Bachelor's Thesis

A Comparison of Tree-Based Methods for Multiple Imputation

Department of Statistics

Ludwig-Maximilians-Universität München

Michael Speckbacher

Munich, December 23rd, 2024

Submitted in partial fulfillment of the requirements for the degree of B. Sc.

Supervised by Prof. Dr. Christian Heumann

Abstract

This thesis evaluates the performance of two tree-based imputation methods, Classification and Regression Trees (CART) and Random Forest (RF), for handling missing data. The study involves generating datasets of varying dimensions, simulating different missingness mechanisms, and applying multiple levels of missing data, resulting in diverse scenarios with a count target variable. Missing values are imputed, and model parameters are estimated using the `mice` package, employing 10 rounds of imputation in accordance with the principles of multiple imputation. The evaluation focuses on key metrics such as bias, coverage rate, and confidence interval width.

The findings demonstrate that RF often surpasses CART, particularly in delivering higher coverage rates and narrower confidence intervals. Nevertheless, a universal recommendation for either method is not feasible, as their performance varies depending on the dataset's characteristics and the type of missingness mechanism. This thesis provides detailed guidance on selecting the most appropriate method for specific scenarios, enabling informed, context-dependent decision-making.

Keywords: missing data, multiple imputation, simulation study, Random Forest, CART

Contents

List of Figures

List of Tables

List of Notations

Notation	Definition
n	Sample size
p	Number of covariates
y	Target variable
Y_{obs}	Observed data
Y_{mis}	Missing data
m	Number of imputations
ϕ	Parameters describing the missingness mechanism
K	Number of datasets for one scenario

1 Introduction

In many scientific disciplines, data plays a fundamental role in deriving conclusions and advancing knowledge. However, it is common for datasets to contain missing values, which pose a major challenge for researchers. Missing data can arise from various sources, such as respondents skipping questions in surveys, failures in measurement equipment, or the loss of data during processing (Altman and Bland, 2007). When the proportion of missing data is small and these missing values arise completely at random, a complete case analysis might be a straightforward and reasonable solution with little impact on the results. Unfortunately, this scenario is rare in real-world datasets. Performing a complete case analysis regardless of the missing data mechanism often results in a loss of information and biased estimates (Faisal and Tutz, 2021).

To address this challenge, researchers often resort to imputation methods to fill in the missing values based on observed data. However, generating valid imputations that closely reflect the true values and allow for reliable statistical inference is not trivial. Ignoring the uncertainty inherent in imputed values by treating them as equivalent to real observed values leads to overly optimistic precision estimates (de Goeij et al., 2013). Multiple imputation, first introduced by Donald Rubin, tackles this problem by generating multiple plausible datasets. This approach captures the uncertainty of missing data, ensuring valid conclusions. The so-called Rubin's Rules formalize this process, providing a systematic framework to combine results from multiple imputed datasets.

Over time, multiple imputation has gained popularity and is applied across various fields, including medical research (Hayati Rezvan et al., 2015), social sciences (see, e.g., Jenkins et al., 2011), and political science (Lall, 2016). Recent advancements have simplified the practical implementation and use of multiple imputation. Tools such as the `mice` package in `R` offer a straightforward and reproducible framework for conducting multiple imputations. Within multiple imputation, a range of models can be used to generate imputed datasets, provided they are appropriately specified for the data and the prediction task. However, the choice of imputation model can affect the quality of imputations (see, e.g., Akande et al., 2017; Chhabra et al., 2017). Thus, careful evaluation and thoughtful selection of imputation methods are crucial when executing multiple imputation. Among

the popular choices for imputation methods are non-parametric tree-based approaches, such as Classification and Regression Trees (see, e.g., Burgette and Reiter, 2010) and Random Forests (see, e.g., Shah et al., 2014).

This thesis aims to give a guideline which of these methods deliver better imputation results and to identify scenarios where one method may be preferred over the other. To achieve this, a simulation study is conducted, creating various scenarios that imitate real-world data complexities.

2 Study Overview

To assess the performance of different imputation methods, multiple datasets were generated. Specifically, datasets with 1000 observations and 500 observations were created, each containing either 5 or 10 covariates alongside a target variable (see Section 3). For each configuration, 1000 datasets were generated to reduce the effects of randomness. Missing values were introduced based on three different mechanisms, which are discussed in further detail in Section 4 and each of them with three different missing percentages. Each dataset was then subjected to multiple imputation using two methods: Classification and Regression Trees and Random Forests, described in Section 5.2. This led to

$$\underbrace{2}_{\substack{\text{configurations}\\\text{of observations}}} \times \underbrace{2}_{\substack{\text{configurations}\\\text{of covariates}}} \times \underbrace{1000}_{\substack{\text{datasets per}\\\text{configuration}}} \times \underbrace{3}_{\substack{\text{missing data}\\\text{mechanisms}}} \times \underbrace{3}_{\substack{\text{missing}\\\text{percentages}}} \times \underbrace{2}_{\substack{\text{imputation}\\\text{methods}}} = \underbrace{72000}_{\substack{\text{total number of}\\\text{multiple imputations}}} \tag{1}$$

multiple imputation rounds in total. For each round of multiple imputation, comprising $m = 10$ individual imputations, a pooled model was subsequently fitted to estimate the model parameters and their corresponding standard errors for each variable. The results were then analyzed in terms of bias, coverage rate, and the width of confidence intervals (CIs) to evaluate the accuracy and reliability of the imputation methods.

3 Data Generation

To address the research question, a simulation study was performed across various scenarios, featuring a conditionally Poisson-distributed count target variable. Two types of datasets were generated: the first containing the variables $x_1, \ldots, x_5$ and y, and the second including $x_1, \ldots, x_{10}$ and y. Each dataset was simulated with two sample sizes: $n = 500$ and $n = 1000$ observations.

The x-variables are generated in a stepwise procedure as follows:

$$
\begin{aligned}
\boldsymbol{x_1} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu = 0, \sigma^2 = 5), \\
x_{i,2} &= x_{i,1} + t_{\nu=10} \cdot 0.15, \\
x_{i,3} &= x_{i,2} + \text{Exp}(\lambda = 1) \cdot 0.15, \\
x_{i,4} &= x_{i,3} + \text{Poisson}(\lambda = 1) \cdot 0.1, \\
x_{i,5} &= x_{i,4} + \chi^2(\text{df} = 2) \cdot 0.1, \\
x_{i,6} &= x_{i,5} + \text{Gamma}(\alpha = 2, \beta = 5) \cdot 0.3, \\
x_{i,7} &= x_{i,6} + \text{Unif}(a = -0.2, b = 0.2), \\
x_{i,8} &= x_{i,7} + \text{Binom}(n = 1, p = 0.5) \cdot 0.3, \\
x_{i,9} &= x_{i,8} + \sqrt{\text{Exp}(\lambda = 0.5)} \cdot 0.3, \\
x_{i,10} &= x_{i,9} + \log(1 + \chi^2(\text{df} = 1)) \cdot 0.3, \\
&\text{with } i = 1, 2, \ldots, n.
\end{aligned}
\tag{2}
$$

Each $\boldsymbol{x_j}$ is rescaled as

$$
\frac{\boldsymbol{x_j} - \min(\boldsymbol{x_j})}{\max(\boldsymbol{x_j}) - \min(\boldsymbol{x_j})}, \quad \text{with } j = 1, 2, \ldots, p \tag{3}
$$

before being used in the calculation of the next variable.

This stepwise process generates each x-variable by adding a differently distributed component to the preceding x-variable, followed by standardization. This approach creates intercorrelations among the variables, leading to a complex correlation pattern, which is essential for the imputation later on. A visualization of the different x-variables for a dataset with a sample size of $n = 1,000,000$ can be found in Appendix A.1.

The coefficient vector $\boldsymbol{\beta}$ is defined as:

$$\boldsymbol{\beta} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^{p \times 1}, \quad \text{where } p \text{ is the number of } x\text{-variables.} \tag{4}$$

The linear predictor $\boldsymbol{\eta}$ is given by:

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}, \quad \text{where } \boldsymbol{X} \text{ is the design matrix } (n \times p). \tag{5}$$

Then, $\boldsymbol{\lambda}$ is calculated as:

$$\boldsymbol{\lambda} = \exp(\boldsymbol{\eta}). \tag{6}$$

Finally, the conditionally Poisson-distributed target variable $\boldsymbol{y}$ was generated as follows:

$$y_i \sim \text{Poisson}(\lambda_i). \tag{7}$$

To illustrate the approximate distribution of the target variable y in the data-generating process, a dataset with 1,000,000 observations and 10 covariates was generated and is shown in Figure 1.

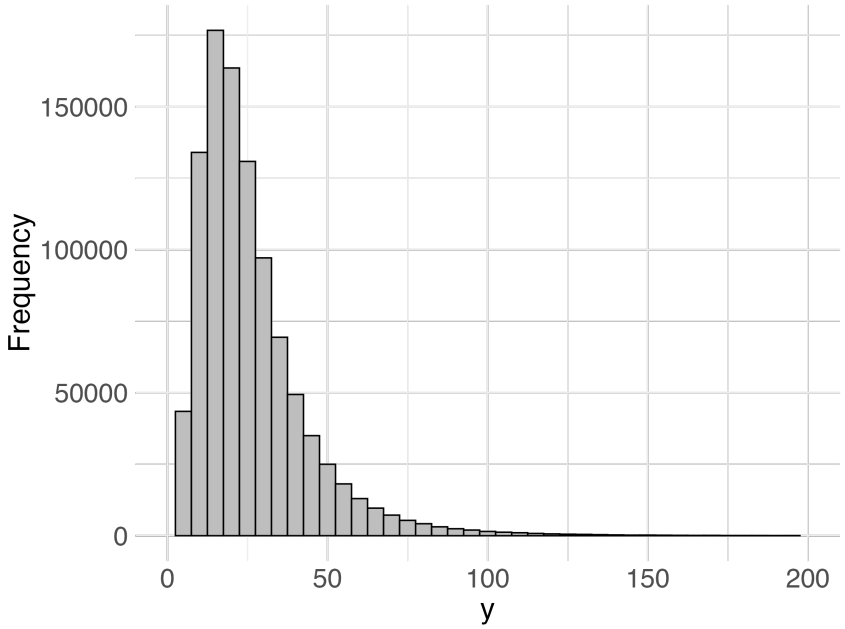


Figure 1: Histogram of the Poisson-distributed target variable y for a dataset with $n =$ 1,000,000 and $p = 10$

To examine the relationships among the generated variables, a Spearman correlation matrix was computed. The resulting correlation plot, presented in Figure 2, provides insight into the degree and direction of association between each x-variable and the target variable y.

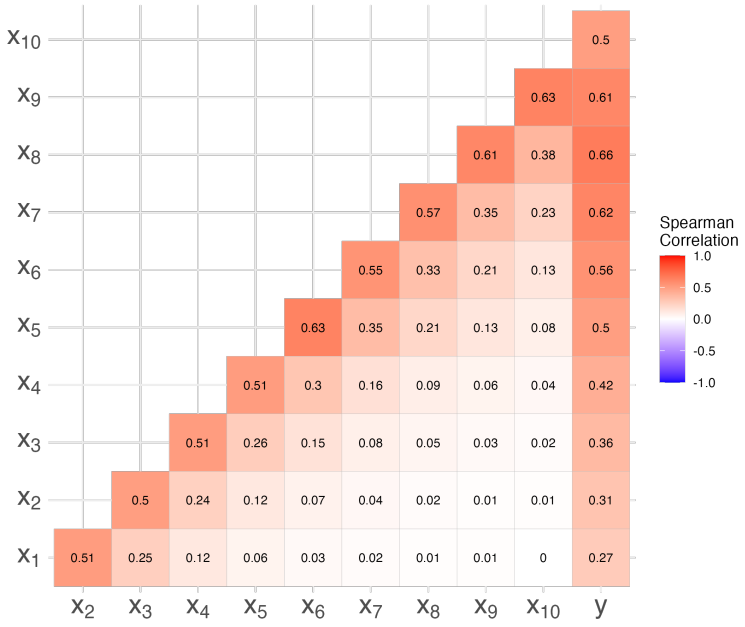


Figure 2: Spearman Correlation Matrix for a dataset with n = 1,000,000 and p = 10

To ensure that the random generation has minimal influence on the study, 1000 datasets were created for each scenario, as done before, for instance, by Lee and Carlin (2012) and Huque et al. (2018).

4 Missing Values

With the datasets generated, the next step involves the systematic introduction of missing values. The following section introduces different mechanisms of missingness. Subsequently, a brief description is provided on how these types of missingness are practically introduced into the synthetically generated data.

4.1 Missing Mechanisms

Various approaches exist to classify missing values in datasets. A common classification distinguishes between Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR), as described, for instance, in Enders (2022). These three mechanisms were also employed in this study to generate different datasets.

A data mechanism is classified as **MCAR** if the probability of a data point being missing is independent of both observed and unobserved data values. This implies that missingness is purely random and unrelated to any specific data values.

The formal definition of MCAR is given by the conditional distribution of the missingness indicator M, which denotes whether data is missing (1 if missing, 0 if observed). For MCAR, the probability of $M = 1$ is defined as:

$$P\left(M = 1 \mid \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \boldsymbol{\phi}\right) = P\left(M = 1 \mid \boldsymbol{\phi}\right) \tag{8}$$

where $\mathbf{Y}_{\text{obs}}$ represents observed data, $\mathbf{Y}_{\text{mis}}$ represents missing data, and $\boldsymbol{\phi}$ is a set of parameters describing the missingness mechanism. This equation shows that in an MCAR mechanism, the probability of missingness depends only on the parameters in $\boldsymbol{\phi}$ and is unrelated to the values of observed or missing data.

A data mechanism is defined as **MAR** if the probability of a data point being missing depends only on the observed data and not on the missing values themselves. This implies that after conditioning on observed data, missingness is independent of the unobserved data.

The formal definition for MAR is:

$$P\left(M = 1 \mid \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \boldsymbol{\phi}\right) = P\left(M = 1 \mid \mathbf{Y}_{\text{obs}}, \boldsymbol{\phi}\right). \tag{9}$$

This equation demonstrates that the probability of missingness is conditional only on observed data and does not depend on missing values, given $\mathbf{Y}_{\text{obs}}$. Under MAR, inference can be accurately performed by conditioning on the observed data, allowing methods like multiple imputation or maximum likelihood estimation to be applied.

A data mechanism is classified as **MNAR** if the probability that data is missing depends on the missing values themselves, even after accounting for observed data. In this case, missingness is directly related to the unobserved values and cannot be explained solely by the observed data. The formal definition for MNAR is given by the conditional distribution of M as:

$$P\left(M = 1 \mid \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \boldsymbol{\phi}\right). \tag{10}$$

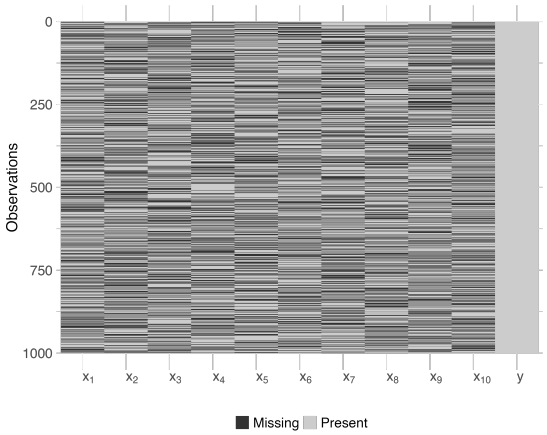
This expression does not simplify further since it depends on both observed data $\mathbf{Y}_{\text{obs}}$ and missing data $\mathbf{Y}_{\text{mis}}$. This indicates that even after accounting for observed values, the probability of missingness is influenced by the missing values themselves.

