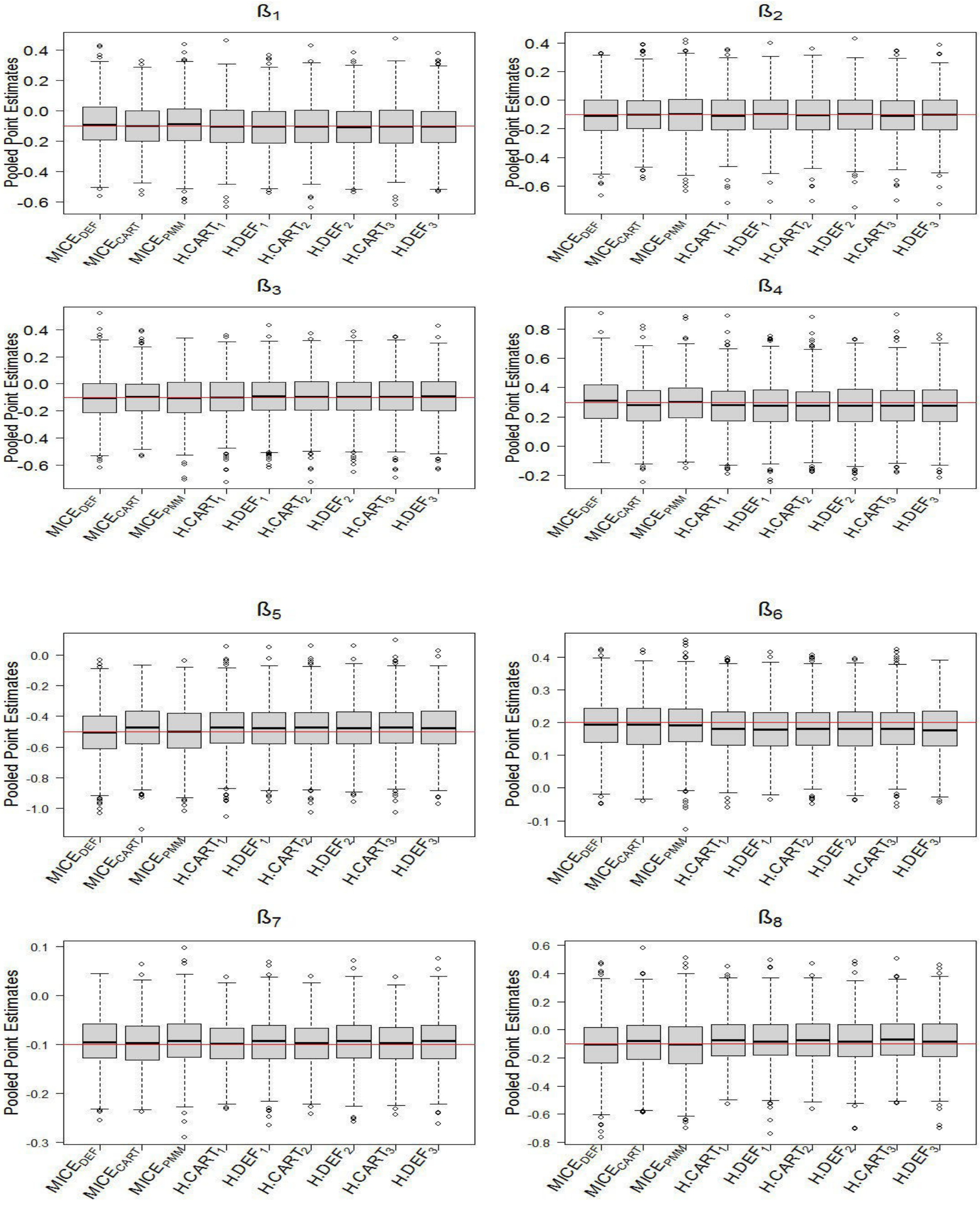


Figure 1. Simulated data: Boxplots of the pooled point estimates for 10% MAR (1000 simulations).

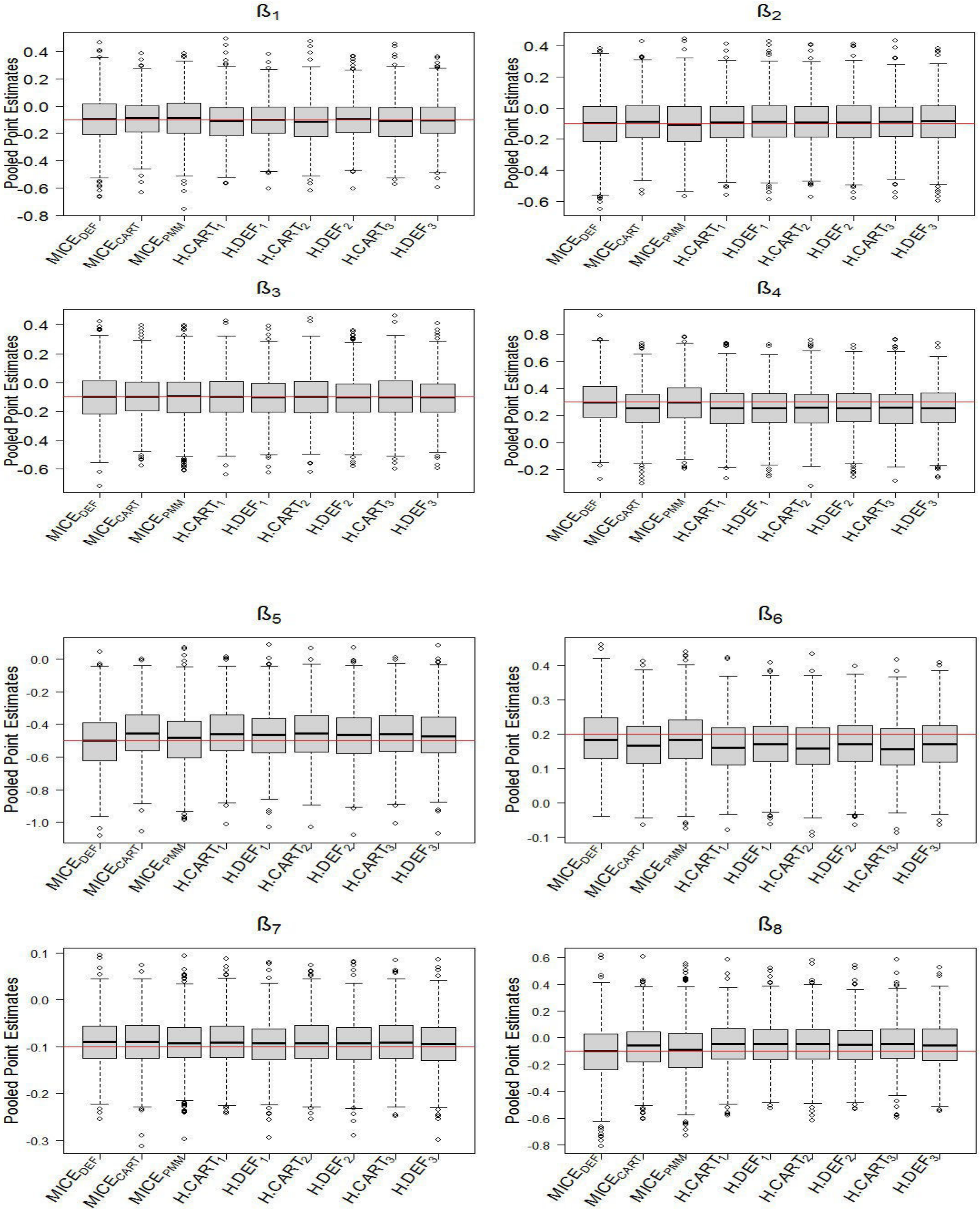


Figure 2. Simulated data: Boxplots of the pooled point estimates for 15% MAR (1000 simulations).

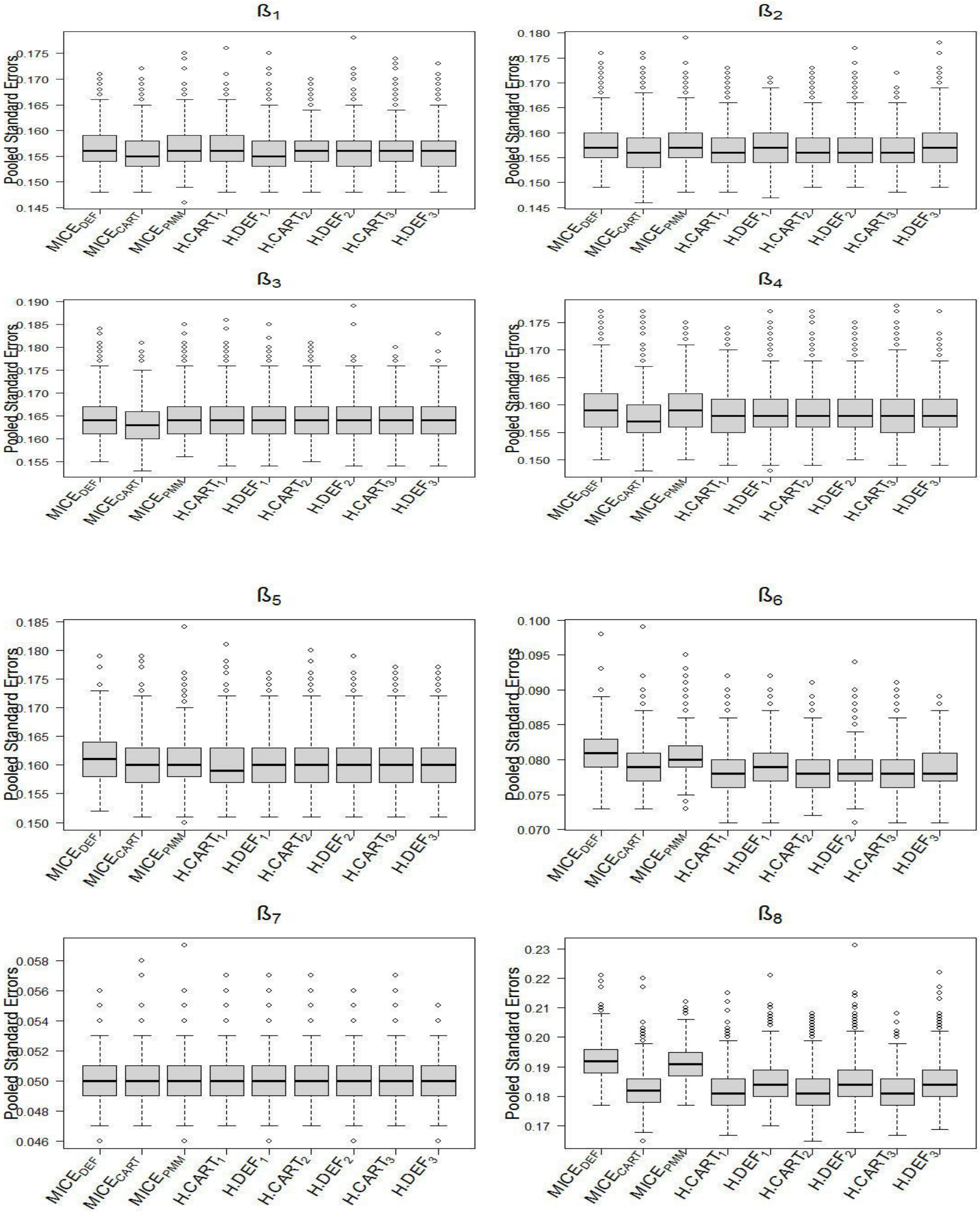


Figure 3. Simulated data: Boxplots of the pooled standard errors for 10% MAR (1000 simulations).

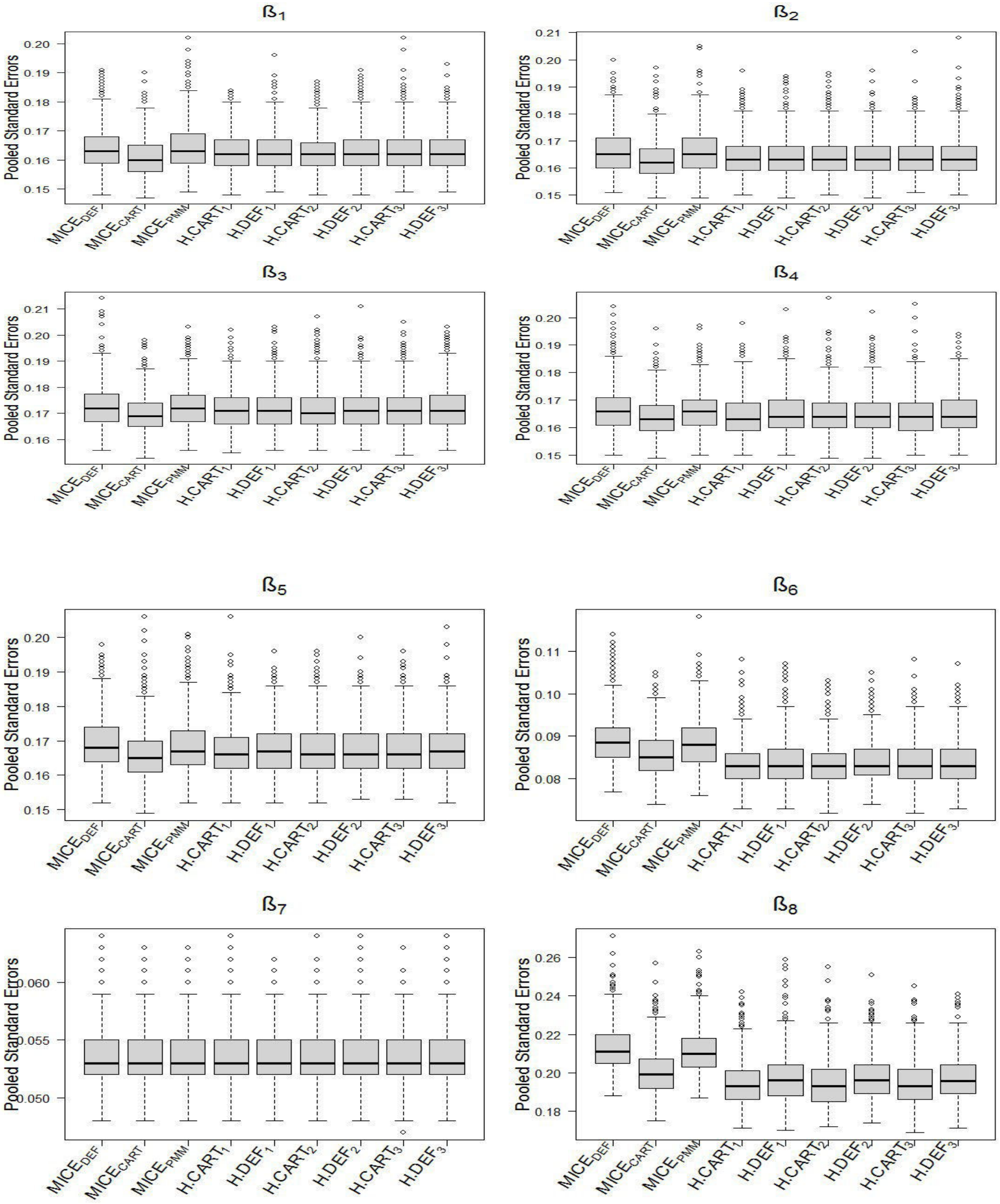


Figure 4. Simulated data: Boxplots of the pooled standard errors for 15% MAR (1000 simulations).

6 Real data-based example

6.1 Motivation

The Bureau of Statistics Punjab has conducted the Multiple Indicator Cluster Survey (MICS) Punjab, 2014, Pakistan, in collaboration with the United Nations Children's Fund (UNICEF). The Government of the Punjab has provided the major funding through the Annual Development Program 2014-15 and UNICEF has provided the annual report. The documents related to MICS Punjab consisting of the final report, key findings, survey plan, list of indicators and questionnaires can be found on the MICS website (www. http://bos.gop.pk). UNICEF in the 1990s has developed the global MICS program as an international household survey program. MICS provides support to the countries in gathering universal comparable data consisting of a wide range of indicators on the health and socio economic situation of children and women. We have used the MICS 2014 women's data that comprises more than *200* background variables on 61286 observations from 36 districts of Punjab. The data contains information of women's background characteristics like demographics, age, education, motherhood and recent births etc. Most of the background variables are categorical with lots of categories whereas few variables like age are numeric. The health benefits of breastfeeding are no longer in doubt (WHO 2003). Breastfeeding does not only contribute to the early development of a child but is also crucial for the wellbeing of the mother as well. MICS 2014 women's data can be used to determine the effect of various factors affecting the feeding practices in Punjab. This analysis could be very helpful in decision making policies regarding women and child health.

6.2 Imputation of MICS background variables

Since MICS data for women contains data with a possibly complex dependency structure, the application of the package “mice” can become problematic due to various limitations, e.g. non-convergence of the Gibbs sampler in special cases, large amount of missing values, tedious work required for specification of imputation models and interaction terms in presence of large data bases with hundreds of variables and multicollinearity problems (van Buuren and Oudshoorn, 1999). It was not possible to have a proper comparison of the proposed and existing MI approaches in such cases. Therefore, it was decided to select a subset containing *7* continuous and *37* categorical variables. The selection of variables is made according to the evidence from demographical and behavioral risk factors effecting inclination towards breastfeeding. Some of the selected categorical variables, i.e. district, has lots of categories *(k=36),* hence keeping the analysis comparable and challenging at the same time. Among these *43* variables, *5* variables have less than *14%* missing values; *16* variables have between *32* to *68* per cent missing values; *20* variables have between *80* to *95* per cent missing values. Only *2* variables are completely observed. All variables are included in the imputation model. The reasons of missing observations in MICS data are typical, i.e. nonresponse, don't know, not reached, etc. For the sake of multiple imputations, all reasons for item nonresponse are treated as MAR.

The whole process of creating imputations is repeated twenty times and *M=10* completed datasets are generated for each MI method. The binary response (Ever Breastfeed), which compromises two categories (Yes / No), is finally modeled using a GLM analysis model depending on four categorical variables (Mother Ever Attended School: two categories, Delivery by C Section: two categories, Satisfaction from Health: two categories, Area: two categories) and two continuous covariates (Age of Mother and Freq. of Mother Reads New). The R package "VIM" (Templ et al., 2012) is utilized to explore the pattern of missing values. Figure 5 displays the proportion of missing values and the missing data pattern for the variables used 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 time taken by each MI method is shown in Table 6. Boxplots of the pooled point estimates and standard errors for the real data are shown in Figures 6 and 7 respectively.

6.3 Results

Figure 5 in the real data example displays the bar plot on the left side which shows the proportions of missing values in the predictors. The categorical predictor "Delivery By C Section" has the highest amount of missing values (i.e. more than 80%) followed by "Ever Breastfeed" (about 80%), "Satisfaction From Health" (about 60%) and "Freq. of Mother Reads New" (about 40%). The amount of missing values is rather small for "Mother Ever Attended School" and "Age" (i.e. less than 20%). The categorical predictor "Area" has no missing values. An aggregation plot on the right side shows all existing combinations of missing (red) and imputed observed (blue) values. The frequencies of different combinations can be seen by a small bar plot on the right side (Templ et al., 2012). The aggregation plot reveals that if missing values occur in the variable "Ever Breastfeed", they most often also occur in the variables "Satisfaction From Health", "Freq. of Mother Reads New" and "Delivery By C Section". We note, that the standard errors for most of the coefficients are smaller relative to the (absolute) point estimates under all MI methods (see Figures 6-7). We noticed that point estimates in MICE.$_{\text{CART}}$ are nearer to the estimates in complete case analysis for most of the cases as compared to the hybrid methods (see Table 5). In the real data example, the HMI methods tend to show smaller pooled standard errors for most of the co-variates as compared to the MICE methods. We see, that when HMI MI methods are applied to the real data set, the pooled standard errors are comparatively smaller for all covariates as compared to the "MICE.$_{\text{DEF}}$" MI method and smaller for most the covariates (i.e. "Age", "Freq. of Mother

Reads New", "Delivery By C Section" and "Area") as compared to the "MICE.$_{PMM}$" MI method. "H.CART" tends to show smaller pooled standard errors for the covariates (i.e. "Age" and "Delivery by C Section") as compared to its counterparts. For the rest of the covariates, the differences are also not so high, which suggests a reasonable performance compared to MICE, see Figures 6-7. The computational burden is significantly reduced for most of the settings using the proposed HMI methods, see Table 6.

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

Variables	est	se
Age	0.14	0.06
Mother Attended School	-0.59	0.77
Freq. Mother Reads News	-0.09	0.15
Delivery by C Section	0.43	0.25
Satisfaction From Health	0.27	0.27
Area	0.16	0.25

The "est" and "se" denote the point estimates and standard errors of the coefficients of the GLM, respectively.

Table 6. Real data: Time taken by various MI methods.

Method	Time
MICE.$_{CART}$	4.20$_d$
MICE.$_{PMM}$	3.52$_d$
MICE.$_{DEF}$	3.14$_d$
H.DEF$_1$	1.70$_d$
H.CART$_1$	1.62$_d$
H.DEF$_2$	1.68$_d$
H.CART$_2$	1.64$_d$
H.DEF$_3$	1.82$_d$
H.CART$_3$	1.77$_d$

Note: Time = the time to complete 10 multiple imputation by variants of MI across 20 simulations and d = days.

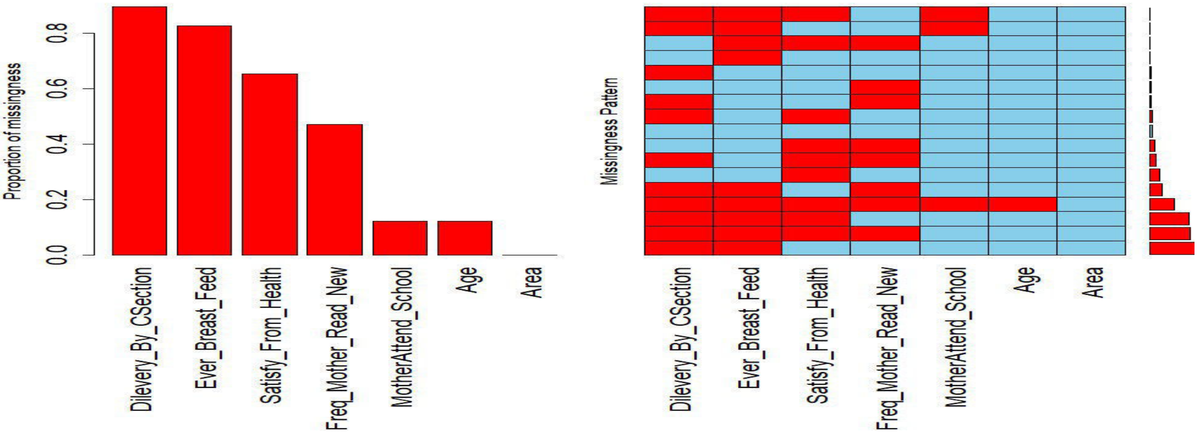


Figure 5. Aggregation graphic of the incomplete covariates.

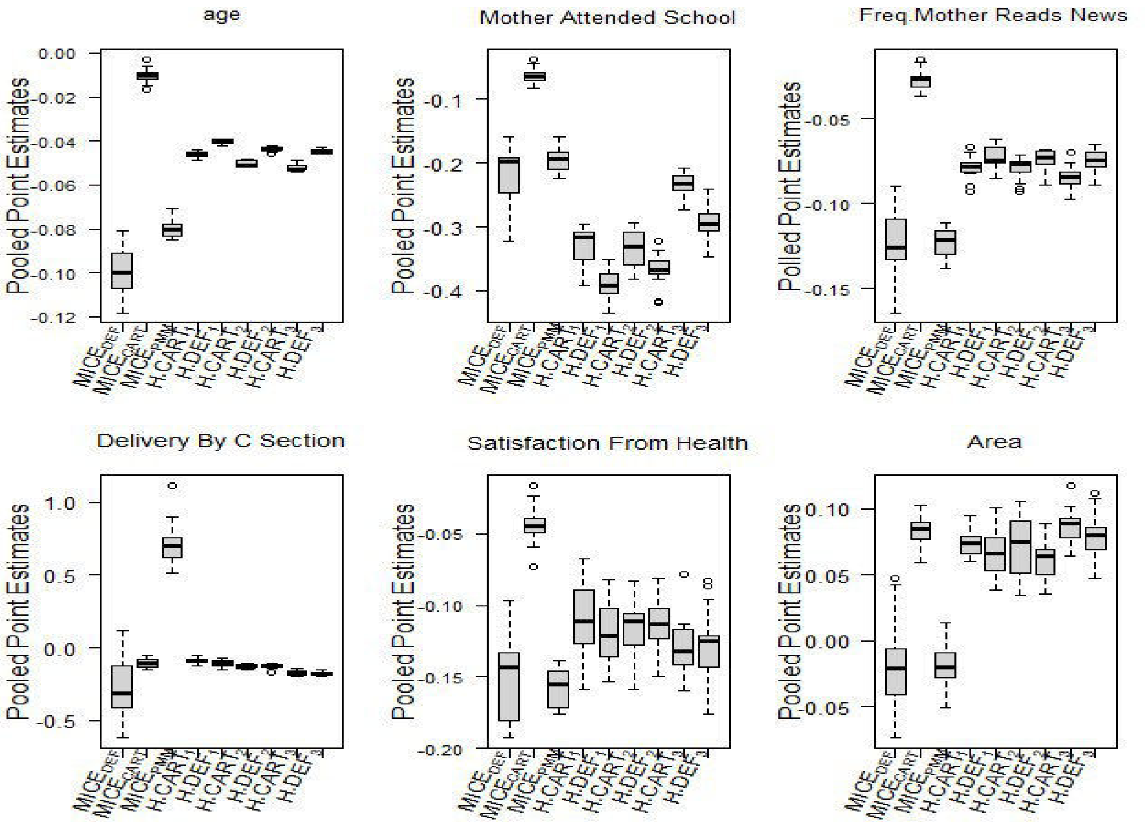


Figure 6. Real data: Boxplots of the pooled point estimates.

