UNIVERSITÄT TRIER

Comparing Methods for Tree-Based Multiple Imputation

ARTEM SMIRNOV, ARTEM TUKHBATULLIN
STATISTICAL ANALYSIS OF INCOMPLETE DATA, TRIER UNIVERSITY

Introduction

Burgette and Reiter (2010) in the "Multiple Imputation for Missing Data via Sequential Regression Trees" paper propose a solution to handle certain challenges in model-based multiple imputation via chained equations if missing values are associated with the target variable in a way that introduces leakage.[1].

Implementation of the packages such as **mice** and **miceRanger** can help to address and resolve these issues.

Objectives

Missing data reflects in a significant challenge in data analysis. Multiple imputation approach is capable to handle missing data [2].

In this study, we conduct a comparative analysis of tree-based imputation methods between the **miceRanger** and traditional tree-based methods in **mice**.

Methods in R

Package **mice** involves iterative imputation of missing values based on observed data in other variables, whereas **miceRanger** extends this approach by incorporating tree-based methods for enhanced accuracy[3].

While both methods aim to deal with missing data in statistical analyses, **miceRanger** can make use of a procedure called predictive mean matching (PMM) to select which values are imputed.

PMM involves selecting a datapoint from the original, nonmissing data which has a predicted value close to the predicted value of the missing sample. Therefore, it results in improved performance compared to the standard **mice** approach[4].

Hypotheses

H1: miceRanger can produce more accurate results than **mice** when dealing with datasets that contain linear relations between independent variables and a response variable.

H2: CART imputation engine handles with less precision than **RF**. As a simpler model, **CART** seems to be more prone to underfitting.

Simulation Study

The linear equation is given by:
$n = 1000, x_i \sim N(0, 1^2), \epsilon \sim N(0, 1^2)$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{20} x_{20} + \epsilon \qquad (1)$$