Since missingness under MNAR is directly related to the unobserved values, statistical analysis in this context requires specialized methods that explicitly model the missing data process. Unlike MCAR or MAR, standard approaches such as maximum likelihood estimation and multiple imputation produces biased results when used under MNAR conditions.

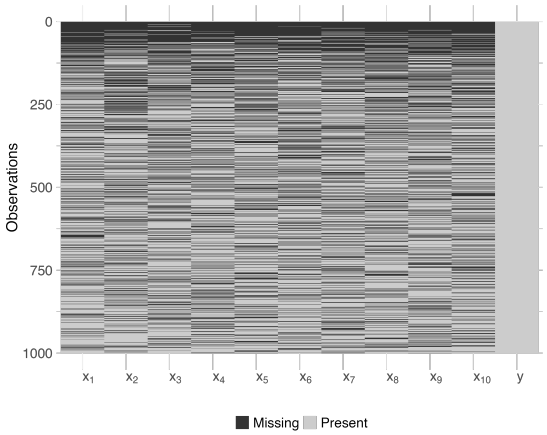
4.2 Simulating Missing Values

The datasets created in Section 3 were used to apply three distinct missing data mechanisms as described in Section 4.1. Missing values were introduced in the x-variables at three levels: 20%, 40%, and 60%. The y-variable was kept fully observed. This process was conducted in `R` using the `produce_NA` function from the `R-miss-tastic` platform, as documented in Mayer et al. (2024). This function makes use of the `ampute` function from the `mice` package introduced by van Buuren and Groothuis-Oudshoorn (2011).

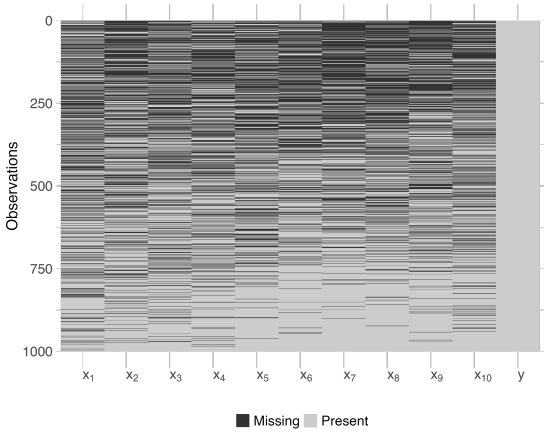
For the MCAR mechanism, a simple random generator determined which cells in the x-variables would be missing, independently of the values in those cells or any others.
In the MAR mechanism, missingness in the x-variables was generated based on values in the y-variable. A standard right-tailed logistic distribution function was used to calculate a weighted sum, creating a higher probability of missingness in the x-variables for larger values of y, as described in Schouten et al. (2018).
For the MNAR mechanism, missingness depended solely on the values within each cell of the respective x-variables. A right-tailed logistic distribution function was again used to calculate a weighted sum, such that higher values within the x-variables led to a higher probability of being replaced with NA (Schouten et al., 2018).
An illustration is provided for an example dataset with 1000 observations, 10 x-variables, and 40% missing values, sorted in descending order by y. The results are shown for MCAR in Figure 3a, for MAR in Figure 3b, and for MNAR in Figure 3c.



(a) MCAR: Missing values sorted by descending y-value ($n = 1000$, 40% missing)



(b) MAR: Missing values sorted by descending y-value ($n = 1000$, 40% missing)



(c) MNAR: Missing values sorted by descending y-value ($n = 1000$, 40% missing)

Figure 3: Comparison of missing mechanisms: MCAR, MAR, and MNAR, with missing values sorted by descending y-value for $n = 1000$ and 40% missing data.

5 Imputation Procedure

Once the datasets have been generated and missing values have been introduced, the next step involves their imputation. This chapter outlines the theoretical framework and practical implementation of the imputation process.

5.1 Multiple Imputation

This section gives a theoretical foundation to the concept of multiple imputation used in this study and is, unless otherwise specified, based on the content of van Buuren (2018). Early approaches to handling missing data often relied on single imputation methods, where each missing value is replaced with a single estimate, such as the mean or a predicted value based on other variables. While straightforward, this approach has an important limitation: it fails to account for the uncertainty associated with missing data. A single imputed value cannot capture the variability of the unobserved information and simply treats the newly imputed values as if they were never missing. Consequently, for instance, p-values are more likely to appear significant than they are in reality. Recognizing this flaw in missing value imputation, early concepts of multiple imputation were introduced in the 1970s, for instance by Rubin (1978), to better reflect this uncertainty.
In general, the goal is to estimate Q, which shall be a population-level quantity of interest. However, Q can only be observed if there is access to the full data-generating process or the entire population from which the estimand is derived. Consequently, Rubin (1996) states, that the primary aim of multiple imputation is to derive an estimate, denoted as $\hat{Q}$, that is both unbiased and provides valid CIs.
Unbiasedness can be formalized as follows, with Y being a random sample:

$$\mathbb{E}(\hat{Q} \mid Y) = Q. \tag{11}$$

To achieve valid CIs, extra conditions must hold. Let U represent the estimated variance-covariance matrix of $\hat{Q}$, then the following condition must hold:

$$\mathbb{E}(U \mid Y) \geq \mathrm{Var}(\hat{Q} \mid Y), \tag{12}$$

where $\text{Var}(\hat{Q} \mid Y)$ describes the variability introduced by the sampling procedure. These properties ensure that the procedure yields estimates that are both accurate and appropriately account for the uncertainty inherent in data with missing values. (Rubin, 1996)
The uncertainty in the estimate $\hat{Q}$ depends on the information available about the missing data (Y_{mis}). If Y_{mis} could be perfectly reconstructed, Q could be calculated with certainty. However, such perfect reconstruction is rarely possible. Instead, the distribution of Q given the observed data Y_{obs} is summarized by the posterior distribution $P(Q \mid Y_{\text{obs}})$. This posterior distribution can be written as:

$$P(Q \mid Y_{\text{obs}}) = \int P(Q \mid Y_{\text{obs}}, Y_{\text{mis}}) P(Y_{\text{mis}} \mid Y_{\text{obs}}) dY_{\text{mis}}. \tag{13}$$

This interprets as drawing plausible values for Y_{mis} from $P(Y_{\text{mis}} \mid Y_{\text{obs}})$, generating imputations denoted as $\dot{Y}_{\text{mis}}$. For each imputation, Q is calculated based on the completed dataset $(Y_{\text{obs}}, \dot{Y}_{\text{mis}})$ using $P(Q \mid Y_{\text{obs}}, \dot{Y}_{\text{mis}})$. By repeating this process for multiple imputations, the posterior distribution $P(Q \mid Y_{\text{obs}})$ is approximated as the average across all imputations. This iterative procedure simplifies the complex computation of $P(Q \mid Y_{\text{obs}})$ into easier steps.
Then, the posterior mean of $P(Q \mid Y_{\text{obs}})$ can be expressed as:

$$\mathbb{E}(Q \mid Y_{\text{obs}}) = \mathbb{E}\left(\mathbb{E}[Q \mid Y_{\text{obs}}, Y_{\text{mis}}] \mid Y_{\text{obs}}\right). \tag{14}$$

This means that the posterior mean of Q, given the observed data, is the average of the posterior means computed for each hypothetical completion of the dataset.
The combined estimate $\bar{Q}$ can be calculated by repeatedly drawing Y_{mis} from $P(Y_{\text{mis}} \mid Y_{\text{obs}})$, calculating Q for each completed dataset, and combining the results. The average of these estimates over m imputations provides the combined estimate of $\bar{Q}$, where ℓ denotes each individual imputation:

$$\bar{Q} = \frac{1}{m} \sum_{\ell=1}^{m} \hat{Q}_{\ell}. \tag{15}$$

The posterior variance of $P(Q \mid Y_{\text{obs}})$ can then be written as:

$$\text{Var}(Q \mid Y_{\text{obs}}) = \mathbb{E}\left[\text{Var}(Q \mid Y_{\text{obs}}, Y_{\text{mis}}) \mid Y_{\text{obs}}\right] + \text{Var}\left[\mathbb{E}(Q \mid Y_{\text{obs}}, Y_{\text{mis}}) \mid Y_{\text{obs}}\right]. \tag{16}$$

The term, $\mathbb{E}[\text{Var}(Q \mid Y_{\text{obs}}, Y_{\text{mis}}) \mid Y_{\text{obs}}]$, represents the expected value of the variances of the completed data also known as the within-variance. The other term, $\text{Var}[\mathbb{E}(Q \mid Y_{\text{obs}}, Y_{\text{mis}}) \mid Y_{\text{obs}}]$, quantifies the variability among the posterior means of Q calculated across different imputations, also known as the between variance. Together, these two components encapsulate the uncertainty in Q arising from both observed and missing data.

For an infinite number of imputations ($m = \infty$), the within and between variances are denoted as $\bar{U}_\infty$ and B_∞, respectively. The total variance in this hypothetical case is:

$$T_\infty = \bar{U}_\infty + B_\infty. \tag{17}$$

For real-world applications with a finite number of imputations m, the within-variance is approximated as the average variance across all imputations:

$$\bar{U} = \frac{1}{m}\sum_{\ell=1}^{m} \bar{U}_\ell, \tag{18}$$

where $\bar{U}_\ell$ represents the variance-covariance matrix of $\hat{Q}_\ell$ for the ℓ-th imputed dataset. The between-variance, which captures the variability among the imputed estimates, is calculated as:

$$B = \frac{1}{m-1}\sum_{\ell=1}^{m} (\hat{Q}_\ell - \bar{Q})(\hat{Q}_\ell - \bar{Q})^\top, \tag{19}$$

with $\bar{Q}$ computed as in Formula 15.

While it might appear intuitive to combine $\bar{U}$ and B directly to estimate the total variance T, this approach neglects the fact that $\bar{Q}$ itself is only an approximation of Q_∞. To account for this source of additional variability, the total variance can be adjusted using the following formula:

$$T = \bar{U} + B + \frac{B}{m}. \tag{20}$$

Without this correction, CIs would be too narrow, and therefore p-values would be underestimated. This equation 20 and equation 15 are the two key formulas for combining imputation results in the context of multiple imputation. Together, they are referred to as **Rubin's Rules**.

With these procedures described before it is relatively straightforward to conduct statistical tests for each parameter. Tests for one component can be derived using the standardized equation:

$$\frac{Q - \bar{Q}}{\sqrt{T}} \sim t_\nu \tag{21}$$

with ν denoting the degrees of freedom for the t-distribution given by:

$$\nu = \frac{\frac{m-1}{\lambda^2}\frac{n-p+1}{n-p+3}(n-k)(1-\lambda)}{\frac{m-1}{\lambda^2} + \frac{n-p+1}{n-p+3}(n-k)(1-\lambda)} \tag{22}$$

with $\lambda \in [0, 1]$ representing the proportion of variation due to the missing data. The parameter λ is defined as:

$$\lambda = \frac{B + \frac{B}{m}}{T}. \tag{23}$$

A formula for the CIs of Q is then given by:

$$\text{CI} = \bar{Q} \pm t_{\nu,1-\alpha/2}\sqrt{T}, \tag{24}$$

where $t_{\nu,1-\alpha/2}$ is the critical value from the t-distribution with ν degrees of freedom.

In conclusion, multiple imputation addresses the problems of single imputation by creating several ($m > 1$) complete datasets, each containing different values for the missing data. This process not only introduces variability in the imputed values but also allows for more robust statistical inference. Each of these m datasets is then analyzed separately, producing multiple estimates for each estimand. Finally, these estimates are pooled to generate a single result, including an adjustment for the imputation uncertainty. This method of imputing missing values and then pooling results has become a foundational tool in statistical analysis, providing reliable inferences in the presence of missing data.

5.2 Tree-Based Methods

The previous section introduced the theoretical background of multiple imputation, motivating its use in missing value scenarios. Multiple imputation can be conducted using different methods, each with different advantages and limitations. This study focuses specifically on tree-based approaches. This section provides a brief theoretical overview of the principles underlying these approaches.

5.2.1 Classification and Regression Trees

Classification and Regression Trees (CART) are one of the methods employed for multiple imputation in this study. CART is a machine-learning method originally developed by Breiman et al. (1984). The method recursively binarily partitions the feature space $\mathcal{X}$ into subsets and provides predictions for the target variable $\mathcal{Y}$ based on these partitions. A decision tree T is formally defined as a mapping function:

$$f_T : \mathcal{X} \to \mathcal{Y}, \tag{25}$$

which assigns a predicted output to each input observation. (Hastie, 2009)
It can be differentiated between regression and classification tasks. In regression tasks, CART partitions data to minimize prediction error within each region, commonly measured by the mean squared error between observed and predicted values. This produces a piecewise constant model, segmenting the predictor space into regions with similar response values. Figure 4 illustrates a basic regression tree where each terminal node (rectangle) represents a constant prediction value within that region. (Loh, 2011)
In classification tasks, CART divides the predictor space into regions, assigning a class to each region based on majority rule or a cost-minimization approach. Each split is made on a predictor variable, resulting in two branches that proceed until reaching terminal nodes. These terminal nodes mark the final class predictions. Figure 5 depicts a schematic for a classification tree with three classes. (Loh, 2011)
To prevent the growth of excessively large trees and avoid overfitting, it is essential to introduce stopping criteria. Several approaches are commonly employed for this purpose.