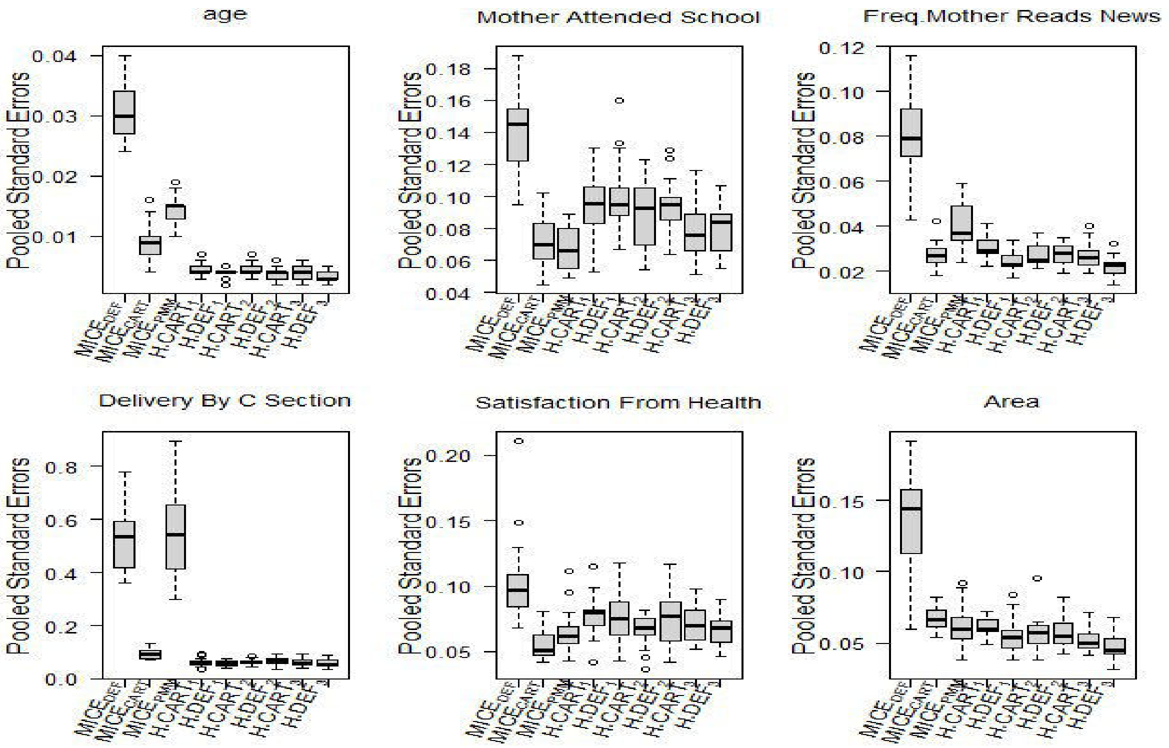


Figure 7. Real data: Boxplots of the pooled standard errors.

7 Concluding remarks

The superiority of CART and the JM technique DPMPM over the default MI methods in MICE is already established in Akande et al. (2017). Results from simulations and a real data example show that for most of the cases, hybrid techniques tend to perform better not only than the default MI methods in MICE, but also than the remaining MICE options in the presence of mixed type variables, at least for the considered GLM analysis model with binary response. The statistical properties of the proposed approach can be further studied for continuous response with mixed type covariates. In this method, chained equations used to multiply impute continuous variables are made dependent on categorical variables which have been already multiply imputed by DPMPMs through a conceptually simple method. The user can choose a set of incomplete categorical covariates that the regular MICE can sometimes fail to impute due to various restrictions, i.e. large datasets, complex dependencies, high percentage of missing data, specification of higher order interactions, multicolinearity and other instability problems. Missing values in categorical variables can be imputed by a nonparametric MI approach called DPMPMs. After filling the categorical variables, these variables are replaced in the original dataset in order to perform regular MICE. This method combines MI by chained equations and mixtures of multinomial distributions. This approach could be very appropriate for a large number of variables with complex association structures, especially coming from large sample surveys. To implement this method, no knowledge of complicated models is required. Various combinations of prior specifications and the maximal number of mixture components can be chosen together with the appropriate MICE algorithms to achieve better coverage rates and point estimates. We have observed that increasing the maximal number of mixture components tends to result in better coverage rates compared to most of the MICE methods in many cases. The proposed method is more flexible in specifying higher order interactions in the model. It also eliminates the use of predictor selection beforehand. Further comparisons can be made for data with ordinal nature and more categories with large values of prior specifications. Our proposed method is also computationally inexpensive and requires less time even when performed with a large number of iterations. Since most of the educational and health surveys contain lots of categorical and comparatively less continuous variables, various organizations can use this imputation method to create completed datasets without understanding the complexity of the dependency and model structures. However, of note, one limitation of the proposed method is, that the information available in the continuous variables is not used for imputing the categorical variables. Therefore, it is too early to make any final conclusion until unless experiments with diversity of settings are conducted.

References

Arnold, B. C. and Press, S. J. 1989. “Compatible Conditional Distributions”. *Journal of the American Statistical Association* 84:152-156.

Allison P. D. 2002. *Missing Data*. Thousand Oaks. CA: Sage Publications.

Abdella, M. and Marwala, T. 2005. “The use of genetic algorithms and neural networks to approximate missing data in database”. In Proceedings of the IEEE 3rd International Conference on Computational Cybernetics, 2005. 24: 207-212.

Ankaiah, N. and Ravi, V. 2011. “A novel soft computing hybrid for data imputation”. In Proceedings of the 7th International Conference on Data Mining (DMIN). Las Vegas. USA.

Akande, O., Li, F. and Reiter, J. 2017. “An empirical comparison of multiple imputation methods for categorical data”. *The American Statistician* 71: 162–170.

Andridge, R.R. and Little, R.J.A. 2017. “A Review of Hot Deck Imputation for Survey Non-response”. *International statistical review* 78(1): 40-64.

Bengio, Y. and Gingras, F. 1995. “Recurrent neural networks for missing or asynchronous data. In Touretzky, D.S., Mozer, M.C. and Hasselmo, M.E. editors”. *Advances in Neural Information Processing Systems* 8: 95–401. MIT Press, Cambridge, MA.

Barnard, J. and Meng, X. 1999. “Applications of multiple imputation in medical studies: From AIDS to NHANES”. *Statistical Methods in Medical Research* 8:17-36.

Breiman, L. 2001. “Random Forests”. *Machine Learning* 45(1): 5-32.

Batista, G. and Monard, M.C. 2003. *Experimental comparison of K-nearest neighbour and mean or mode imputation methods with the internal strategies used by C4.5 and CN2 to treat missing data*. University of Sao Paulo.

Chung, D. and Merat, F.L. 1996. Neural network based sensor array signal processing. In: Proc Int Conf Multisens Fusion Integr Intell Syst. Washington. USA 757–764.

Chandra, A., Martinez, G.M., Mosher, W.D., Abma, J.C. and Jones, J. 2005. “Fertility, family planning, and reproductive health of U.S. women: data from the 2002 National Survey of Family Growth”. *Vital Health Stat* 23: 1-160.

Corsi, D.J., Perkins, J.M., Subramanian, S.V. 2017. “Child anthropometry data quality from Demographic and Health Surveys, Multiple Indicator Cluster Surveys, and National Nutrition Surveys in the West Central Africa region: are we comparing apples and oranges?”. *Global Health Action* 10:1328185.

Dunson, D. B. and Xing, C. 2009. “Nonparametric Bayes modeling of multivariate categorical data”. *Journal of the American Statistical Association* 104:1042-1051.

Gelman, A. and Speed, T. P. 1993. “Characterizing a joint probability distribution by conditionals”. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 55: 85-188.

Graham, J. W. and Schafer, J. L. 1999. “On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle (Ed.)”. *Statistical strategies for small sample research* 1-29.

Gulliford, M.C., Ukoumunne, O.C. and Chinn, S. 1999. “Components of Variance and Intra class Correlations for the Design of Community-based Surveys and Intervention Studies: Data from the Health Survey for England”. *American Journal of Epidemiology* 149(9): 876-883.

Harel, O. and Zhou, X.H. 2007. “Multiple imputation: Review of theory, implementation and Software”. *Statistics in Medicine* 26: 3057-3077.

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

Honaker, J., King, G., and Blackwell, M. 2011. “Amelia II: A program for missing data”. *Journal of Statistical Software* 45(7):1-47.

Kohonen, T. 1995. *Self-Organizing Maps.* Springer. Heidelberg.

Koehler, E., Brown, E. and Haneuse, S.J. 2009. “On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses”. *The American Statisticians* 63(2):155-162.

Lazarsfeld, P. F. 1950. *The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen, Studies in social psychology in World War H: Vol. 4. Measurement and prediction (chap. 10, pp. 362-412).* Princeton, NJ: Princeton University Press.

Little, R. J. A. 1988. “A Test of Missing Completely at Random for Multivariate Data with Missing Values”. *Journal of the American Statistical Association* 83(404): 1198-1202.

Little, R. J. A. and Rubin, D. B 2002. *Statistical analysis with missing data (2nd ed.)*. New York: Wiley.

Li, F., Yu, Y., Rubin, D. B. 2012. *Imputing missing data by fully conditional models: some cautionary examples and guidelines*. Duke University Department of Statistical Science Discussion Paper 11–24.

McLachlan, G. J. and Peel, D. 2000. *Finite mixture models*. New York: Wiley.

Marseguerra, M. and Zoia, A. 2005. "The autoassociative neural network in signal analysis. II. Application to on-line monitoring of a simulated BWR component". *Annals of Nuclear Energy* 32(11):1207–1223.

Marwala, T. and Chakraverty, S. 2006. "Fault classification in structures with incomplete measured data using auto associative neural networks and genetic algorithm". *Current Science India* 90(4):542–548.

Morris, T.P., Ian, R.W. and Patrick, R. 2014. "Tuning Multiple Imputation by Predictive Mean Matching and Local Residual Draws. *BMC Medical Research Methodology* 14 (1): 75.

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.

Narayanan, S., Vian, J. L., Choi, J., El-Sharkawi, M. and Thompson, B.B. 2002. *Set constraint discovery: missing sensor data restoration using auto-associative regression machines*. In Proceedings of the international Joint Conference on Neural Networks (IJCNN). 2872–2877. Honolulu.

Oja E. and Kaski, S. 1999. *Kohonen Maps.* Elsevier. Amsterdam.

Pyle, D. 1999. *Data preparation for data mining*. Morgan Kauf- mann Publishers Inc. San Francisco.

Pérez, A., Dennis, R.J., Gil, J.F., Rondón, M.A. and López, A. 2002. "Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia". Statistics in Medicine 21:3885-3896.

Quanli, W., Danial, M.V., Reiter, J.P. and 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. 1976. "Inference and Missing Data". *Biometrika* 63: 581-590.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Roth, P. L. 1994. "Missing data: A conceptual review for applied psychologysts". *Personnel Psychology* 47: 537-560.

Rubin, D.B. 1996. "Multiple imputation after 18+ years". *Journal of the American Statistical Association* 91: 473 - 489.

Raghunathan, T.W., Lepkowksi, J.M., Van Hoewyk, J. and Solenbeger, P. A. 2001. "Multivariate technique for multiply imputing missing values using a sequence of regression models". *Survey Methodology* 27: 85-95.

Reiter, J. P., Raghunathan, T. E. and Kinney, S. 2006. "The importance of modeling the survey design in multiple imputation for missing data". *Survey Methodology* 32: 143-149.

Royston, P. and White, I.R. 2011. "Multiple imputation by chained equations (mice): Implementation in Stata". *Journal of Statistical Software* 45(4): 1-20.

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

Sharpe, P. K. and Solly, R. J. 1995. "Dealing with missing values in neural network-based diagnostic systems". *Neural Computing and Applications* 3(2):73–77.

Schafer, J. L. 1997. *Analysis of incomplete multivariate data.* London: Chapman and Hall.

Schafer, J. L. and Graham, J. W. 2002. "Missing data: Our view of the state of the art". *Psychological methods* 7:147-177.

Contribution 3:
Razzak, H. and Heumann, C. (2019d). Hybrid multiple imputation in a large scale complex survey. *Statistics in transition new series*, 20(4): 33-58.

The original publication is available under: https://doi.org/10.21307/stattrans-2019-033

STATISTICS IN TRANSITION new series, December 2019
Vol. 20, No. 4, pp. 155–170, DOI 10.21307/stattrans-2019-XXX
Submitted – XXX.2019; Paper ready for publication – XXXX.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 $g_i \epsilon \{1, \dots, k\}$ indicate the class of individual *i*, and let π_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

$$y_j \ \epsilon \ \{1, \dots, d_j\}, \text{let } \phi_{k_cj}^{(j)} = Pr(y_{ij} = y_j \,|g_i = k).$$

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

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

$$y_{ij}|\mathrm{g}_i\ ,\phi \overset{ind}{\sim} \text{Multinomial } (\phi_{g_i1}^{(j)}, \dots, \phi_{g_id_j}^{(j)}) \text{ for all } i \text{ and } j \tag{4.1}$$

$$\mathrm{g}_i|\ \pi \sim \text{Multinomial } (\pi_1, \dots, \pi_K) \text{ for all } i \tag{4.2}$$

For prior distributions on ϕ and π , we have

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

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

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

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

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

5. Proposed hybrid architecture

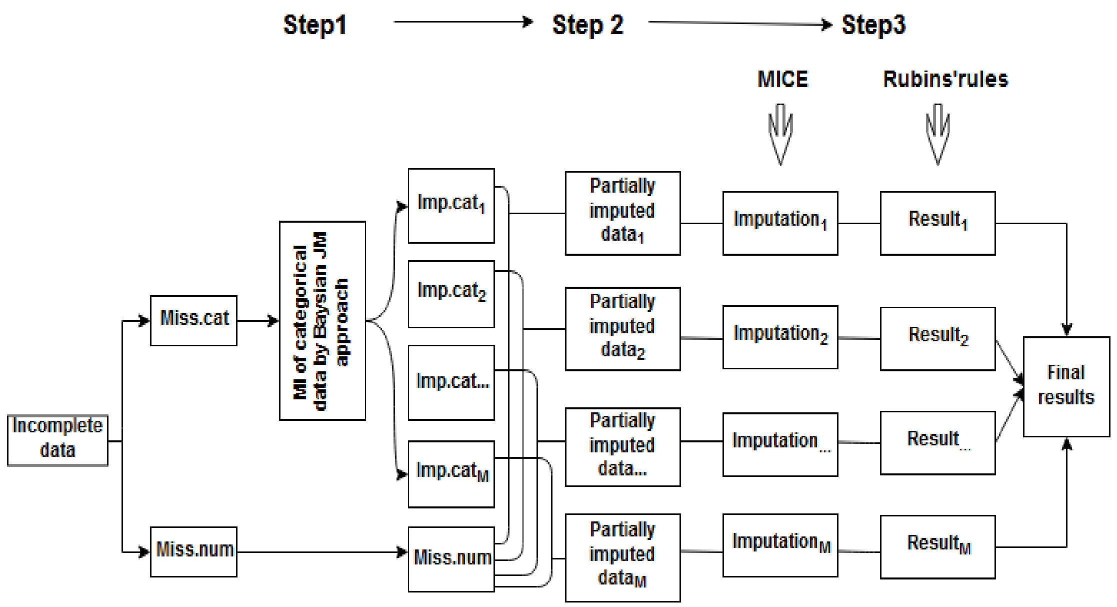


Figure 1. Schematic diagram illustrating the proposed hybrid architecture

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 “MiceCART” (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 ~ *N (0, 0.5),* where *i={1,…,31}.*The correlation structure is given as:

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

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

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

where *i*={*1,…,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}~ $N\ (\mu_1; \sqrt{2}),\ \ C_{33}$ ~ $N\left(\mu_2; \sqrt{2}\right)$, where μ_1 and μ_2 are described as:

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

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

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

$$logit\ Pr\ (X_{b_{32}}) = 0.001 - 0.01X_{b_1} - 0.09X_{b_2} - 0.09X_{b_3} - 0.09X_{b_4} + 0.05X_{b_5} + 0.08X_{b_6} - 0.02\,X_{b_7} + 0.08\,X_{b_8} + 0.01X_{b_9} + 0.01\,X_{b_{10}} - 0.02\,X_{b_{11}} + 0.01X_{b_{i12}} - X_{b_{13}} + 0.02X_{b_{14}} - 0.01X_{b_{15}} + 0.02\,X_{b_{16}} - 0.03X_{b_{17}} - 0.02X_{b_{18}} - 0.07X_{b_{19}} + 0.08X_{b_{20}} + 0.08X_{b_{21}} + 0.01X_{b_{22}} + 0.09X_{b_{23}} + 0.09X_{b_{24}} + 0.05X_{b_{25}} + 0.08X_{b_{26}} - 0.02X_{b_{27}} + 0.08X_{b_{28}} + 0.08X_{b_{29}} - 0.01X_{b_{30}} + 0.09\,X_{b_{31}} + 0.02\,C_{32} + 0.02C_{33} + 0.02\,X_{b_{12}}X_{b_{29}} - 0.02X_{b_{15}}X_{b_{18}}X_{b_{29}}\,. \qquad (6.3$$

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

$$X_{n_1}\text{~}\ N\ \left(\mu_3; \sqrt{0.5}\right).$$

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

$$X_{n_2}\text{~}\ N\ \left(\mu_4; \sqrt{0.5}\right). \qquad (6.5)$$

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

$$\mu_y = 1 + \sum_{i=1}^{32} X_{b_i} + \sum_{i=1}^{4} X_{n_i} + \sum_{i=2}^{4} X_{m_{1\_i}} + \sum_{i=2}^{6} X_{m_{2\_i}} + X_{b_9}X_{b_{15}} + X_{b_1}X_{b_{17}} + X_{b_{14}}X_{b_{20}} + \epsilon. \qquad (6.7)$$

Additionally, a random component ϵ ~ $N\ (\ 0;\ 0.5)$ 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:

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

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

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	MICECART	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)
$\bar{q}(\sqrt{T})$	$X_{b_{23}}$	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.

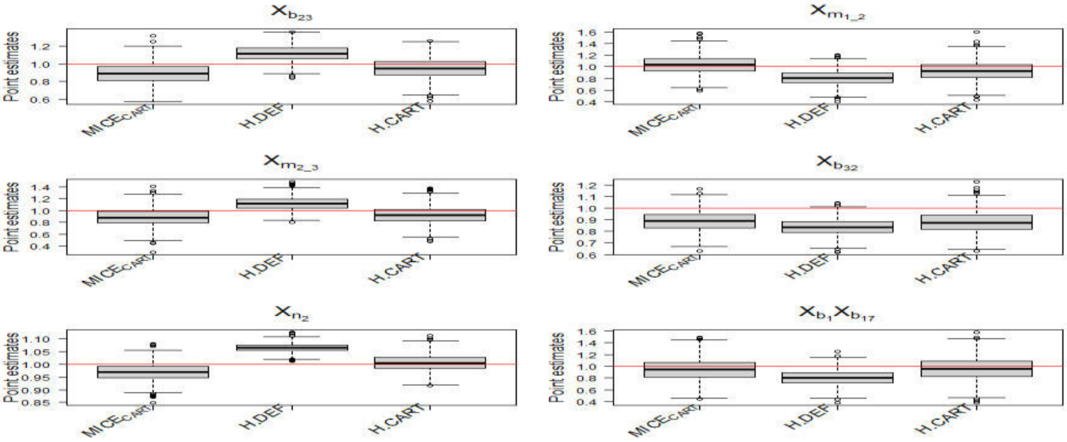


Figure 2. Simulated data: Boxplots for the point estimates $(\overline{q}_M)$ across 1000 simulations by imputation methods under Missing at Random (MAR) and ten imputations. Point estimates are shown for only six regression coefficients, i.e. for variables $X_{b_{23}}$, $X_{m_{1\_2}}$, $X_{m_{2\_3}}$,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

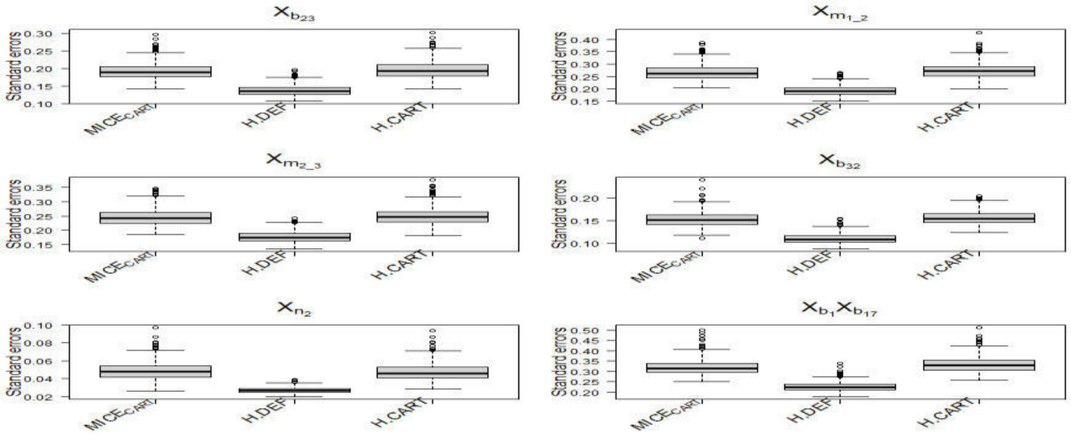


Figure 3. Simulated data: Boxplots for standard errors ($\sqrt{T_M}$) across 1000 simulations by imputation methods under Missing at Random (MAR) and ten imputations. Standard errors are shown for only six regression, i.e. $X_{b_{23}}$, $X_{m_{1\_2}}$, $X_{m_{2\_3}}$, $X_{b_{32}}$, X_{n_2} , $X_{b_1}X_{b_{17}}$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

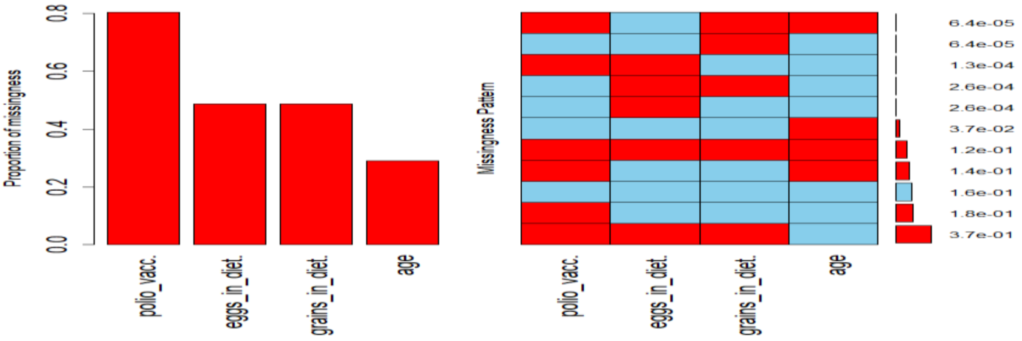


Figure 4. Real data: Aggregateplot in R, graphics of incomplete predictors. For purposes of displaying the graphical depiction, only four variables with proportions of missing values ranges between 18-28 were selected

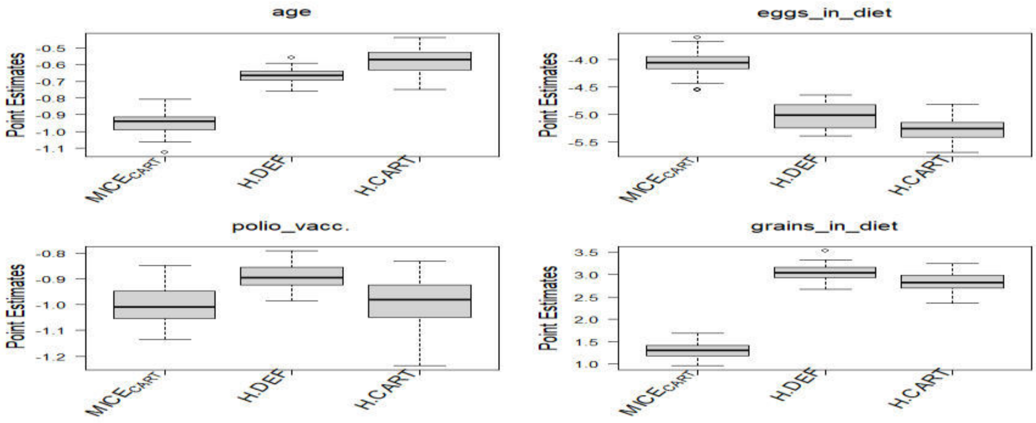


Figure 5. Real data: Boxplots for point estimates ($\overline{q}_M$) across 200 simulations by imputation methods under Missing at Random (MAR) and ten imputations

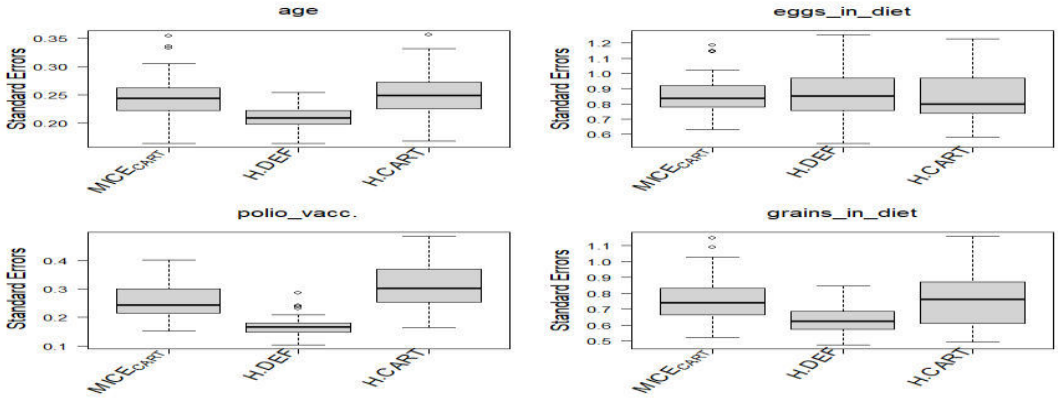


Figure 6. Real data: Boxplots for standard errors ($\sqrt{T_M}$) across 200 simulations by imputation methods under Missing at Random (MAR) and ten imputations.

