# Master Research Report: Multilevel Multivariate Imputation by Chained Equations through Bayesian Additive Regression Trees

*Methodology and Statistics for the Behavioural, Biomedical and Social Sciences*

*Heleen Brüggen*

| | |
|---|---|
| **Word count:** | ... |
| **Candidate Journal:** | Computational Statistics & Data Analysis |
| **FETC Case Number:** | 23-1778 |
| **Supervisors:** | |
| MSc. T. Volker | Utrecht University |
| Dr. G. Vink | Utrecht University |
| MSc. H. Oberman | Utrecht University |

# 1 Introduction

Incomplete data sets are a common challenge in many different fields. Unlike quick fixes like mean imputation or listwise deletion, multiple imputation is considered a valid method for dealing with incomplete data [Mistler and Enders, 2017, van Buuren, 2018]. With multiple imputation, each missing value is filled in more than once, thereby considering necessary variation associated with the missingness problem. The multiply imputed data sets are analyzed, and the corresponding inferences are pooled [van Buuren, 2018, Austin et al., 2021]. Generally, multiple imputation operates under two frameworks: joint modeling and fully conditional specification [Mistler and Enders, 2017, van Buuren, 2018]. Joint modeling (JM) employs a multivariate data distribution and regression model to impute missing values [van Buuren, 2018, Enders et al., 2018a]. Fully conditional specification (FCS), or chained equations, iteratively imputes one variable with missing values at a time through conditional univariate distributions [Enders et al., 2018a, van Buuren, 2018]. The JM and FCS approaches are extended to a multilevel imputation context, where data is structured in a hierarchical way (students nested within classes) [Mistler and Enders, 2017].

Currently, the specifications of the imputation models in a multilevel context are quite complex [van Buuren, 2018]: they should at least be as general as the analysis model [Grund et al., 2018b] and preferably all-encompassing. However, the complexity of the analysis model is built step-wise with non-linearities [Hox et al., 2017] and a very complex model might not converge [van Buuren, 2018]. Bayesian Additive Regression Trees (BART) model non-linearities well and automatically through recursive binary partitioning of the predictor space often outperforming other machine learning approaches [Hill et al., 2020]. Recursive binary partitioning doesn't assume a specific data form; it divides the predictor space to maximize variance explanation by automatically identifying best fitting splits [Hastie, 2017, James et al., 2021, Salditt et al., 2023]. In a single-level context, the use of tree-based models like regression trees, random forests or BARTs simplified imputation models and performed better than parametric methods: the estimates showed better confidence interval coverage of the population parameters, lower variance and lower bias [Burgette and Reiter, 2010, Xu et al., 2016]. Also in a multilevel prediction context, BART provides better estimates with a lower Mean Squared Error (MSE) and lower relative bias compared to the standard multilevel models [Wagner et al., 2020, Chen, 2020]. However, their use in multiple imputation in a multilevel context is yet to be implemented, even though their performance in a single-level context seems promising [Burgette and Reiter, 2010, Xu et al., 2016]. Thus, my research question will be: *Can multivariate imputation by chained equations through a multilevel bayesian additive regression trees model improve the bias, variance and coverage of the estimates in a multilevel context compared to current practices?* Given the success of non-parametric models in single-level multiple imputation, I anticipate that employing multilevel BART models in a multilevel missing data context will reduce bias, accurately model variance, and improve estimate coverage compared to classical multilevel imputation through *21.pmm* in MICE.

# 2 Method

We conduct a simulation study in which five factors are varied:

1. *Intraclass Correlation* (ICC = 0, .05, .3 and .5)
2. *Number of clusters* (J = 30 and 50)
3. *Within-cluster sample size* ($n_j$ = 5, 15, 25 and 50)
4. *The Missing At Random (MAR) and Missing Completely At Random (MCAR) data rate* (0%, 25% and 50%)
5. *Within-group effect size:* (size of the regression coefficients $\beta$ = .2, .5 and .8)

All these values are realistic in practice and/or previously proposed [Gulliford et al., 1999, Murray and Blitstein, 2003, Hox et al., 2017, Grund et al., 2018b, Enders et al., 2018b, Enders et al., 2020]. The ICC can be interpreted as the expected correlation between two randomly sampled individuals from the same group or the proportion of the total variance at the cluster level [Gulliford et al., 2005, Shieh, 2012, Hox et al., 2017]. The simulation study will be performed in R with the package MICE [Buuren and Groothuis-Oudshoorn, 2011] to perform the FCS imputations, which I will enchance by integrating BART. The classical, *21.pmm* in