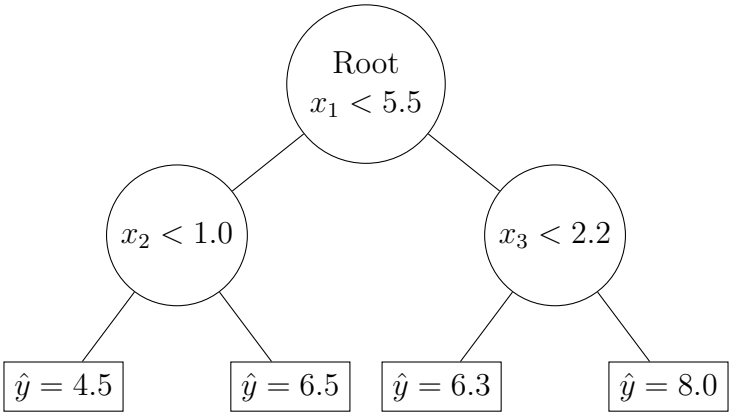


Figure 4: Schematic of a regression tree

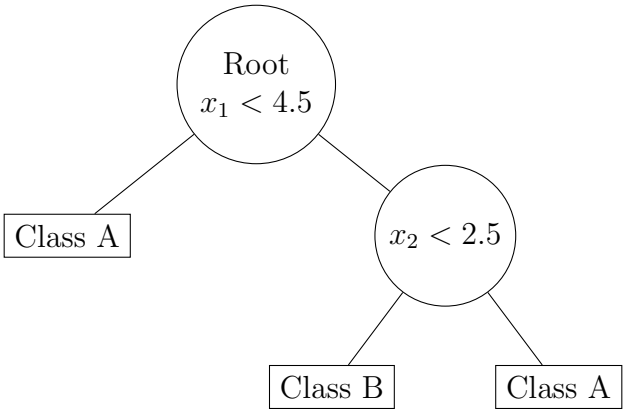


Figure 5: Schematic of a classification tree

Frequently used methods include a minimum decrease in the sum of squared errors, a minimum reduction in misclassification error rate, or cost-complexity pruning, which balances goodness of fit against tree size. (Hastie, 2009)

A key advantage of CART for imputation is its ability to handle both continuous and categorical variables. Additionally, it can handle irrelevant input variables and it is robust against outliers – the latter making it a more reliable choice for datasets with irregular or extreme values. (Hastie, 2009)

5.2.2 Random Forest

Apart from CART, Random Forest (RF) is a popular tree-based method for imputing missing values. Developed by Breiman (2001), RF is an ensemble machine learning method, which combines multiple trees into one model.

Same as with CART, a RF F is formally defined as a mapping function:

$$f_F : \mathcal{X} \to \mathcal{Y}, \tag{26}$$

which assigns a predicted output to each input observation.

The RF algorithm consists of two main steps. First, B decision trees are built using bootstrapped samples of the data. This means generating a random sample with replacement from the original dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ to create a new training set $\mathcal{D}_b$ for each tree T_b. Formally, for each $b \in \{1, \ldots, B\}$:

$$\mathcal{D}_b = \{(x_{i_b}, y_{i_b}) \mid i_b \sim \text{Uniform}(1, n)\}, \tag{27}$$

where i_b is sampled independently n times with replacement. This results in $\mathcal{D}_b$ containing approximately 63% unique observations from $\mathcal{D}$, with the rest being duplicates due to the sampling process. Second, for each tree in the ensemble, a random subset of the predictors is selected at each node to determine the best split. This process continues recursively until a predefined minimum node size is reached. The result is an ensemble of trees, denoted as $\{T_b\}_{b=1}^B$. (Hastie, 2009)

For prediction, the method differs depending on the task. In regression, the prediction for a new observation $\mathbf{x}$ is obtained by averaging the predictions of all trees:

$$\hat{f}_{\mathrm{F}}^B(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} T_b(\mathbf{x}). \tag{28}$$

In classification, each tree votes for a class label, and the final prediction is the majority vote:

$$\hat{C}_{\mathrm{F}}^B(\mathbf{x}) = \text{majority vote}\{\hat{C}_b(\mathbf{x})\}_{b=1}^B. \tag{29}$$

(Hastie, 2009)

The approach of RF combines bagging (bootstrap aggregation) with random feature selection to reduce overfitting and improve generalization performance. Additional advantages for imputation with RFs include their ability to model complex dependency structures, their strong performance even without hyperparameter tuning, and their robustness to

outliers and extreme values. (Hastie, 2009)

5.3 `mice` package

The `mice` package (**M**ultivariate **I**mputation by **C**hained **E**quations) by van Buuren and Groothuis-Oudshoorn (2011) has all these theoretical steps described in Section 5.1, Section 5.2.1 and Section 5.2.2 implemented in `R` and has become a standard choice for multiple imputation in `R`. The package consists of three main functions described in the following. An overview of the classic workflow of the `mice` package can be found in Figure 6.

5.3.1 `mice` function

This section introduces the `mice` function based on the package description by van Buuren and Groothuis-Oudshoorn (2011) for the respective package. In accordance with descriptions from multiple imputation in Section 5.1, the `mice` function is designed to handle missing data by iteratively imputing plausible values.
Let Y be the hypothetically complete dataset, which is only partially observed. This data follows a p-variate distribution $P(Y \mid \theta)$, where θ is a vector of unknown parameters that fully specifies the distribution of Y. The challenge lies in estimating the posterior distribution of θ as it is hard to determine.
The mice algorithm achieves this by breaking down the problem into conditional distributions of the form:

$$P(Y_1 \mid Y_{-1}, \theta_1), \quad P(Y_2 \mid Y_{-2}, \theta_2), \quad \ldots, \quad P(Y_p \mid Y_{-p}, \theta_p). \tag{30}$$

Each parameter, denoted as $\theta_1, \ldots, \theta_p$, represents a component of the conditional distributions that is used in the sampling process. These parameters are not required to directly correspond to a factorization of the joint probability distribution $P(Y \mid \theta)$. Instead, the algorithm simplifies the problem by focusing on the conditional distributions individually. This approach enables an iterative process where values for each parameter are sampled one at a time from their respective conditional distributions.

The algorithm begins by using initial estimates obtained from the observed marginal distributions. Subsequently, at each iteration t, the `mice` function applies a Gibbs sampling approach to iteratively update the imputations. For each variable Y_j with missing values, a predictive model $P(Y_j \mid Y_{-j}, \theta_j)$ is constructed using the observed data $Y_{\text{obs},j}$ and the most recent imputations for the missing data $Y_{-j}^{(t-1)}$. The model, parameterized by θ_j, is designed to approximate the conditional distribution $P(Y_j \mid Y_{-j})$. Methods such as linear regression, predictive mean matching, or as in this study CART and RF can be employed to estimate this conditional distribution. At this stage, the parameters θ_j are drawn from their posterior distribution, as given by:

$$\theta_j^{*(t)} \sim P(\theta_j \mid Y_{\text{obs},j}, Y_1^{(t-1)}, \ldots, Y_p^{(t-1)}). \tag{31}$$

Using these parameters, the missing values for Y_j are imputed by sampling from the conditional distribution:

$$Y_j^{*(t)} \sim P(Y_j \mid Y_{\text{obs},j}, Y_1^{(t-1)}, \ldots, Y_p^{(t-1)}, \theta_j^{*(t)}). \tag{32}$$

This process is iteratively applied to all variables with missing data, cycling through them until the imputations converge. Convergence is assessed by comparing the imputations $Y_j^{*(t)}$ from the current iteration with those from the previous iteration, $Y_j^{*(t-1)}$. Typically, the algorithm converges quickly, as the imputations stabilize after only a few iterations.
The term "chained equations" reflects the structure of the algorithm, where missing values for one variable are imputed at a time, assuming that all other variables are either observed or have already been imputed. For instance, when imputing Y_1, the algorithm relies on $Y_2, Y_3, \ldots, Y_p$, which are either observed or were updated during the previous iteration. This sequential process of imputing each variable in turn, with dependencies on the updated values of other variables, forms a chain of equations, giving the algorithm its name.
The implementation of the `mice` function in `R` executes this algorithm m times independently, resulting in m imputed datasets. Due to the inherent stochasticity of the sampling process, the imputed values differ across these datasets.

5.3.2 `with` function

After creating m datasets with imputed values, parameter estimation is performed independently for each dataset. This can be achieved using the `with` function from the `mice` package, applying an appropriate formula and model class. Consequently, m separate estimates are obtained for each coefficient. In this model-fitting process, the imputed datasets are treated as complete and as if the imputed data were observed. (van Buuren and Groothuis-Oudshoorn, 2011)

5.3.3 `pool` function

The `pool` function combines the individual results from each imputed dataset and the corresponding models into one final set of results. This process is carried out according to Rubin's Rules, as outlined in Formula 15 and Formula 20 in Section 5.1. The pooled results provide a single estimate for each parameter. These final results can then be assessed and evaluated. This marks the last step of the imputation process. (van Buuren and Groothuis-Oudshoorn, 2011)

5.3.4 Imputation with CART

Imputation in the MICE algorithm is performed within the `mice` function and allows for the use of various imputation methods to model the conditional distribution of the missing values given the observed data. To impute data using CART, the method argument is set to `"cart"`. The underlying function used for this process is `mice.impute.cart`, which is also part of the `mice` package. This implementation relies on the `rpart` package developed by Therneau et al. (2015). The default settings for each tree specify a minimum bucket size of 5, meaning that each terminal node must contain at least 5 observations. Additionally, the complexity parameter is set to 1×10^{-4}, which requires any potential split to reduce the overall lack of fit by at least this factor to be considered. For regression trees, this implies that each split must improve the R^2-value by at least 1×10^{-4} to be included in the model. (van Buuren and Groothuis-Oudshoorn, 2011)

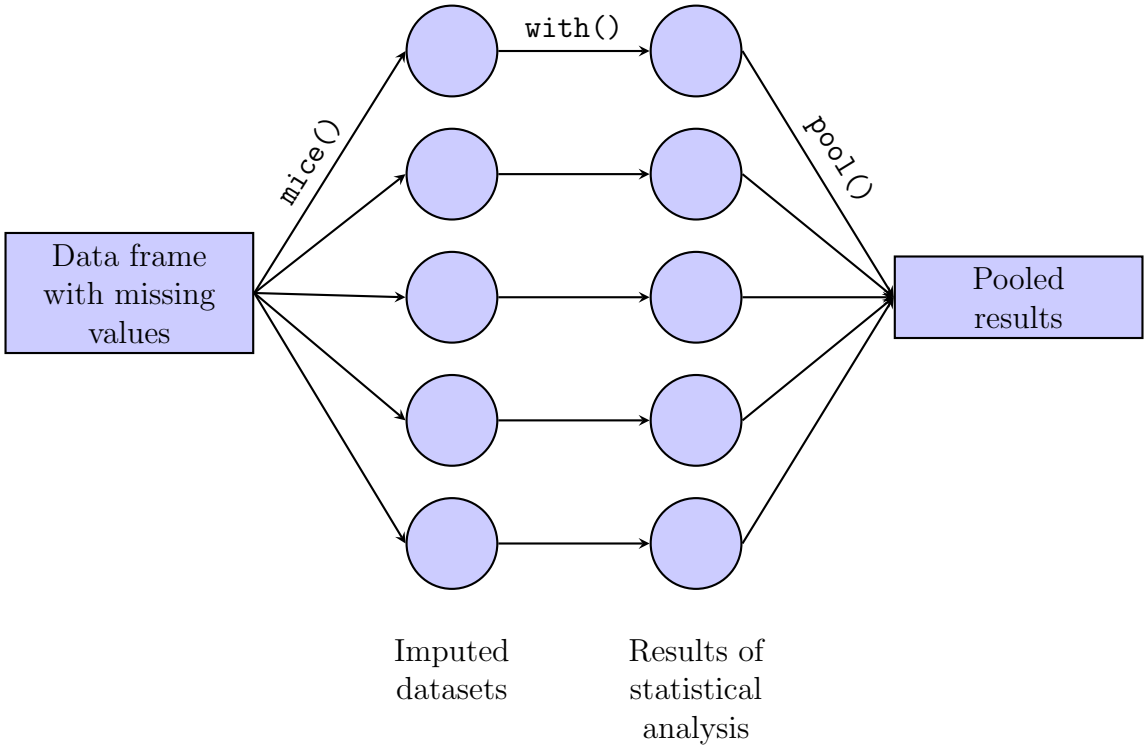


Figure 6: Flowchart of the multiple imputation process using the MICE algorithm, adapted from Zhang (2016) with $m = 5$

5.3.5 Imputation with Random Forest

Another method used for imputation in this study is RF, which models missing data through ensemble learning. To impute using RF, the method argument in the `mice` function is set to `"rf"`. The underlying function for this process is `mice.impute.rf`, which is part of the `mice` package. By default, this implementation relies on the `ranger` package developed by Wright et al. (2019) for the creation of the random forests. The default settings include a maximal number of trees of 10 and a minimal number of predictors considered for each split set to $\sqrt{p}$, where p is the total number of predictors. (van Buuren and Groothuis-Oudshoorn, 2011)

5.4 Implementation

The datasets created in Section 4.2 were imputed using multiple imputation within the `mice` framework, as described in Section 5.3. For each imputation method, ten imputations ($m = 10$) were performed, as done by Harel and Zhou (2007) and Lüdtke et al.

(2017) for instance. The default settings were used for both imputation methods: CART as described in Section 5.3.4 and RF as described in Section 5.3.5. As demonstrated by Shah et al. (2014), increasing the number of trees beyond 10 does not necessarily result in a meaningful improvement in the RF method.

After imputation, a Poisson regression generalized linear model was fit to the imputed datasets using the `with` function, as the data is conditionally Poisson-distributed (see Section 3). The results from the regression models were then combined using Rubin's Rules via the `pool` function. Once the results are pooled, the imputation process is finished and the results can be evaluated.

6 Analysis Metrics

To evaluate when to use which imputation mechanism three different metrics were analyzed for each scenario. In general, an effective imputation method aims to minimize bias, achieve coverage of at least 95%, and produce narrow CIs. These are common metrics for evaluating multiple imputation results (see, e.g., Slade and Naylor (2020); Carpenito and Manjourides (2022); Deng and Lumley (2024)).

Firstly, the bias for each parameter, each imputation method, and each scenario was calculated. The bias quantifies the systematic deviation of the estimated parameter values from the true parameter value and is computed using the following formula:

$$\text{Bias}_j = \frac{1}{K}\sum_{k=1}^{K}(\hat{\theta}_{k,j} - \theta_j), \tag{33}$$

where $\hat{\theta}_{k,j}$ is the estimated value of parameter j in the k-th imputed dataset, θ_j denotes the true value of the parameter ($\theta_0 = 0$ and $\theta_j = 1$ for $j = 1, \ldots, 10$; see Section 3), and K is the total number of imputed datasets for a single scenario ($K = 1000$).

Secondly, the coverage rate for each parameter, each imputation method, and each scenario was calculated. The coverage rate indicates the proportion of times the true parameter value is contained within the 95% CIs of the estimates across the imputed datasets. It is calculated using the following formula:

$$\text{Coverage Rate}_j = \frac{1}{K}\sum_{k=1}^{K}\mathbb{I}\left(\theta_j \in \text{CI}_{k,j}\right), \tag{34}$$

where $\text{CI}_{k,j}$ represents the CI for parameter $\hat{\theta}_{k,j}$ and $\mathbb{I}\left(\theta_j \in \text{CI}_{k,j}\right)$ is an indicator function that that equals 1 if θ_j lies within CI_k, and 0 otherwise. A desirable coverage rate is close to 95%.

Finally, the width of the CIs was calculated for each parameter, scenario, and imputation method. This measure is particularly useful because very wide CIs are likely to include the true parameter value, leading to higher coverage rates. However, this also means that coverage rate alone is not sufficient to assess the quality of the imputation method; it must be considered in combination with the CI width. The width of the CIs reflects the

precision of the parameter estimates. It is calculated using the following formula:

$$\text{CI Width}_{k,j} = \text{Upper Bound}_{k,j} - \text{Lower Bound}_{k,j} \tag{35}$$

where $\text{Upper Bound}_{k,j}$ and $\text{Lower Bound}_{k,j}$ represent the upper and lower limits of the CI, respectively, for parameter $\hat{\theta}_{k,j}$ in the k-th imputed dataset. Narrower CIs indicate higher precision and are generally preferred, provided that the coverage rate remains adequate.

