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:	2488
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 is a common challenge in many fields of research. Frequently used ad hoc strategies to deal with missing data, as listwise or mean imputation often lead to erroneuous inferences in realistic situations, due to biased estimates and inaccurate variance estimates because these strategies don't consider the multivaraite nature of the data: the missingness can relate to observed values (Austin et al., 2021; Enders, 2017; Kang, 2013; van Buuren, 2018). Rubin defined three of such missing data mechanisms: Missing Completely At Random (MCAR) where the cause of the missing data is unrelated to the data, Missing At Random (MAR) where the missing data is related to the observed data, and Missing Not At Random (MNAR) where the missing data may also be related to unobserved data (Rubin, 1976).

Multiple imputation (MI) (Rubin, 1987) is considered a valid method for dealing with incomplete data, it allows us to separate the missing data problem from the analysis problem (Audigier et al., 2018; Austin et al., 2021; Burgette and Reiter, 2010; Enders, 2017; Grund et al., 2021; Hughes et al., 2014; Mistler and Enders, 2017; Van Buuren, 2007; van Buuren, 2018). MI is used to impute each missing value in the dataset more than once given the observed data, considering necessary variation associated with the missingness problem. The multiply imputed datasets are analyzed, and the corresponding inferences are pooled according to Rubin's rules (Austin et al., 2021; Carpenter and Kenward, 2013; Rubin, 1987; van Buuren, 2018). However, specifying the imputation models, the models used to impute the missing data, can be challenging. The concept of congeniality dictates that the imputation models should be at least as general as the analysis model and preferably all-encompassing (Bartlett et al., 2015; Enders et al., 2018a; Grund et al., 2016, 2018a; Meng, 1994). Otherwise, it will not capture every aspect of the data and the analysis model estimates may be biased. So, when the complexity of data increases, specifying the imputation models becomes more difficult (Grund et al., 2018a; van Buuren, 2018).

Congeniality-issues become more pronounced when MI is used in a multilevel data context (Audigier et al., 2018; Dong and Mitani, 2023; Enders et al., 2016, 2018a,b, 2020; Grund et al., 2016, 2018a,b, 2021; Lüdtke et al., 2017; Mistler and Enders, 2017; Quartagno and Carpenter, 2022; Resche-Rigon and White, 2018; Taljaard et al., 2008; van Buuren, 2018). Multilevel data is hierarchically structured, where, for example, students are nested within schools (Hox and Roberts, 2011; Hox et al., 2017). When analysing multilevel data, the hierarchical structure should be taken into account, which can be done using multilevel models (MLM) (Hox and Roberts, 2011; Hox et al., 2017; Lüdtke et al., 2017). Since ignoring the multilevel structure will underestimate the intra-class correlation (ICC) (Hox and Roberts, 2011; Lüdtke et al., 2017; Taljaard et al., 2008; van Buuren, 2018), which can be interpreted the proportion of the total variance at level-2 (Gulliford et al., 2005; Hox and Roberts, 2011; Shieh, 2012). MLMs can contain both level-1, and level-2 variables, relating to the individual and class respectively, random intercepts, random slopes, and cross-level interactions (Hox and Roberts, 2011; Hox et al., 2017). The complexity of the multilevel analysis model is built step-wise with non-linearities, meaning the analysis model is not determined beforehand (Hox and Roberts, 2011; Hox et al., 2017). Thus, including the hierarchical structure, along with the complicated non-linearities from cross-level interactions in imputation models can be quite challenging (Burgette and Reiter, 2010; Hox and Roberts, 2011; van Buuren, 2018) and a very complex model might not converge (van Buuren, 2018).

One of the two MI frameworks, fully conditional specification (FCS), otherwise known as chained equations, is believed to be more flexible than its counterpart, joint modeling (JM) (Audigier et al., 2018; Burgette and Reiter, 2010; Grund et al., 2018b; Van Buuren, 2007). FCS iteratively imputes each incomplete variable conditional on complete and previously imputed variables (Enders et al., 2016, 2018a,b; Grund et al., 2018b; Hughes et al., 2014; Mistler and Enders, 2017; van Buuren, 2018). In a multilevel context, FCS employs univariate linear mixed models to account for the hierarchical structure (Enders et al., 2018a; Mistler and Enders, 2017; Resche-Rigon and White, 2018). Furthermore, FCS can be used to impute non-linearities, such as cross-level interactions, by using 'passive imputation' or defining a separate imputation model for the non-linearities (Grund et al., 2018a; van Buuren, 2018). On the other hand, JM doesn't account for cross-level interactions at all (Grund et al., 2018a; van Buuren, 2018). Still, imputation models including cross-level interaction or non-linear terms in FCS is still very complicated (Grund et al., 2018a, 2021) and, thus, researchers' focus has predominantly been on the inclusion of random intercepts and slopes, but not of cross-level interactions (Enders et al., 2016, 2018a,b, 2020; Grund et al., 2016, 2018b).