MICE, FCS multilevel imputation method [Lüdtke et al., 2017, Enders et al., 2018b, Enders et al., 2020] and complete case analysis will serve as a benchmark. The population data-generating mechanism will be

$$y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \epsilon_{ij}, \tag{1.1}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + \upsilon_{0j}, \tag{1.2}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}Z_j + \upsilon_{1j}, \tag{1.3}$$
$$\beta_{2j} = \gamma_{20} + \upsilon_{2j}, \tag{1.4}$$

where $y_{ij}$ is a continuous level 1 outcome variable for person $i$ in group $j$ and $Z_j$ is a continuous level 2 variable. The random intercept $\beta_{0j}$ is determined by the grand mean $\gamma_{00}$, the group effect $\gamma_{01}Z_j$ and the group-level random residuals $\upsilon_{0j}$. The regression coefficient $\beta_{1j}$ for the continuous variable $X_{1ij}$ depends on the the intercept $\gamma_{10}$, the cross-level interaction $\gamma_{11}Z_j$ and the random slopes $\upsilon_{1j}$. For the ordinal variable $X_{2ij}$ (treated as continuous), $\beta_{2j}$ is determined by the intercept $\gamma_{20}$ and the random slopes $\upsilon_{2j}$. The residuals and random slopes $\upsilon_{0j}$, $\upsilon_{1j}$, $\upsilon_{2j}$, and $\epsilon_{ij}$, and random slopes follow a zero-mean normal distribution. $X_1$, $X_2$ and $Z$ are multivariate normally distributed. The estimates will be evaluated on their relative bias (the difference between the average estimate and the true value), modeled variance and the 95% confidence interval coverage.

# 3 Results

## References


[Audigier et al., 2018] Audigier, V., White, I. R., Jolani, S., Debray, T. P. A., Quartagno, M., Carpenter, J., Van Buuren, S., and Resche-Rigon, M. (2018). Multiple Imputation for Multilevel Data with Continuous and Binary Variables. *Statistical Science*, 33(2).

[Austin et al., 2021] Austin, P. C., White, I. R., Lee, D. S., and Van Buuren, S. (2021). Missing Data in Clinical Research: A Tutorial on Multiple Imputation. *Canadian Journal of Cardiology*, 37(9):1322–1331.

[Burgette and Reiter, 2010] 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.

[Buuren and Groothuis-Oudshoorn, 2011] Buuren, S. V. and Groothuis-Oudshoorn, K. (2011). **Mice** : Multivariate Imputation by Chained Equations in *R*. *Journal of Statistical Software*, 45(3).

[Chen, 2020] Chen, S. (2020). *A New Multilevel Bayesian Nonparametric Algorithm and Its Application in Causal Inference*. PhD thesis, Texas A&M University.

[Chipman et al., 2010] Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1).

[Dong and Mitani, 2023] Dong, M. and Mitani, A. (2023). Multiple imputation methods for missing multilevel ordinal outcomes. *BMC Medical Research Methodology*, 23(1):112.

[Doove et al., 2014] Doove, L., Van Buuren, S., and Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72:92–104.

[Enders, 2017] Enders, C. K. (2017). Multiple imputation as a flexible tool for missing data handling in clinical research. *Behaviour Research and Therapy*, 98:4–18.

[Enders et al., 2020] Enders, C. K., Du, H., and Keller, B. T. (2020). A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. *Psychological Methods*, 25(1):88–112.

[Enders et al., 2018a] Enders, C. K., Hayes, T., and Du, H. (2018a). A Comparison of Multilevel Imputation Schemes for Random Coefficient Models: Fully Conditional Specification and Joint Model Imputation with Random Covariance Matrices. *Multivariate Behavioral Research*, 53(5):695–713.

[Enders et al., 2018b] Enders, C. K., Keller, B. T., and Levy, R. (2018b). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods*, 23(2):298–317.

[Enders et al., 2016] Enders, C. K., Mistler, S. A., and Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods*, 21(2):222–240.

[Grund et al., 2016] Grund, S., Lüdtke, O., and Robitzsch, A. (2016). Multiple imputation of missing covariate values in multilevel models with random slopes: A cautionary note. *Behavior Research Methods*, 48(2):640–649.

[Grund et al., 2018a] Grund, S., Lüdtke, O., and Robitzsch, A. (2018a). Multiple Imputation of Missing Data at Level 2: A Comparison of Fully Conditional and Joint Modeling in Multilevel Designs. *Journal of Educational and Behavioral Statistics*, 43(3):316–353.