7 Results

There are a total of 36 scenarios analyzed in this study, leading to 36 comparisons across the three metrics: bias, coverage rate, and CI width.

To provide a concise summary of the results, an overview metric is introduced to evaluate the relative performance of CART and RF across three criteria: bias, coverage rate, and median CI width. For each parameter, CART is considered superior if it shows a lower bias, a higher coverage rate, or a smaller median CI width compared to RF. For each parameter where CART is superior, a score of 1 is added. If both methods produce identical values for a metric, a score of 0.5 is assigned for that parameter. The total score is then calculated by summing the scores across all parameters where CART outperforms RF. The maximum possible score depends on the number of parameters p, including the intercept. For scenarios with $p = 5$, the maximum score is 6, while for $p = 10$, the maximum score is 11. For instance, a score of 1 in a given metric, such as coverage rate, indicates that CART outperforms RF for one parameter, while RF achieves better results for the remaining parameters. Conversely, a score of 0 across all metrics indicates that RF consistently outperforms CART for all parameters, suggesting that RF is the preferred method for similar data scenarios. However, this overview metric should only serve as a general guide. To form a comprehensive conclusion, consult the detailed tables of the individual metrics. The specific values for bias, coverage rate, and median CI width provide deeper insights and can be found in Appendix A.2, A.3, and A.4, respectively. The results of the overview metric are summarized in Table 1 for $p = 5$ and Table 2 for $p = 10$. In these tables, the cells are sequentially color-coded to reflect performance: orange indicates that CART outperforms RF for more than half of the parameters in a given metric, while blue indicates that RF outperforms CART for more than half of the parameters.

Table 1: Overview of performance metrics for CART and RF with $p = 5$. Orange indicates that CART outperforms RF for more than half of the parameters in a given metric, while blue indicates that RF outperforms CART for more than half of the parameters.

Missing Mech.	**n**	**p**	**Missing Perc.**	**Bias**	**Coverage Rate**	**CI Width**
MCAR	500	5	20	5.5	1	4.5
MCAR	500	5	40	5	2	0.5
MCAR	500	5	60	5	3	0
MCAR	1000	5	20	5.5	2	3.5
MCAR	1000	5	40	5	2	0.5
MCAR	1000	5	60	5	4	0
MAR	500	5	20	5	0.5	0.5
MAR	500	5	40	4.5	2	0
MAR	500	5	60	4.5	4	0
MAR	1000	5	20	5	2	0
MAR	1000	5	40	5	4	0
MAR	1000	5	60	5	4	0
MNAR	500	5	20	2.5	0	0
MNAR	500	5	40	2.5	1	0
MNAR	500	5	60	1.5	3	0
MNAR	1000	5	20	3	2	0
MNAR	1000	5	40	3	3	0
MNAR	1000	5	60	3	3	0

For $p = 5$, Table 1 shows that CART tends to produce lower bias values for more parameters in scenarios with MCAR and MAR missingness mechanisms compared to RF. In contrast, no clear superiority is observed between the two methods for coverage rate, suggesting that neither method consistently outperforms the other in this metric. It is crucial to interpret the coverage rate alongside the CI width, as wider CIs naturally increase the chance of containing the true parameter. In this context, a consistent trend emerges: RF generally produces narrower median CI widths compared to CART across most scenarios and parameters.

Table 2: Overview of performance metrics for CART and RF with $p = 10$. Orange indicates that CART outperforms RF for more than half of the parameters in a given metric, while blue indicates that RF outperforms CART for more than half of the parameters.

Missing Mech.	n	p	Missing Perc.	Bias	Coverage Rate	CI Width
MCAR	500	10	20	5.5	0	11
MCAR	500	10	40	4	0	4
MCAR	500	10	60	1.5	2.5	0
MCAR	1000	10	20	7.5	0	11
MCAR	1000	10	40	5.5	2	2
MCAR	1000	10	60	4.5	6.5	0
MAR	500	10	20	7.5	2.5	11
MAR	500	10	40	5.5	4.5	8
MAR	500	10	60	2	2	0
MAR	1000	10	20	5	2	9
MAR	1000	10	40	4.5	4.5	0
MAR	1000	10	60	4.5	4	0
MNAR	500	10	20	4	2.5	9
MNAR	500	10	40	3	2.5	1
MNAR	500	10	60	3	3.5	1
MNAR	1000	10	20	3	2	2.5
MNAR	1000	10	40	3	2	0
MNAR	1000	10	60	4	4	0

Table 2 presents the results for scenarios with $p = 10$. Since there are 11 parameters, the maximum possible score is 11, representing the case where CART outperforms RF for all parameters in a given metric. While no clear overall pattern emerges, it is notable that RF tends to perform better across all three metrics in scenarios characterized by a MNAR missingness mechanism. Unlike the observations for $p = 5$, where CART demonstrated lower bias for more parameters under MCAR and MAR mechanisms, such a clear trend is not visible for $p = 10$.

As these two tables are only intended to provide an overview, four scenarios are analyzed in greater depth.

First, the basic scenario with an MCAR missingness mechanism, $n = 500$, $p = 5$, and 20% missingness proportion is analyzed. The bias for this scenario is presented in Table 3. Overall, the bias is relatively low, with no signs of systematic bias being introduced.

In this aspect, CART appears to perform slightly better though. The coverage rate for this scenario is shown in Table 4, where it can be observed that coverage is high for both methods across all parameters, with RF achieving slightly higher coverage for almost all parameters. Turning to the CI width, Figure 7 presents the CI widths in a boxplot (excluding outliers). The distributions of CI widths for CART and RF are nearly identical, although RF exhibits a marginally higher median CI width for more parameters. In conclusion, for this basic scenario under the strong MCAR assumption, neither method demonstrates clear superiority. For similar cases, the computationally less expensive and faster CART imputation can be a practical choice.

Table 3: Bias for $n = 500$ with $p = 5$ for MCAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.03	-0.08	0.03	0.02	0.02	-0.07
RF	0.03	-0.09	0.05	0.03	0.05	-0.10

Table 4: Coverage Rate (%) for $n = 500$ with $p = 5$ for MCAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	91.1	92.9	95.6	96.2	96.4	94.3
RF	93.5	93.6	97.2	97.5	97.4	93.3

Another scenario analyzed in detail involves the same conditions as the previous scenario, with an MCAR missingness mechanism, $n = 500$, and 20% missingness, but with $p = 10$ parameters. The bias, as shown in Table 5, is very low overall, with minimal differences between CART and RF. However, as presented in Table 6, RF achieves consistently higher coverage rates across all parameters compared to CART. Conversely, Figure 8 illustrates that CART produces smaller median CI widths for all parameters. Given this trade-off between coverage rate and CI width, it is hardly possible to definitively recommend one method over the other for this scenario. The choice of method ultimately depends on the relative importance of these metrics in the context of the specific study.

Furthermore, a detailed analysis is provided for a scenario with a MAR missingness mechanism, $n = 500$, $p = 10$, and a missing percentage of 60%. Table 7 indicates that RF imputation yields lower bias for most parameters compared to CART. Table 8 highlights

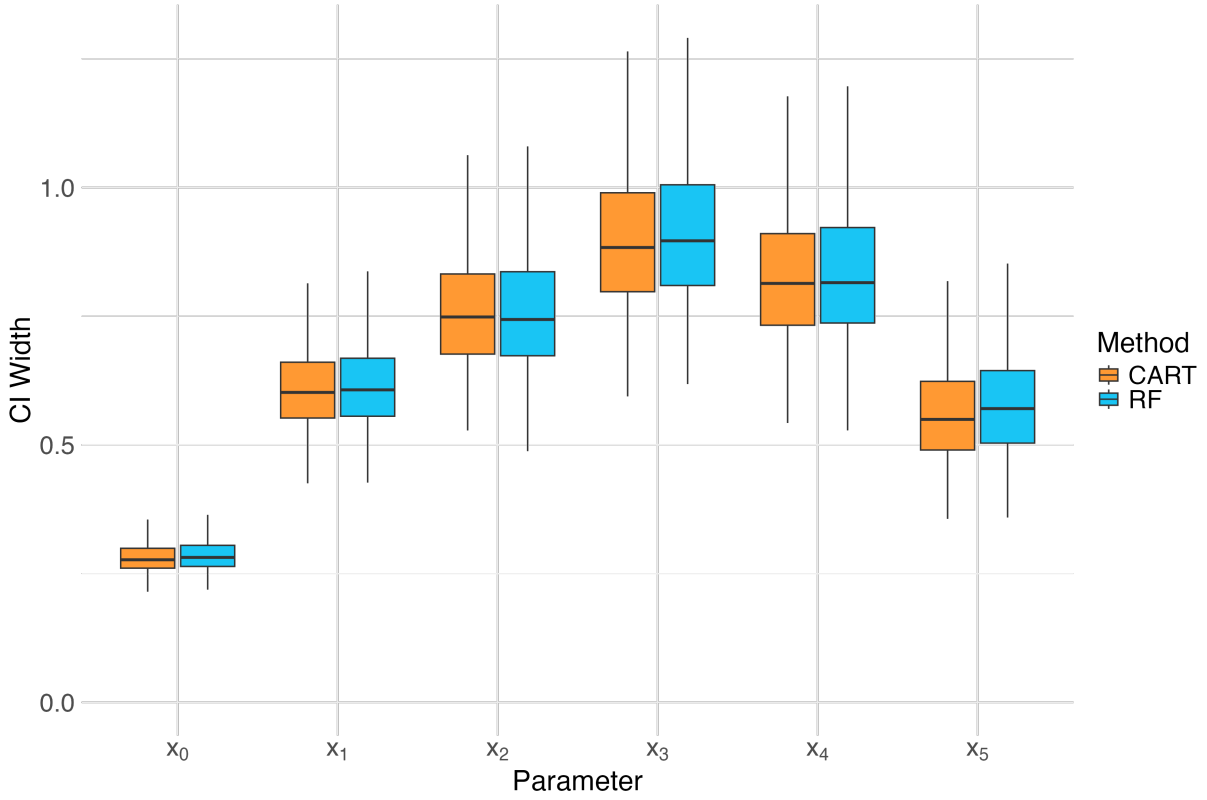


Figure 7: Width of CIs for MCAR mechanism, $n = 500$, $p = 5$ and 20% missing values

Table 5: Bias for $n = 500$ with $p = 10$ for MCAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.11	-0.14	0.00	-0.03	0.00	-0.00	0.04	0.03	0.02	-0.02	-0.09
RF	0.09	-0.15	0.02	-0.02	0.01	-0.00	0.07	0.03	0.01	-0.01	-0.09

Table 6: Coverage Rate (%) for $n = 500$ with $p = 10$ for MCAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	73.3	90.1	98.0	97.2	95.9	95.8	95.6	96.6	95.6	97.2	96.0
RF	86.0	92.2	98.2	99.0	98.0	97.3	97.4	98.7	98.7	99.0	98.1

that the coverage rate is extremely high for both methods. As depicted in Figure 9, the CI widths are very large for both methods, as consequence of the high percentage of missing data and the missing mechanism. Nonetheless, RF consistently produces slightly smaller median CI widths compared to CART. For scenarios like this, RF imputation is the preferable method, though it is important to acknowledge that the resulting CIs for the parameters are very large.

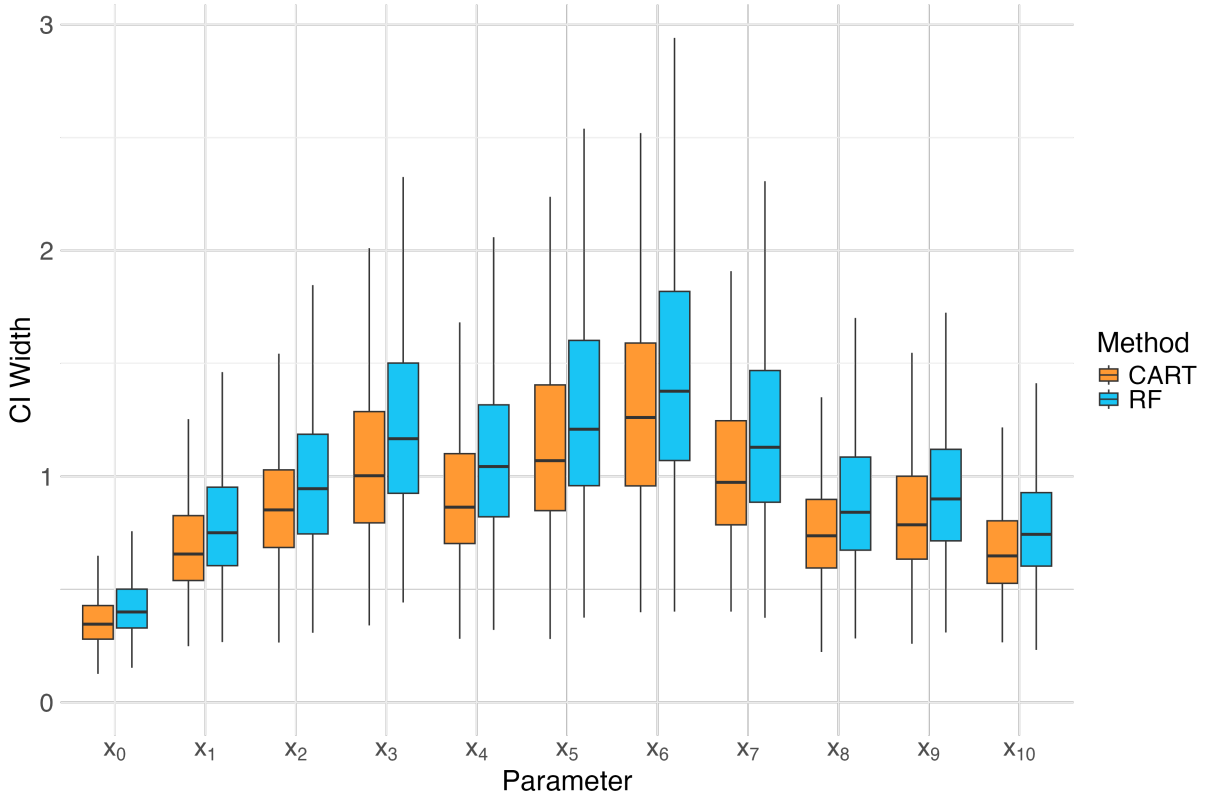


Figure 8: Width of CIs for MCAR mechanism, $n = 500$, $p = 10$ and 20% missing values

Table 7: Bias for $n = 500$ with $p = 10$ for MAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.58	-0.42	-0.27	-0.08	0.06	0.13	-0.02	0.40	0.22	-0.31	-0.40
RF	0.47	-0.40	-0.22	0.07	0.05	0.25	0.21	0.34	0.08	-0.23	-0.31