Using non-parametric tree-based models might solve this problem because they do not assume a specific data distribution and, thus, implicitly model non-linear relationships and interactions between the predictor variables, and handle continuous and categorical variables simultaneously (Breiman et al., 1984; Burgette and Reiter, 2010; Chipman et al., 2010; Hill et al., 2020; James et al., 2021; Lin and Luo, 2019; Salditt et al., 2023). In a single-level imputation context, the use of tree-based, non-parametric models like regression trees, random forests, or Bayesian Additive Regression Trees (BART) simplified imputation models and performed better than parametric methods: the imputations showed better confidence interval coverage of the population parameters, lower variance and lower bias, especially in non-linear and interactive contexts (Burgette and Reiter, 2010; Silva and Gutman, 2022; Xu et al., 2016). Waljee et al. (2013) also found lower imputation error when imputing with a random forest algorithm compared to `MICE`, KNN and mean imputation.

BART models have been implemented in a multilevel prediction context. However, multilevel-BART models (M-BART) have predominantly been implemented with only random intercepts (Chen, 2020; Tan et al., 2016; Wagner et al., 2020; Wundervald et al., 2022). In a prediction context, Wagner et al. (2020) have found that this random intercept M-BART model provided better predictions with a lower mean squared error (MSE) compared to a parametric MLM, Tan et al. (2016) found higher area under the curve (AUC) values compared to a single level BART model and linear random intercept model, and Chen (2020) found better predictions and better coverage of the estimates compared to parametric models and a single-level BART model. Other researchers modeled the random intercept as an extra split on each terminal node and found a lower MSE compared to a standard BART model and parametric MLMs (Wundervald et al., 2022). Dorie et al. (2022) developed a multilevel BART model that included random intercepts and random slopes by combining BART with the Stan algorithm. Where the random parts are modeled by Stan (Lee et al., 2017). Their results showed that their algorithm `stan4bart` showed better coverage of the population values and lower root mean squared error (RMSE) compared to BART models with varying intercept, BART models ignoring the multilevel structure, bayesian causal forests (BCF), and parametric MLMs.

In spite of these promising findings, M-BART models have yet to be implemented in a multilevel multiple imputation context. Thus, my thesis 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 MI, I anticipate that employing M-BART models in a multilevel missing data context will reduce bias, accurately model variance, and improve estimate coverage compared to classical multilevel imputation through `2l.pmm, 2l.lmer, 2l.pan, 2l.jomo, rf` and single-level `pmm` and complete case analysis in the R-package `MICE` (Buuren and Groothuis-Oudshoorn, 2011). However, in this research report, I will only focus on the implementation of M-BART models in a prediction context and asses their performance in terms of relative bias and MSE. The research question is: *Can M-BART models improve the relative bias and MSE of the predictions in a multilevel context compared to a single-level BART model?.*

The research report's sections will cover theoretical background, methods for evaluating M-BART models, preliminary results, and discussion of next steps.

2 Method

2.1 Theoretical background

2.1.1 Bayesian Additive Regression Trees (BART)

BART is a sum-of-trees model proposed by Chipman et al. (2010) that has regression trees as its building blocks (Chipman et al., 2010; Hill et al., 2020; James et al., 2021). Regression trees divide the data into subgroups by recursively splitting the data into binary subgroups based on the predictors minimizing variability within the subgroups (Hastie, 2017; James et al., 2021; Salditt et al., 2023). Recursive binary partitioning of the predictor space doesn't assume a specific data form, making this a non-parametric model (Hastie, 2017; James et al., 2021; Salditt et al., 2023) and allows regression trees to model non-linearities well and automatically (Burgette and Reiter, 2010; Hill et al., 2020). Chipman et al. (2010) define the BART model as:

