# Master Thesis Proposal: 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:** | 750 |
| **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 occurrence in many different fields. Nowadays, multiple imputation is considered the best general method for imputing incomplete data sets [34, 25]. Multiple imputation completes an incomplete data set multiple times, conducts the statistical analysis of interest on each completed set and after, pools the results together [2, 34]. In general, there are two broad frameworks of multiple imputation: joint modelling and fully conditional specification [34, 25]. Joint modeling (JM) uses a multivariate distribution of the data through a multivariate regression model from which imputations are drawn for the variables with missing values [11, 34]. Fully conditional specification (FCS), or chained equations, iteratively imputes the variables with missing values one at a time through conditional univariate distributions [11, 34]. The JM and FCS approaches are extended to a multilevel imputation context, where data is structured in a hierarchical way (students nested within classes) [25].

Currently, the specifications of the imputation models in a multilevel context are quite complex due to the hierarchical structure of the data [34]. In a single-level context, the use of tree-based models like regression trees, random forests or Bayesian Additive Regression Trees (BART) not only simplified the specification of the imputation models, they also performed better than the classical specification: the estimates showed better confidence interval coverage of the real estimates, lower variance and lower bias [3, 36]. These models are able to capture complicated relationships by creating a hierarchical tree-like structure through recursive binary partitioning of predictor space [22, 18]. BART excel at this, often outperforming other machine learning approaches [19]. BART relies on regression trees as fundamental building blocks and combines multiple to form an overall fit. To prevent overfitting, a regularization prior imposes constraints on the size and fit of each individual tree. The backfitting algorithm iteratively cycles through each tree until convergence [19, 6]. Considered in a prediction context, BART provides better estimates with a lower Mean Squared Error (MSE) and lower relative bias compared to the standard multilevel models [35, 5]. However, the use of tree-based models in multiple imputation in a multilevel context is yet to be implemented, even though their performance in a single-level context seems promising [3, 36]. Thus, my research question will be: *How can multilevel multivariate imputation by chained equations through a bayesian additive regression trees model improve the bias, variance and coverage of the estimates in a multilevel context?*. Considering the succes of non-parametric models in multiple imputation in a single-level context, I expect that the use of BART models in a multilevel missing data context will decrease the bias, accurately model the variance and increase the coverage of the estimates when compared to the classical multilevel imputation through chained equations.

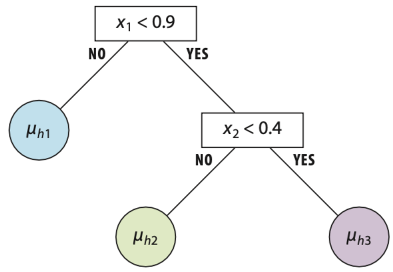


Figure 1: An example of a regression tree [19]

# 2 Analytic strategy

A simulation study will performed. Five factors will be varied in the study:

1. *Intraclass Correlation* (ICC = .05, .20 and .50)
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%, 5%, 15% and 25%)
5. *The effect size* ($\delta$ = .2, .5 and .8)

The ICC can be interpreted as the expected correlation between two randomly sampled individuals from the same group or the variance at the cluster level [31, 15, 20]. These values are realistic values in practice and/or previously proposed [16, 26, 12, 10, 20]. The simulation study will be performed in R with the package MICE [4] to perform the FCS imputations. The classical FCS multilevel imputation method [24, 12, 10] will serve as a benchmark. 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. 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 $j$. $\beta_{0j}$ is random intercept defined by the grand mean $\gamma_{00}$, the group effect $\gamma_{01}Z_j$ and the group-level random residuals $\upsilon_{0j}$ with a normal distribution $\upsilon_{0j} \sim \mathcal{N}(0, \sigma^2)$. $\beta_{1j}$ is the regression coefficient for the continuous variable $X_{1ij}$, which is defined by the intercept $\gamma_{10}$, the cross-level interaction $\gamma_{11}Z_j$ and the random slopes $\upsilon_{1j}$ with a normal distribution $\upsilon_{1j} \sim \mathcal{N}(0, \sigma^2)$. $\beta_{2j}$ is a the regression coefficient for $X_{2ij}$, which is an ordinal variable with 7 categories and defined by the intercept $\gamma_{20}$ and the random slopes $\upsilon_{2j}$ ($\upsilon_{2j} \sim \mathcal{N}(0, \sigma^2)$). $X_{2ij}$ will be treated as continuous. $\epsilon_{ij}$ are the normally distributed residuals $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

# References


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

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

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

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

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

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

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

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

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

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

[11] Enders, C. K., Hayes, T., and Du, H. 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 (Sept. 2018), 695–713.

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

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

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

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

[16] Gulliford, M. C., Ukoumunne, O. C., and Chinn, S. 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 (May 1999), 876–883.

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

[18] Hastie, T. J., Ed. *Statistical Models in S*, 1st ed. Routledge, 2017.

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

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

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

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

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

[24] Lüdtke, O., Robitzsch, A., and Grund, S. Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods 22*, 1 (Mar. 2017), 141–165.

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

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

[27] Pellagatti, M., Masci, C., Ieva, F., and Paganoni, A. M. 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 (June 2021), 241–257.

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

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

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

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

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

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

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

[35] Wagner, J., West, B. T., Elliott, M. R., and Coffey, S. 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 (Dec. 2020), 907–931.

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