Table 8: Coverage Rate (%) for $n = 500$ with $p = 10$ for MAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	90.1	99.8	99.9	100.0	100.0	99.3	99.8	99.7	99.6	99.8	99.5
RF	94.2	99.3	100.0	100.0	100.0	100.0	100.0	99.9	100.0	99.9	99.8

The final scenario analyzed in detail involves an MNAR mechanism, with $n = 1000$, $p = 10$, and 40% missing data. Table 9 shows that the bias is higher compared to other scenarios, as anticipated given the MNAR mechanism. However, RF imputation yields less biased estimates for more parameters. Table 10 presents the coverage rates for the parameters. While both methods exhibit relatively small differences, RF demonstrates a slightly higher coverage rate for most parameters. Notably, the coverage rate for β_0 is close

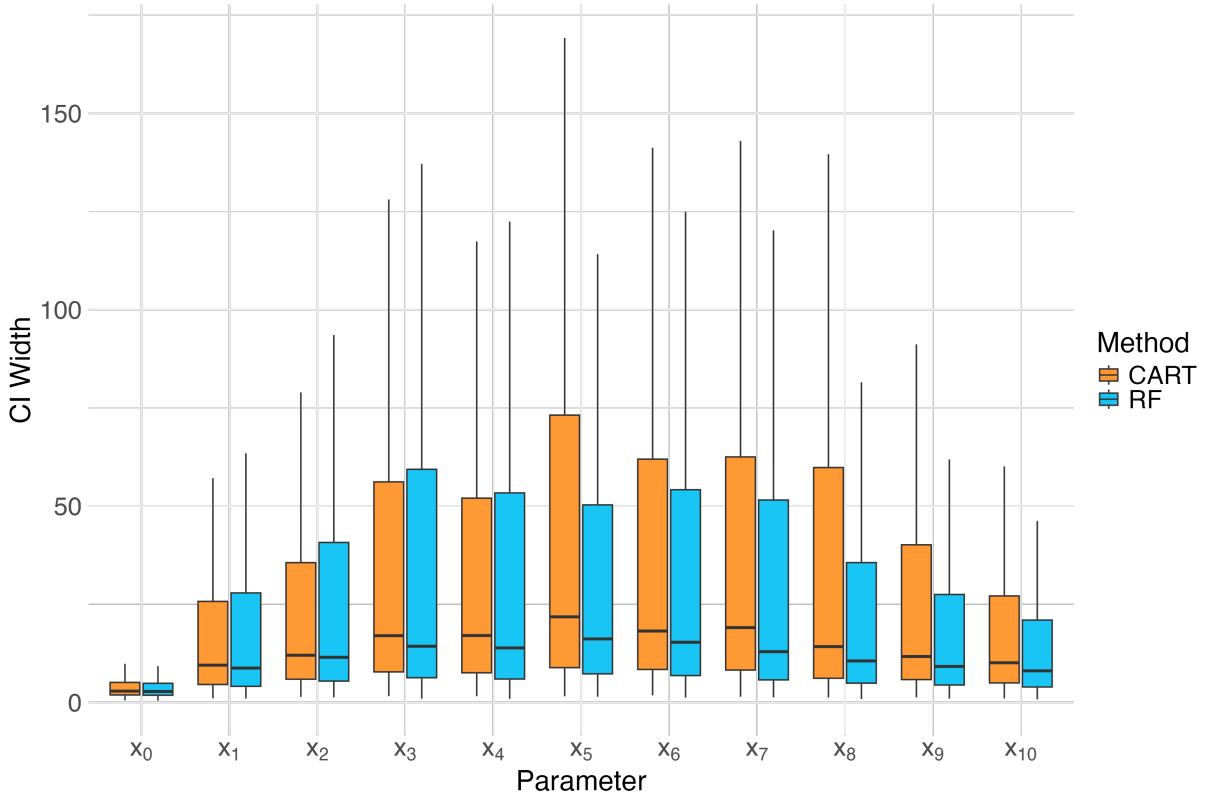


Figure 9: Width of CIs for MAR mechanism, $n = 500$, $p = 10$ and 60% missing values

to 0, which is expected under the MNAR mechanism. This occurs because the models estimate an intercept to adjust for the MNAR mechanism, even though such an intercept does not exist in the true model. Figure 10 illustrates the CI widths, showing that CART consistently produces higher median CI widths for all parameters compared to RF. This scenario represents a clear case where RF imputation outperforms CART. Nevertheless, caution is required when interpreting the pooled results, as the MNAR mechanism leads to heavily biased estimates.

Table 9: Bias for $n = 1000$ with $p = 10$ for MNAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	1.14	-0.43	-0.29	-0.27	-0.18	0.18	-0.05	0.05	0.12	-0.18	-0.42
RF	0.99	-0.40	-0.25	-0.12	-0.15	0.22	0.10	0.08	0.04	-0.11	-0.27

As outlined in Section 2, the analysis was conducted across a wide range of scenarios. The detailed results for bias are provided in Appendix A.2, for coverage rate in Appendix A.3, and for CI widths in Appendix A.4. Several general observations can be made.

Table 10: Coverage Rate (%) for $n = 1000$ with $p = 10$ for MNAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	2.0	90.0	97.9	98.8	99.3	97.4	99.0	99.5	98.5	98.4	93.4
RF	3.7	90.3	97.8	99.4	99.2	99.0	99.2	99.6	99.4	99.0	97.6

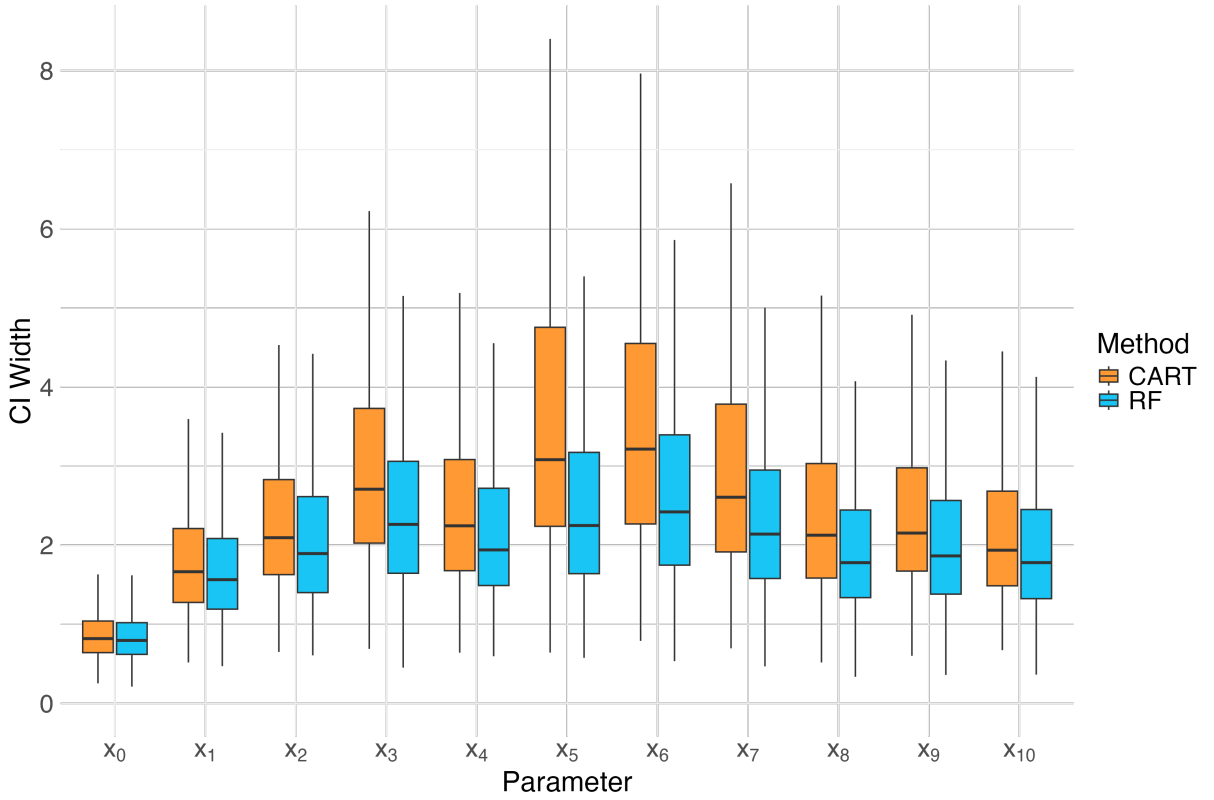


Figure 10: Width of CIs for MNAR mechanism, $n = 1000$, $p = 10$ and 40% missing values

RF frequently produces narrower CIs. This is consistent with the structure of RF as an ensemble model, which reduces variability in individual imputations, resulting in a smaller combined variance and consequently narrower CIs. The coverage rate is generally very high across most scenarios, indicating that both imputation methods reliably capture the true parameter values. For scenarios involving an MNAR mechanism, a relatively high bias is observed for many parameters, which can be attributed to the nature of the data removal process. Another notable observation is that the intercept is estimated away from its true value in MNAR scenarios, resulting in low coverage rates and high bias for this parameter.

Overall, RF appears to be the better choice in many of the analyzed scenarios. However,

making a universal recommendation is difficult, as each scenario requires individual assessment. In certain cases, the less computationally intensive CART imputation may still serve as a reasonable alternative.

8 Outlook and Conclusion

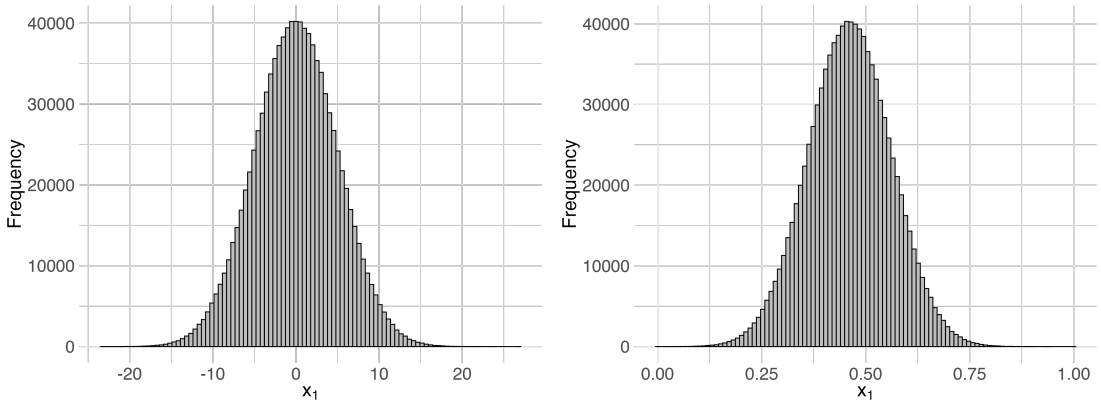
This study has limitations that should be acknowledged and could be used to build on further research. First, the number of imputations (m) was fixed at 10 across all scenarios. While this choice aligns with existing literature, it may not fully account for the variability in imputation results, particularly in datasets with more complex missing data patterns or higher proportions of missingness. Allowing for variations in the number of imputations could lead to different, potentially clearer, results. Additionally, this study focused on comparing CART and RF for multiple imputation using their default settings. While these default parameters generally provide good performance, they may not be optimal for all datasets. Incorporating computationally intensive hyperparameter tuning could enhance the imputation results and offer a deeper understanding of the methods' capabilities. Future research could investigate the impact of such optimization on imputation performance to identify the conditions under which these methods achieve the best results. Other possibilities to build on this study include considering a broader variety of scenarios, such as different missing data patterns or target variable types which could provide a more comprehensive evaluation. Moreover, the approach used here for CART and RF could be readily extended to other imputation methods, offering an opportunity to further generalize and compare the performance of diverse techniques.

Despite its limitations and the scope for further research, this study provides valuable insights. Through a comparative analysis of two tree-based imputation methods, CART and RF, it addresses missing data within the framework of multiple imputation under varying conditions. An extensive simulation study was conducted to evaluate key metrics, such as bias, coverage rate, and CI width, across a wide range of scenarios. The results indicate that RF generally outperforms CART in terms of producing narrower CIs and achieving higher coverage rates, particularly under scenarios characterized by the MNAR missing mechanism. However, CART performs competitively in certain scenarios, such as MCAR with lower percentages of missing data. For scenarios with high missingness, inference from the data becomes challenging due to the generation of excessively wide CIs. Overall, it is not feasible to recommend one method universally, as the performance of both methods is dependent on specific data conditions. The metrics show variation even

with minor changes in the data. Consequently, the choice between CART and Random Forest should be guided by the characteristics of the dataset and the specific requirements of the analysis. A general statement favoring one method over the other is not supported by the findings of this study.

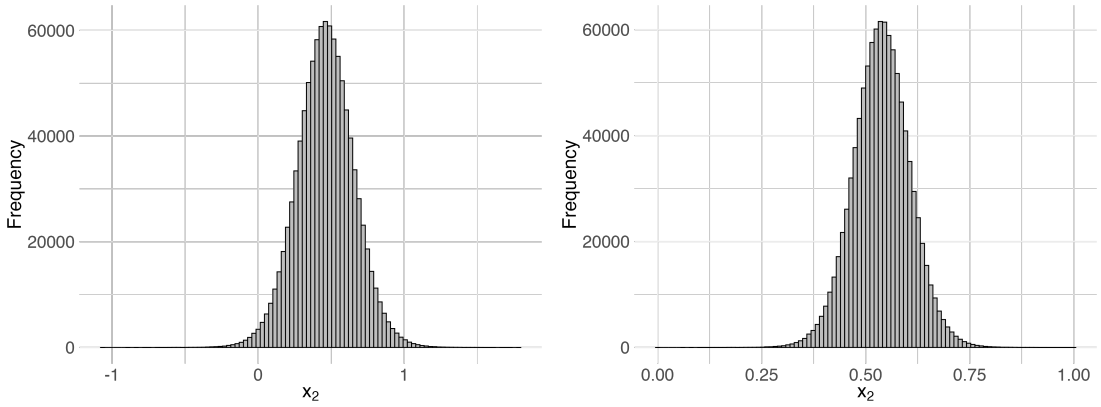
A Appendix

A.1 Distributions of x-variables



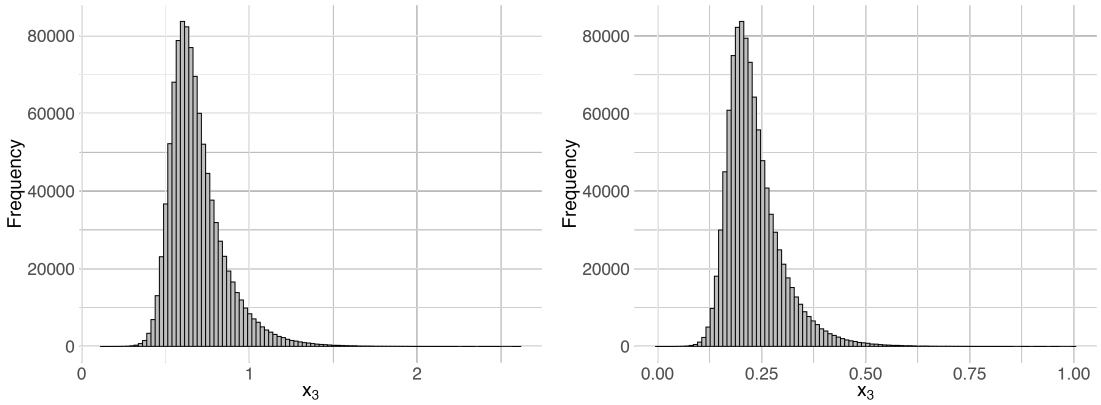
(a) Original distribution of x_1 (b) Standardized distribution of x_1