$$f(\mathbf{x}) = \sum_{k=1}^{m} g(\mathbf{x}; T_k, M_k), \tag{1}$$

where $f(\mathbf{x})$ is the overall fit of the model: the sum of m regression trees, $\mathbf{x}$ are the predictor variables, T_k is the k^{th} tree and M_k is the collection of leaf parameters within the k^{th} tree (Chipman et al., 2010; Hill et al., 2020; James et al., 2021). The data are assumed to arise from a model with additive normally distributed errors: $Y = \sum_{k=1}^{m} g(\mathbf{x}; T_k, M_k) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$. Next to the sum-of-trees model, BART also includes a regularization prior that constrains the size and fit of each tree so that each contributes only a small part to prevent overfitting (Chipman et al., 2010; Hill et al., 2020; James et al., 2021). BARTs are estimated using the Bayesian back-fitting Markov Chain Monte Carlo (MCMC) algorithm. It updates individual trees, considering the remaining trees, their associated parameters, and the residual standard deviation (σ). It fits a new tree to the partial residuals, r_i, treating them as the data, by perturbing the tree from the previous iteration. Perturbations entail either *growing*, *pruning*, or *changing* a tree. *Growing* means adding additional splits, *pruning* removes splits, and *changing* changes decision rules. The algorithm stops after the specified amount of iterations. The partial residuals are defined as:

$$r_i = y_i - \sum_{k'<k} \hat{f}_{k'}^{b}(x_i) - \sum_{k'>k} \hat{f}_{k'}^{b-1}(x_i), \text{with } i = 1, \ldots, N \tag{2}$$

where $\hat{f}_k^b(x_i)$ is the prediction of the k^{th} tree in the b^{th} iteration for person i and sample size N.

2.1.2 Multilevel-BART (M-BART)

Chen (2020); Wagner et al. (2020) and Tan et al. (2016) define a M-BART model including a random intercept building on the work of Lin and Luo (2019). The M-BART algorithm breaks down the observed variable into fixed and random components. The fixed components are modeled by BART and the random components are modeled by a linear mixed effects model (Chen, 2020; Tan et al., 2016; Wagner et al., 2020). The BART model (1) can be extended to include a random intercept by:

$$f(\mathbf{x}) = \sum_{k=1}^{m} g(\mathbf{x}; T_k, M_k) + \alpha_j, \tag{3}$$

where, now, $f(\mathbf{x})$ is the overal fit of the model incorporating random intercept α_j for cluster j.

2.2 Simulation study

For this research report, I conduct a simulation study to examine the performance of three M-BART models in a multilevel prediction context compared to a sinlge-level BART model.

2.2.1 Data generating mechanism

The population data-generating mechanism will be based on the following MLM:

$$y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \beta_{3j}X_{3ij} + \beta_{4j}X_{4ij} + \beta_{5j}X_{5ij} + \beta_{6j}X_{6ij} + \beta_{7j}X_{7ij} + \epsilon_{ij}, \tag{4}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_{1j} + \upsilon_{0j}, \tag{4.1}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}Z_{1j} + \upsilon_{1j}, \tag{4.2}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}Z_{1j} + \upsilon_{2j}, \tag{4.3}$$

$$\beta_{3j} = \gamma_{30} + \gamma_{32}Z_{2j} + \upsilon_{3j}, \tag{4.4}$$

$$\beta_{4j} = \gamma_{40} + \upsilon_{4j}, \tag{4.5}$$

$$\beta_{5j} = \gamma_{50} + \upsilon_{5j}, \tag{4.6}$$

$$\beta_{6j} = \gamma_{60} + \upsilon_{6j}, \tag{4.7}$$

$$\beta_{7j} = \gamma_{70}, \tag{4.8}$$

where y_{ij} is a continuous level-1 outcome variable for person i in group j and Z_{1j} and Z_{2j} are continuous level-2 variables. The random intercept β_{0j} is determined by the grand mean γ_{00}, the group effect $\gamma_{01}Z_{1j}$ and the group-level random residuals υ_{0j}. The regression coefficients β_{1j}, β_{2j}, and β_{3j} for the continuous variables X_{1ij}, X_{2ij}, and X_{3ij} depend on the the intercepts γ_{10}, γ_{20}, and γ_{30}, the cross-level interactions $\gamma_{11}Z_{1j}$, $\gamma_{21}Z_{1j}$, and $\gamma_{32}Z_{2j}$, and the random slopes υ_{1j}, υ_{2j}, and υ_{3j}. The regression coefficients β_{4j}, β_{5j} and β_{6j} are determined by the intercepts γ_{40}, γ_{50} and γ_{60} and the random slopes υ_{4j}, υ_{5j} and υ_{6j}. The regression coefficient β_{7j} is determined by the intercept γ_{70}. The residuals and random slopes υ_{0j}, υ_{1j}, υ_{2j}, υ_{3j}, υ_{4j}, υ_{5j}, and υ_{6j} and ϵ_{ij} follow a zero-mean normal distribution. The variance of υ_{0j}, the group-level random residuals, were scaled such that the specified ICC value was obtained. υ_{1j}, υ_{2j}, υ_{3j}, υ_{4j}, υ_{5j}, and υ_{6j} all have a variance of 1. ϵ_{ij} had a variance of 25. X_1, X_2, X_3, X_4, X_5,