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.

Contribution 4:
Razzak, H. and Heumann, C. (2019e):The Ability of Different Imputation Methods to Capture Complex Dependencies in High Dimensions. Article under review at *Romanian Statistical Review*, since 29th March 2019.

As the paper is still under review, the following technical report that is identical to the submitted revision is included in Chapter 3 of this thesis instead:

Razzak, H. and Heumann, C. (2020b):The Ability of Different Imputation Methods to Capture Complex Dependencies in High Dimensions (LMU Munchen): *Technical Reports*, Nr. 230, last updated 7. January 2020.

Available under: https://doi.org/10.5282/ubm/epub.70080

LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Humera Razzak, Christian Heumann

THE ABILITY OF DIFFERENT IMPUTATION METHODS TO CAPTURE COMPLEX DEPENDENCIES IN HIGH DIMENSIONS

Technical Report Number 230, 2019
Department of Statistics
University of Munich

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

THE ABILITY OF DIFFERENT IMPUTATION METHODS TO CAPTURE COMPLEX DEPENDENCIES IN HIGH DIMENSIONS

Humera Razzak (Humera.Razzak@stat.uni-muenchen.de)
Department of Statistics, LMU Munich.

Christian Heumann (chris@stat.uni-muenchen.de)
Department of Statistics, LMU Munich.

ABSTRACT

Multiple-imputation (MI) is a method for treating the problem of missing data. There are various competing computational algorithms available in the *R* environment to address missing data problems of categorical and continuous variables. In the case of a high amount of missing information, large sample sizes and complex dependency structures among categorical variables, the utility of the provided *R* packages is somewhat limited. A computationally expedient, fully Bayesian, joint modeling (JM) approach known as "Dirichlet process mixtures of multinomial distributions" (DPMD), automatically models complex dependencies among variables. But this approach is limited to categorical variables only. We propose a simple and easy to implement combining algorithm which imputes continuous variables using various algorithms and uses the JM approach to detect complex dependency structures among categorical variables. We review, describe and evaluate software packages commonly available in *R* and compare the results with the proposed MI method by using as example an artificial data set. The results suggest that the MI approach which combines the JM approach and various algorithms based on generalized linear models dominates various algorithms when applied solely.

Keywords: Survey data; Multiple Imputation; Complex dependencies; Hybrid; Dirichlet process prior distributions, *R* - project.

1. INTRODUCTION

Item non response is a main problem in large scale surveys. Such surveys usually have a large number of categorical variables as compared to the number of continuous variables. Using only the available data results in decreased efficiency and possibly biased inference. Rubin (1987) has proposed multiple-imputation (MI), a method for handling missing data, more than 40 years ago. For more details, see Rubin (1987) and Schafer (1997).

MI requires random draws from the posterior distribution of the missing data given the observed data. Although this method is conceptually simple but can lead to potentially unsound imputations when there are mixed type variables (i.e. continuous and categorical variables with many categories). There exist various competing computational algorithms to impute data. There is a need to investigate which of these algorithms outperform the others with respect to MI in the presence of complex dependencies among categorical variables in large scale surveys. A fully Bayesian, joint modeling approach called "Dirichlet process mixtures of multinomial distributions" (DPMD) for multiple imputation (MI) for categorical data (Si and Reiter, 2013) in large scale surveys automatically models complex dependencies while being computationally efficient at the same time. Akande et al. (2017) have compared repeated sampling properties of various MI methods for categorical data. They found that chained equations using Classification and Regression Trees (CART), and a fully Bayesian approach based on Dirichlet Process mixture models dominate the default chained equations approaches based on Generalized Linear Models (GLM's). The DPMD MI approach is limited to categorical variables; but it is possible to impute categorical variables with complex dependencies and high dimensions using DPMD and continuous variables with existing MI methods by combining two approaches. In this paper we propose a hybrid MI (HMI) approach which combines DPMD and existing MI approaches by imputing categorical variables with DPMD and use various imputation techniques to impute the continuous variables. In this paper, we compare the performance of existing and proposed MI methods in the presence of complex dependency structures among categorical variables. The judgment about the performance will be based on various dimensions, such as accuracy in comparison with the true values, point estimates and standard errors for the fitted GLM's and coverage rates of 95% confidence intervals.

2 NOTATIONS AND ASSUMPTIONS FOR THE MISSING MECHANISMS

Let D denote the incomplete data with sample size n and p variables. The distribution of D is an arbitrary multivariate distribution.
Also assume i and j refer to observations where $i=1,...,n$ and variables $j=1,...,p$, respectively. There are two components of the data set $D= \{D^{obs}, D^{miss}\}$. A response indictor matrix with same dimensions as D is

$$\mathrm{R}_{ij} = \begin{cases} 0 \; if \; v_{ij} \, is \; missing \\ 1 \; if \; v_{ij} \, is \; observed \end{cases}$$

Note that we use R in atelic for the R environment in this article. Missing Completely At Random (MCAR) is one possible assumption where $Pr(\mathrm{R}|D^{miss}, D^{obs})= Pr(\mathrm{R})$. The second possible assumption is Missing At Random (MAR) where $Pr(\mathrm{R}|D^{miss}, D^{obs})= Pr(\mathrm{R}|D^{obs})$. Missing Not At Random (MNAR) is another possible assumption where $Pr(\mathrm{R}|D^{miss}, D^{obs}) \neq Pr(\mathrm{R}|D^{obs})$ and depends on D^{miss}. The third assumption is also called non-ignorable (NI) (Little and Rubin, 2002) and not further used in the paper.

3 IMPUTATION SOFTWARE

Various imputation algorithms are implemented in a variety of statistical packages to handle missing data and to perform MI. Many standard statistical packages i.e., *R*, S-Plus, SAS, SPSS, and STATA not only implement standard algorithms but also offer user-written programs to facilitate a variety of more elaborated methods to handle missing data. Readers who are interested in the comparison of the performances of these packages are suggested to read Yu et al. (2007) or Horton and Kleiman (2007). We take *R* under consideration in this paper due to its open source character and its popularity. NA's are used to indicate missing values in *R*. There are various statistical packages that use *R* environment to impute missing data. For example "Amelia II" implements MI by bootstrapping and Expectation Maximization (EM) algorithm, "Hmisc" implements MI using additive regression and bootstrapping, *R* package "mi" offers various features (e.g. choice of predictors, models, and transformations for chained imputation models etc.) for imputations, "mice" algorithm can impute mixed type data and offer various diagnostic functions to inspect the quality of the imputations,"yaImpute" performs nearest neighbor-based imputation, "mix" performs MI for mixed categorical and continuous data, "NPBayesImpute"

impute categorical data by using Dirichlet process mixtures of multinomial distributions, “norm” uses multivariate normal model for imputations, “pan” is a MI technique for multivariate panel or clustered data. The “mitools” is a useful package to combine the results from MI whereas the package “VIM” can be utilized for exploring data and the pattern of missing values. We use “Amelia II”, “Hmisc”, “mice” and "NPBayesImpute” in our examples.

4 REVIEW OF EXISTING APPROACHES

There is a wide range of imputation models available which are based on the missingness patterns. These approaches can be categorized according to the data types. In case of a monotone missing pattern, simple methods, i.e. “propensity” (Rosenbaum and Rubin, 1983) or “Predictive Mean Matching” (PMM) (Little, 1988), are used for continuous variables. Markov Chain Monte Carlo (MCMC) approaches use Markov chains to generate random draws from multidimensional probability distributions. One can obtain a sample of the desired distribution by recording states from the chain (Gilks, 1995). MCMC approaches are suggested for complicated missingness patterns. The MCMC approach has few downsides; it is complicated and usually requires more time. Statistical packages “SAS”, “S-Plus” and “*R*” etc. use the MCMC approach. Multivariate normality assumptions apply to both the predictive mean matching and MCMC approaches (Horton and Lipsitz, 2001). According to Schafer (1997), inferences based on this normality assumption can be robust for minor departures.

Discriminant analysis or logistic regression are preferred for discrete variables for monotone missing pattern. There are a variety of imputation methods for categorical data in high dimensions. For details, see Vermunt et al. (2008). Log-linear models may be the preferred method for discrete variables, since arbitrary complex dependency structures can be modeled. But the implementation of this approach becomes difficult or impossible in high dimensions (Erosheva, et al., 2002). Naturally, there are a large number of possible models in high dimensions which makes model selection very challenging and makes it also impossible to select a model from all possible log-linear models as well. In this situation, implementation of automated model selection procedures becomes unavoidable. Moreover, model selection procedures become more complicated with missing data. Maximum likelihood estimates of the log-linear model coefficients can be biased in high dimensions (Bishop et al., 1975).

Imputation methods like fully normal (FN) imputation (Rubin and Schenker, 1986) convert categorical data to multivariate normal or continuous by applying rounding techniques. But there are evidences that the performance of these methods is limited. For example, an imputed value when made “plausible” using rounding, can tend to generate a bias and the method can fail 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). Below we discuss in detail the MI algorithms we used for comparison purposes. Advantages and disadvantages of the algorithms are discussed as well.

4.1 EXPECTATION-MAXIMIZATION WITH BOOTSTRAPPING (EMB) USED BY AMELIA II

R package called ‘Amelia II’ by Honaker et al. (2011) implements imputation method. Amelia assumes that all variables in data set are distributed multivariate normally. ‘Amelia II’ combines the bootstrap (Efron, 1979) with the EM algorithm (Dempster, Laird, and Rubin, 1977). The combination of the expectation-maximization algorithm and bootstrapping is called the Expectation-Maximization with Bootstrapping (EMB) algorithm. The bootstrapping method works by utilizing the observed sample as the pseudo-population and randomly drawing a subsample of size *n* with replacement from this observed sample. The EMB algorithm consists of the following steps: First: assuming a data set with q observed and n - q missing values**,** bootstrap samples of size *n* are drawn from incomplete data *M* times by applying bootstrapping method. Second: M point estimates of μ and Σ are calculated by applying the EM algorithm to each of these M bootstrap samples. Maximization steps are iterated until estimates converge. Finally, M multiply-imputed data sets are constructed by repeating this process M times (Wooldridge, 2002). For more details on the expectation maximization with bootstrapping (EMB) algorithm see (Schafer, 1997; Watanabe and Yamaguchi, 2000; Little and Rubin, 2002). Although EMB is computationally more efficient as compared to MCMC methods but is only an approximate Bayesian procedure (Lin, 2008).

4.2 MIXTURE MODELS FOR MULTIPLE IMPUTATION

To impute high-dimensional categorical data with significant item non-response, one has to face the challenges of model selection and estimation of log-linear models. Moreover, log-linear models and sequential regression techniques become computationally inefficient and potentially biased when the number of possible models becomes very large. Therefore, a MI technique is preferred that not only addresses these difficulties but also has a theoretical grounding as a coherent Bayesian joint model and tackles all sources of uncertainty, including parameter estimation and inference, see Rubin (1987). According to Si and Reiter (2013), Bayesian models incorporate such uncertainty automatically. They propose to use the fully Bayesian, joint modeling (JM) approach known as "Dirichlet process mixtures of products of multinomial distributions model" (DPMPM) which was originally proposed by Dunson and Xing (2009). DPMPM is a nonparametric Bayesian model for multivariate unordered categorical data. Below we describe categorical data imputation using DPMPM. A brief description is given how this approach can be combined with existing approaches through a flexible and easy to implement architecture.

Assume, we have item non-response in *n* individuals with *p* variables C_{ij} i.e. (value of variable *j* for individual i, where each i belongs to exactly one of $K < \infty$ latent classes). Further assume for *i = 1,…, N*, we have the class z_i of individual i i.e. $z_i \epsilon \{1, \dots, K\}$ with probability π_k $=Pr\ (z_i = k)$. Let π $= \{\pi_1, \dots, \pi_k\}$ be the same for all individuals. We suppose that within any class, each of the *p* variables independently follows a class-specific multinomial distribution. For any value $c_j \epsilon \{1, \dots, d_j\}$, let $¥_{klj}^{(j)} = Pr\left(C_{ij} = c_j | z_i = k\right)$. We can express the finite mixture model mathematically as $C_{ij}|z_i, ¥ \ \widetilde{ind}\ Multinomial\left(¥_{z_i1}^{(j)}, \dots, ¥_{z_id_j}^{(j)}\right)$ for all *i* and *j* and $z_i|\pi$ $\sim\ Multinomial\ (\pi_1, \dots, \pi_k)$ for all i. For prior distributions on $¥$ and π, we have $\pi_k = V_k\ (\prod_{l<k} 1 - V_l)$ for *k =1,…, K* and $V_k \sim Beta\ (1,\ \alpha)$ for *k=1,…,*$K - 1, V_k = 1$. Finally we have $\alpha \sim Gamma\ (a_\alpha\ ,\ b_\beta\ )$ and $\left(¥_{k1}^{(j)}, \dots, ¥_{kd_j}^{(j)}\right) \sim Dirichlet\ (\ a_{j1}, \dots, a_{jd_j}\ )$. In order to get complete data sets, first the latent class indicator for each individual is drawn from the full conditional and then, second, each missing C_{ij} is drawn from class-specific, independent categorical distributions.

This approach is consistent (i.e. any multivariate categorical data distribution can be approximated by DPMPM for a sufficiently large number of mixture (Dunson and Xing, 2009)), is computationally efficient and easy to code. The *R* package, "NPBayesImpute" by Manrique-Vallier et al. (2014) implements this approach. Shortcoming of this package is that it only takes categorical variables into account.

4.3 FULLY CONDITIONAL SPECIFICATION (FCS): CHAINED EQUATIONS

The FCS approach is another approach to multiple imputation. Multivariate missing data is imputed on a variable-by-variable basis. We specify a multivariate distribution $Pr(D, \text{R} \,|\, \theta)$ using a series of conditional densities $Pr(D_j \,|\, D_{-j}, \text{R}, \lambda_j)$ where λ is the unknown parameter of the imputation model. An imputation model is specified for each variable, depending on the observed values in the dataset and the response mechanism, i.e $Pr\left(D^{mis}|D^{obs}, \text{R}\right)$ in our setting. A simple draw is made using the marginal distributions first. Then imputation is repeated over the conditionally specified imputation models (van Buuren, 2012). Imputations are created for each variable iteratively. Multivariate Imputation by Chained Equations (MICE) is a prominent conditionally specified imputation model. MICE works as follows.

1 Specify an imputation model for each variable D_j

$$Pr(D_{j,miss} \,|\, D_{j,obs}, D_{-j}, \text{R}).$$

2 Let $\widetilde{D_{j,0}}$ be the starting imputation for each variable *j*. This value is e.g. obtained by making random draws from the observed values $D_{j,obs}$.

3 Repeat this process for $t=1,\dots,T$ and $j=1,\dots,p$ as well.

4 Draw $\widetilde{\lambda_{j,t}} \sim Pr(\lambda_{j,t} \,|\, D_{j,obs}, \widetilde{D_{-j,t}}, \text{R})$.

5 At the end draw imputations

$$\widetilde{D_{j,t}} \sim Pr(D_{j,miss} \,|\, D_{j,obs}, \widetilde{D_{-j,t}}, \text{R}, \widetilde{\lambda_{j,t}}).$$

MICE uses logistic or multinomial logistic regression models for categorical variables. Similar to log-linear models, these conditional models suffer from model selection and estimation problems in high dimensions. Moreover, it is very time consuming to specify many conditional models when the number of variables is large. This can result in biased estimates if default settings are used for chained equations, i.e. when we are ignoring interaction effects in the conditional models and hence fail to capture complex dependencies (Vermunt et al., 2008). The *R* Package, "mice" 2.13 (van Buuren and Groothuis-Oudshoorn, 2011) implements the FCS algorithm.

4.4 ADDITIVE REGRESSIONS, BOOTSTRAPPING AND PREDICTIVE MEAN MATCHING TECHNIQUES

Additive regressions, bootstrapping and predictive mean matching techniques for MI are implemented in the "Hmisc" package using "aregImpute" functions. A brief summary of the steps used by the "aregImpute" algorithm is as follow:

Consider *p* variables containing *m* missing observations (NAs)

1 Initial values are assigned to the NAs by drawing a random sample of size *m* from observed values. Random samples are drawn with replacement if there exist a sufficient number of NAs.

2 The observations from the variable already imputed, i.e. having no missings, are used to draw a sample with replacement for a variable containing any missing value.

3 After transforming the variable, a flexible additive model is fitted to predict this target variable.

4 This semi-parametric fitted model is used to predict the target variable in all of the original observations.

5 The target variable can be imputed either by using the observed value whose predicted transformed value is closest to the predicted transformed value of the missing value or a drawn from a multinomial distribution with probabilities derived from distance weights.

6 Repeat this process whenever predicting other missing variables with current target variable by using random draws from imputations obtained.

This approach has few downsides. Many of the multiple imputations for an observation will be identical when the predicted transformed value is closest to the predicted transformed value of the missing value. This happens when less than three variables are used to predict the target variable and implementation of PMM fails. Moreover, PMM and Bayesian predicted values will always match to same donor observation when only monotonic transformations of left and right-side variables are allowed e.g., every bootstrap resample will give predicted values of the target variable that are monotonically related to predicted values from every other bootstrap resample.

5 MI METHOD FOR COMBINING ESTIMATES

For $m = 1,\ldots, M,$ assume q and u are complete-data estimates θ and its covariance matrix Σ. Let $q^{(m)}$ and $u^{(m)}$ be respectively the point estimates of quantity of interest, Q and variance estimates of $q^{(m)}$. Valid inferences for scalar Q by combining the $q^{(m)}$ and $\mathrm{u}^{(m)}$, by Rubin (1987) are as follow.

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

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

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

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

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

with degrees of freedom $v_M = (M-1)(r^{-1})$, where $r = \frac{(1+M^{-1})b_M}{\overline{u}_M}$ represents the relative increase in the conditional variance due to the missing data (see Rubin, 1987). Confidence intervals can be constructed using standard multiple imputation confidence interval construction rules, possibly based on a t-distribution. For more details see Rubin (1996), Barnard and Meng (1999).

6 HYBRID MI (HMI) APPROACH

Implementations of fully conditional MI methods to tackle missing data can become problematic for high missing rates or when there exist complex dependencies structures among variables. For example, implementation of MICE MI become challenging when incompatibility issue arises due to high dimensions in large scale complex data (White et al., 2011; Razzak and Heumann, 2019). Such complex structures are common in high dimension household surveys where categorical variables have lots of categories i.e. District, Country etc. Moreover these methods are computationally expensive and, in some cases, less accurate as compared to full Bayesian joint models for MI (Si and Reiter, 2013). Many MI algorithms are specific for categorical variables, only, and cannot be implemented with continuous variables or require transformations (other tricks) for continuous variables (Si and Reiter, 2013). Murray and Reiter (2016) implement Bayesian mixture models with local dependence to impute both categorical and continuous values. However, combining the Dirichlet process for multinomial (discrete) mixes with the ones for multivariate (continuous) normal mixes involves knowledge of complicated models to create the dependence structure between the continuous and the categorical variables. These limitations create serious problems for researchers to obtain complete datasets with mixed type variables. We propose an easy to implement hybrid MI (HMI) approach to handle incomplete complex datasets with mixed type variables. HMI combines full Bayesian joint models (JM) MI for categorical data with various MI algorithms commonly implemented in the *R* environment.

The proposed method consists of three stages: Firstly, data instances are separated into two different groups i.e. G_{cat} and G_{num}. All categorical variables are assigned to G_{cat} and numeric ones to G_{num}. Both groups may have missing information. We impute G_{cat} using the DPMPM MI method implemented in *R* package, “NPBayesImpute” (Manrique-Vallier et al., 2014) in the second stage. Then, we combine G_{cat} and G_{num} again but this time we have missing information in $G_{num,}$ only. Lastly, we apply different algorithms to impute G_{num} based on values already imputed by DPMPM. We investigate the ability of various approaches to detect complex dependency structures in high dimensions using the HMI approach. Algorithm 1 explains HMI in detail. To assess the efficiency, we applied three well known MI methods (*R*-packages: “mice”, “Amelia” and “Hmisc”) to both groups and contrast the results with our HMI methods (“H.Amelia”, “H.MICE”, “H.Hmics”). Details of all methods are already provided in section 4 of this article. However, short descriptions of existing and hybrid methods can be seen in Table 1 and Table 2 respectively.

Table 1. Basic information: Multiple Imputation in *R*

#Method	Acronym	Description
1	Amelia II	Uses a bootstrap + EM algorithm
2	Hmisc	Uses Additive Regression, Bootstrapping and PMM algorithms
3	NPBayesImpute	Uses a fully Bayesian, joint modeling approach to multiple imputations for categorical data based on latent class models with structural zeros.
4	mice	MI using FCP

Source: Based on Manuals available on http://www.r-project.org/

Table 2 .Basic information: Hybrid Multiple Imputation (HMI) in *R*

#Method	Acronym	Description
1	H.Amelia	Amelia+NPBayesImpute
2	H.Hmisc	Hmisc+NPBayesImpute
3	H.MICE	Mice+NPBayesImpute

Source: Self-prepared.

Algorithm 1: Hybrid MI

```
Require: $n \times p$ matrix with incomplete data.
$G_{cat}, G_{num}$ ← Initial division of p variables into two factor and numeric groups
for z= 1, … ,Z do
    for $m$= 1, … ,$M$ do
        $G^{z}_{cat_m}$ ← Imputation using NPBayesImpute.
        $G^{z}_{cat_m}$ $G^{z}_{num_m}$ ← Combining $G^{z}_{cat_m}$ imputed and $G^{z}_{num_m}$ missing to generate partially imputed dataset.
        $G^{z}_{m}$ ← Imputing $G^{z}_{num_m}$missing using mice |Amelia |Hmisc |i.e. $f($ $G^{z}_{num_m}$missing | $G^{z}_{cat_m}$imputed).
        $G^{z}_{m}$ ← Final imputed data set.
        $\bar{q}^{z}_{M}$ ← $\sum_{m=1}^{M} \frac{q^{(m)}}{M}$    Pooled point estimates[1].
        $b^{z}_{M}$ ← $\sum_{m=1}^{M} \frac{(q^{(m)}-\bar{q}_{M})^2}{M-1}$
        $\bar{u}^{z}_{M}$ ← $\sum_{m=1}^{M} \frac{u^{(m)}}{M}$
        $T^{z}_{M}$ ← $\left(1+\frac{1}{M}\right) b^{z}_{M} + \bar{u}^{z}_{M}$    Pooled variances[2].
    end for
    $\bar{q}$ ← $\sum_{z=1}^{Z} \frac{\bar{q}^{z}_{M}}{Z}$    Average of pooled point estimate[3].
    $\bar{T}$ ← $\sum_{z=1}^{Z} \frac{T^{z}_{M}}{Z}$    Average of pooled variance[4].
end for
```

[1] $\bar{q}^{z}_{M}$ are pooled point estimates over *M* imputed datasets across *z* simulations.

[2] T^{z}_{M} are pooled variances over *M* imputed datasets across *z* simulations.

[3] $\bar{q}$ is an average of pooled point estimates *(*$\bar{q}^{z}_{M}$ *)* across *z* simulations.

[4] $\bar{T}$ is an average of pooled variances *(*T^{z}_{M}*)* across *z* simulations.

7 SIMULATION STUDIES

The simulation studies are inspired by Si and Reiter (2013). The data consists of $N = 1000$ observations. First, five binary variables (X_1, X_2, X_3, X_4, and X_5) are generated from a multivariate normal (MVN) distribution, followed by a categorization. The marginal distributions of X_1, X_2, X_3, X_4, X_5 are normal and we set the mean of each variable at 0 and the variance of each variable at 0.5. The correlation structure is given as:

$$\mathrm{H} = \begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}$$

Where $\rho = 0.5$. Random variates are transformed into binary values using the following threshold:

$$X_i = \begin{cases} 0 & if \quad X_i \leq 0.5 \\ 1 & if \quad X_i > 0.5 \end{cases}$$

Here $i=1, 2, 3,4,5.$

We than define $\mu_6 = 5\,X_1 - 3X_2 + 5X_3 - 4\,X_4 + X_5$ and $\mu_7 = -2+\mu_6$. Outcomes for two continuous covariates are generated from a normal distribution (ND) as described below:

$$X_6 \sim N(\mu_6; \sqrt{2}),$$
$$X_7 \sim N(\mu_7; \sqrt{2}).$$

We generate X_8 from Bernoulli distributions with probabilities governed by the logistic regression with

$logit\ Pr\ (X_8) =-1 - 1.5X_1 -1.15X_2+1.25X_3+1.6X_4 + 2.88X_5 +1.11X_6 - 1.5\ X7 -1.9\ X_2X_3 + 2.3X_1X_3 - 1.5X_2X_6 -2X_5X_6\,X7 +1.21\ X_1X_5 -2.7X_1X_2 +1.2X_1X_2\,X_3 +3X_6X_7.$

We then define a co-variate dependent binary response generated from Bernoulli distributions with probabilities governed by the logistic regression as follow:

$logit\ Pr\ (y) = 0.5 - 0.1X_1 - 0.1\ X_2 -0.1X_3 + 0.9X_4 - 0.5X_5 + 0.2\ X_6 - 0.1\ X7 - 0.5\ X_8$ and $\phi = \beta true = (0.5;-0.1;-0.1;-0.1;0.9;-0.5;0.2;-0.1;-0.5)$. We suppose that values in all covariates are MAR with the following probability

$$p = 1 - \frac{e^{(-0.001-X7)}}{(1+e^{(-0.001-X7)})}.$$

This provides around 10% of the observations in *Xi* to be missing (at random). Since Si and Reiter (2013) did not observe noticeable differences in the posterior distributions of θ for higher values of prior specifications, we set relatively small prior specification values i.e. (a_α= 0.05, b_α= 0.01) in *R* package "NPBayesImpute" version 0.6 (Manrique-Vallier et al., 2014). Akande et al. (2017) suggest that the latent classes (*k*) less than 40 can appear sufficiently large based on tuning with initial runs. However, we follow Dunson and Xing (2009) who suggest that large enough *k* can make the latent class model consistent for any joint probability distribution in case of dependencies among variable. Therefore, we set the sufficiently large number of latent classes (*k*) 80 and run each MCMC chain for 1000 iterations using the first 200 as burn-in. We implement a default version of chained equations using the "mice" software package in *R* version 2.12 (van Buuren and Oudshoorn, 1999). We implement bootstrap and PMM MI methods using 13 iterations (for convenience) with the "aregImpute" function in the "Hmisc" software package in *R* version 4.1 (Harrell, 2010). We also use the *R* package "Amelia II" version 1.6.1 (Honaker et al., 2011) with defaults as basic command. Various imputations are generated for each MI method. Five thousand sampling simulations are run.