Figure 11: Histograms for x_1 for a dataset with $n = 1,000,000$: Original and Standardized



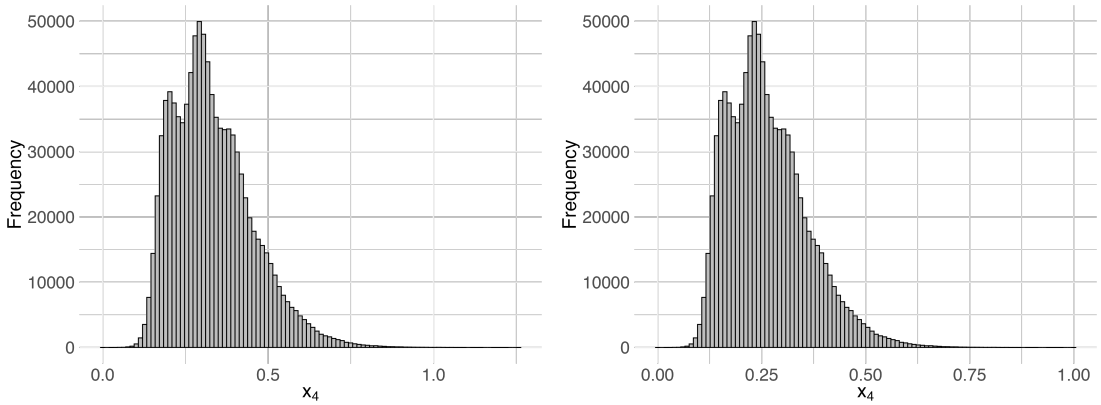
(a) Original distribution of x_2 (b) Standardized distribution of x_2

Figure 12: Histograms for x_2 for a dataset with $n = 1,000,000$: Original and Standardized



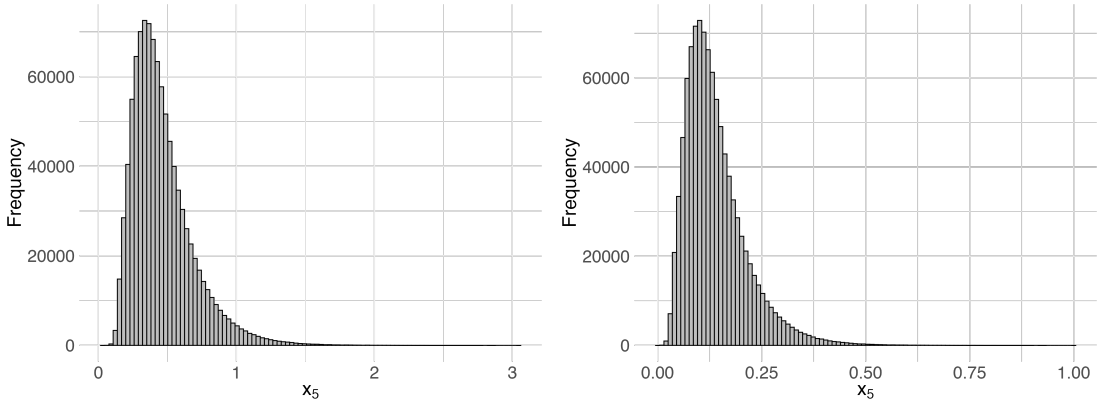
(a) Original distribution of x_3 (b) Standardized distribution of x_3

Figure 13: Histograms for x_3 for a dataset with $n = 1,000,000$: Original and Standardized



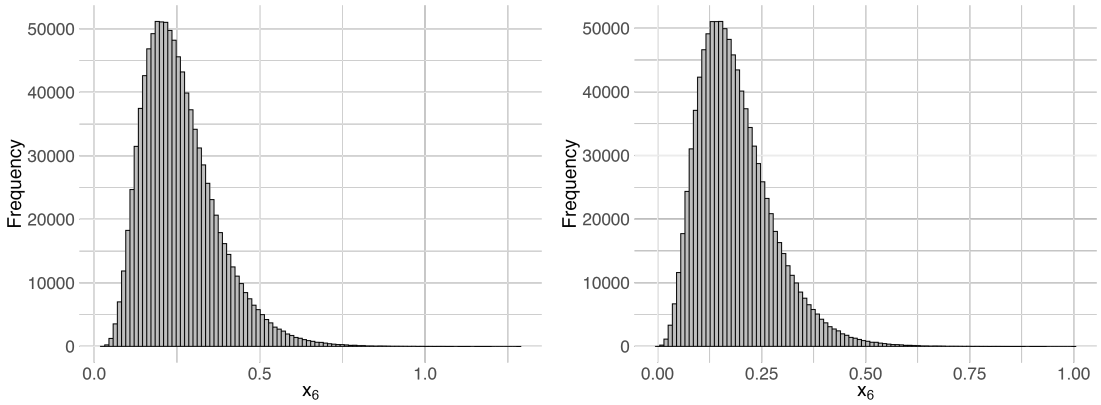
(a) Original distribution of x_4 (b) Standardized distribution of x_4

Figure 14: Histograms for x_4 for a dataset with $n = 1,000,000$: Original and Standardized



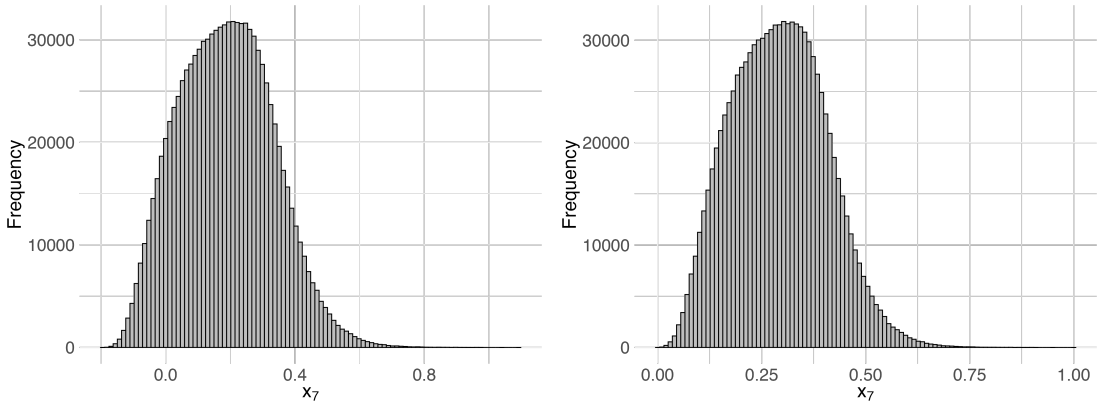
(a) Original distribution of x_5 (b) Standardized distribution of x_5

Figure 15: Histograms for x_5 for a dataset with $n = 1,000,000$: Original and Standardized



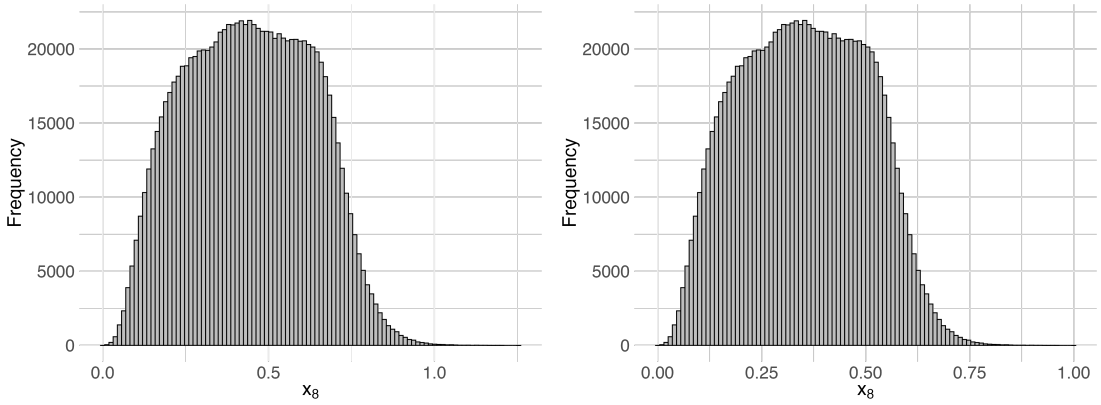
(a) Original distribution of x_6 (b) Standardized distribution of x_6

Figure 16: Histograms for x_6 for a dataset with $n = 1,000,000$: Original and Standardized



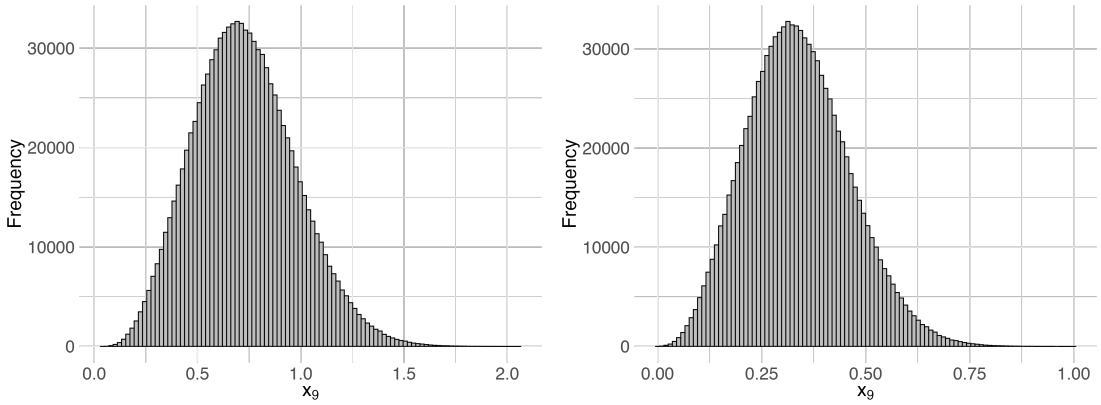
(a) Original distribution of x_7 (b) Standardized distribution of x_7

Figure 17: Histograms for x_7 for a dataset with $n = 1,000,000$: Original and Standardized



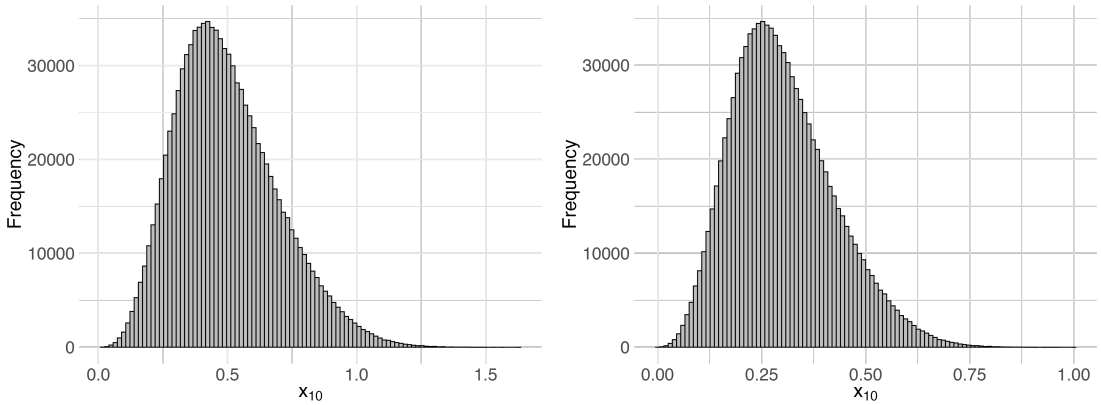
(a) Original distribution of x_8 (b) Standardized distribution of x_8

Figure 18: Histograms for x_8 for a dataset with $n = 1,000,000$: Original and Standardized



(a) Original distribution of x_9 (b) Standardized distribution of x_9

Figure 19: Histograms for x_9 for a dataset with $n = 1,000,000$: Original and Standardized



(a) Original distribution of x_{10} (b) Standardized distribution of x_{10}

Figure 20: Histograms for x_{10} for a dataset with $n = 1,000,000$: Original and Standardized

A.2 Tables Bias

Table 11: Bias for $n = 500$ with $p = 5$ for MCAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.03	-0.08	0.03	0.02	0.02	-0.07
RF	0.03	-0.09	0.05	0.03	0.05	-0.10

Table 12: Bias for $n = 500$ with $p = 5$ for MCAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.08	-0.14	0.04	0.04	0.01	-0.12
RF	0.05	-0.16	0.06	0.08	0.07	-0.17

Table 13: Bias for $n = 500$ with $p = 5$ for MCAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.14	-0.23	0.05	0.10	-0.03	-0.19
RF	0.09	-0.24	0.07	0.18	0.05	-0.24

Table 14: Bias for $n = 1000$ with $p = 5$ for MCAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.03	-0.05	0.01	-0.00	0.01	-0.05
RF	0.03	-0.07	0.04	0.01	0.03	-0.07

Table 15: Bias for $n = 1000$ with $p = 5$ for MCAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.07	-0.10	0.01	0.01	0.00	-0.10
RF	0.06	-0.13	0.04	0.05	0.04	-0.14

Table 16: Bias for $n = 1000$ with $p = 5$ for MCAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.14	-0.18	0.03	0.03	-0.01	-0.17
RF	0.10	-0.21	0.06	0.12	0.05	-0.24

Table 17: Bias for $n = 500$ with $p = 5$ for MAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.01	-0.06	0.03	0.09	0.04	-0.07
RF	-0.01	-0.08	0.03	0.14	0.11	-0.09

Table 18: Bias for $n = 500$ with $p = 5$ for MAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.08	-0.13	0.04	0.11	0.03	-0.13
RF	0.04	-0.16	0.04	0.22	0.13	-0.16

Table 19: Bias for $n = 500$ with $p = 5$ for MAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.18	-0.21	0.05	0.11	0.02	-0.24
RF	0.10	-0.23	0.06	0.29	0.12	-0.24

Table 20: Bias for $n = 1000$ with $p = 5$ for MAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.03	-0.06	0.02	0.06	0.02	-0.05
RF	0.01	-0.08	0.03	0.09	0.08	-0.07

Table 21: Bias for $n = 1000$ with $p = 5$ for MAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.10	-0.12	0.04	0.04	-0.00	-0.11
RF	0.05	-0.16	0.06	0.14	0.08	-0.13

Table 22: Bias for $n = 1000$ with $p = 5$ for MAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.20	-0.20	0.05	0.03	-0.03	-0.19
RF	0.11	-0.24	0.08	0.20	0.07	-0.21

Table 23: Bias for $n = 500$ with $p = 5$ for MNAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.12	-0.16	0.04	0.17	-0.01	0.04
RF	0.09	-0.15	0.04	0.21	0.07	0.02

