# Missing Data: Tree-based Multiple Imputation Methods

Mühlbauer David, Pedros Philipp, Pfuderer Tristan

Institut für Statistik, Universität Bamberg

## Motivation: Valid Inference under Missingness

Proper multiple Imputation (MI) aims at valid inference with incomplete data by accounting for total variance. This mitigates information loss due to missing Data. Flexible methods such as PMM, CART, and miceRanger are widely used in complex, non-linear settings, instead of standard MICE[1]
In our simulation study, we examine whether these flexible methods are proper.[2]

**Good predictions ≠ Valid inference**

## What is "proper" Multiple Imputation (MI)

MI aims at **valid inference→correct total uncertainty** $T_j$ for each Parameter $j$:

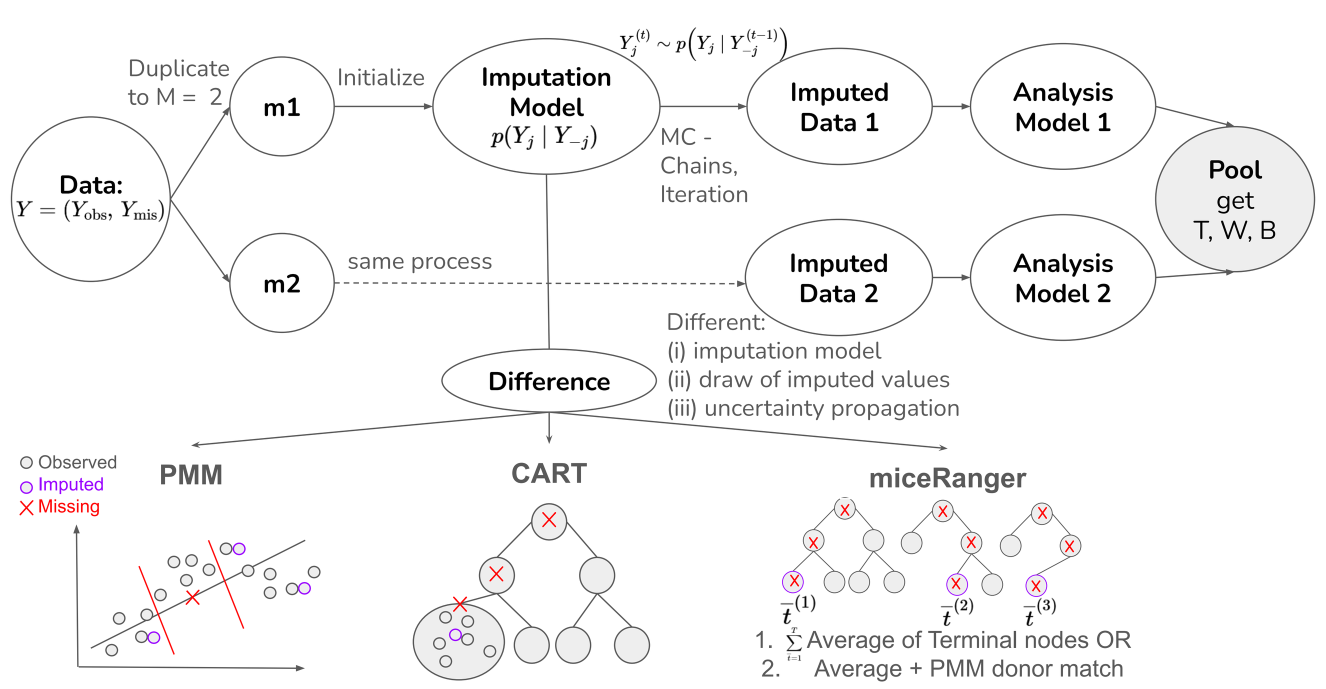
$$T_j = W_j + (1 + 1/M)B_j \tag{1}$$

$W_j =$ within-imputation variance (analysis model)
$B_j =$ between-imputation variance (missing-data uncertainty).