Pooled point estimates and standard errors for the fitted GLM's with binary response are presented in figures 1, 2 ,3 and 4 for 10 and 20 imputed data sets, respectively. In order to get insight into the performance of the imputation algorithms, we make comparisons of different imputation methods using the root mean square error (RMSE) and empirical standard errors (ESE) indices, which are calculated using the following formulas:

$$\text{RMSE}_{\bar{q}_m} = \sqrt{\frac{\sum_{z=1}^{Z}\left(\bar{q}_M^z - \theta\right)^2}{Z}},$$

$$\text{ESE}_{\bar{q}_m} = \sqrt{\frac{\sum_{z=1}^{Z}\left(\bar{q}_M^z - \bar{q}\right)^2}{Z}}$$

where $\bar{q}_m$ and θ denote the estimated parameter pooled over *M* imputed data sets and original parameters, respectively. The average values of the pooled estimated parameters over the 5000 simulations are presented by $\bar{q}$. The coverage rates of at least 95% are calculated as:

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

where $1\left[\theta \in \text{CI}\left(\bar{q}_M^z,\ T_M^z\right)\right]$ is an indicator function whose value is equal to one when the confidence interval based on $\bar{q}_M^z$ and T_M^z contains θ and equal to zero otherwise.

8 SIMULATION RESULTS

As discussed, we used three software package in *R* i.e. ("Amelia","MICE" and "Hmisc") for comparison with our proposed HMI methods, i.e. ("H.Amelia","H.MICE" and "H.Hmisc"). We limited the simulation study to low missingness rates and consider 10% of values MAR, only. We also increased the number of imputations from *M=10* to *M=20* for eventually better estimates. Table 3 shows the performance of various MI methods based on estimated means RMSEs, ESEs (top) and coverage rates of 95% confidence intervals (bottom) over 5000 simulation runs. The estimated amount of bias and between imputations variation can be assesed by RMSEs and ESEs respectively. Overall, "MICE" tends to result in the most mean coverage rates concentrated around 95% and fewest high rates. The mean coverage rates for "H.MICE" tend to be larger than the mean coverage rates for "MICE", although both tend to be close to 95%. Standard "Amelia" results in coverage rates above 95% for most of the covariates. Sometimes it reaches very high rates for categorical covariates (i.e. *M* = 10: $ß_2$ and $ß_3$= 98) except one binary covariate where it reaches very low rates (i.e. *M* = 10, 20: $ß_4$ = 92). "H. Amelia" results in mean coverage rates for all covariates that are concentrated slightly above

95%, but its lower and upper tails are comparable to that of “Amelia”. “Hmisc” results in the mean coverage rates for most of the covariates that are concentrated very above 95%, it has the longest upper tail, sometimes reaching very high rates (i.e. M = 20: $ß_2$ = 98). Across the simulations, the mean coverage rates for “H.Hmisc” tend to be similar to the mean coverage rates for “Hmisc” but its upper tail is comparable to that of “Hmisc” (i.e. M = 20: $ß_2$ = 97). We observe that the estimated mean ESEs for “H.MICE” MI method are smaller for all types of covariates as compared to “MICE”, whereas “H.Hmisc” shows similar or smaller mean ESEs as compared to “Hmisc” and “H. Amelia” shows similar or slightly higher mean ESEs as compared to “Amelia” for most of the covariates. The estimated mean RMSEs for “H.MICE” MI method are smaller for most of the covariates as compared to “MICE”, whereas “H.Hmisc” have similar or slightly higher mean RMSEs as compared to “Hmisc” and Amelia” have the similar or smaller mean RMSEs as compared to Amelia” for most of the covariates. There seem to be similarities in structure among all MI methods i.e. all methods are slightly upward biased for most of the binary covariates e.g. $ß_1$, $ß_2$, $ß_3$, $ß_5$, $ß_8$ and downward biased for continues covariates and one binary covariates e.g. $ß_4$. The point estimates based on “MICE” and “H.MICE” methods are closer to the corresponding true values as compared to other methods (see Figures 1-2). Hybrid MI methods (i.e. “H.MICE”, “H.Hmisc”, “H. Amelia”) tend to have smaller standard errors as compared to their counterparts (i.e. “MICE”, “Hmisc”, “Amelia”) for most of the covariates except three binary covariates i.e. $ß_2$, $ß_5$, $ß_8$ where “H.Amelia” shows similar or slightly higher standard errors as compared to “Amelia” (see Figures 3-4).

Table 3**.** The performance of methods for MI

M =10

	Coef.	H. Hmics	Hmics	H.Amelia	Amelia	H.MICE	MICE
RMSEs(ESEs)	β_1	0.19(0.17)	0.18(0.18)	0.19(0.17)	0.18(0.16)	0.19(0.18)	0.20(0.20)
	β_2	0.17(0.17)	0.17(0.16)	0.17(0.17)	0.16(0.16)	0.18(0.17)	0.18(0.18)
	β_3	0.19(0.18)	0.19(0.18)	0.19(0.18)	0.18(0.17)	0.19(0.18)	0.20(0.20)
	β_4	0.19(0.18)	0.19(0.17)	0.19(0.17)	0.21(0.16)	0.19(0.18)	0.19(0.19)
	β_5	0.17(0.16)	0.16(0.16)	0.17(0.16)	0.16(0.15)	0.17(0.16)	0.17(0.17)
	β_6	0.27(0.23)	0.27(0.26)	0.27(0.23)	0.28(0.24)	0.28(0.24)	0.30(0.30)
	β_7	0.47(0.46)	0.47(0.47)	0.47(0.46)	0.46(0.46)	0.50(0.49)	0.51(0.51)
	β_8	0.17(0.17)	0.16(0.15)	0.17(0.17)	0.16(0.15)	0.17(0.17)	0.18(0.18)
Coverage %	β_1	96	97	96	97	96	96
	β_2	97	97	97	98	97	95
	β_3	96	96	96	98	96	95
	β_4	94	94	94	92	94	95
	β_5	96	96	96	96	96	96
	β_6	95	97	95	96	95	96
	β_7	97	97	97	97	96	95
	β_8	96	97	96	96	96	95

M = 20

	Coef.	H. Hmics	Hmics	H.Amelia	Amelia	H.MICE	MICE
RMSEs(ESEs)	β_1	0.18(0.17)	0.18(0.18)	0.18(0.17)	0.18(0.16)	0.19(0.18)	0.20(0.20)
	β_2	0.17(0.17)	0.16(0.16)	0.17(0.17)	0.16(0.16)	0.17(0.17)	0.18(0.18)
	β_3	0.18(0.18)	0.18(0.18)	0.18(0.18)	0.18(0.17)	0.19(0.18)	0.20(0.20)
	β_4	0.19(0.18)	0.19(0.17)	0.19(0.17)	0.20(0.16)	0.19(0.18)	0.19(0.19)
	β_5	0.16(0.16)	0.16(0.16)	0.17(0.16)	0.16(0.15)	0.16(0.16)	0.17(0.17)
	β_6	0.28(0.23)	0.27(0.26)	0.27(0.23)	0.28(0.24)	0.28(0.25)	0.30(0.30)
	β_7	0.46(0.46)	0.47(0.47)	0.46(0.46)	0.46(0.46)	0.49(0.49)	0.51(0.51)
	β_8	0.17(0.17)	0.16(0.15)	0.17(0.17)	0.17(0.15)	0.17(0.17)	0.18(0.18)
coverage %	β_1	96	97	96	97	96	95
	β_2	97	98	96	98	97	96
	β_3	97	97	97	97	96	95
	β_4	94	95	94	92	94	96
	β_5	96	96	96	96	96	96
	β_6	95	97	96	96	95	95
	β_7	97	97	97	97	96	95
	β_8	96	96	96	96	96	96

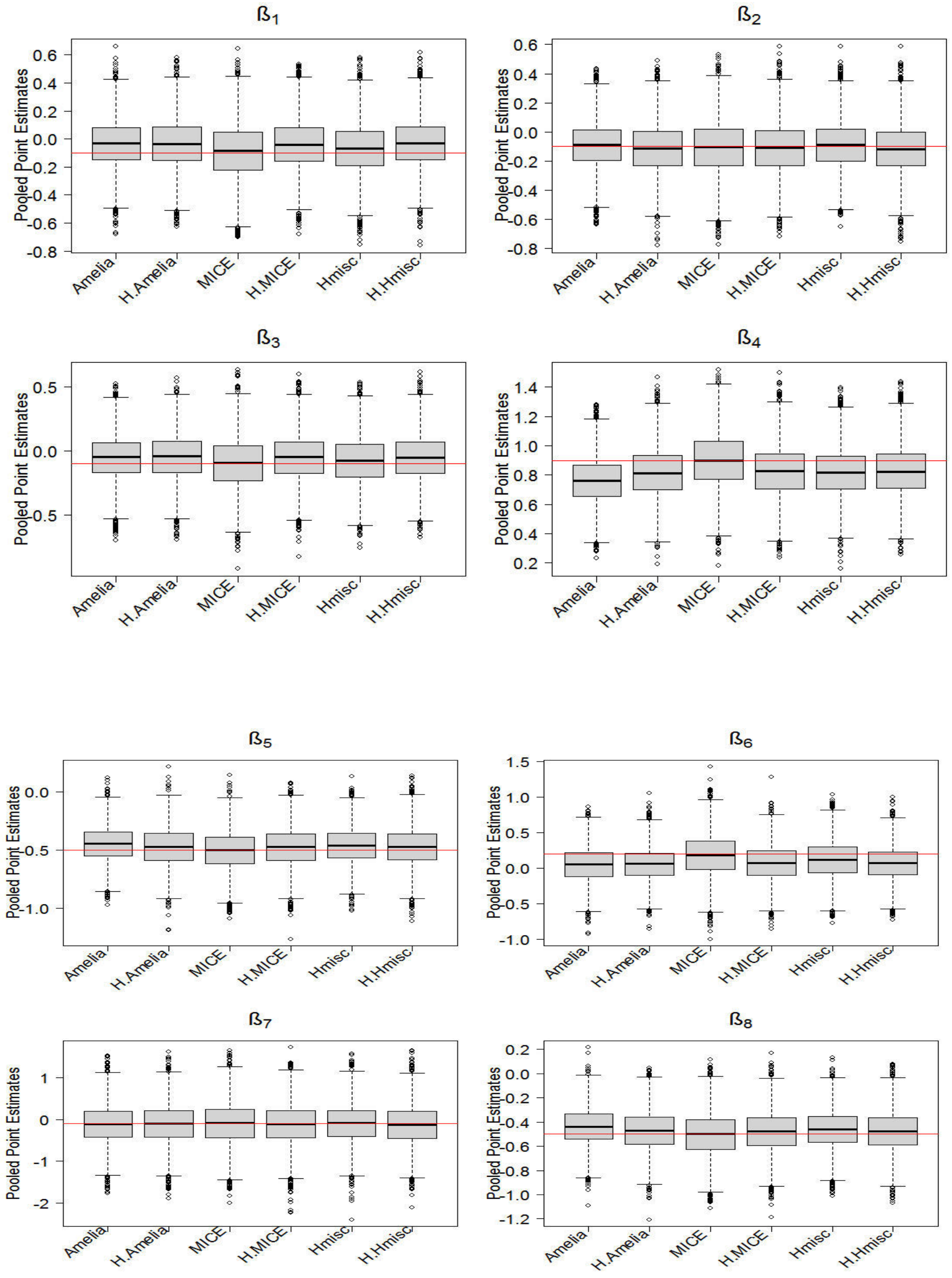


Figure 1. Boxplots for the point estimates across 5000 simulations and 10 imputations by various imputation methods.

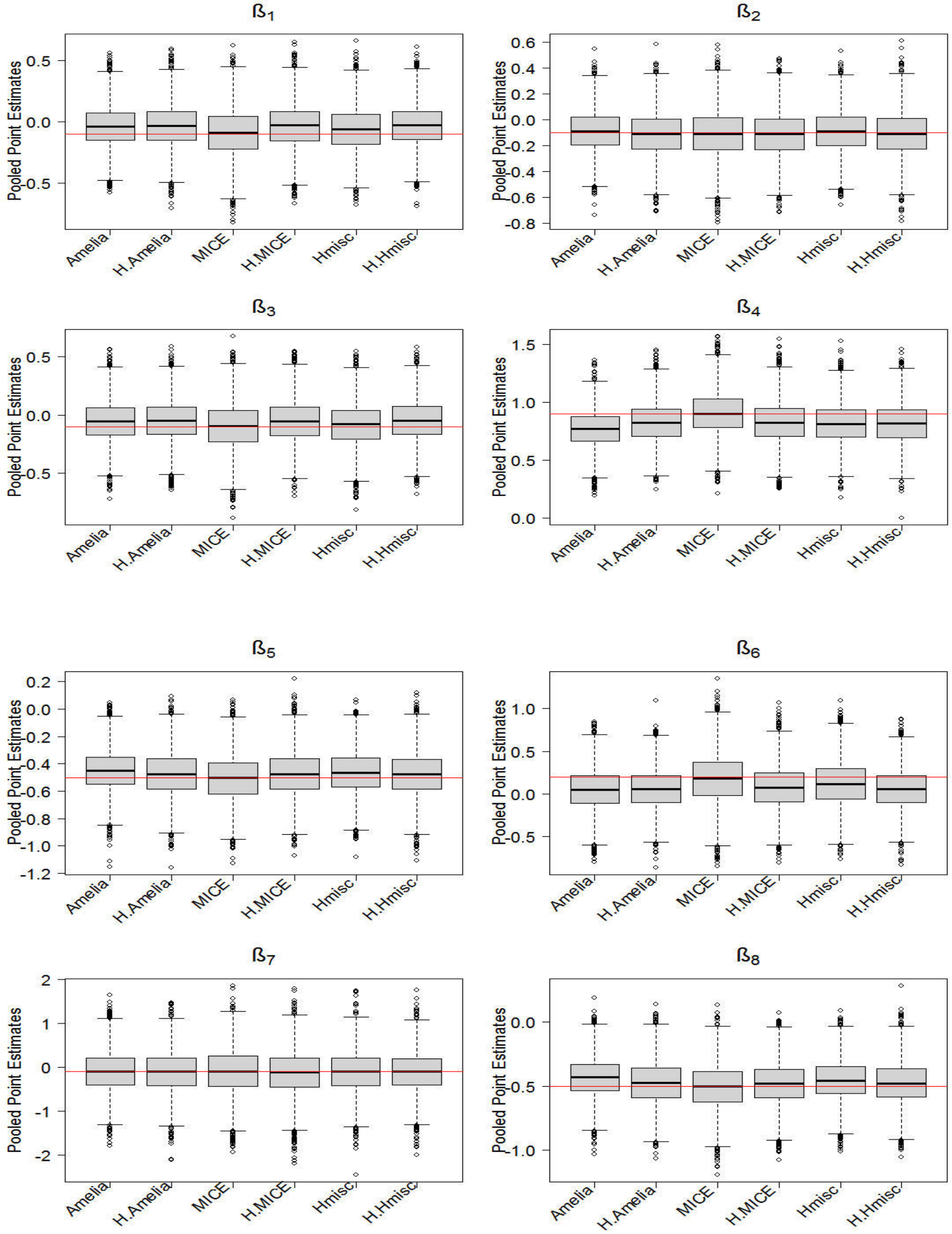


Figure 2. Boxplots for the point estimates across 5000 simulations and 20 imputations by various imputation methods.

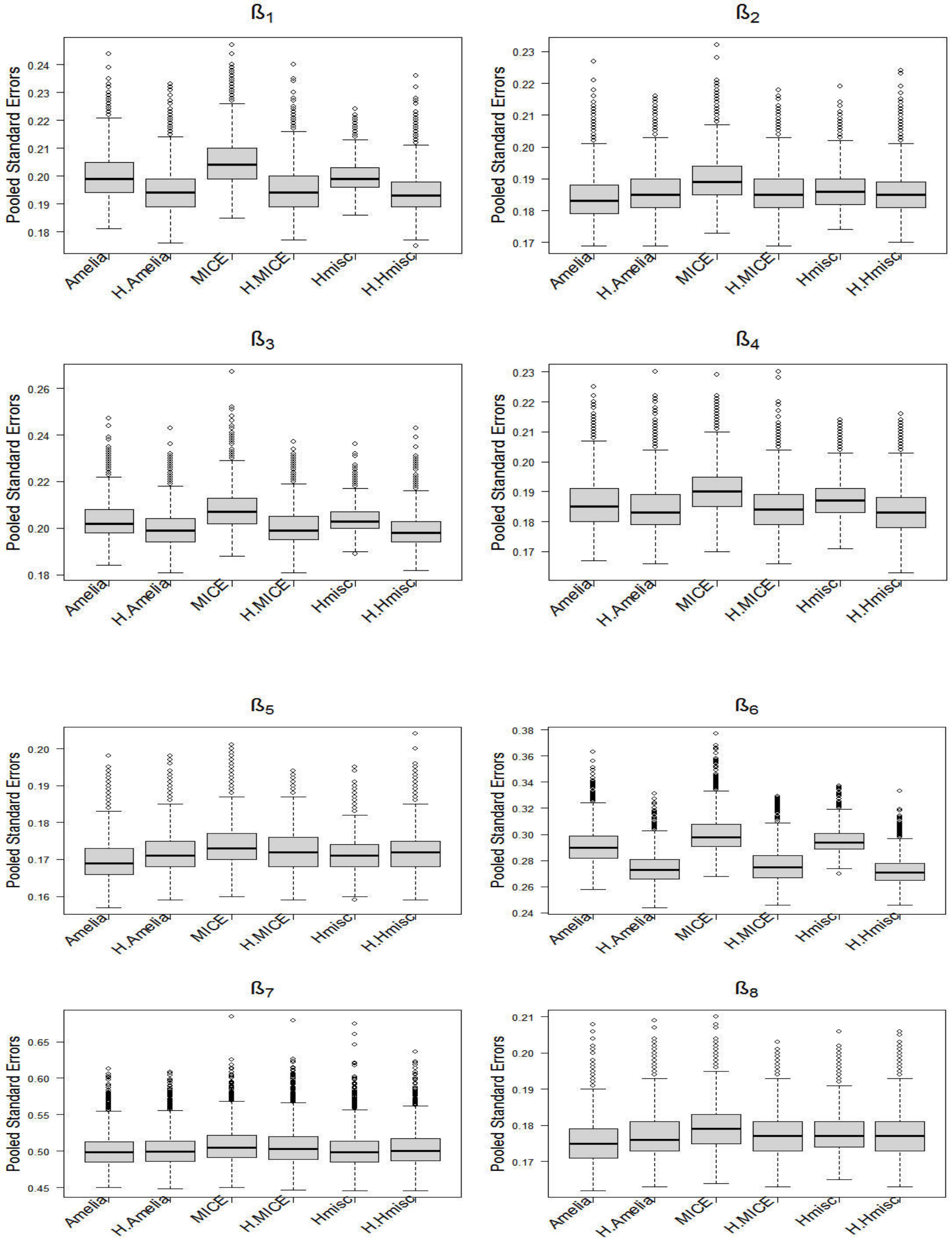


Figure 3. Boxplots for the standard errors across 5000 simulations and 10 imputations by various imputation methods.

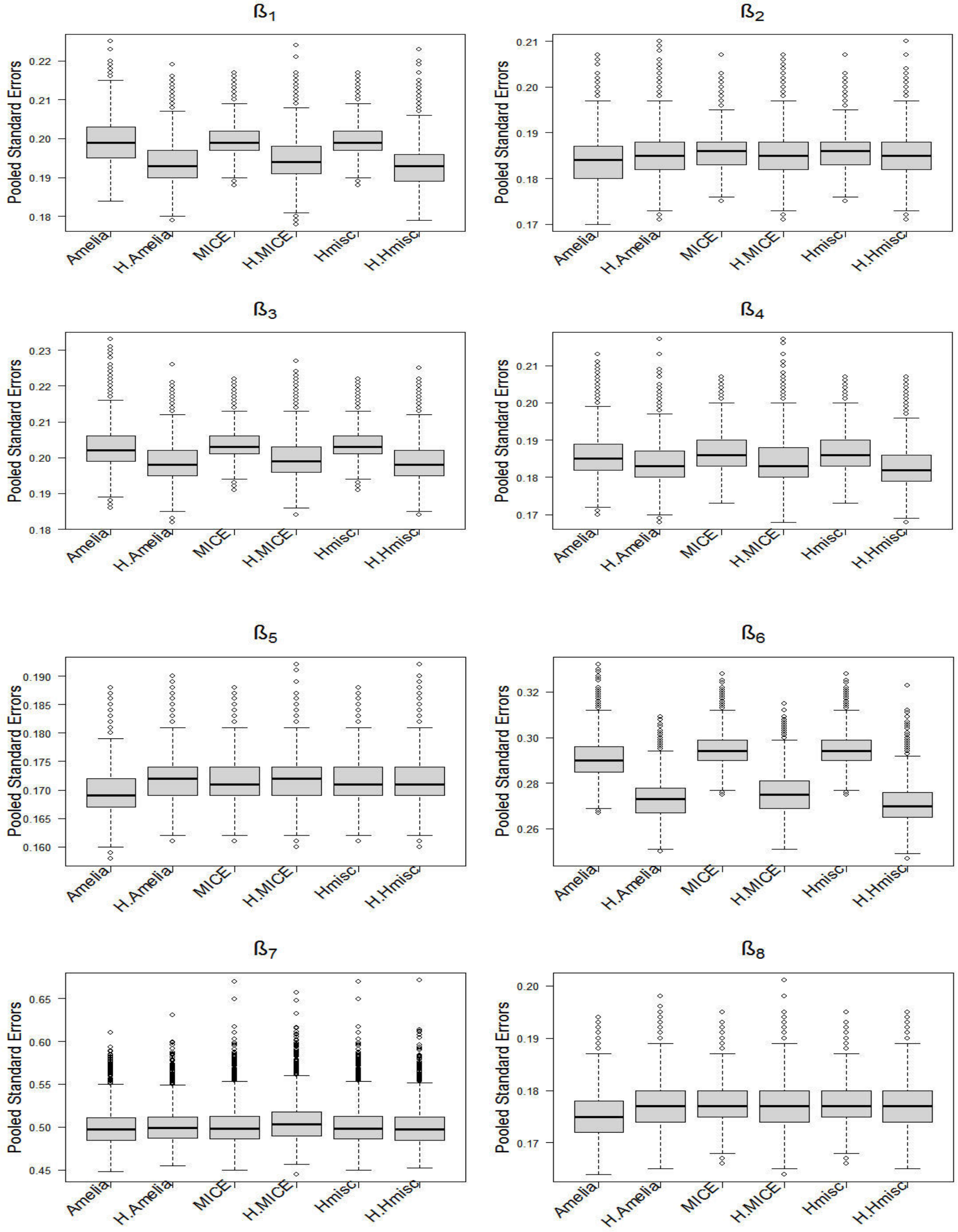


Figure 4. Boxplots for the standard errors across 5000 simulations and 20 imputations by various imputation methods.

9 CONCLUDING REMARKS

Based on results obtained by simulations, we can make several general conclusions about various MI procedures. First, the default application of "MICE", appears to be inferior to "H.MICE", overall. "H.MICE" utilizes the JM approach to identify complex dependency structures among categorical variables where missing continuous variables are imputed using the PMM technique. Of course, one could use various applications offered by MICE, (e.g. CART). Second, analysts may prefer "H.Amelia" for high coverage rates for most estimands with slight bias and due to its fastness[5]. Third, identification of a clear winner between "Hmisc" and "H.Hmisc" is little difficult. "H.Hmisc" tends to result in slightly higher mean RMSEs than "Hmisc" does, but its coverage rates are comparable that of "Hmisc". Based on results obtained by simulations, we can also make some general conclusions about three HMI procedures. Analysts concerned with getting at least nominal coverage rates for most estimands at the expense of some high mean RMSEs and ESEs, may prefer "H.MICE" over "H.Hmisc" and "H.Amelia". Simulation studies indicate that "H.Hmisc" and "H.Amelia" tend to perform in most cases. Further evaluations with diversity of experimental settings will undoubtedly be needed to account for this behavior. Increasing the number of imputed data sets improves results by reducing RMSEs. Since now, we have considered small numbers of prior specifications (a_α , b_β) and mixture components (*k)* in simulations, extensive comparisons are required for increased levels of a_α , b_β and *k.* We considered only binary response with binary and continuous covariables. Of course, statistical properties of the HMI approach can be studied for continuous response with mixed type covariates, also. Additionally, data with ordinal nature and more categories can be included for further comparisons. Real data applications can prove to be useful to see potential of proposed methods.

References

Allison, P.D. 2000. *Multiple imputation for missing data: A cautionary tale,* Sociological Methods and Research, 28, 301-309.

Ake, C.F. 2005. Rounding after multiple imputation with non-binary categorical covariates. In *Proceedings of the 13th Annual SAS Users Group International Conference.* SAS Institute Inc.