[Grund et al., 2018b] Grund, S., Lüdtke, O., and Robitzsch, A. (2018b). Multiple Imputation of Missing Data for Multilevel Models: Simulations and Recommendations. *Organizational Research Methods*, 21(1):111–149.

[Gulliford et al., 2005] Gulliford, M., Adams, G., Ukoumunne, O., Latinovic, R., Chinn, S., and Campbell, M. (2005). Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. *Journal of Clinical Epidemiology*, 58(3):246–251.

[Gulliford et al., 1999] Gulliford, M. C., Ukoumunne, O. C., and Chinn, S. (1999). Components of Variance and Intraclass Correlations for the Design of Community-based Surveys and Intervention Studies: Data from the Health Survey for England 1994. *American Journal of Epidemiology*, 149(9):876–883.

[Hajjem et al., 2011] Hajjem, A., Bellavance, F., and Larocque, D. (2011). Mixed effects regression trees for clustered data. *Statistics & Probability Letters*, 81(4):451–459.

[Hastie, 2017] Hastie, T. J., editor (2017). *Statistical Models in S*. Routledge, 1st edition.

[Hill et al., 2020] Hill, J., Linero, A., and Murray, J. (2020). Bayesian Additive Regression Trees: A Review and Look Forward. *Annual Review of Statistics and Its Application*, 7(1):251–278.

[Hox et al., 2017] Hox, J. J., Moerbeek, M., and Van De Schoot, R. (2017). *Multilevel Analysis: Techniques and Applications*. Routledge, Third edition. — New York, NY : Routledge, 2017. —, 3 edition.

[Hughes et al., 2014] Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K., and Sterne, J. A. (2014). Joint modelling rationale for chained equations. *BMC Medical Research Methodology*, 14(1):28.

[James et al., 2021] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics. Springer US, New York, NY.

[Lin and Luo, 2019] Lin, S. and Luo, W. (2019). A New Multilevel CART Algorithm for Multilevel Data with Binary Outcomes. *Multivariate Behavioral Research*, 54(4):578–592.

[Lüdtke et al., 2017] 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):141–165.

[Mistler and Enders, 2017] Mistler, S. A. and Enders, C. K. (2017). A Comparison of Joint Model and Fully Conditional Specification Imputation for Multilevel Missing Data. *Journal of Educational and Behavioral Statistics*, 42(4):432–466.

[Murray and Blitstein, 2003] Murray, D. M. and Blitstein, J. L. (2003). Methods To Reduce The Impact Of Intraclass Correlation In Group-Randomized Trials. *Evaluation Review*, 27(1):79–103.

[Pellagatti et al., 2021] Pellagatti, M., Masci, C., Ieva, F., and Paganoni, A. M. (2021). Generalized mixed-effects random forest: A flexible approach to predict university student dropout. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 14(3):241–257.

[Quartagno and Carpenter, 2022] Quartagno, M. and Carpenter, J. R. (2022). Substantive model compatible multilevel multiple imputation: A joint modeling approach. *Statistics in Medicine*, 41(25):5000–5015.

[Resche-Rigon and White, 2018] Resche-Rigon, M. and White, I. R. (2018). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research*, 27(6):1634–1649.

[Salditt et al., 2023] Salditt, M., Humberg, S., and Nestler, S. (2023). Gradient Tree Boosting for Hierarchical Data. *Multivariate Behavioral Research*, pages 1–27.

[Shieh, 2012] Shieh, G. (2012). A comparison of two indices for the intraclass correlation coefficient. *Behavior Research Methods*, 44(4):1212–1223.

[Sparapani et al., 2021] Sparapani, R., Spanbauer, C., and McCulloch, R. (2021). Nonparametric Machine Learning and Efficient Computation with Bayesian Additive Regression Trees: The **BART** *R* Package. *Journal of Statistical Software*, 97(1).

[Van Buuren, 2007] Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3):219–242.

[van Buuren, 2018] van Buuren, S. (2018). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Interdisciplinary Statistics Series. CRC Press, Taylor & Francis Group, Boca Raton London New York, second edition edition.

[Wagner et al., 2020] Wagner, J., West, B. T., Elliott, M. R., and Coffey, S. (2020). Comparing the Ability of Regression Modeling and Bayesian Additive Regression Trees to Predict Costs in a Responsive Survey Design Context. *Journal of Official Statistics*, 36(4):907–931.

[Xu et al., 2016] Xu, D., Daniels, M. J., and Winterstein, A. G. (2016). Sequential BART for imputation of missing covariates. *Biostatistics*, 17(3):589–602.