X_6 and X_7 are multivariate normally distributed: $\mathbf{X}_{ij} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = (0, 0, 0, 0, 0, 0, 0)$ and $\boldsymbol{\Sigma} = \text{diag}(6.25, 9, 4, 11.56, 4, 2.5, 19.36)$ with no co-variances. The level-2 variables Z_1 and Z_2 are also multivariate normally distributed: $\mathbf{Z}_j \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = (0, 0)$ and $\boldsymbol{\Sigma} = \text{diag}(1, 2.56)$. The group-level effects (γ_{01} and $\gamma 02$) were set to .5, the cross-level interactions (γ_{11}, γ_{21}, and γ_{32}) were set to .35, and the overall intercept (γ_{00}) was set to 10. The within-group effect sizes (γ_{10}, γ_{20}, γ_{30}, γ_{40}, γ_{50}, γ_{60}, and γ_{70}) were varied in the simulations.

2.2.2 Simulation design

Table 1 shows the variations considered in the simulation study. All these values are realistic in practice and/or previously proposed (Enders et al., 2018b, 2020; Grund et al., 2018a; Gulliford et al., 1999; Hox et al., 2017; Murray and Blitstein, 2003). For each combination of varying parameters, 6 datasets are simulated for every scenario to reduce computational time. 4 different models are compared: a single level BART, single level BART with groups modelled using dummy-variables, a multilevel BART model incorporating a random intercept (Chen, 2020; Tan et al., 2016; Wagner et al., 2020; Wundervald et al., 2022), and a multilevel BART model combined with Stan to model the random parts of the models (Dorie et al., 2022). The first three models are fitted with the package `dbarts` (Dorie, 2023a) and the last with the package `stan4bart` (Dorie, 2023b) in R (R Core Team, 2023). The default arguments from the function `rbart_vi` are used for all models, as well as the default priors.

Table 1: Simulation design

Parameter	**Values**
Number of clusters (j)	30, 50
Within-cluster sample size ($\mathrm{n_j}$)	5, 15, 35, 50
Intraclass Correlation (ICC)	0, .05, .3, .5
Within-group effect size (γ)	.2, .5, .8

2.2.3 Evaluation

The fitted models are evaluated in terms relative bias and Mean Squared Error (MSE) of the predictions (Morris et al., 2019):

$$Bias = \frac{1}{n_{\text{sim}}} \sum_{t=1}^{n_{\text{sim}}} (\hat{\theta}_t - \theta), \tag{5}$$

$$MSE = \frac{1}{n_{\text{sim}}} \sum_{t=1}^{n_{\text{sim}}} (\hat{\theta}_t - \theta)^2, \tag{6}$$

$$\tag{5a}$$

where $\hat{\theta}_t$ is the estimated parameter in simulation t, θ is the true value, and n_{sim} is the number of simulated datasets and smaller is better.

3 Results

Figure 1 shows the average relative bias over the simulations for all models, simulated datasets, every *ICC* value, and within-group effect size. On the x-axis we can see the different simulated datasets with their names specifying the total sample size, with respectively number of groups and group sizes within parentheses. We can see that when there is no multilevel structure in the dataset, $ICC = 0$, the models perform similarly in terms of relative bias: overall, the bias is around zero. Overall, we can see a slight increase in uncertainty when the total sample size is small for all γ and *ICC*. This effect is examplified when γ and *ICC* increase. However, when considering the `stan4bart` model, which models the random parts of the model in Stan and the fixed parts in BART, the uncertainty stays considerably constant when increasing the *ICC*: the relative bias is higher when the total sample size is small, but the uncertainty does not seem to significantly increase with higher *ICC* or higher γ. Overall, `stan4bart` has the lowest bias for all γ and *ICC* compared to the other models.

Figure 1: Bias of the estimates for all simulated datasets over six simulations with ICC values in the rows and within group effect sizes in the columns for four models: single-level BART (`bart`), single-level BART with group dummies (`gbart`), random intercept multilevel BART (`rbart`) and random intercept random slope multilevel BART (`stan4bart`). The x-axis denotes the total sample size (number of groups, group size).