Akande, O., Li, F. and Reiter, J.P. 2017. *An Empirical Comparison of Multiple Imputation Methods for Categorical Data,* The American Statistician, 71, 162-170.

Bishop, Y., Feinberg, S. and Holland, P. 1975. *Discrete multivariate analysis: Theory and practice.* Cambridge, MA: MIT Press.

Barnard, J. and X, Meng. 1999. *Applications of multiple imputation in medical studies: From aids to nhanes*. Statistical Methods in Medical Research, 8(1), 17–36.

[5] The time taken by hybrid methods may vary depending on number of iterations and mixture components assigned i.e. it takes more time for large values of *k* and iterations. Therefore, "H.Amelia" is slower than "Amelia" but fastest then all the remaining MI methods used in analysis.

Bernaards, C.A., Belin, T.R. and Schafer, J.L. 2007. *Robustness of a multivariate normal approximation for imputation of binary incomplete data,* Statistics in Medicine, 26, 1368-1382.

Dempster, A.P., Laird, N.M. and Rubin D.B. 1977. *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society, series B, 39, 1–38.

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

Efron, B. 1979. *Bootstrap Methods: Another Look at the Jackknife*, The Annals of Statistics, 7, 1–26.

Erosheva, E. A. Fienberg, S. E. and Junker, B. W. 2002. *Alternative statistical models and representations for large sparse multi-dimensional contingency tables*, Annales de la Faculte´ des Sciences de Toulouse, 11, 485-505.

Finch, W.H. 2010. *Imputation methods for missing categorical questionnaire data: A comparison of approaches*, Journal of Data Science, 8, 361-378.

Gilks, W., Richardson, S., Spiegelhalter, D. 1996. *Markov Chain Monte Carlo in Practice.* New York: Chapman and Hall/CRC.

Graham, J.W. and 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 (pp. 1-29). Thousand Oaks, CA: Sage.

Horton, N.J. and Lipsitz, S.R. 2001. *Multiple imputation in practice: comparison of software packages for regression models with missing variables*, The American Statistician, 55(3), 244-254.

Horton, N.J., Lipsitz, S.P. and Parzen, M. 2003. *A potential for bias when rounding in multiple imputation,* The American Statistician, 57, 229-232.

Horton, N.J. and Kleinman, K.P. 2007. *Much Ado About Nothing:A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models*, The American Statistician, 6 (1), 79-90.

Honaker, J. and Gary, K. 2010. *What to do About Missing Values in Time Series Cross-Section Data*, American Journal of Political Science, 54(2), 561-581.

Honaker, J., King, G. and Blackwell, M. 2011. *Amelia II: A Program for Missing Data*, Journal of Statistical Software, 45(7), 1-47.

Little, R.J.A. 1988. *Missing-Data Adjustments in Large Surveys*, Journal of Business and Economic Statistics, 6, 287-296.

Little, R.J.A. and Rubin, D.B. 2002. *Statistical Analysis with Missing Data.* Hoboken, NJ: John Wiley Sons.

Lin, T.H. 2008. *A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data*, Quality & Quantity, 44(2), 277–287.

Manrique-Vallier, D., Reiter, J.P., Hu, J. and Quanli, W. 2014. *NPBayesImpute: Non-parametric Bayesian multiple imputation for categorical data*, The Comprehensive *R* Archive Network.

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 American Statistical Association, 111, 1466–1479.

Oba, S., Sato, M.A., Takemasa, I., Monden, M., Matsubara, K.I. and Ishii, S.A. 2003. *A bayesian missing value estimation method for gene expression profile data*, Bioinformatics, 19, 2088-96.

Rosenbaum, P.R. and Rubin, D.B. 1983. *Assessing Sensitivity to an Un-observed Binary Covariate in an Observational Study with Binary Outcome*, Journal of the Royal Statistical Society, Series B, 45, 212-218.

Rubin, D.B. and 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: Wiley.

Rubin, D.B. 1996. *Multiple Imputation After 18+ Years*. Journal of the American Statistical Association, 91(434), 473–89.

Razzak, H. and Heumann, C. 2019. *Hybrid multiple imputation in a large scale complex survey,* Statistics in Transition New Series, 20(4), 33-58.

Schafer, J.L. 1997. *Analysis of Incomplete Multivariate Data*, London: Chapman & Hall/CRC.

Si, Y. and Reiter, J. P. 2013. *Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys,* Journal of Educational and Behavioral Statistics, 38, 499-521.

van Buuren, S. and Oudshoorn, C. 1999. *Flexible multivariate imputation by MICE (Tech. rep. TNO/VGZ/PG 99.054)*, Leiden: TNO Preventie en Gezondheid.

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

van Buuren S., and Groothuis-Oudshoorn K. 2011. MICE: Multivariate Imputation by Chained Equations in R, Journal of Statistical Software, in press.

van Buuren, S. 2012. *Flexible Imputation of Missing Data*, London: Chapman & Hall/CRC.

Watanabe, M. and Kazunori Y. 2000. *EM Algorithm to Fukanzen Data no Shomondai (EM Algorithm and the Problems of Incomplete Data)*, Tokyo: Taga Shuppan.

Wooldridge, J.M. 2002. E*conometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press.

White, I.R., Royston, P. and Wood, A.M. 2011. Multiple imputation using chained equations: issues and guidance for practice, *Statistics in Medicine,* 30(4), 377–399.

Yu, L.-M., Burton, A. and Rivero-Arias, O. 2007. *Evaluation of software for multiple imputation of semi-continuous data,* Statistical Methods in Medical Research, 16, 243-258.

Yucel, R.M., He, Y. and Zaslavsky, A.M. 2011. *Gaussian-based routines to impute categorical variables in health surveys*, Statistics in Medicine, 30, 3447-3460.

Contribution 5:
Razzak, H. and Heumann, C. (2019f):Application of Iterative Hybrid MI Approach to Household Survey Data with Complex Dependency Structures. Article under review at *Communications in Statistics - Simulation and Computation.*

As the paper is still under review, the following technical report that is identical to the submitted revision is included in Chapter 3 of this thesis instead:

Razzak, H. and Heumann, C. (2020c):Application of Iterative Hybrid MI Approach to Household Survey Data with Complex Dependency Structures (LMU Munchen): *Technical Reports*, Nr. 231, last updated 7. January 2020.

Available under: https://doi.org/10.5282/ubm/epub.70081

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 $i=1,\dots, n$ and $j=1,\dots, p$. 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.

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

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

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

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

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 $z$= $1, …,Z$ do
                  for m= $1, …,M$ do
        $Imp^{z}_{num\_cat_m}$ ← Categorizing $Miss_{num}$.
        $Imp^{z}_{cat_m}$ ← imputation using JM approach for $Miss._{cat} | Imp^{z}_{num\_cat_m}$.
        $Imp^{z}_{num_m}$ ← imputation using FCS approach for $Miss._{num} | Imp^{z}_{cat_m}$.
              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 $Miss_{num} \,|\, Imp_{cat\_i}$.
    (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 $Miss._{cat} \,|\, Imp_{num\_cat\_i}$.
    (b) $Imp^{z}_{num_m}$ ← imputation using FCS approach for $Miss._{num} \,|\, Imp^{z}_{cat_m}$.
    (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:

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

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: $C_{32} \sim N\ (\mu_1; \sqrt{2}),\ \ C_{33} \sim N\ \left(\mu_2; \sqrt{2}\right)$, where μ_1 and μ_2 are described as:

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

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

·.

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

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

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

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

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

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$_{\text{CART}}$" (classification and regression trees (CART)) and "Mice$_{\text{DEF}}$" (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_α= *0.25,* b_α = *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:

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

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

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).

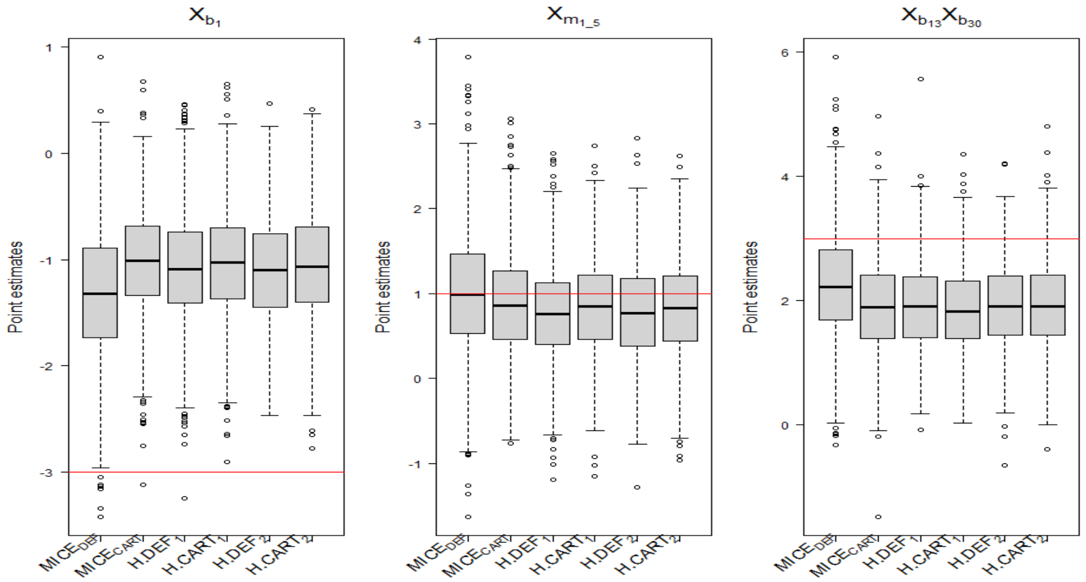


Figure1. Simulated data: Boxplots for the point estimates across 1000 simulations by imputation methods under Missing at Random (MAR) and ten imputations with 10% of missing data. Point estimates are shown for only three regression coefficients, i.e. for variables X_{b_1}, $X_{m_{1\_5}}$, $X_{b_{13}}X_{b_{30}}$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

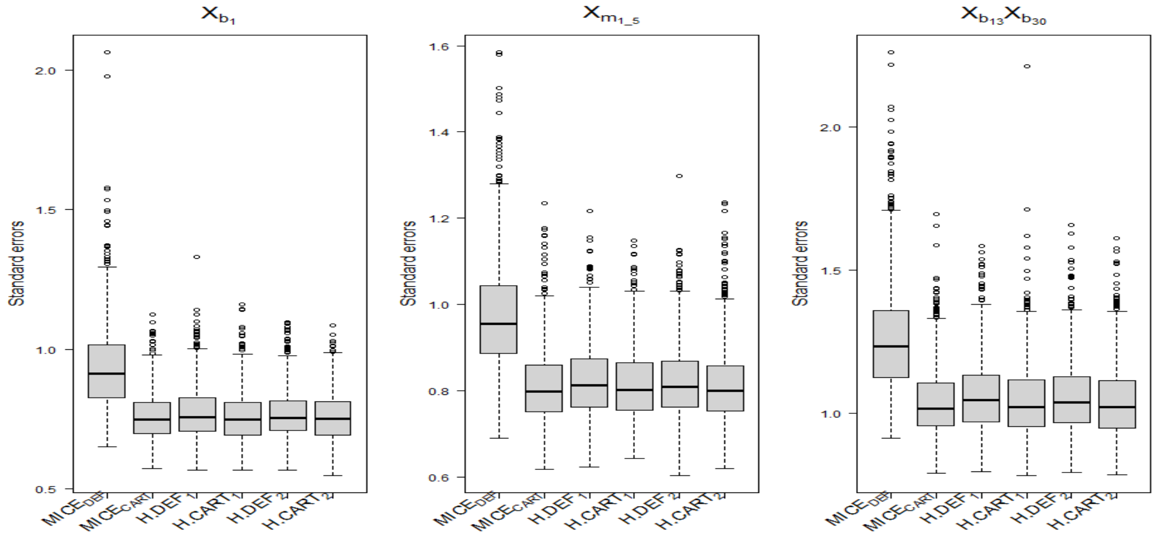


Figure2. Simulated data: Boxplots for the standard errors across 1000 simulations by imputation methods under Missing at Random (MAR) and ten imputations with 10% of missing data. Standard errors are shown for only three regression coefficients, i.e. for variables X_{b_1}, $X_{m_{1\_5}}$, $X_{b_{13}}X_{b_{30}}$.

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$_{\text{DEF}}$	MICE$_{\text{CART}}$	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,

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

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

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

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.

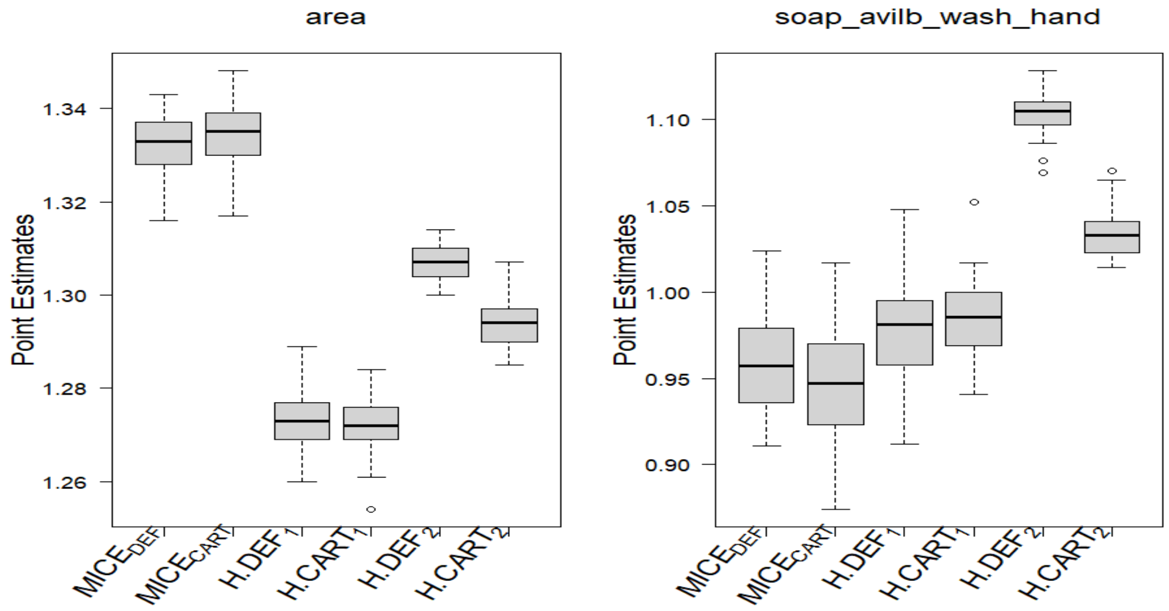


Figure3. Real data: Boxplots for point estimates across 50 simulations by imputation methods under Missing at Random (MAR) and ten imputations.

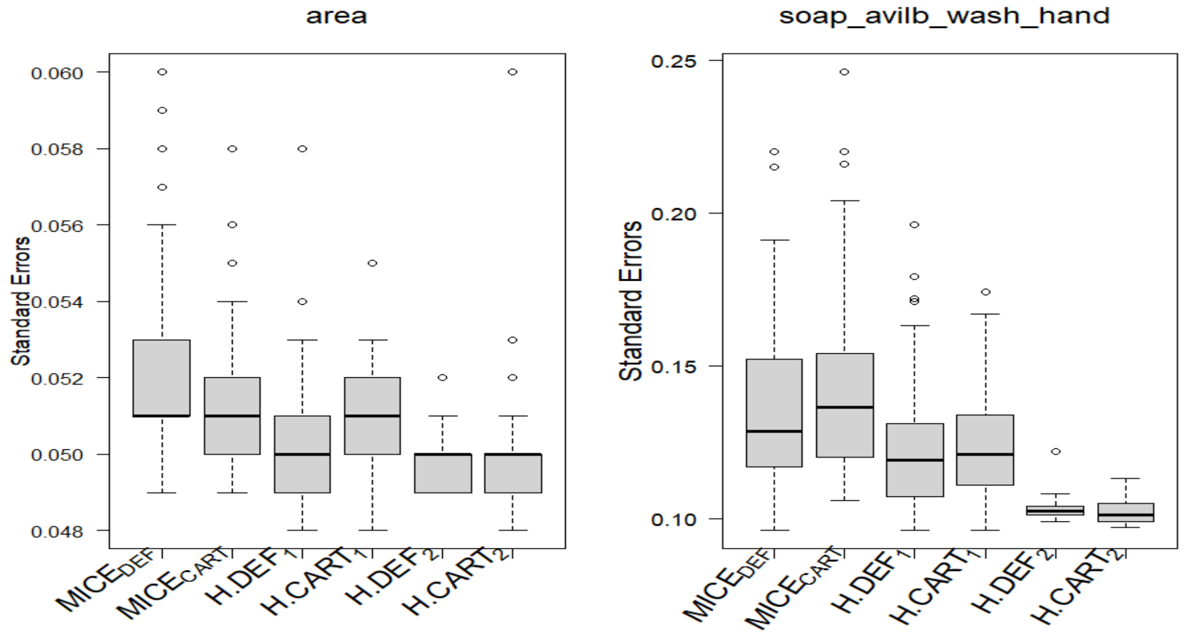


Figure4. Real data: Boxplots for standard errors across 50 simulations by imputation methods under Missing at Random (MAR) and ten imputations.

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. *Survey 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$_{\mathrm{DEF}}$**	**MICE$_{\mathrm{CART}}$**	**H.DEF$_1$**	**H.CART$_1$**	**MICE$_{\mathrm{DEF}}$**	**MICE$_{\mathrm{CART}}$**	**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$_{\text{DEF}}$**	**MICE$_{\text{CART}}$**	**H.DEF$_1$**	**H.CART$_1$**	**MICE$_{\text{DEF}}$**	**MICE$_{\text{CART}}$**	**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$_{\text{DEF}}$**	**MICE$_{\text{CART}}$**	**H.DEF$_2$**	**H.CART$_2$**	**MICE$_{\text{DEF}}$**	**MICE$_{\text{CART}}$**	**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

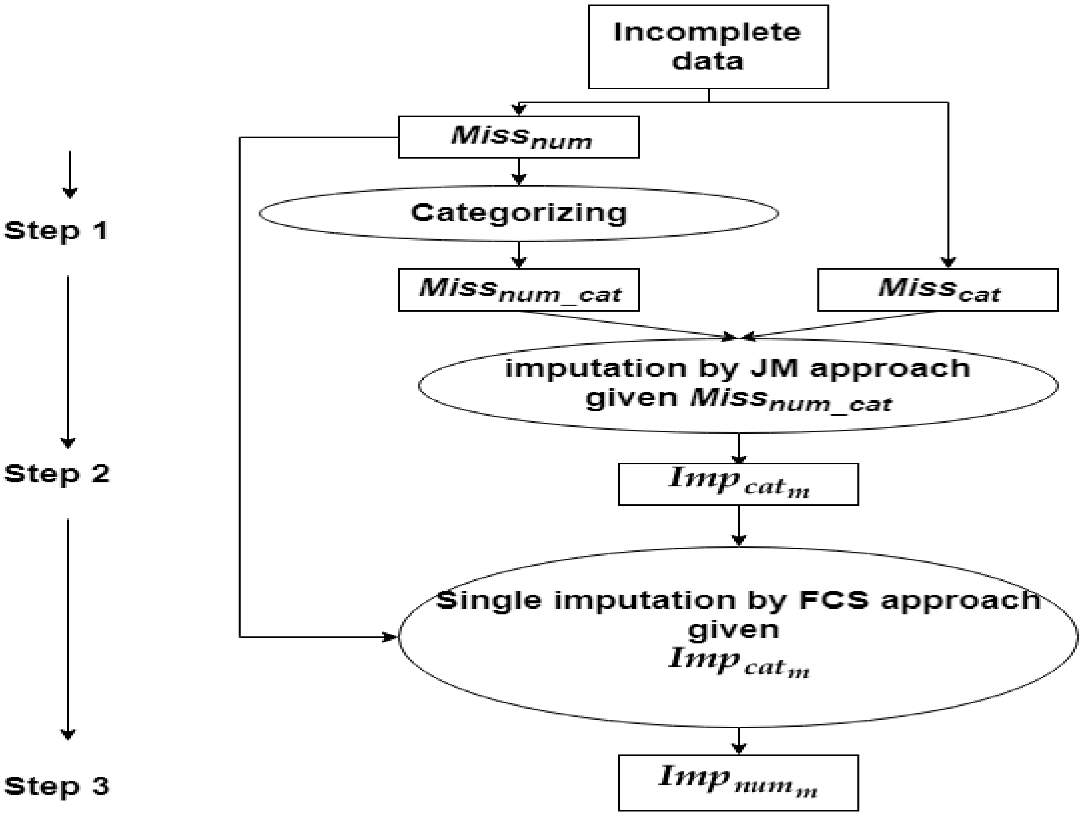
47	HH_own_land_agri	Any household member own land that can be used for agriculture	2	7
48	Type.of.toilet.facility	Type of toilet facility	13	7
49	T.V.	Television	2	7

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

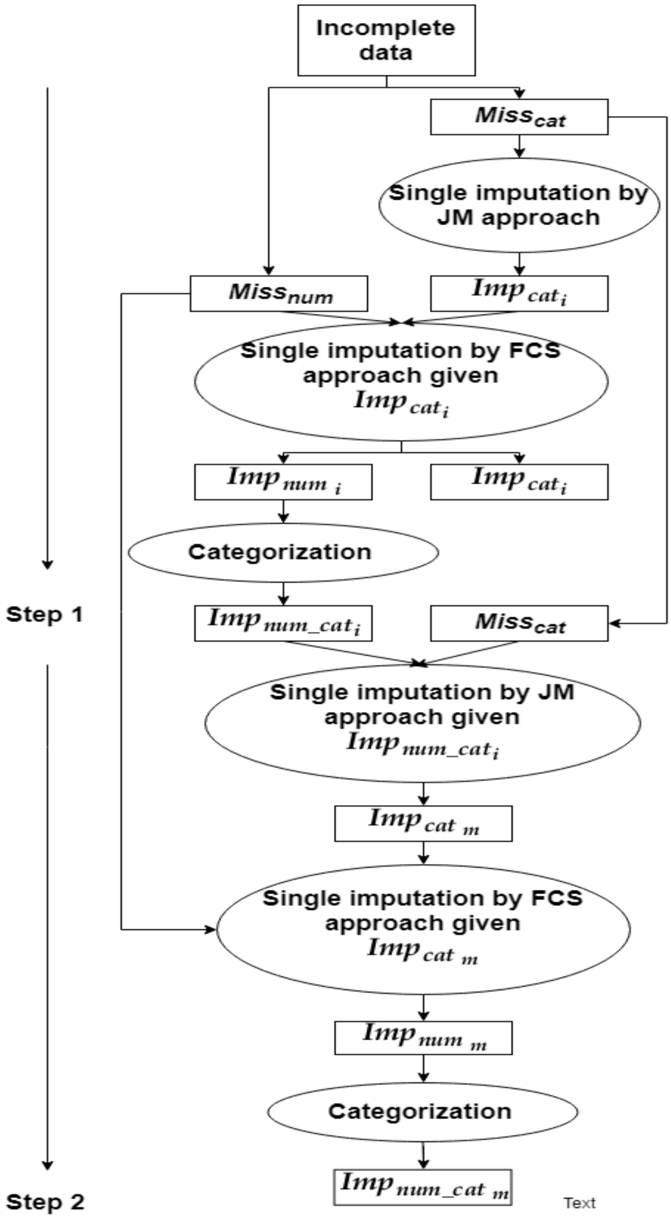
TableS6. Real data: Summary of all continuous variables

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

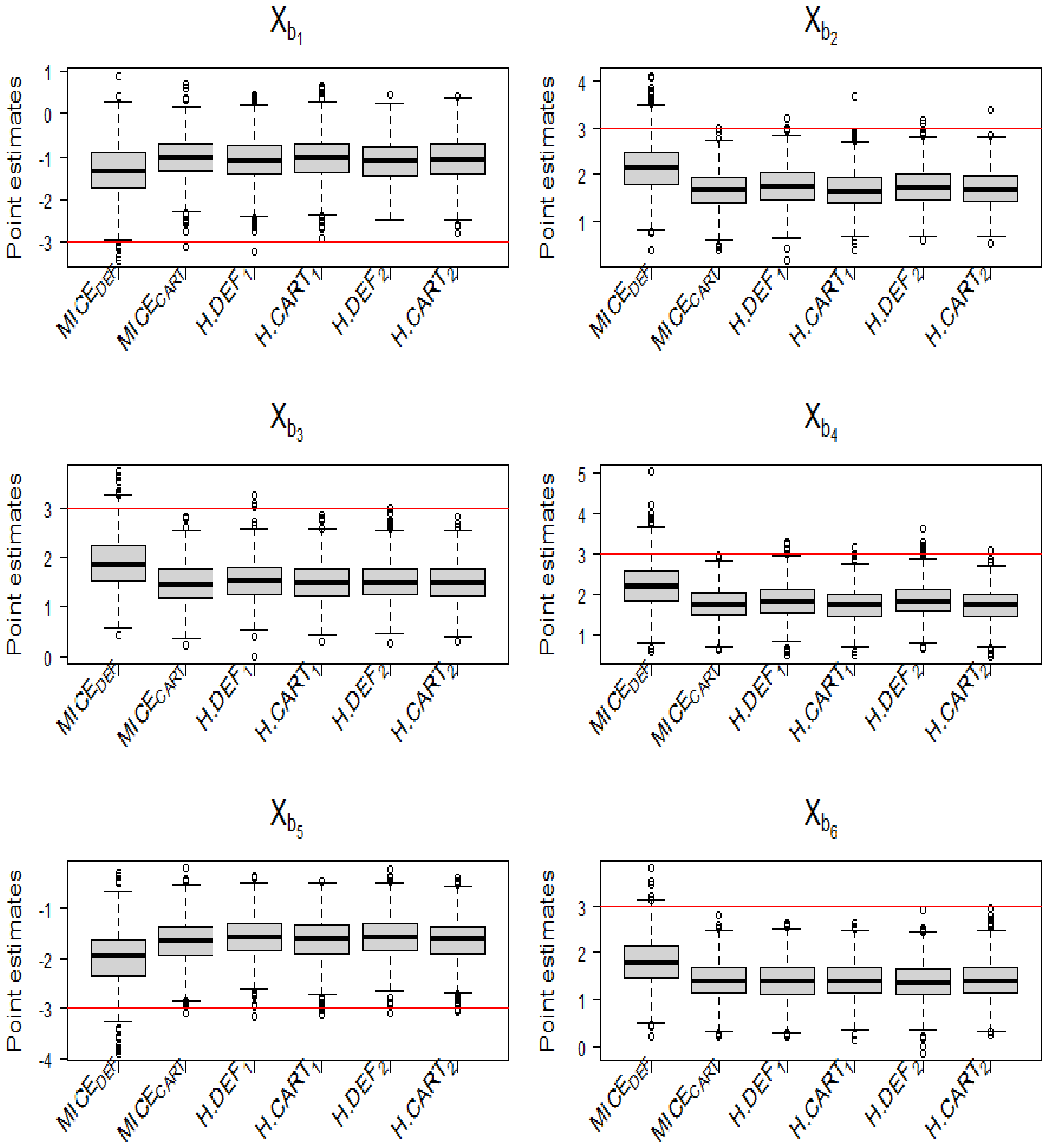
“% mis” indicates percentage of missing observations in all variables