**One sentence:** MI is *proper* if it captures the correct total uncertainty $T_j$, so pooled SEs and confidence intervals are valid.

## Models and Methodological Differences

MICE follows a Markov chain (FCS) procedure with the idea of "perturb everything" to reflect missing-data uncertainty. The graphic below displays the process of imputing with standard mice and where the methods we investigate differ.[2][3][1]




**PMM, CART, RF :** more flexible model assumptions. Easier + better nonlinear.
**Distinction:** Imputation Model ≠ Analysis Model

- **PMM:**[4][6][7] Fit $Y_{\text{mis}} \sim Y_{\text{obs}} \to$ compute $\hat{y}_i \to$ select the $k$ donors minimizing $d_{i\ell} = |\hat{y}_i - \hat{y}_\ell| \to$ randomly draw observed $y_\ell$.
- **CART (only the differing step):** Fit tree $Y_{\text{mis}} \sim Y_{\text{obs}} \to$ route case $i \to$ sample observed $y$ from the terminal node where it lands (not from other nodes).
- **miceRanger:**[4][5]
  - Works alongside `mice`; pooling via Rubin's rules.
  - Fit RF $Y_{\text{mis}} \sim Y_{\text{obs}} \to$ average tree predictions $\to$ directly impute ensemble prediction (no donor sampling by default).

## Data Generating Process

The focus lays on a multivariate Normaldistribution with $\forall k : X_k \sim \mathcal{N}(0,1); \varepsilon \sim \mathcal{N}(0,1)$. The model is based on that of [6] and [7].

$$Y = 0.5 \cdot \sum_{i=1}^{10} X_i + 0.5 \cdot X_3^2 + X_1X_2 + X_4X_5 + \varepsilon \tag{2}$$

with correlation $\rho$ given as:

$$\rho(X_i, X_j) = \begin{cases} 0.5, & \text{falls } i \neq j; \quad i,j \in \{1;...;5\} \\ 0.3, & \text{falls } i \neq j; \quad i,j \in \{6;...;10\} \end{cases} \tag{3}$$

The simulation runs with following parameters: $n = 1000$; $M = 10$; $R = 300$ and 25 iterations

## Missingness Model

The generation of the Missing Data follows a Missing at Random (MAR)-mechanism. The missingness indicator $R_l$ is determined by a Bernoulli function, which probability depends on a logit function with arguments $X_9$ and $X_{10}$.

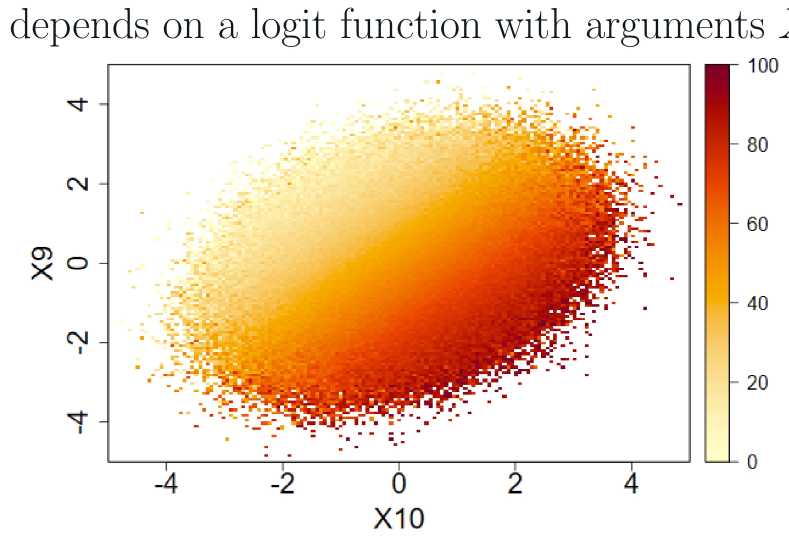



Figure 1: Heat map of the average missing entries depending on $X_9$ and $X_{10}$. An $i$-th grid square contains $a_i$ values, of which $b_i$ are NA. The colour visualises the averaged $\frac{b_i}{a_i}$ over several NA simulations. The existence of a color gradient implies a MAR-process.