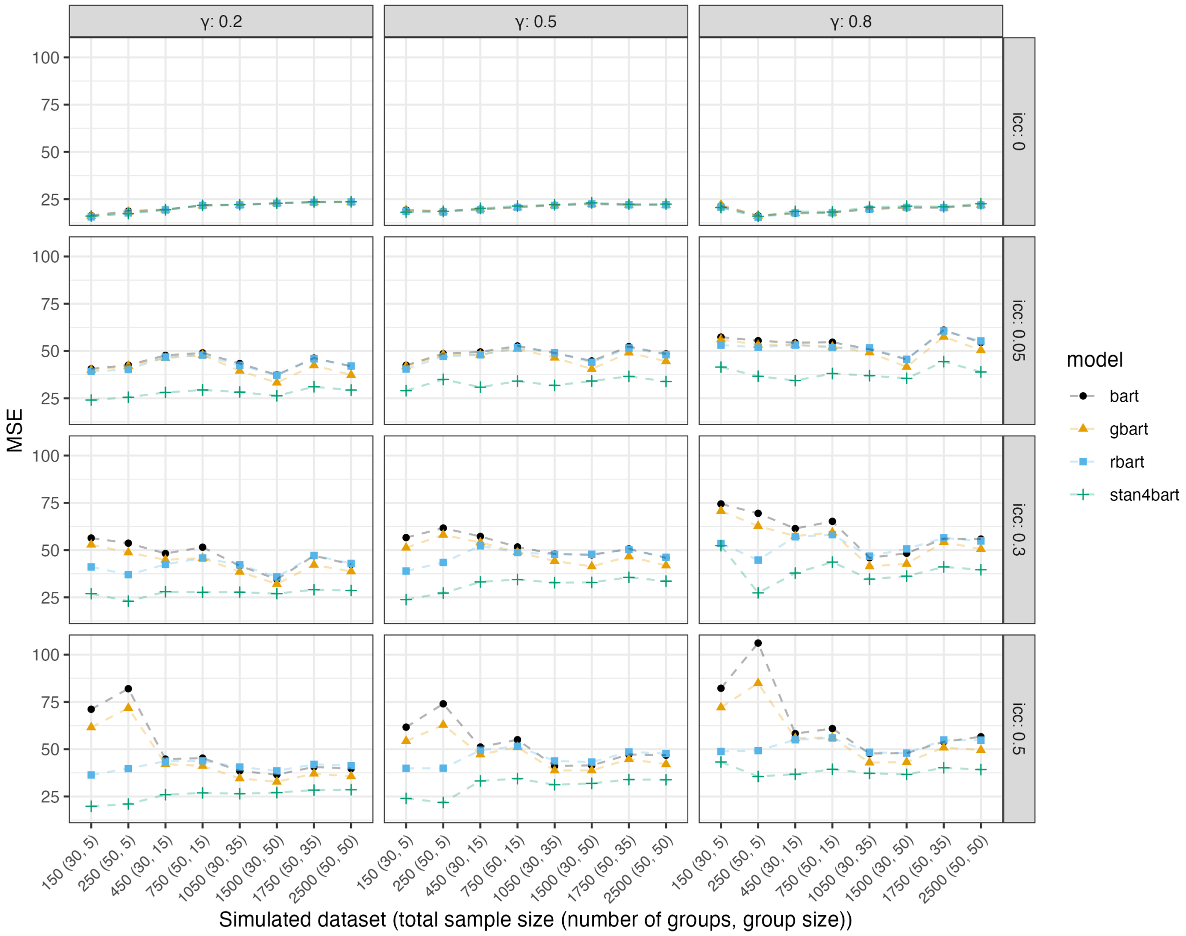
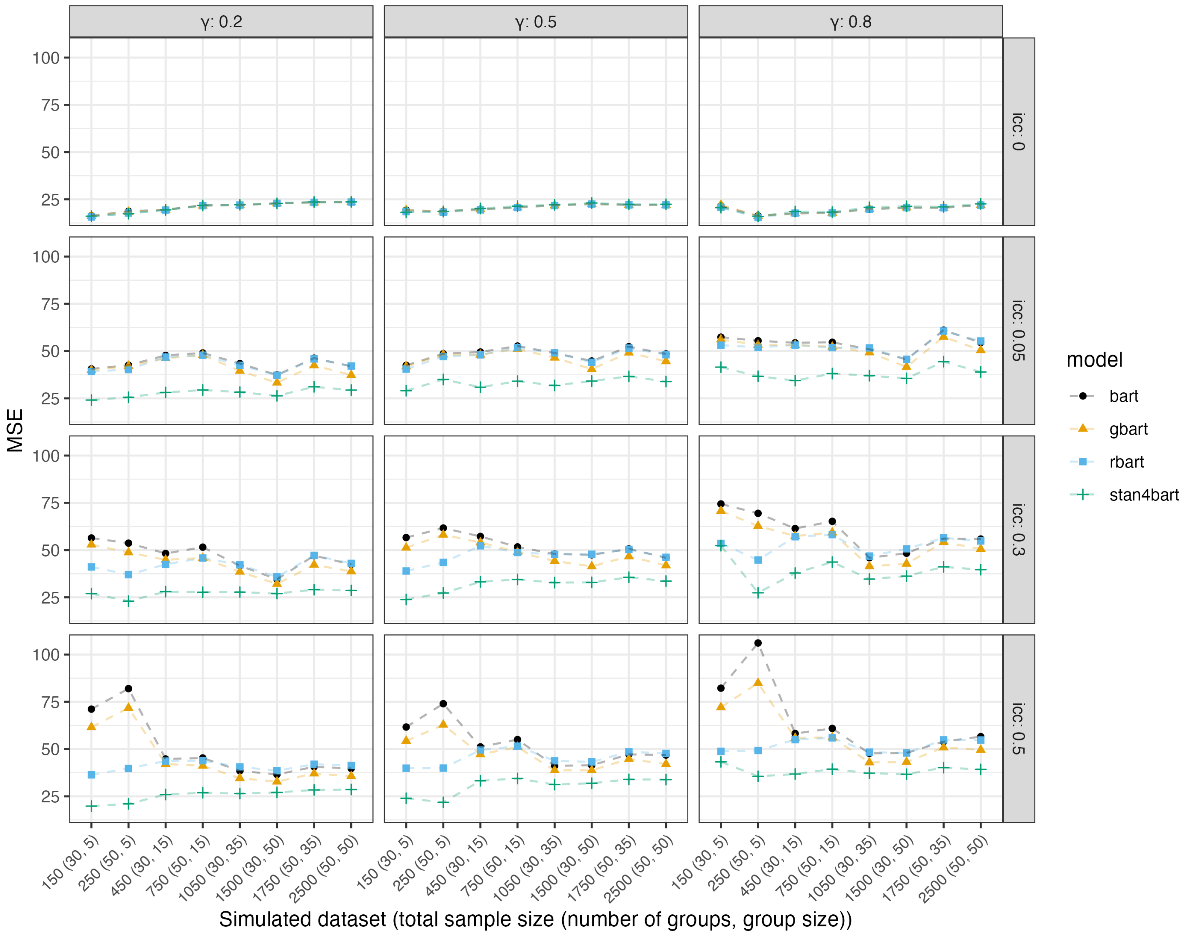