Completely observed variables: $Y, x_{11}, x_{16}, x_{18}$

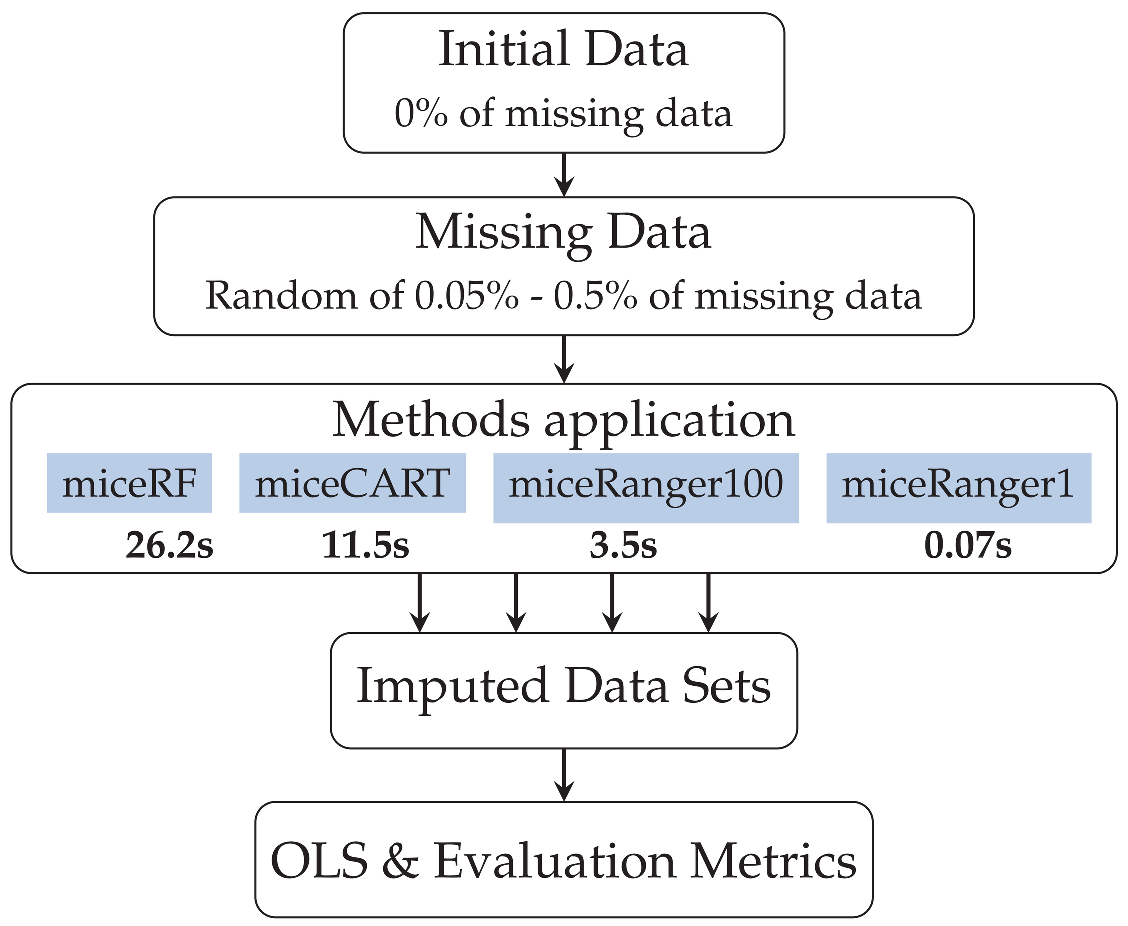

Figure 1: Data Analysis Framework.

Imputation Pseudocode

Algorithm 1 RF (via CART) Imputation in MICE[5]

1: *input*: dataset $X = \{X^{obs}, X^{mis}\}$, p partially observed variables, column j, currently imputed $\dot{X}$, I iterations.
2: *output*: dataset $X_{imputed}$
3: **for** $j = 1$ to p **do** imputations $\dot{X}_j^0$ by random draws from X_j^{obs}, update $\dot{X}$.
4: **end for**
5: **for** $j = 1$ to p **do** ▷ replacing $\dot{X}_j^0$
6: Draw k bootstrap samples from $\dot{X}$, restricted to items in X_j^{obs}.
7: Build k CART by fitting each on a bootstrap sample from 6 to find the best split at each node. Each tree has leaves containing a subset of X_j^{obs}.
8: For X_j^{mis} items, find leaf in each of the k trees from 7. Hence k leaves with donors $\forall x \in X_j^{mis}$.
9: For X_j^{mis}, randomly select one X^{obs} from donors in the k leaves of 8. Replace missing $\dot{X}_j^{obs}$ values, and add the complete $\dot{X}_j$ to $\dot{X}$.
10: **end for**
11: Repeat for-loop 4 I times.
12: Repeat steps 2 and 10 m times to get m sets.
13: **return** pooled dataset $X_{imputed}$

Graphical Results

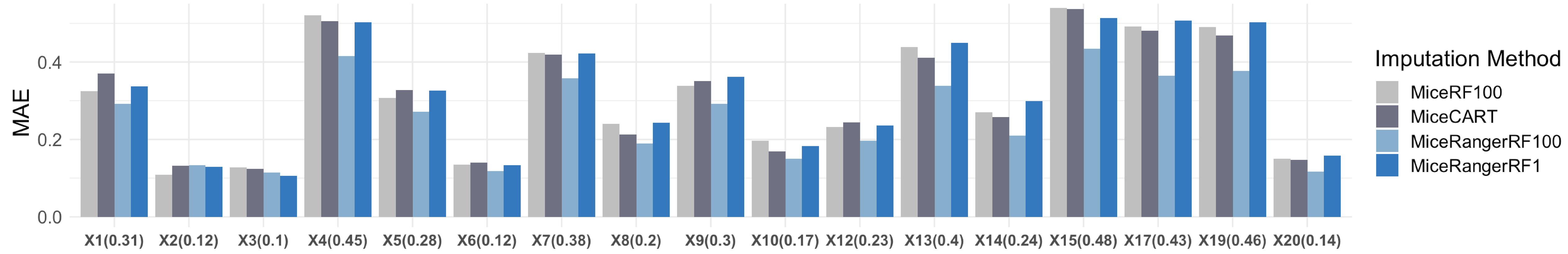

Figure 2: Mean absolute error for the imputed variables. miceRangerForest yielded the smallest error among the tested methods. This becomes even more evident when a significant percentage of data is missing, which is shown in the brackets.