$X_{i,l}$ depicts a value of the $i^{th}$ column and $l^{th}$ row. As recommended in [8], the parameters were chosen such that it results in a missingness of around 34% in the data.

$$z_l = -0.85 + 0.5 \cdot (X_{4;l} - X_{5;l}) \tag{4}$$

$$\sigma(z_l) = \frac{1}{1 - \exp(z_l)} \tag{5}$$

$$R_l \sim \text{Ber}(\sigma(z_l)) \tag{6}$$

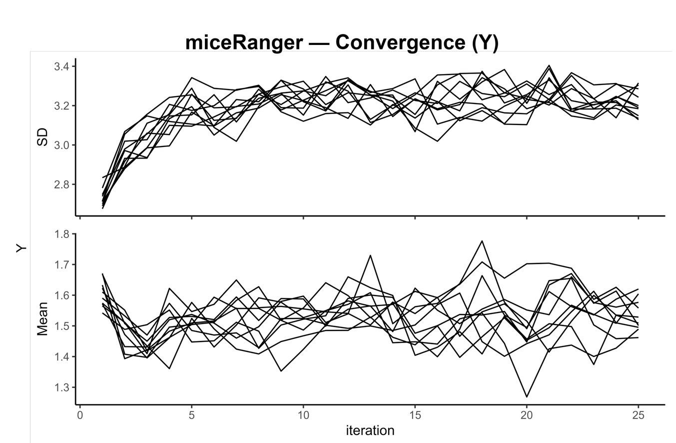



Figure 2: Convergence diagnostic of Y gained with miceranger for one simulation. After a small Burn-in of around 5 iterations, the chains dont show systematic divergence indicating visual evidence for meaningful results. PMM and CART depict similar behaviour.

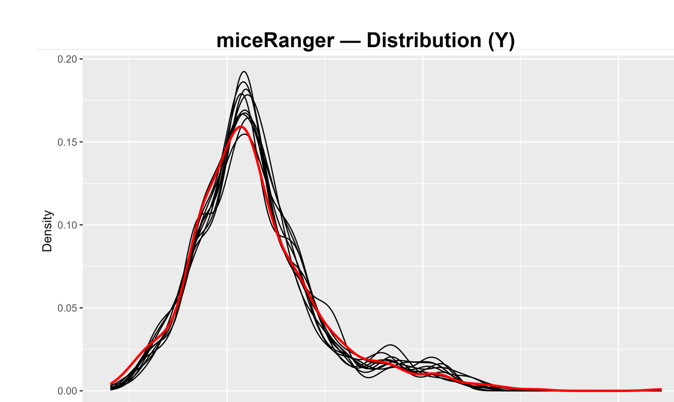



Figure 3: Density distribution of Y gained with miceranger. The red line indicates the true distribution and the black lines indicate the distribution after implementation of the NAs and imputation. Although some slight deviation are visible, particular at the peak, the imputed values generally follow the distribution of the true values

## Results

- As you can see in figure 5 CART and RF less biased for interactions (still biased) and quadratic terms than PMM
- CART overestimates Between-Variance

$$B = \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\theta}^{(m)} - \bar{\theta}_{\text{MI}}\right)^2$$

- RF underestimates B

MI Performance and Properness Summary

Predictive Accuracy and Rubin Diagnostics

| Metric | rf_ranger | cart | pmm |
|---|---|---|---|
| ↓ RMSE | 2.106 | 2.247 | 2.412 |
| ↑ R² | 0.663 | 0.608 | 0.537 |
| Mean SE | 0.064 | 0.108 | 0.116 |
| λ | 0.301 | 0.578 | 0.547 |
| FMI | 0.244 | 0.406 | 0.391 |
| DF | 118.700 | 31.100 | 34.700 |
| Var (B) | 0.001 | 0.007 | 0.007 |