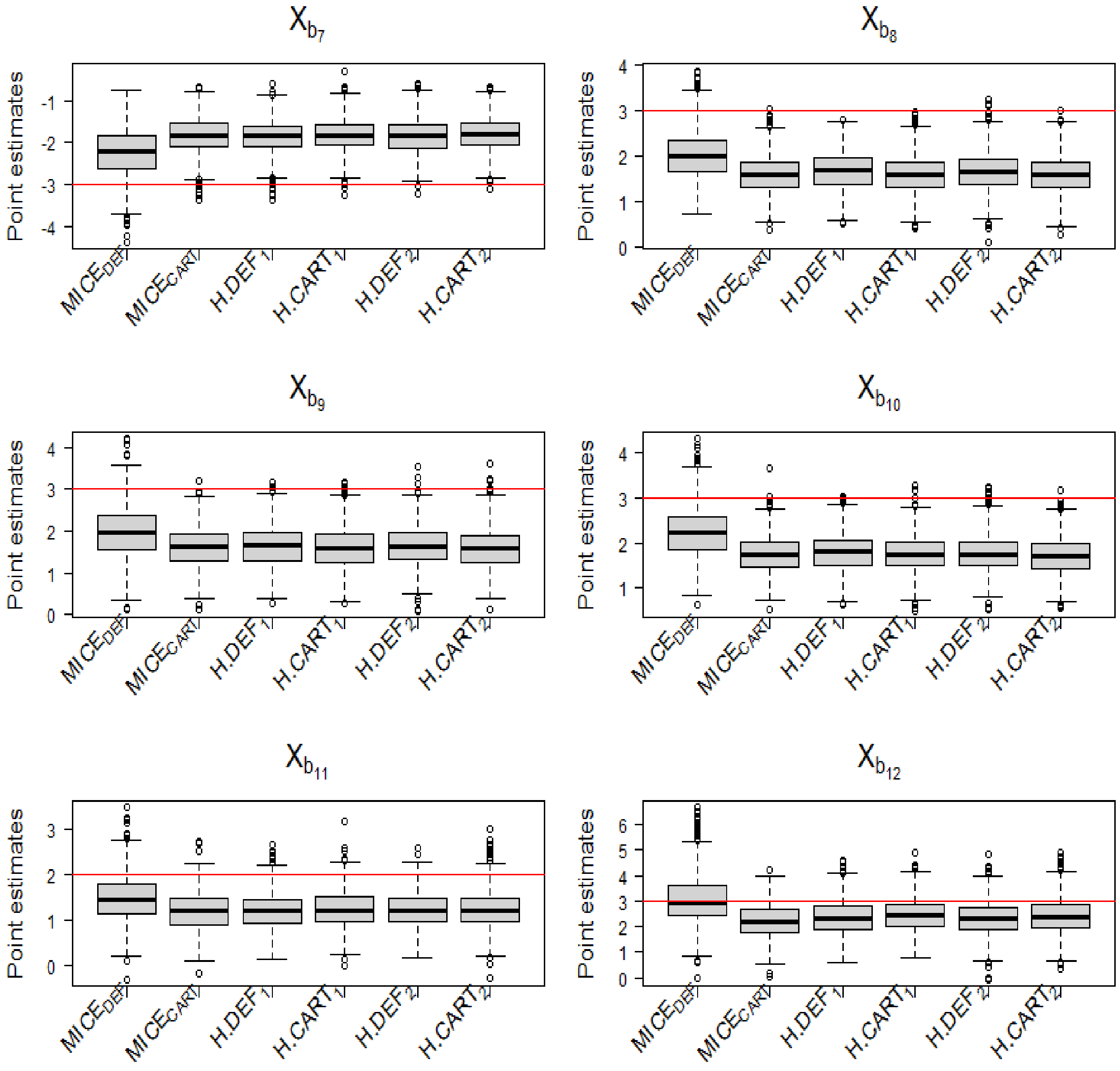
FigureS1. Schematic diagram illustrating the proposed hybrid architecture 1



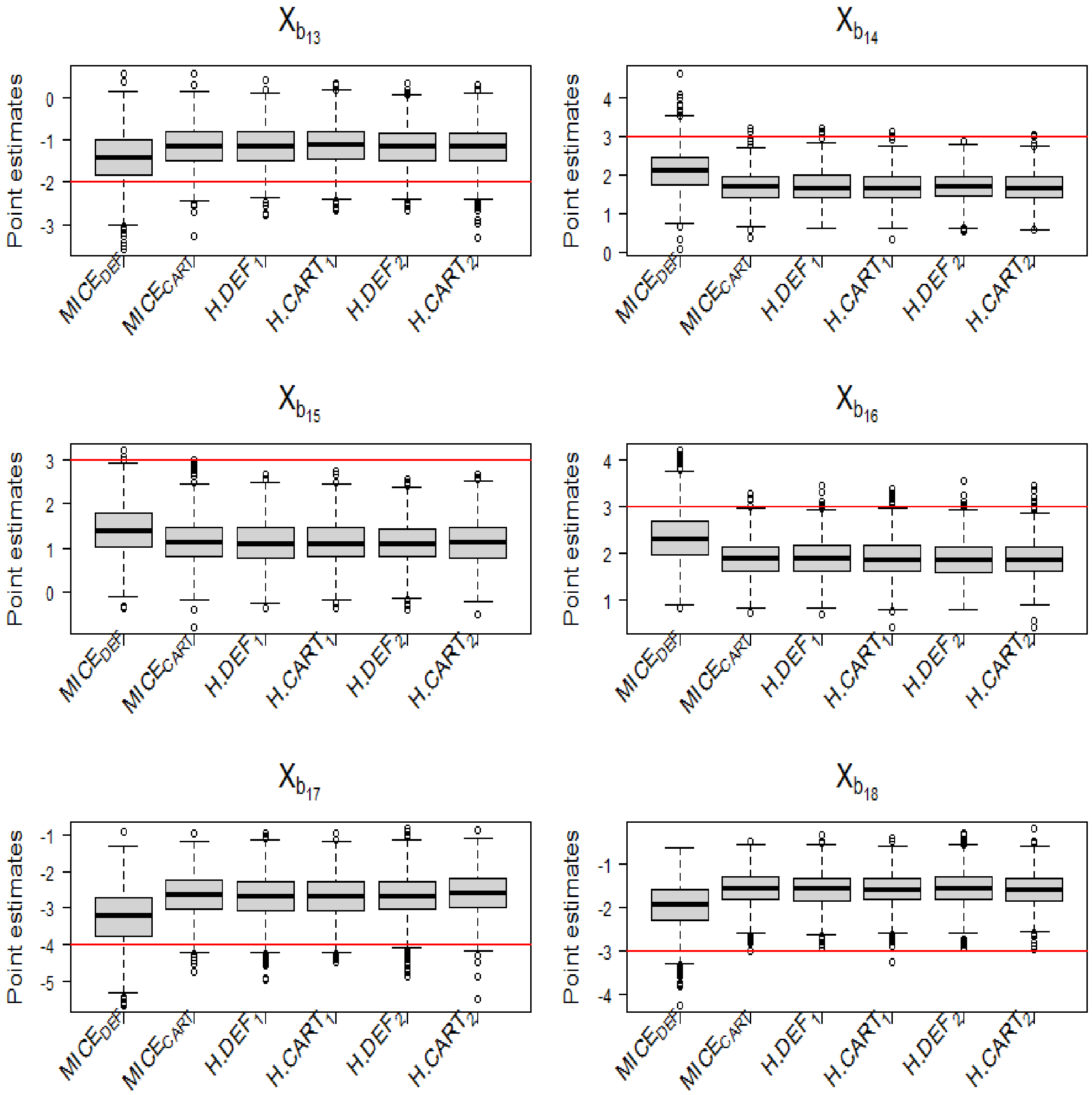
FigureS2. Schematic diagram illustrating the proposed hybrid architecture 2



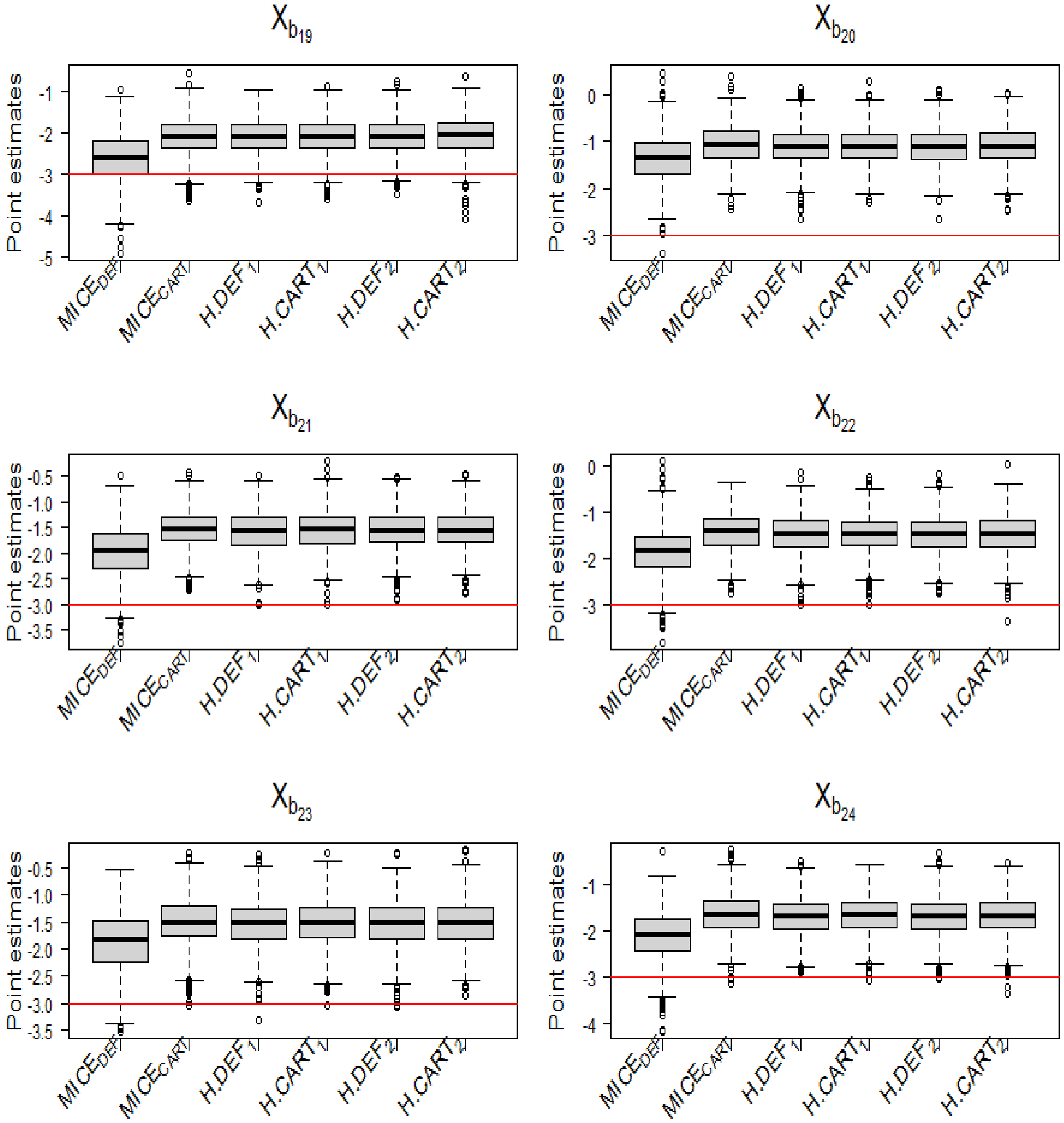
FigureS3. Simulated data: Boxplots of point estimates for coefficients X_{b_1}, X_{b_2} , X_{b_3}, X_{b_4}, X_{b_5}, X_{b_6}under various MI methods over 1000 simulations



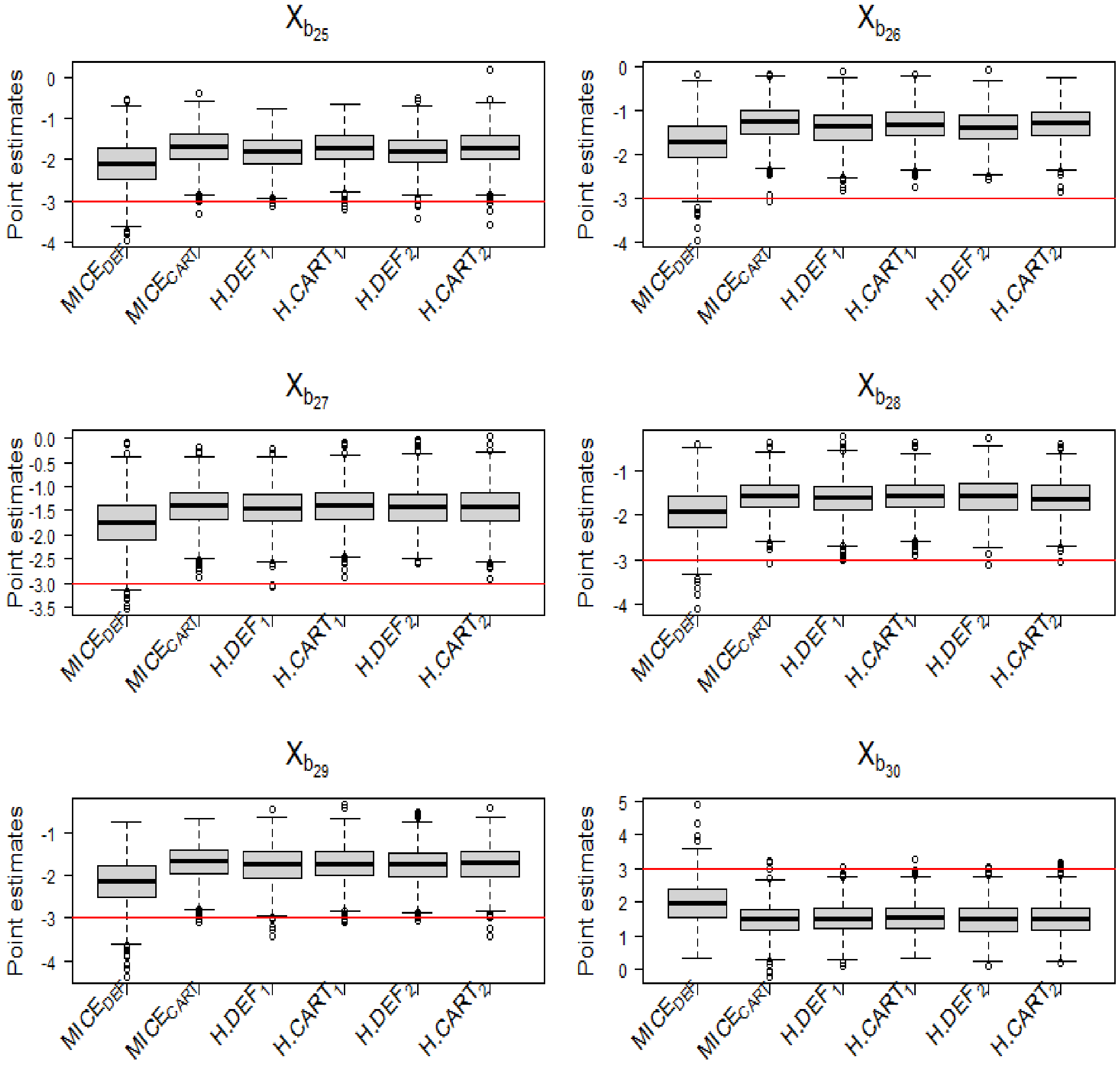
FigureS4. Simulated data: Boxplots of point estimates for coefficients X_{b_7}, X_{b_8} , X_{b_9}, $X_{b_{10}}$, $X_{b_{11}}$, $X_{b_{12}}$under various MI methods over 1000 simulations



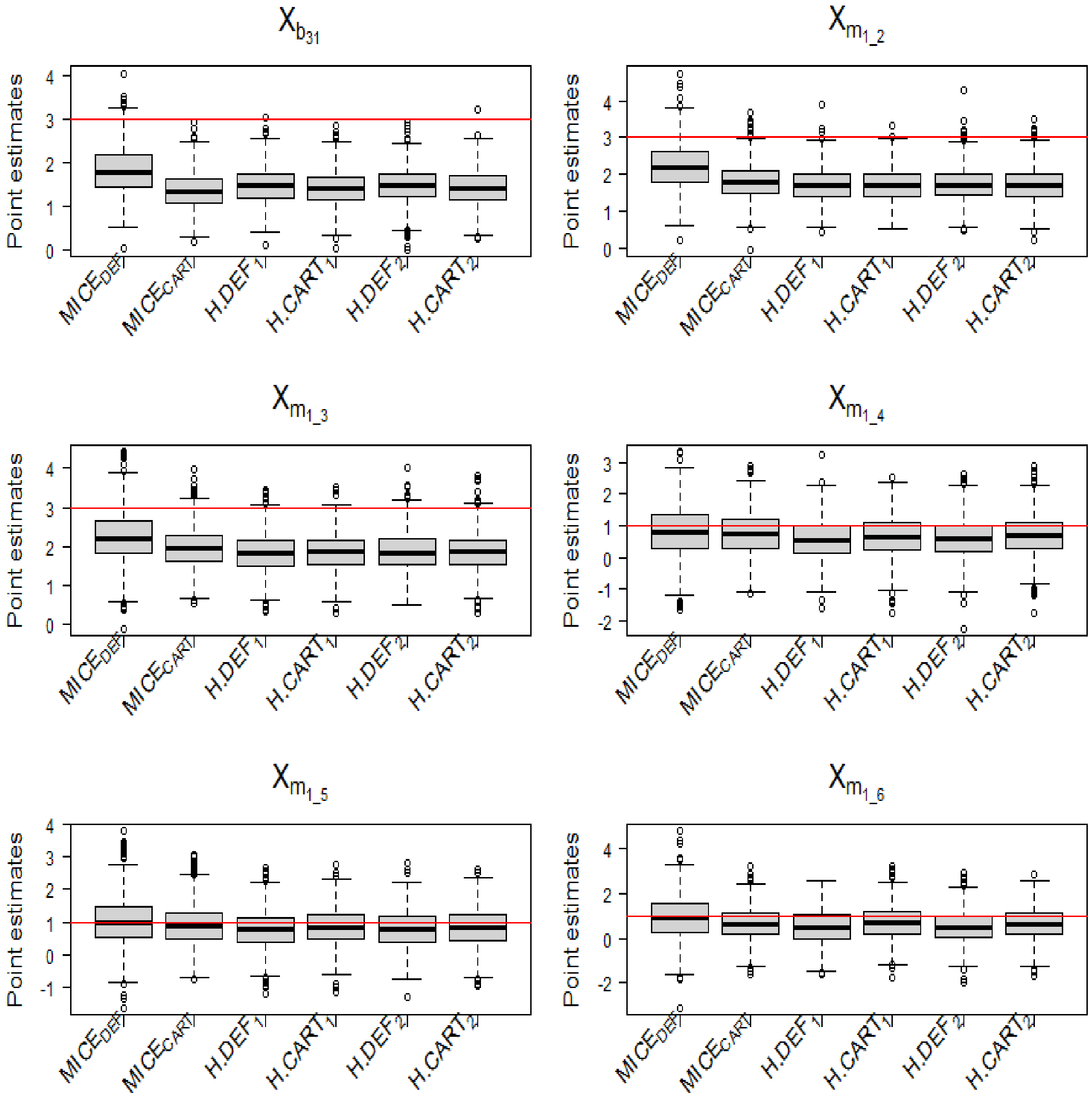
FigureS5. Simulated data: Boxplots of point estimates for coefficients $X_{b_{13}}$, $X_{b_{14}}$, $X_{b_{15}}$, $X_{b_{16}}$, $X_{b_{17}}$, $X_{b_{18}}$under various MI methods over 1000 simulations



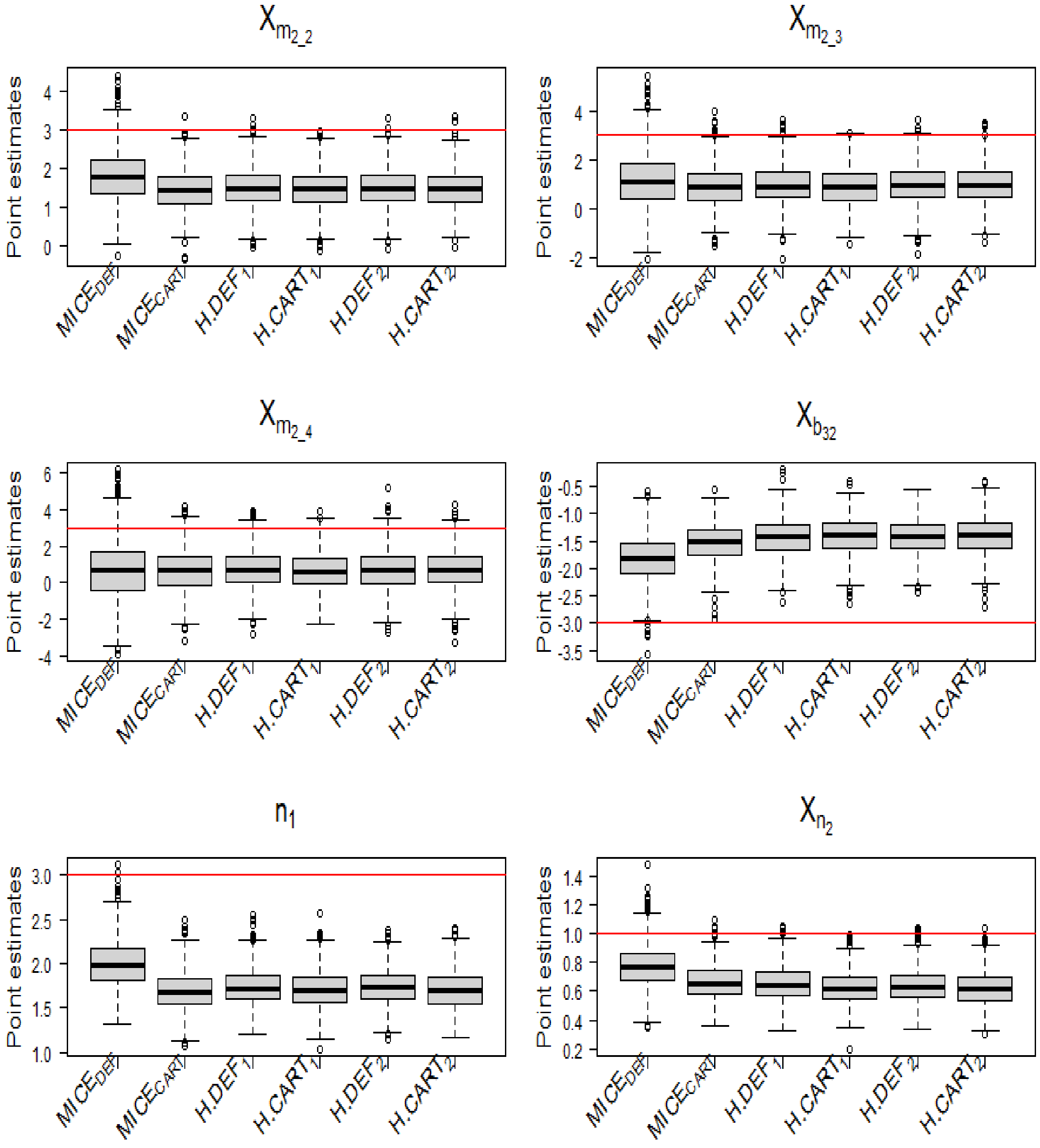
FigureS6. Simulated data: Boxplots of point estimates for coefficients $X_{b_{19}}$, $X_{b_{20}}$, $X_{b_{21}}$, $X_{b_{22}}$, $X_{b_{23}}$, $X_{b_{24}}$ under various MI methods over 1000 simulations