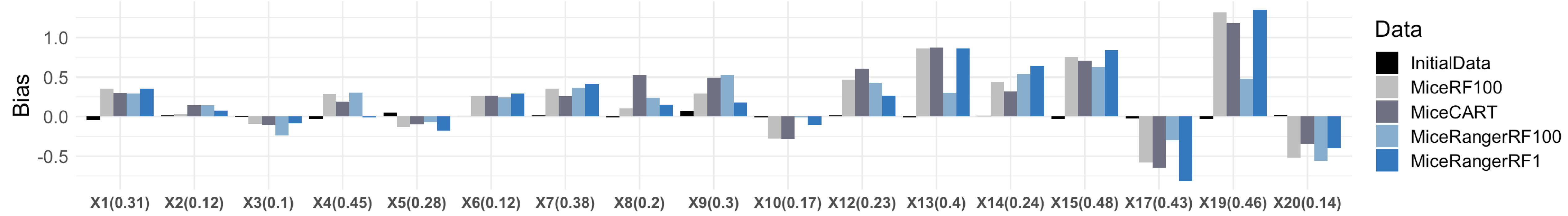

Figure 3: Bias for β.

Tabular Summary

	β	M_RF	M_CART	MR_100	MR_1
β_1	0.5	0.739	0.831	0.662	0.773
β_2	0.5	0.399	0.471	0.459	0.456
β_3	0.5	0.487	0.492	0.440	0.420
β_4	0.5	0.956	0.943	0.771	0.924
β_5	-0.5	0.742	0.772	0.641	0.762
β_6	1.0	0.470	0.508	0.430	0.470
β_7	1.0	0.862	0.863	0.722	0.867
β_8	1.0	0.663	0.606	0.522	0.667
β_9	1.0	0.778	0.812	0.672	0.821
β_{10}	-1.0	0.581	0.529	0.463	0.568
β_{12}	2.0	0.612	0.628	0.520	0.617
β_{13}	2.0	0.864	0.811	0.672	0.889
β_{14}	2.0	0.681	0.661	0.538	0.741
β_{15}	2.0	0.969	0.953	0.766	0.924
β_{17}	-2.0	0.929	0.913	0.706	0.972
β_{19}	3.0	0.913	0.855	0.695	0.913
β_{20}	-3.0	0.516	0.485	0.389	0.531

Table 1: RMSE for β estimates of miceRF, miceCART, miceRanger100, and miceRanger1.

	M_RF	M_CART	MR_100	MR_1
AIC	5789.3	5585.6	5121.6	5812.9
BIC	5897.2	5693.6	5229.6	5920.9
ARMSE	4.279	3.865	3.064	4.330

Table 2: ICs and ARMSE for the OLS models. Residuals of all models are homoscedastic and normally distributed at $\alpha = 0.01$.

Conclusion and Future Scope

H1 is not rejected. **miceRanger** showed better results on the simulated data than **mice** in case of **RF**. It is worth noting that the miceRanger1, i.e. **RF** with a single tree, is less precise than the **CART** implemented in mice.
H2 is not rejected. Indeed, **RF** copes better with missing values than **CART**. However, **RF** cannot be considered as the best option for any case. This method requires more training time due to the computational complexity. Which is also important, **RF** is a black box ML model, so a researcher cannot interpret results directly.

The future scope of this topic involves interpreting results of **RF** through Explainable AI, such as LIME and SHAP. Additionally, it would be interesting to analyze other R packages, such as missForest.

References

[1] Jerome P. Reiter Lane F. Burgette. Multiple imputation for missing data via sequential regression trees. In *American Journal of Epidemiology*, page 172:1070–1076, 2010.

[2] Rubin DB. Multiple imputation for nonresponse in surveys. In *Hoboken, NJ: Wiley-IEEE*, 1987.

[3] Karin Groothuis-Oudshoorn Stef van Buuren. mice: Multivariate imputation by chained equations. 2023.

[4] S Wilson. miceranger: Multiple imputation by chained equations with random forests. 2020.

[5] Lisa L Doove, Stef Van Buuren, and Elise Dusseldorp. Recursive partitioning for missing data imputation in the presence of interaction effects. volume 72, pages 92–104. Elsevier, 2014.