Figure 4

Confidence Intervals

$$CI_{\text{MI}} = \bar{\theta}_{\text{MI}} \pm t_{\nu,1-\alpha/2}\sqrt{W + \left(1 + \frac{1}{M}\right)B} \tag{7}$$

of RF too narrow
As you can see in figure 6 CART and RF achieve higher coverage rates

$$\widehat{\text{Coverage}} = \frac{1}{R}\sum_{r=1}^{R} \mathbb{1}_{\theta \in CI_{\text{MI}}^{(r)}} \tag{8}$$

or non-linear parameter, but still under-coverage

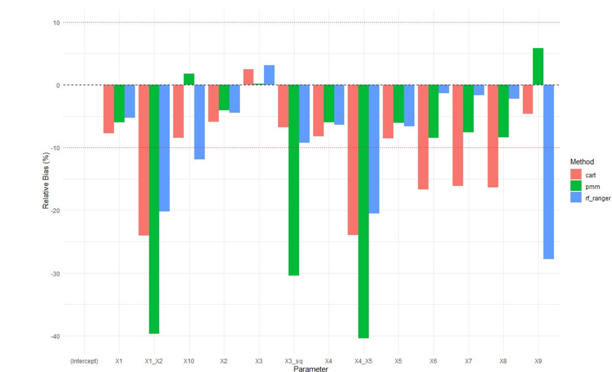


Figure 5: Comparison of rel. bias by parameter

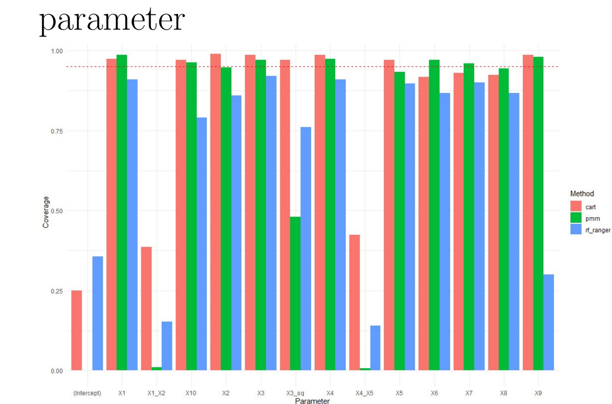


Figure 6: Comparison of coverage rates

## Conclusion

CART and RF can lead to lower bias and higher coverage rates for imputations of non-linear parameter. However, in this case, PMM, CART and RF are not 'proper' MI methods.

## References

[1] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, June 1987.

[2] Dr. Florian Meinfelder. *Part II: Theoretical Background*. Lecture slides, University of Bamberg. 2025.

[3] Stef van Buuren. "mice: Multivariate Imputation by Chained Equations in R". In: *Journal of Statistical Software* 45 (Dec. 2011).

[4] Sam Wilson. *miceRanger: Multiple Imputation by Chained Equations with Random Forests*. 2021. URL: `https://cran.r-project.org/web/packages/miceRanger/index.html` (visited on 02/19/2026).

[5] Farrell Day. *miceRanger*. 2022. URL: `https://github.com/farrellday/miceRanger` (visited on 02/19/2026).

[6] Lane F. Burgette* and Jerome P. Reiter. "Multiple Imputation for Missing Data via Sequential Regression Trees". In: *American Journal of Epidemiology* 172 (Jan. 2010).

[7] Lane F. Burgette* and Jerome P. Reiter. "Recursive partitioning for missing data imputation in the presence of interaction effects". In: *Computational Statistics & Data Analysis* 72 (Jan. 2014), pp. 92–104.

[8] Xijuan Zhang. "How to generate missing data for simulation studies". In: *The Quantitative Methods for Psychology* 19 (Feb. 2023), pp. 100–122.