Table 24: Bias for $n = 500$ with $p = 5$ for MNAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.42	-0.28	-0.16	0.13	-0.00	-0.11
RF	0.36	-0.27	-0.13	0.19	0.09	-0.11

Table 25: Bias for $n = 500$ with $p = 5$ for MNAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.65	-0.39	-0.29	0.10	-0.08	-0.25
RF	0.59	-0.39	-0.23	0.17	-0.02	-0.23

Table 26: Bias for $n = 1000$ with $p = 5$ for MNAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.12	-0.11	0.03	0.07	0.03	0.03
RF	0.09	-0.11	0.03	0.14	0.09	0.02

Table 27: Bias for $n = 1000$ with $p = 5$ for MNAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.41	-0.28	-0.09	0.05	-0.06	-0.04
RF	0.35	-0.29	-0.07	0.14	0.03	-0.07

Table 28: Bias for $n = 1000$ with $p = 5$ for MNAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.64	-0.39	-0.25	-0.04	-0.00	-0.28
RF	0.59	-0.41	-0.21	0.06	0.04	-0.25

Table 29: Bias for $n = 500$ with $p = 10$ for MCAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.11	-0.14	0.00	-0.03	0.00	-0.00	0.04	0.03	0.02	-0.02	-0.09
RF	0.09	-0.15	0.02	-0.02	0.01	-0.00	0.07	0.03	0.01	-0.01	-0.09

Table 30: Bias for $n = 500$ with $p = 10$ for MCAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.26	-0.25	-0.04	-0.07	-0.05	0.03	0.01	0.11	0.07	-0.05	-0.24
RF	0.21	-0.26	-0.02	-0.01	-0.03	0.06	0.09	0.12	0.02	-0.03	-0.23

Table 31: Bias for $n = 500$ with $p = 10$ for MCAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.53	-0.43	-0.20	-0.13	-0.08	0.17	-0.09	0.27	0.11	-0.13	-0.44
RF	0.41	-0.42	-0.17	-0.00	-0.07	0.22	0.09	0.23	0.01	-0.10	-0.37

Table 32: Bias for $n = 1000$ with $p = 10$ for MCAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.12	-0.14	0.03	-0.04	-0.03	-0.02	0.04	0.05	-0.01	-0.00	-0.11
RF	0.11	-0.15	0.04	-0.03	-0.03	-0.01	0.06	0.06	-0.02	0.01	-0.12

Table 33: Bias for $n = 1000$ with $p = 10$ for MCAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.29	-0.22	-0.05	-0.09	-0.06	0.00	0.02	0.09	0.02	-0.04	-0.23
RF	0.26	-0.24	-0.04	-0.05	-0.06	0.02	0.09	0.12	-0.02	-0.04	-0.22

Table 34: Bias for $n = 1000$ with $p = 10$ for MCAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.53	-0.36	-0.18	-0.16	-0.11	-0.02	0.03	0.16	0.08	-0.12	-0.36
RF	0.46	-0.37	-0.17	-0.06	-0.10	0.06	0.18	0.18	-0.02	-0.12	-0.32

Table 35: Bias for $n = 500$ with $p = 10$ for MAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.04	-0.13	-0.03	-0.04	0.03	0.09	0.07	0.14	0.16	0.05	-0.13
RF	-0.05	-0.14	0.00	0.06	0.05	0.14	0.20	0.19	0.06	0.05	-0.10

Table 36: Bias for $n = 500$ with $p = 10$ for MAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.29	-0.27	-0.12	-0.07	0.05	0.09	0.02	0.25	0.21	-0.10	-0.26
RF	0.19	-0.29	-0.09	0.07	0.06	0.18	0.20	0.26	0.08	-0.07	-0.20

Table 37: Bias for $n = 500$ with $p = 10$ for MAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.58	-0.42	-0.27	-0.08	0.06	0.13	-0.02	0.40	0.22	-0.31	-0.40
RF	0.47	-0.40	-0.22	0.07	0.05	0.25	0.21	0.34	0.08	-0.23	-0.31

Table 38: Bias for $n = 1000$ with $p = 10$ for MAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.04	-0.12	-0.03	-0.03	0.05	0.07	0.09	0.11	0.09	0.05	-0.08
RF	-0.02	-0.14	0.00	0.02	0.04	0.11	0.21	0.18	0.05	0.03	-0.09

Table 39: Bias for $n = 1000$ with $p = 10$ for MAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.29	-0.25	-0.10	-0.12	0.03	0.07	0.07	0.16	0.15	-0.04	-0.22
RF	0.20	-0.28	-0.08	-0.00	0.03	0.15	0.24	0.23	0.06	-0.03	-0.19

Table 40: Bias for $n = 1000$ with $p = 10$ for MAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.58	-0.41	-0.20	-0.20	-0.00	0.06	0.05	0.23	0.18	-0.17	-0.37
RF	0.46	-0.42	-0.17	-0.03	-0.01	0.22	0.29	0.23	0.04	-0.12	-0.31

Table 41: Bias for $n = 500$ with $p = 10$ for MNAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.41	-0.22	-0.06	-0.04	-0.09	0.17	0.07	0.16	0.11	-0.05	-0.13
RF	0.27	-0.19	-0.03	0.08	-0.06	0.22	0.19	0.17	0.03	0.02	-0.04

Table 42: Bias for $n = 500$ with $p = 10$ for MNAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	1.15	-0.43	-0.40	-0.25	-0.14	0.16	0.03	0.22	-0.00	-0.17	-0.35
RF	0.96	-0.36	-0.31	-0.09	-0.09	0.19	0.17	0.20	-0.08	-0.06	-0.19

Table 43: Bias for $n = 500$ with $p = 10$ for MNAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	1.67	-0.60	-0.59	-0.27	-0.32	0.07	0.03	0.39	-0.16	-0.48	-0.54
RF	1.54	-0.53	-0.47	-0.16	-0.27	0.09	0.09	0.30	-0.19	-0.32	-0.38

Table 44: Bias for $n = 1000$ with $p = 10$ for MNAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.43	-0.20	-0.05	-0.10	-0.03	0.06	0.08	0.07	0.17	-0.11	-0.12
RF	0.31	-0.19	-0.04	0.02	-0.02	0.12	0.20	0.12	0.10	-0.05	-0.05

Table 45: Bias for $n = 1000$ with $p = 10$ for MNAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	1.14	-0.43	-0.29	-0.27	-0.18	0.18	-0.05	0.05	0.12	-0.18	-0.42
RF	0.99	-0.40	-0.25	-0.12	-0.15	0.22	0.10	0.08	0.04	-0.11	-0.27

Table 46: Bias for $n = 1000$ with $p = 10$ for MNAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	1.63	-0.61	-0.52	-0.41	-0.23	-0.02	0.09	0.04	-0.00	-0.41	-0.52
RF	1.52	-0.57	-0.46	-0.27	-0.22	0.03	0.18	0.06	-0.08	-0.30	-0.39

A.3 Tables Coverage Rate

Table 47: Coverage Rate (%) for $n = 500$ with $p = 5$ for MCAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	91.1	92.9	95.6	96.2	96.4	94.3
RF	93.5	93.6	97.2	97.5	97.4	93.3

Table 48: Coverage Rate (%) for $n = 500$ with $p = 5$ for MCAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	85.0	92.2	96.8	97.1	97.7	92.3
RF	91.1	90.9	97.3	97.9	97.9	90.5

Table 49: Coverage Rate (%) for $n = 500$ with $p = 5$ for MCAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	71.6	88.5	95.9	95.0	96.0	90.4
RF	83.7	85.3	98.3	94.8	98.8	86.3

Table 50: Coverage Rate (%) for $n = 1000$ with $p = 5$ for MCAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	91.5	93.3	95.1	94.8	96.1	93.6
RF	91.9	92.7	96.6	97.3	96.4	92.2

Table 51: Coverage Rate (%) for $n = 1000$ with $p = 5$ for MCAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	76.6	91.2	96.6	96.4	96.0	89.3
RF	84.6	88.3	97.7	97.6	96.8	83.3

Table 52: Coverage Rate (%) for $n = 1000$ with $p = 5$ for MCAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	59.8	88.0	98.0	95.9	96.5	87.3
RF	74.7	78.5	97.0	95.3	97.4	73.2

Table 53: Coverage Rate (%) for $n = 500$ with $p = 5$ for MAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	94.4	97.3	96.6	96.3	98.2	96.2
RF	96.5	97.6	98.3	98.4	98.2	96.8

Table 54: Coverage Rate (%) for $n = 500$ with $p = 5$ for MAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	90.7	95.9	97.6	97.4	97.9	94.3
RF	96.0	95.8	99.5	97.0	98.4	94.9

Table 55: Coverage Rate (%) for $n = 500$ with $p = 5$ for MAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	82.8	94.6	97.3	96.0	97.9	91.7
RF	93.5	91.8	98.3	93.7	97.8	90.7

Table 56: Coverage Rate (%) for $n = 1000$ with $p = 5$ for MAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	94.4	95.9	97.0	96.3	98.0	95.7
RF	96.3	96.1	97.6	97.2	97.7	94.3

Table 57: Coverage Rate (%) for $n = 1000$ with $p = 5$ for MAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	84.9	93.7	96.5	96.8	98.0	93.3
RF	92.8	89.8	97.0	94.6	97.6	92.3

Table 58: Coverage Rate (%) for $n = 1000$ with $p = 5$ for MAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	61.8	90.9	96.8	96.7	98.7	93.4
RF	83.0	83.5	97.4	94.7	97.5	86.9

Table 59: Coverage Rate (%) for $n = 500$ with $p = 5$ for MNAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	74.2	92.0	96.3	94.0	98.2	96.2
RF	84.3	94.4	97.6	94.6	98.9	97.6

Table 60: Coverage Rate (%) for $n = 500$ with $p = 5$ for MNAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	6.7	87.3	96.3	96.5	97.5	94.5
RF	10.1	87.0	97.2	97.3	98.8	97.6

Table 61: Coverage Rate (%) for $n = 500$ with $p = 5$ for MNAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.2	78.8	93.8	96.7	98.0	92.3
RF	0.0	65.1	93.7	97.9	98.9	91.7

Table 62: Coverage Rate (%) for $n = 1000$ with $p = 5$ for MNAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	61.5	92.3	96.9	96.7	97.8	95.9
RF	76.6	93.1	97.6	96.0	97.6	98.2

Table 63: Coverage Rate (%) for $n = 1000$ with $p = 5$ for MNAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.9	75.9	96.2	96.3	98.0	95.5
RF	2.0	66.2	97.5	95.5	98.1	94.7

Table 64: Coverage Rate (%) for $n = 1000$ with $p = 5$ for MNAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.0	63.1	92.3	96.9	97.7	86.8
RF	0.0	37.8	89.5	98.5	97.8	82.8

Table 65: Coverage Rate (%) for $n = 500$ with $p = 10$ for MCAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	73.3	90.1	98.0	97.2	95.9	95.8	95.6	96.6	95.6	97.2	96.0
RF	86.0	92.2	98.2	99.0	98.0	97.3	97.4	98.7	98.7	99.0	98.1

Table 66: Coverage Rate (%) for $n = 500$ with $p = 10$ for MCAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	61.3	91.6	98.9	98.1	97.7	98.1	97.9	97.6	98.4	98.8	93.5
RF	75.2	92.6	99.4	99.0	98.5	98.2	98.2	98.8	99.3	99.0	94.1

Table 67: Coverage Rate (%) for $n = 500$ with $p = 10$ for MCAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	50.3	94.6	98.8	99.6	99.2	98.8	98.8	97.1	98.5	99.0	93.3
RF	64.4	91.4	99.0	99.9	99.1	98.8	99.3	98.3	99.5	99.7	95.7

Table 68: Coverage Rate (%) for $n = 1000$ with $p = 10$ for MCAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	49.9	80.1	96.7	96.0	95.7	95.0	95.3	96.0	96.5	97.1	86.1
RF	65.6	81.9	97.8	98.6	98.1	97.0	96.1	97.6	97.6	98.3	90.6

Table 69: Coverage Rate (%) for $n = 1000$ with $p = 10$ for MCAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	27.1	83.8	98.3	98.3	97.3	97.1	97.1	96.7	97.5	97.5	84.6
RF	37.2	81.3	98.4	99.1	98.5	98.0	97.8	96.6	98.8	98.7	85.1

Table 70: Coverage Rate (%) for $n = 1000$ with $p = 10$ for MCAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	14.3	84.6	97.8	98.7	98.5	98.0	97.7	96.7	97.0	97.4	84.1
RF	21.8	78.8	97.6	99.2	98.2	98.0	97.1	96.0	99.1	98.2	82.2

Table 71: Coverage Rate (%) for $n = 500$ with $p = 10$ for MAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	99.8	100.0	100.0	99.9	100.0	99.8	99.9	99.8	99.9	100.0	100.0
RF	99.9	100.0	100.0	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0

Table 72: Coverage Rate (%) for $n = 500$ with $p = 10$ for MAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	96.5	100.0	99.8	99.9	99.9	100.0	100.0	99.9	99.8	100.0	99.9
RF	97.8	99.9	100.0	99.9	100.0	100.0	100.0	99.9	100.0	100.0	99.8

Table 73: Coverage Rate (%) for $n = 500$ with $p = 10$ for MAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	90.1	99.8	99.9	100.0	100.0	99.3	99.8	99.7	99.6	99.8	99.5
RF	94.2	99.3	100.0	100.0	100.0	100.0	100.0	99.9	100.0	99.9	99.8

Table 74: Coverage Rate (%) for $n = 1000$ with $p = 10$ for MAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	99.0	99.6	99.9	99.9	100.0	99.5	99.9	99.9	99.7	99.8	99.9
RF	99.4	99.9	100.0	100.0	100.0	99.8	99.9	100.0	100.0	99.9	99.8

Table 75: Coverage Rate (%) for $n = 1000$ with $p = 10$ for MAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	87.7	98.7	100.0	99.9	99.8	99.7	99.9	99.6	99.5	99.8	99.3
RF	93.3	98.5	100.0	100.0	100.0	99.8	99.9	99.6	99.8	99.7	99.2

Table 76: Coverage Rate (%) for $n = 1000$ with $p = 10$ for MAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	62.1	97.3	99.9	99.6	99.9	99.7	99.7	99.2	99.2	99.7	98.4
RF	70.5	97.9	99.6	100.0	99.8	99.8	99.4	99.5	100.0	99.9	96.7

Table 77: Coverage Rate (%) for $n = 500$ with $p = 10$ for MNAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	70.9	98.0	99.2	99.5	99.4	99.2	99.4	99.1	99.2	99.6	99.1
RF	87.1	98.7	99.2	99.3	99.8	99.6	99.7	99.4	99.6	99.4	99.7

Table 78: Coverage Rate (%) for $n = 500$ with $p = 10$ for MNAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	26.1	97.9	99.2	99.7	99.7	99.5	99.6	99.1	99.5	99.8	98.4
RF	34.9	98.5	99.6	99.7	99.5	99.5	99.9	99.3	99.9	99.8	99.4