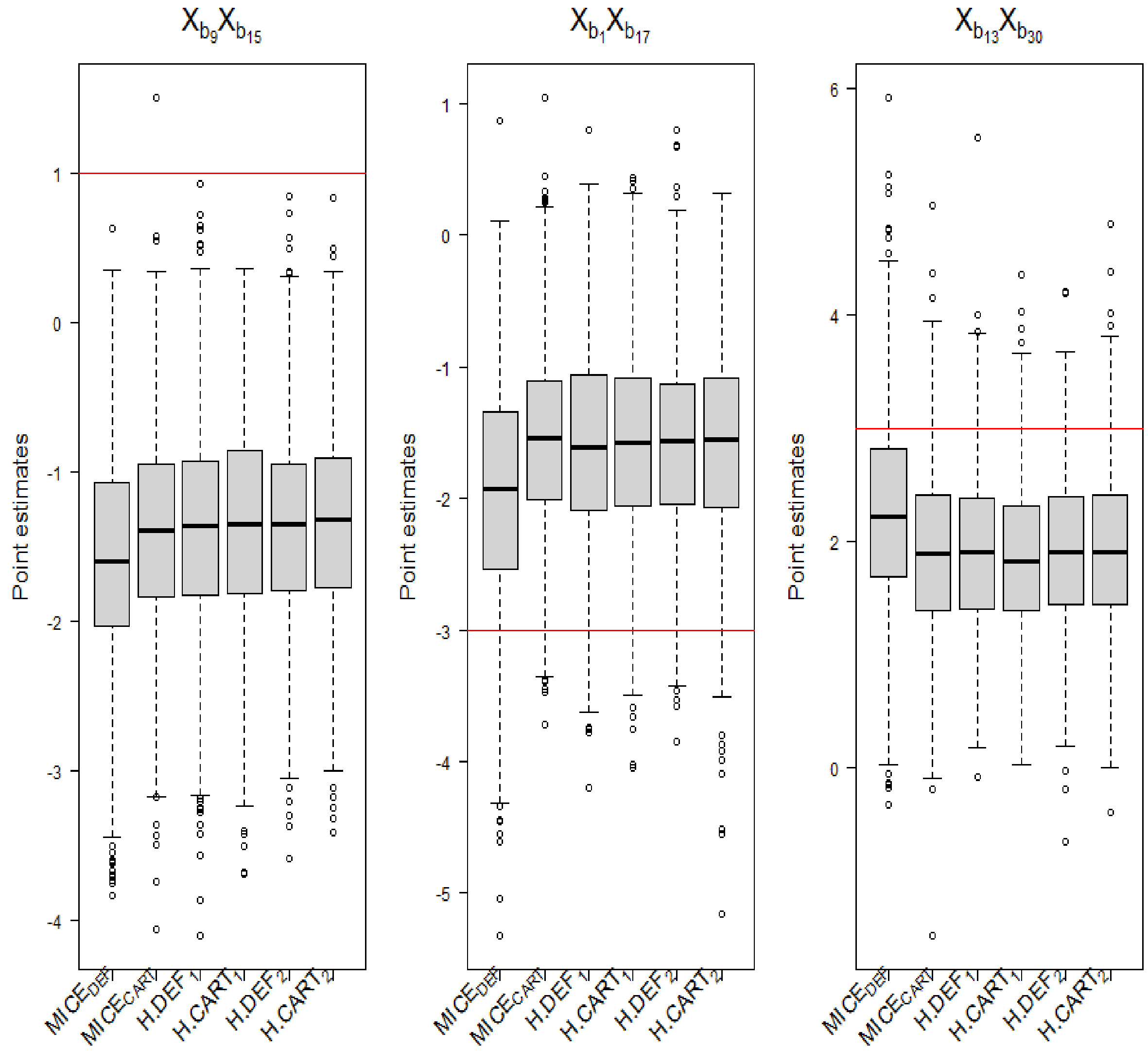
FigureS7. Simulated data: Boxplots of point estimates for coefficients $X_{b_{25}}$, $X_{b_{26}}$,$X_{b_{27}}$, $X_{b_{28}}$, $X_{b_{29}}$, $X_{b_{30}}$ under various MI methods over 1000 simulations



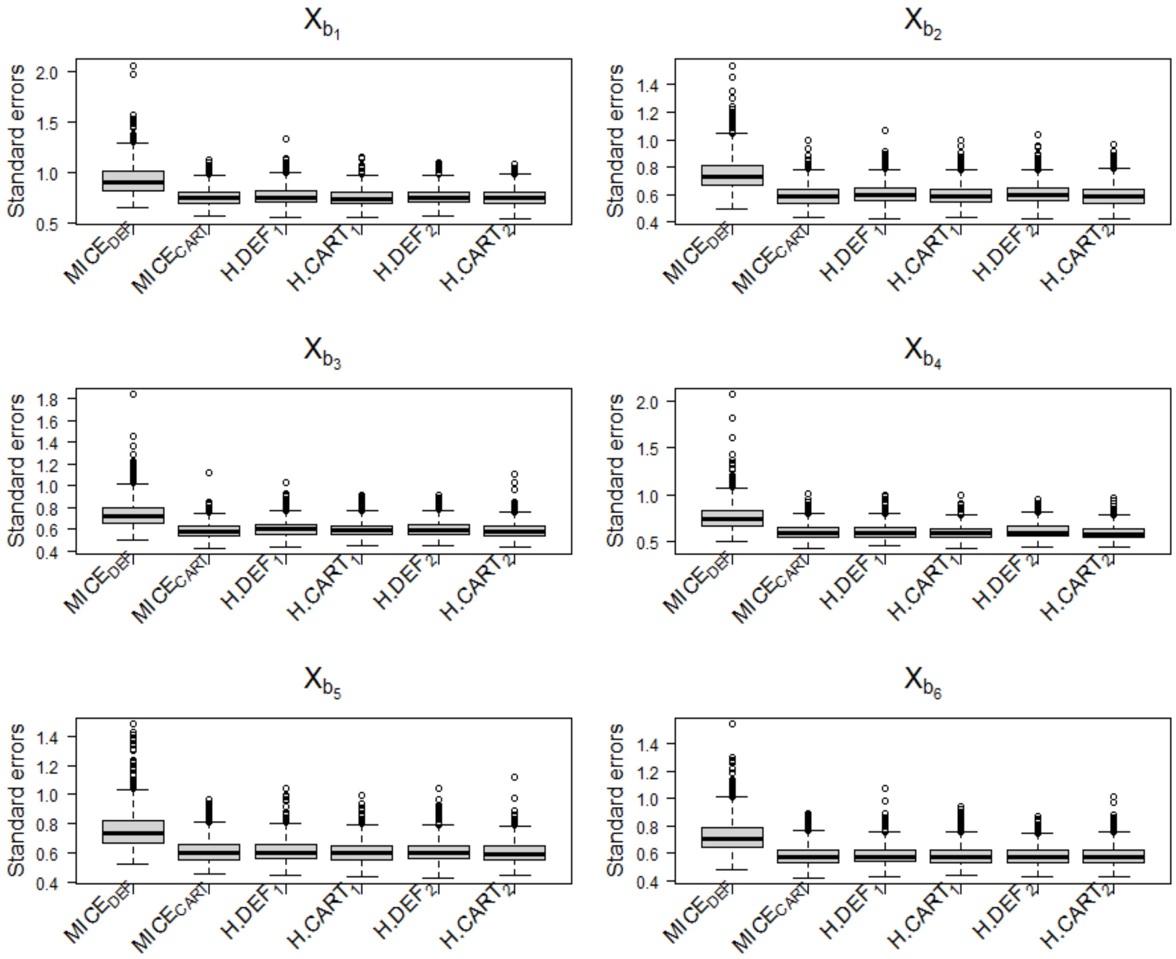
FigureS8. Simulated data: Boxplots of point estimates for coefficients $X_{b_{31}}$, $X_{m_{1\_2}}$, $X_{m_{1\_3}}$, $X_{m_{1\_4}}$,$X_{m_{1\_5}}$, $X_{m_{1\_6}}$ under various MI methods over 1000 simulations



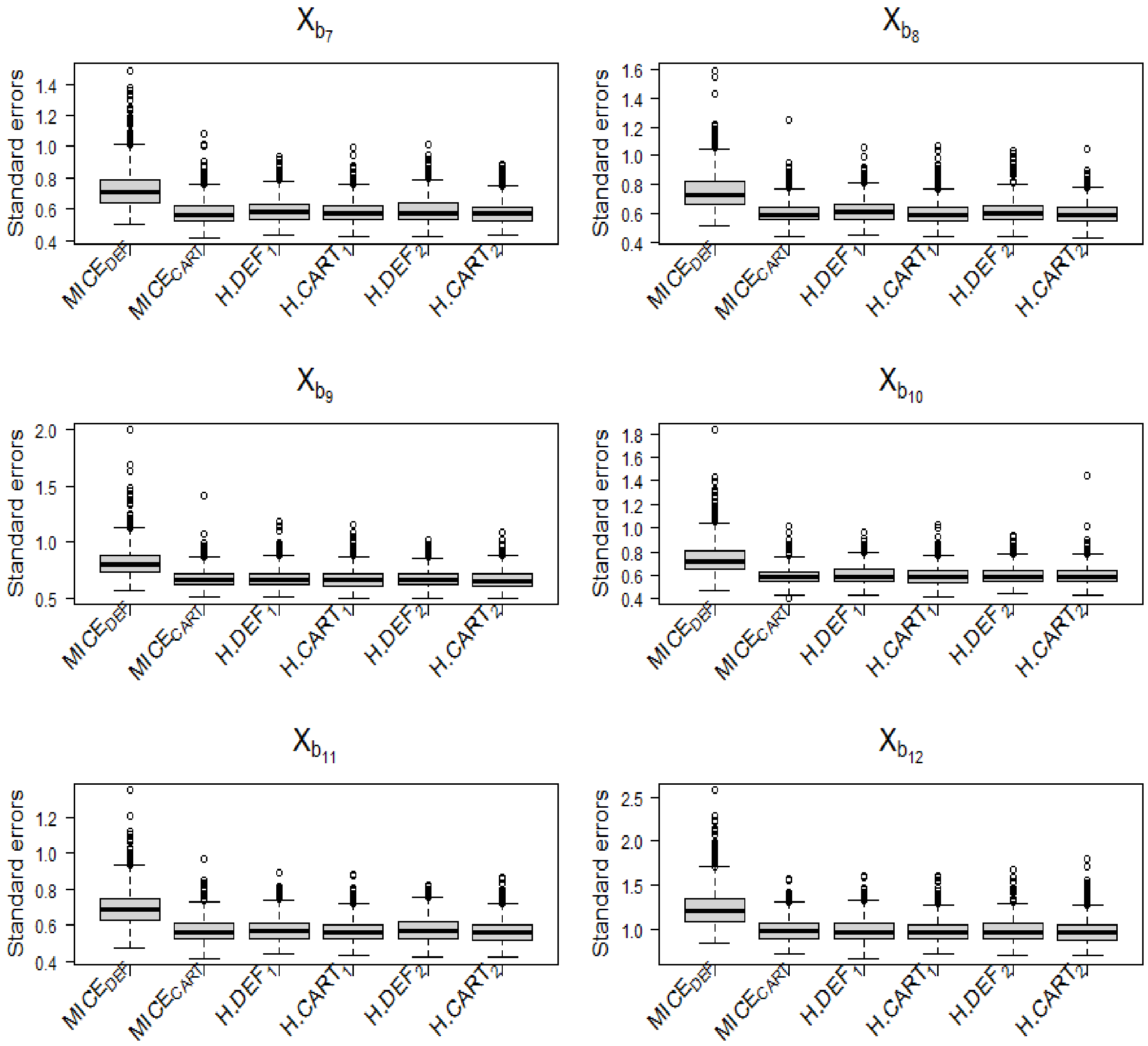
FigureS9. Simulated data: Boxplots of point estimates for coefficients $X_{m_{2\_2}}$,$X_{m_{2\_3}}$,$X_{m_{2\_4}}$,X$_{b32}$, X_{n_1} , X_{n_2} under various MI methods over 1000 simulations



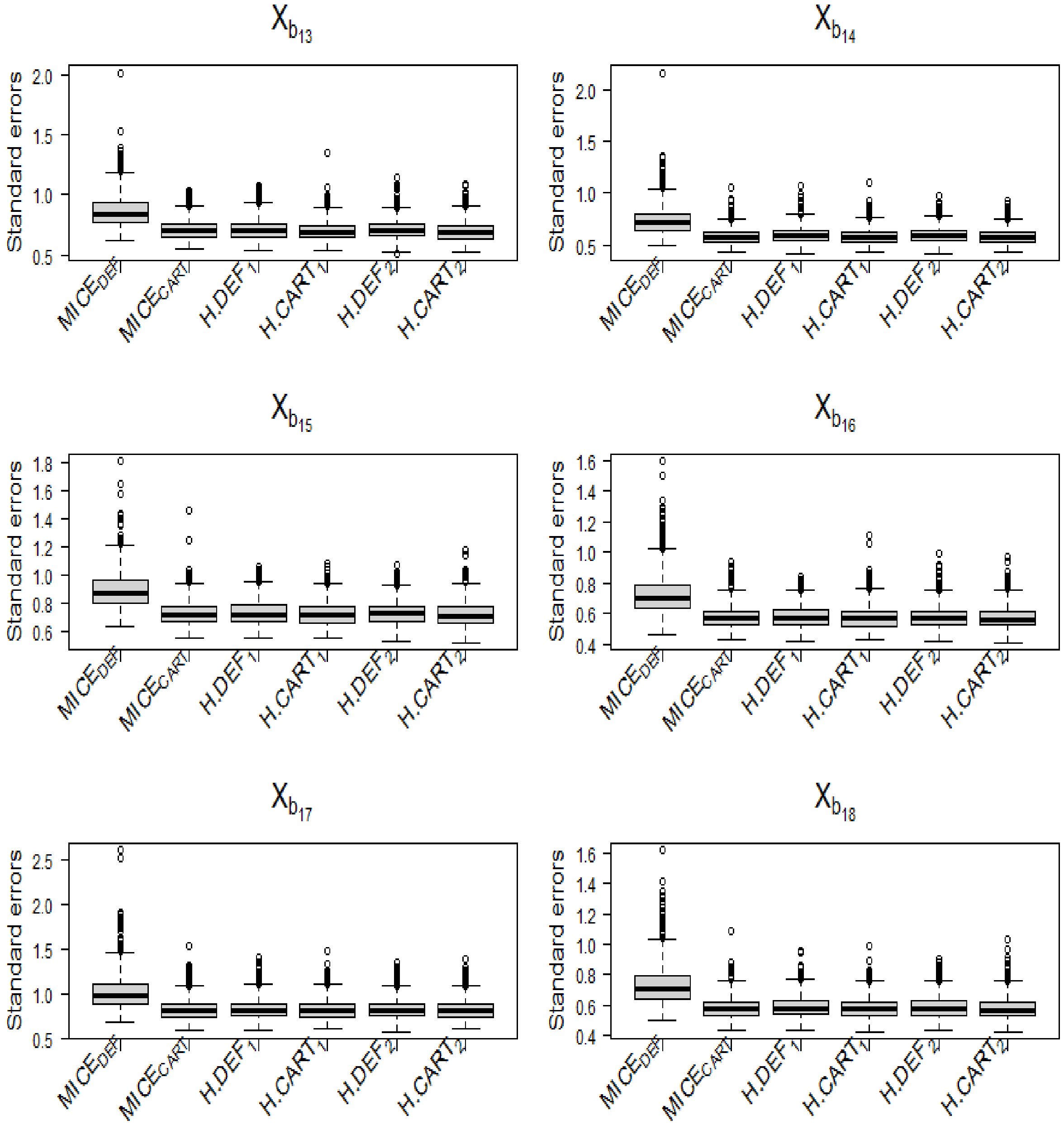
FigureS10. Simulated data: Boxplots of point estimates for coefficients $X_{b_9}\,X_{b_{15}}$, $X_{b_1}\,X_{b_{17}}$ $X_{b_{13}}\,X_{b_{30}}$ under various MI methods over 1000 simulations



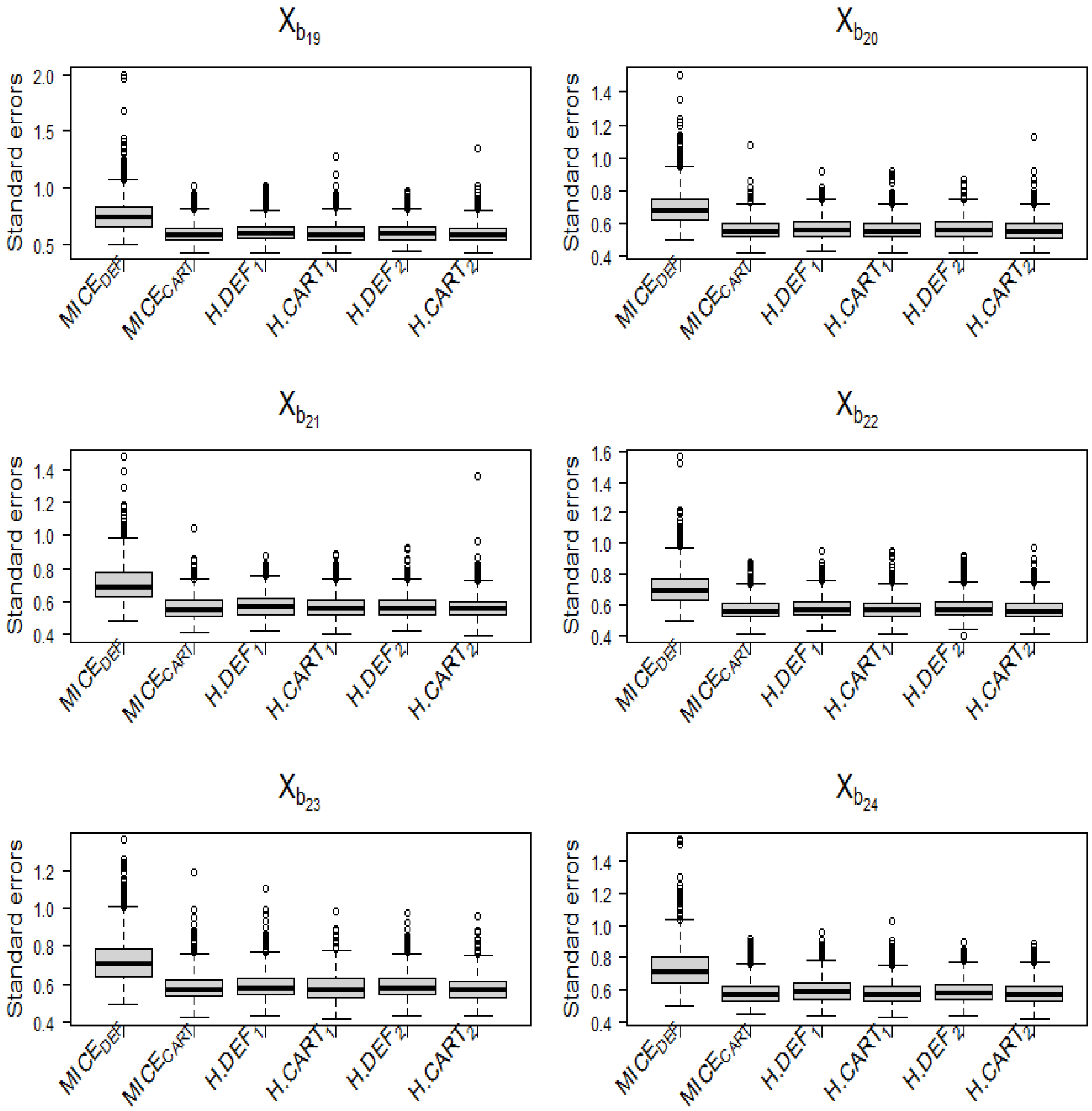
FigureS11. Simulated data: Boxplots of standard errors for coefficients X_{b_1}, X_{b_2} , X_{b_3}, X_{b_4}, X_{b_5}, X_{b_6}under various MI methods over 1000 simulations



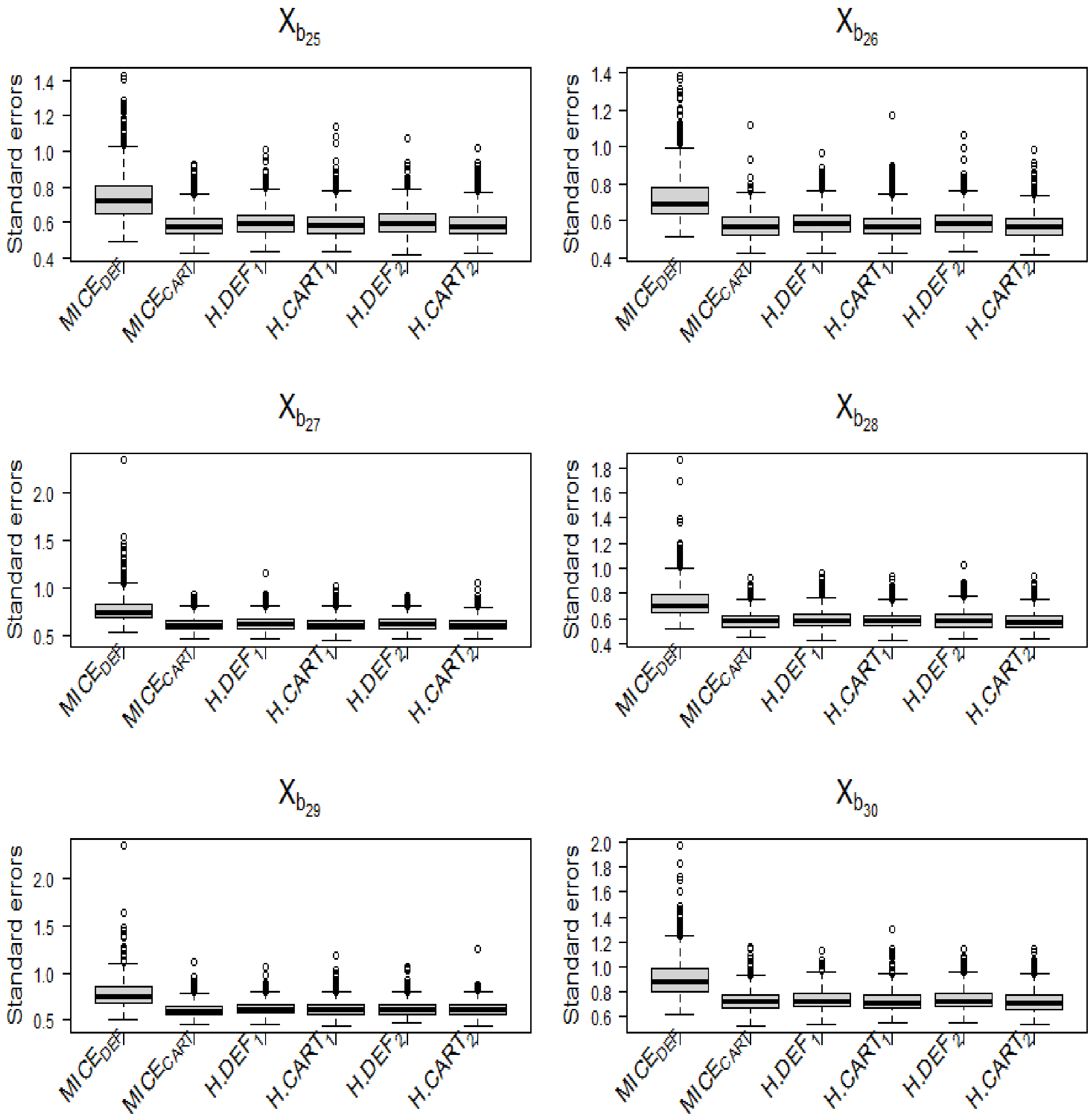
FigureS12. Simulated data: Boxplots of standard errors for coefficients X_{b_7}, X_{b_8} , X_{b_9}, $X_{b_{10}}$, $X_{b_{11}}$, $X_{b_{12}}$under various MI methods over 1000 simulations



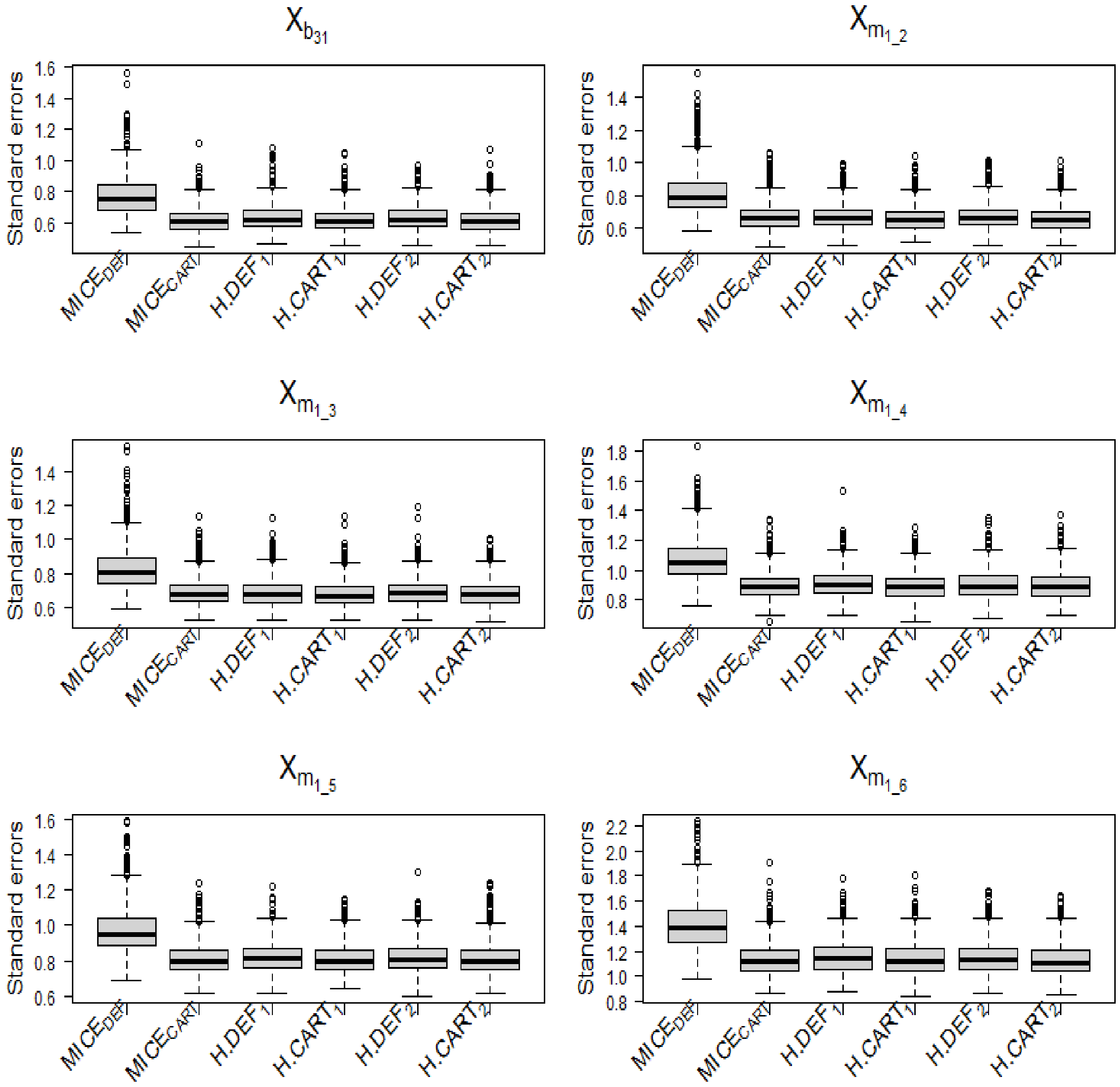
FigureS13. Simulated data: Boxplots of standard errors for coefficients $X_{b_{13}}$, $X_{b_{14}}$, $X_{b_{15}}$, $X_{b_{16}}$, $X_{b_{17}}$, $X_{b_{18}}$under various MI methods over 1000 simulations



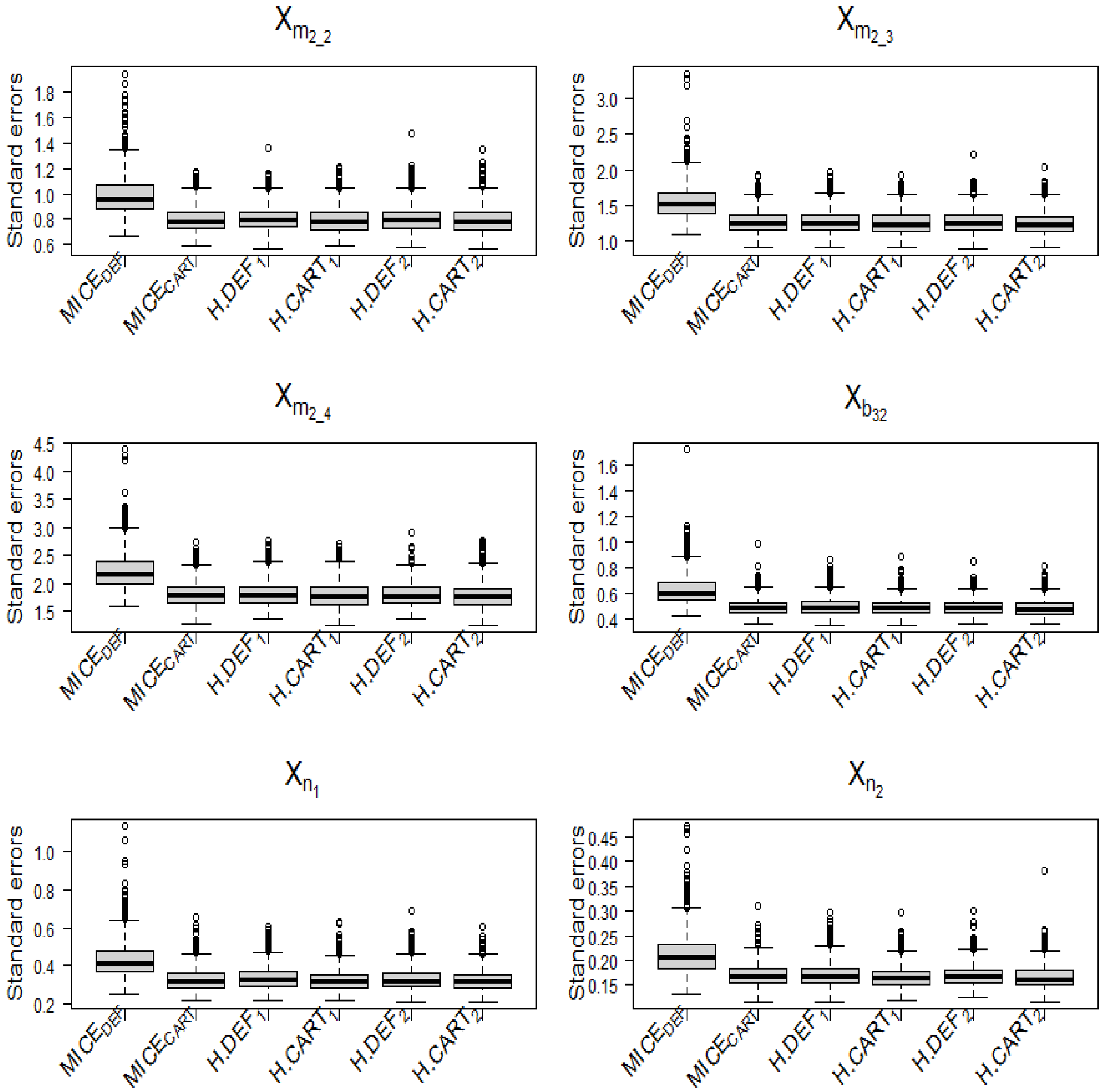
FigureS14. Simulated data: Boxplots of standard errors for coefficients $X_{b_{19}}$, $X_{b_{20}}$, $X_{b_{21}}$, $X_{b_{22}}$, $X_{b_{23}}$, $X_{b_{24}}$ under various MI methods over 1000 simulations



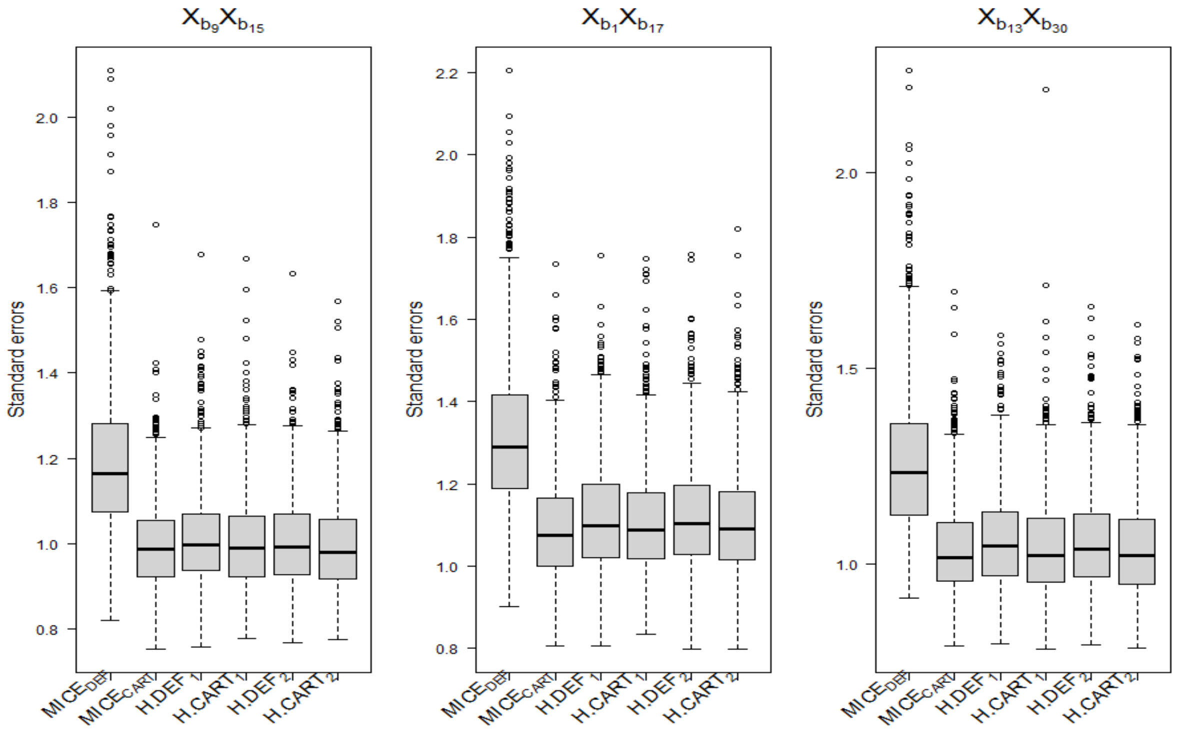
FigureS15. Simulated data: Boxplots of standard errors for coefficients $X_{b_{25}}$, $X_{b_{26}}$,$X_{b_{27}}$, $X_{b_{28}}$, $X_{b_{29}}$, $X_{b_{30}}$ under various MI methods over 1000 simulations



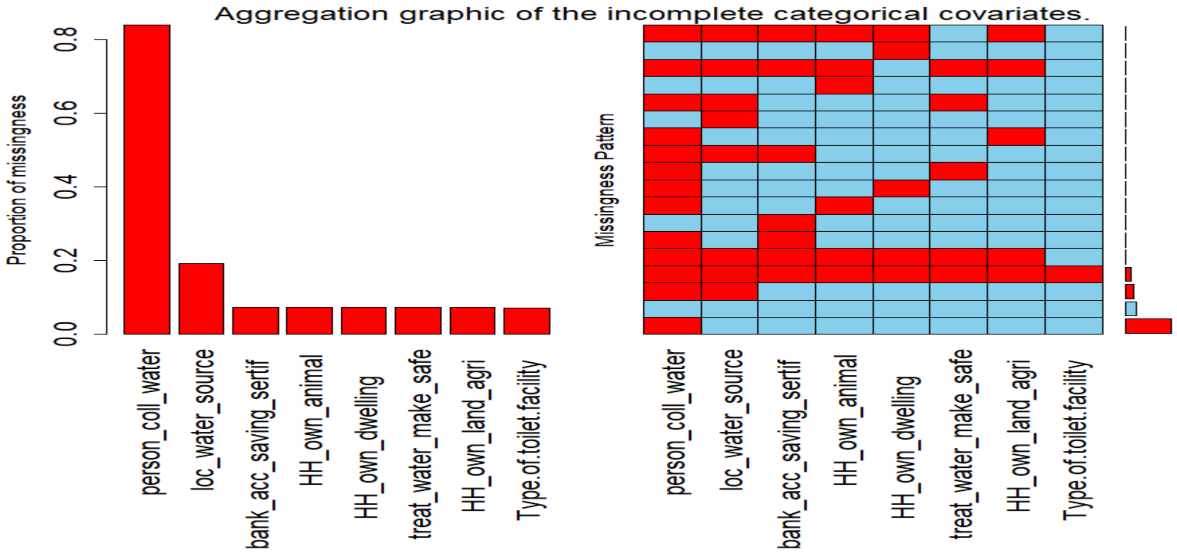
FigureS16. Simulated data: Boxplots of standard errors for coefficients $X_{b_{31}}$, $X_{m_{1\_2}}$, $X_{m_{1\_3}}$, $X_{m_{1\_4}}$,$X_{m_{1\_5}}$, $X_{m_{1\_6}}$ under various MI methods over 1000 simulations



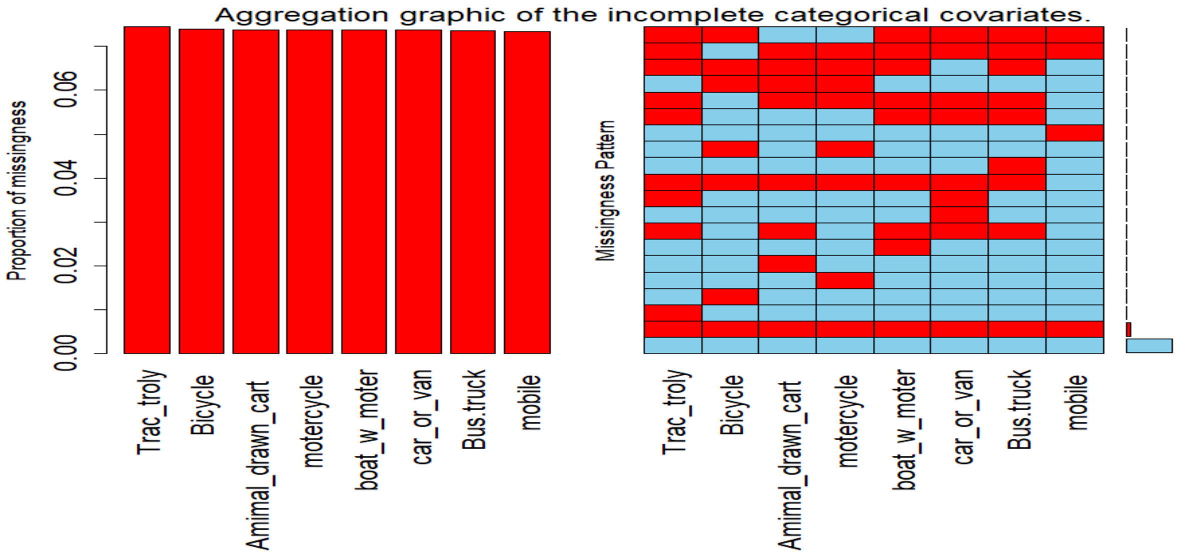
FigureS17. Simulated data: Boxplots of standard errors for coefficients $X_{m_{2\_2}}$,$X_{m_{2\_3}}$,$X_{m_{2\_4}}$,Xb32, X_{n_1} , X_{n_2} under various MI methods over 1000 simulations



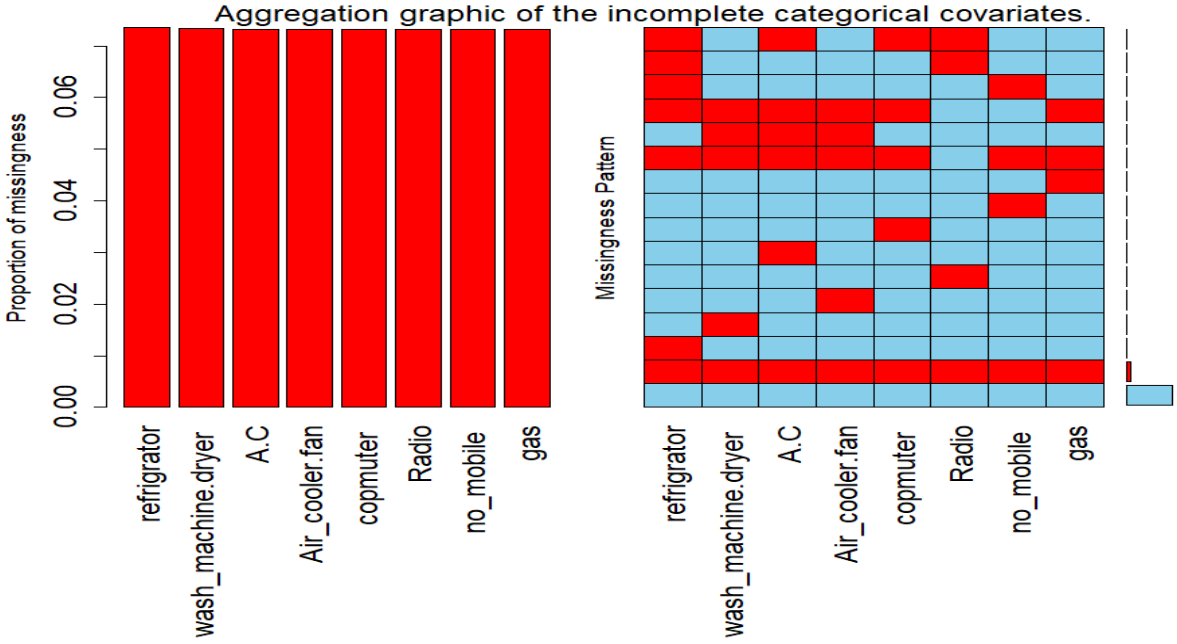
FigureS18. Simulated data: Boxplots of standard errors for coefficients $X_{b_9} X_{b_{15}}$, $X_{b_1} X_{b_{17}}$, $X_{b_{13}} X_{b_{30}}$ under various MI methods over 1000 simulations



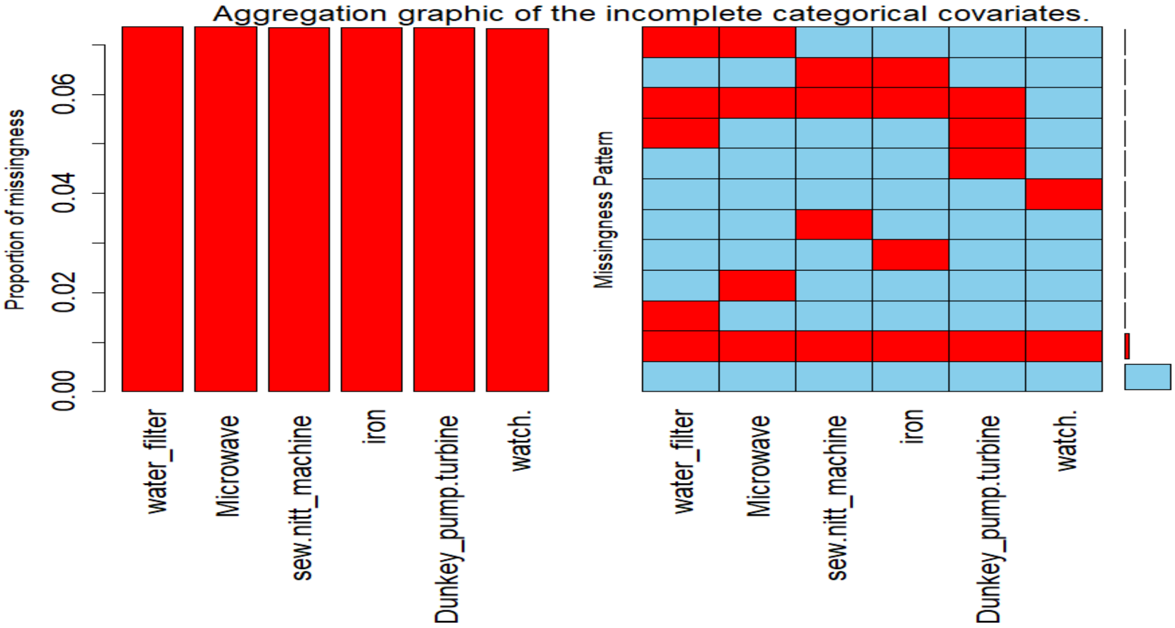
FigureS19. Real data: Aggregate plots in R, graphics of incomplete variables i.e. "HH_own_dwelling,"HH_own_land_agri","Type.of.toilet.facility","HH_own_animal","treat_water_make_safe", "bank_acc_saving_sertif","loc_water_source","person_coll_water"



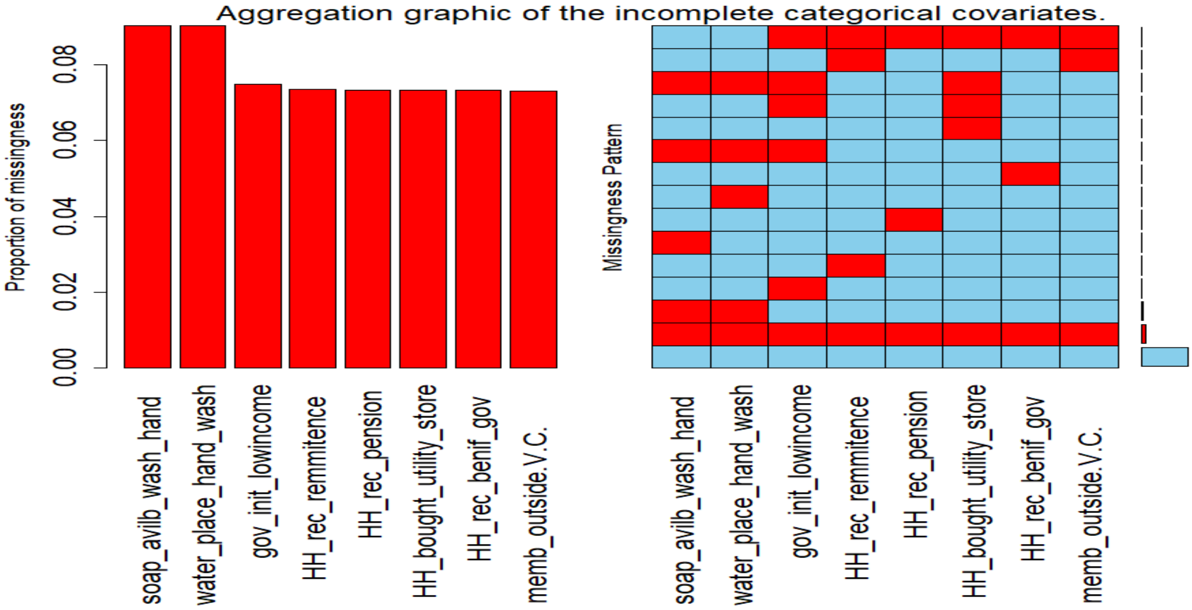
FigureS20. Real data: Aggregate plots in R, graphics of incomplete variables i.e. "mobile ","Bicycle","motercycle","Amimal_drawn_cart","Bus.truck","boat_w_moter","car_or_van","Trac_troly"



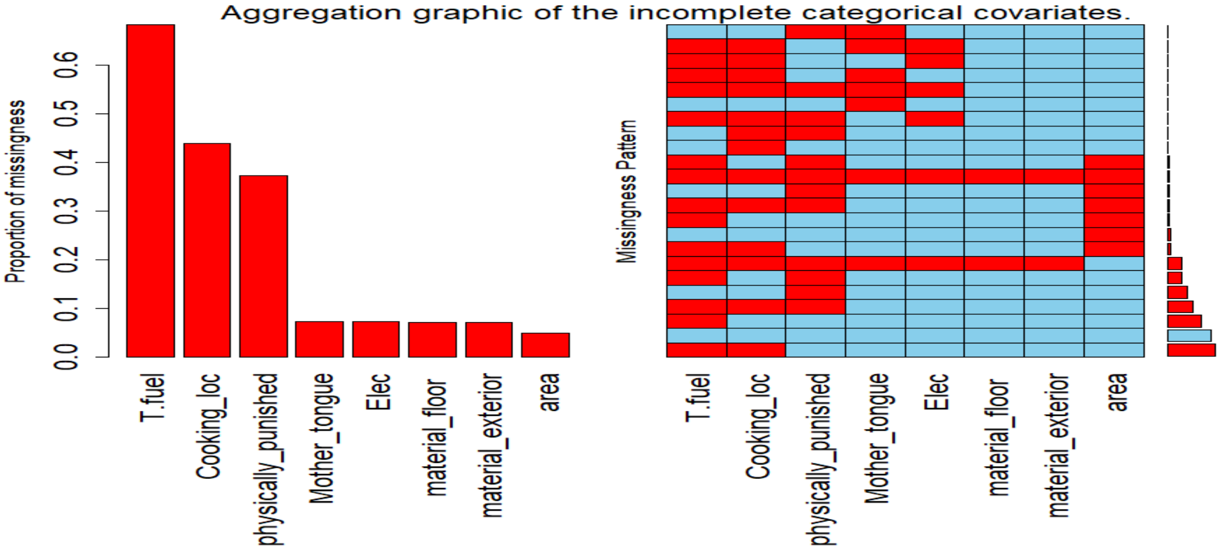
FigureS21. Real data: Aggregate plots in R, graphics of incomplete variables i.e. "Radio", "no_mobile", refrigrator","gas"," copmuter ", "A.C", "wash_machine.dryer ", "Air_cooler.fan"



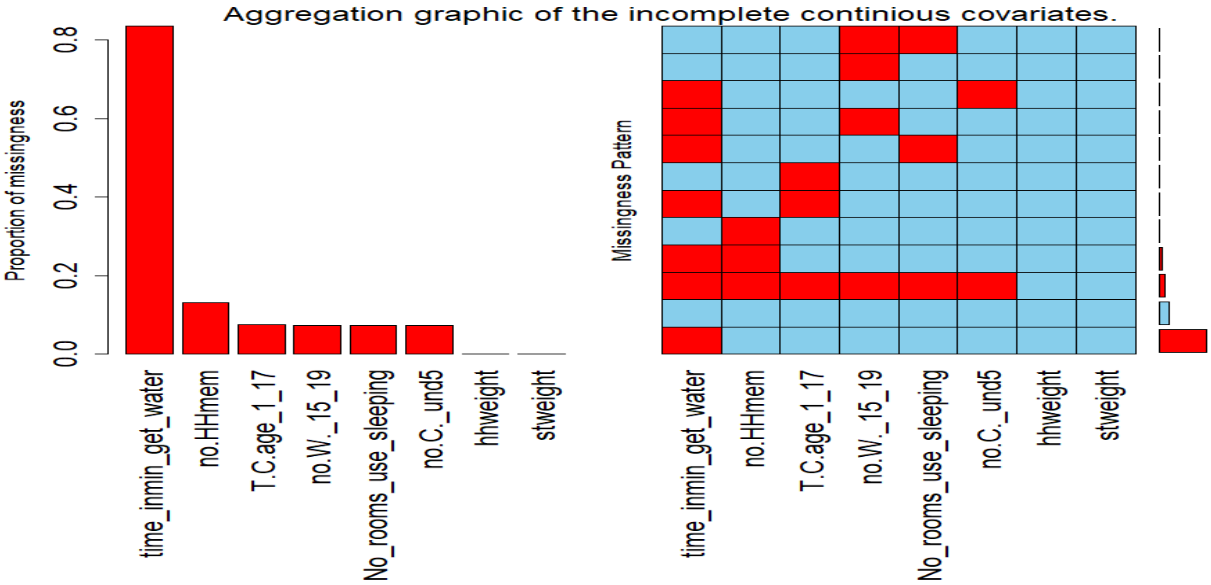
FigureS22. Real data: Aggregate plots in R, graphics of incomplete variables i.e. "Microwave", "sew.nitt_machine ","iron", "water_filter", "Dunkey_pump.turbine ", "watch"



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"

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.2011, § 8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation title 'Multiple imputation of large scale complex surveys' von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 01.04.2020

Humera Razzak