Table 79: Coverage Rate (%) for $n = 500$ with $p = 10$ for MNAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	13.3	96.2	97.9	99.4	99.3	99.8	99.3	98.7	99.8	99.2	98.2
RF	14.7	95.7	97.2	99.5	99.4	99.8	99.9	100.0	99.8	99.5	98.5

Table 80: Coverage Rate (%) for $n = 1000$ with $p = 10$ for MNAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	35.3	95.9	98.7	98.7	99.1	98.4	99.5	98.7	96.3	99.2	98.2
RF	56.4	96.2	98.8	99.2	99.3	98.7	99.1	98.5	98.8	99.5	99.0

Table 81: Coverage Rate (%) for $n = 1000$ with $p = 10$ for MNAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	2.0	90.0	97.9	98.8	99.3	97.4	99.0	99.5	98.5	98.4	93.4
RF	3.7	90.3	97.8	99.4	99.2	99.0	99.2	99.6	99.4	99.0	97.6

Table 82: Coverage Rate (%) for $n = 1000$ with $p = 10$ for MNAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	1.0	80.0	92.9	97.6	98.1	99.0	99.0	98.8	99.0	96.1	91.6
RF	0.2	76.3	90.9	97.9	97.8	99.7	99.3	99.2	99.7	96.7	92.8

A.4 Tables Confidence Intervals

Table 83: Median CI Widths for $n = 500$ with $p = 5$ for MCAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.28	0.60	0.75	0.88	0.81	0.55
RF	0.28	0.61	0.74	0.90	0.82	0.57

Table 84: Median CI Widths for $n = 500$ with $p = 5$ for MCAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.32	0.75	0.94	1.14	1.04	0.73
RF	0.32	0.71	0.87	1.05	0.95	0.71

Table 85: Median CI Widths for $n = 500$ with $p = 5$ for MCAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.40	1.02	1.28	1.48	1.36	1.00
RF	0.38	0.84	1.02	1.16	1.08	0.88

Table 86: Median CI Widths for $n = 1000$ with $p = 5$ for MCAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.21	0.45	0.56	0.67	0.60	0.42
RF	0.21	0.45	0.56	0.67	0.60	0.43

Table 87: Median CI Widths for $n = 1000$ with $p = 5$ for MCAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.24	0.55	0.72	0.86	0.77	0.54
RF	0.24	0.52	0.66	0.78	0.69	0.53

Table 88: Median CI Widths for $n = 1000$ with $p = 5$ for MCAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.30	0.75	0.98	1.14	1.01	0.78
RF	0.29	0.62	0.77	0.89	0.78	0.66

Table 89: Median CI Widths for $n = 500$ with $p = 5$ for MAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.38	0.90	1.23	1.50	1.28	0.87
RF	0.38	0.89	1.19	1.43	1.25	0.86

Table 90: Median CI Widths for $n = 500$ with $p = 5$ for MAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.48	1.18	1.61	1.92	1.63	1.17
RF	0.46	1.02	1.37	1.59	1.39	1.06

Table 91: Median CI Widths for $n = 500$ with $p = 5$ for MAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.63	1.53	2.11	2.39	2.02	1.51
RF	0.55	1.16	1.50	1.75	1.50	1.20

Table 92: Median CI Widths for $n = 1000$ with $p = 5$ for MAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.29	0.66	0.90	1.09	0.91	0.66
RF	0.28	0.62	0.83	1.01	0.86	0.63

Table 93: Median CI Widths for $n = 1000$ with $p = 5$ for MAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.37	0.85	1.20	1.44	1.16	0.89
RF	0.32	0.74	0.98	1.14	0.96	0.76

Table 94: Median CI Widths for $n = 1000$ with $p = 5$ for MAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.47	1.15	1.57	1.86	1.52	1.20
RF	0.40	0.85	1.11	1.31	1.10	0.90

Table 95: Median CI Widths for $n = 500$ with $p = 5$ for MNAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.39	0.87	1.10	1.37	1.24	0.93
RF	0.38	0.83	1.03	1.25	1.16	0.90

Table 96: Median CI Widths for $n = 500$ with $p = 5$ for MNAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.43	1.11	1.44	1.73	1.52	1.29
RF	0.40	0.96	1.20	1.41	1.26	1.14

Table 97: Median CI Widths for $n = 500$ with $p = 5$ for MNAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.43	1.30	1.73	2.01	1.82	1.53
RF	0.37	0.97	1.20	1.41	1.28	1.19

Table 98: Median CI Widths for $n = 1000$ with $p = 5$ for MNAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.29	0.63	0.81	1.05	0.90	0.72
RF	0.28	0.59	0.75	0.94	0.82	0.69

Table 99: Median CI Widths for $n = 1000$ with $p = 5$ for MNAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.32	0.83	1.08	1.37	1.15	0.97
RF	0.30	0.70	0.86	1.03	0.91	0.83

Table 100: Median CI Widths for $n = 1000$ with $p = 5$ for MNAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5
CART	0.32	0.98	1.28	1.60	1.37	1.26
RF	0.28	0.71	0.90	1.06	0.93	0.91

Table 101: Median CI Widths for $n = 500$ with $p = 10$ for MCAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.35	0.66	0.85	1.00	0.86	1.07	1.26	0.97	0.74	0.79	0.65
RF	0.40	0.75	0.95	1.17	1.04	1.21	1.38	1.13	0.84	0.90	0.74

Table 102: Median CI Widths for $n = 500$ with $p = 10$ for MCAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.63	1.25	1.58	1.93	1.68	2.14	2.37	1.92	1.43	1.53	1.21
RF	0.68	1.23	1.55	1.84	1.66	2.10	2.20	1.94	1.50	1.46	1.26

Table 103: Median CI Widths for $n = 500$ with $p = 10$ for MCAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	1.05	2.27	2.91	3.59	3.11	4.16	4.23	3.46	2.83	2.75	2.35
RF	1.04	1.97	2.32	2.98	2.74	3.37	3.39	3.13	2.45	2.36	2.02

Table 104: Median CI Widths for $n = 1000$ with $p = 10$ for MCAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.25	0.45	0.58	0.71	0.62	0.74	0.83	0.66	0.51	0.55	0.44
RF	0.29	0.51	0.64	0.83	0.72	0.83	0.92	0.76	0.55	0.60	0.50

Table 105: Median CI Widths for $n = 1000$ with $p = 10$ for MCAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.42	0.78	1.02	1.25	1.09	1.32	1.47	1.15	0.92	0.96	0.80
RF	0.45	0.77	0.98	1.19	1.07	1.27	1.41	1.18	0.91	0.92	0.77

Table 106: Median CI Widths for $n = 1000$ with $p = 10$ for MCAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.64	1.21	1.53	1.88	1.70	2.06	2.14	1.85	1.49	1.51	1.27
RF	0.62	1.06	1.32	1.55	1.43	1.65	1.67	1.56	1.27	1.21	1.03

Table 107: Median CI Widths for $n = 500$ with $p = 10$ for MAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	1.91	5.57	7.48	9.70	8.42	9.27	10.74	8.79	6.49	6.52	5.04
RF	1.92	6.54	9.16	12.30	9.56	14.46	15.05	11.06	8.11	8.06	6.10

Table 108: Median CI Widths for $n = 500$ with $p = 10$ for MAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	2.29	7.14	9.66	12.46	11.39	14.67	13.14	12.27	11.71	8.92	7.46
RF	2.33	8.14	11.29	13.40	13.20	14.97	16.11	13.63	9.28	8.81	6.67

Table 109: Median CI Widths for $n = 500$ with $p = 10$ for MAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	2.87	9.47	12.00	16.98	17.02	21.78	18.17	19.04	14.20	11.68	10.09
RF	2.77	8.73	11.47	14.29	13.87	16.16	15.31	12.92	10.55	9.15	8.03

Table 110: Median CI Widths for $n = 1000$ with $p = 10$ for MAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.96	2.00	2.62	3.26	2.81	3.25	3.33	2.88	2.27	2.37	1.95
RF	0.92	2.07	2.77	3.24	2.85	3.32	3.62	3.13	2.31	2.38	1.97

Table 111: Median CI Widths for $n = 1000$ with $p = 10$ for MAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	1.11	2.35	3.06	3.63	3.41	4.11	4.19	3.57	2.86	2.84	2.41
RF	1.05	2.17	2.81	3.29	2.99	3.57	3.62	3.23	2.49	2.51	2.17

Table 112: Median CI Widths for $n = 1000$ with $p = 10$ for MAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	1.31	2.81	3.71	4.54	4.15	5.09	4.97	4.74	3.81	3.39	2.83
RF	1.19	2.31	2.86	3.39	3.16	3.79	3.69	3.45	2.60	2.41	2.19

Table 113: Median CI Widths for $n = 500$ with $p = 10$ for MNAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	1.16	2.14	2.69	3.58	3.20	4.06	4.32	3.29	2.65	2.80	2.33
RF	1.17	2.16	2.76	3.62	3.22	4.03	4.10	3.52	2.85	2.93	2.42

Table 114: Median CI Widths for $n = 500$ with $p = 10$ for MNAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	1.50	3.53	4.60	5.68	5.20	8.35	7.86	6.51	5.00	4.79	4.00
RF	1.51	3.27	4.47	5.41	4.56	6.31	6.32	5.90	4.84	4.30	3.80

Table 115: Median CI Widths for $n = 500$ with $p = 10$ for MNAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	1.41	4.71	5.93	8.29	7.04	11.02	11.24	11.10	8.45	6.35	5.80
RF	1.43	4.04	5.04	6.36	6.12	7.55	7.75	6.94	5.96	5.41	4.72

Table 116: Median CI Widths for $n = 1000$ with $p = 10$ for MNAR and 20% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.70	1.19	1.51	1.99	1.75	2.13	2.31	1.80	1.45	1.56	1.34
RF	0.70	1.19	1.54	1.84	1.64	1.90	2.05	1.80	1.40	1.51	1.30

Table 117: Median CI Widths for $n = 1000$ with $p = 10$ for MNAR and 40% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.82	1.66	2.09	2.71	2.24	3.08	3.21	2.61	2.13	2.15	1.94
RF	0.80	1.56	1.89	2.26	1.94	2.25	2.42	2.14	1.78	1.86	1.78

Table 118: Median CI Widths for $n = 1000$ with $p = 10$ for MNAR and 60% missing

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
CART	0.79	1.90	2.35	3.00	2.68	3.88	4.00	3.41	2.68	2.57	2.26
RF	0.69	1.58	1.96	2.26	1.97	2.46	2.47	2.28	1.91	1.90	1.79

B Electronic appendix

The `R` code to reproduce this study is available on GitHub: `https://github.com/mspeckbacher/BT-MultipleImputation`.

References

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

Altman, D. G. and Bland, J. M. (2007). Missing data, *Bmj* **334**(7590): 424–424.

Breiman, L. (2001). Random forests, *Machine learning* **45**: 5–32.

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

Burgette, L. F. and Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees, *American journal of epidemiology* **172**(9): 1070–1076.

Carpenito, T. and Manjourides, J. (2022). Misl: Multiple imputation by super learning, *Statistical methods in medical research* **31**(10): 1904–1915.

Chhabra, G., Vashisht, V. and Ranjan, J. (2017). A comparison of multiple imputation methods for data with missing values, *Indian Journal of Science and Technology* **10**(19): 1–7.

de Goeij, M. C., Van Diepen, M., Jager, K. J., Tripepi, G., Zoccali, C. and Dekker, F. W. (2013). Multiple imputation: dealing with missing data, *Nephrology Dialysis Transplantation* **28**(10): 2415–2420.

Deng, Y. and Lumley, T. (2024). Multiple imputation through xgboost, *Journal of Computational and Graphical Statistics* **33**(2): 352–363.

Enders, C. K. (2022). *Applied missing data analysis*, Guilford Publications.

Faisal, S. and Tutz, G. (2021). Multiple imputation using nearest neighbor methods, *Information Sciences* **570**: 500–516.

Harel, O. and Zhou, X.-H. (2007). Multiple imputation: review of theory, implementation and software, *Statistics in medicine* **26**(16): 3057–3077.

Hastie, T. (2009). The elements of statistical learning: data mining, inference, and prediction.

Hayati Rezvan, P., Lee, K. J. and Simpson, J. A. (2015). The rise of multiple imputation: a review of the reporting and implementation of the method in medical research, *BMC medical research methodology* **15**: 1–14.

Huque, M. H., Carlin, J. B., Simpson, J. A. and Lee, K. J. (2018). A comparison of multiple imputation methods for missing data in longitudinal studies, *BMC medical research methodology* **18**: 1–16.

Jenkins, S. P., Burkhauser, R. V., Feng, S. and Larrimore, J. (2011). Measuring inequality using censored data: a multiple-imputation approach to estimation and inference, *Journal of the Royal Statistical Society Series A: Statistics in Society* **174**(1): 63–81.

Lall, R. (2016). How multiple imputation makes a difference, *Political Analysis* **24**(4): 414–433.

Lee, K. J. and Carlin, J. B. (2012). Recovery of information from multiple imputation: a simulation study, *Emerging themes in epidemiology* **9**: 1–10.

Loh, W.-Y. (2011). Classification and regression trees, *Wiley interdisciplinary reviews: data mining and knowledge discovery* **1**(1): 14–23.

Lüdtke, O., Robitzsch, A. and Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies., *Psychological methods* **22**(1): 141.

Mayer, I., Sportisse, A., Josse, J., Tierney, N. and Vialaneix, N. (2024). R-miss-tastic: a unified platform for missing values methods and workflows.
URL: *https://arxiv.org/abs/1908.04822*

Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse, *Proceedings of the survey research methods sec-*

tion of the American Statistical Association, Vol. 1, American Statistical Association Alexandria, VA, pp. 20–34.

Rubin, D. B. (1996). Multiple imputation after 18+ years, *Journal of the American statistical Association* **91**(434): 473–489.

Schouten, R. M., Lugtig, P. and Vink, G. (2018). Generating missing values for simulation purposes: a multivariate amputation procedure, *Journal of Statistical Computation and Simulation* **88**(15): 2909–2930.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O. and 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**(6): 764–774.

Slade, E. and Naylor, M. G. (2020). A fair comparison of tree-based and parametric methods in multiple imputation by chained equations, *Statistics in medicine* **39**(8): 1156–1166.

Therneau, T., Atkinson, B., Ripley, B. and Ripley, M. B. (2015). Package 'rpart', *Available online: cran. ma. ic. ac. uk/web/packages/rpart/rpart. pdf (accessed on 20 April 2016)* .

van Buuren, S. (2018). *Flexible imputation of missing data*, CRC press.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r, *Journal of Statistical Software* **45**(3): 1–67.

Wright, M. N., Wager, S., Probst, P. and Wright, M. M. N. (2019). Package 'ranger', *Version 0.11* **2**.

Zhang, Z. (2016). Multiple imputation with multivariate imputation by chained equation (mice) package, *Annals of translational medicine* **4**(2).

Declaration of authorship

I hereby declare that the report submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the Thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the report as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future Theses submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.

Munich, December 23rd 2024

Michael Speckbacher