Figure 2 shows the average Mean Squared Error (MSE) for all models, datasets, *ICC* values, and within-group effect sizes (γ). Figure 2 shows that when the $ICC = 0$ the models perform well and almost exactly the same. When increasing the *ICC*, we can start to see a divide in the performance of the models. When $ICC = .05$ the models `bart`, `gbart` and `rbart` perform similarly. Increasing the *ICC* to .3 or .5, the performance of the models seperates when the dataset is small: `bart` now has the highest MSE with `gbart` performing slightly better. `rbart` performs better than `bart` and `gbart`, but when the datasets increase in size, it performs similar to them. `stan4bart` consistently outperforms the other three models in terms of MSE for all *ICC* and γ values.

Figure 2: Mean Squared Error (MSE) of the estimates for all simulated datasets over six simulations with ICC values in the rows and within group effect sizes in the columns for four models: single-level BART (`bart`), single-level BART with group dummies (`gbart`), random intercept multilevel BART (`rbart`) and random intercept random slope multilevel BART (`stan4bart`). The x-axis denotes the total sample size (number of groups, group size).



4 Discussion

In this research report I have investigated the performance of different BART models in terms of relative bias and MSE of the estimates. I considered four different models: a single-level BART model, a single-level BART model including a group-dummy, a multilevel BART model including a random intercept, and a multilevel BART model combining Stan and BART. The results indicate that the `stan4bart` model performs best out of the four models: it shows to lowest relative bias as well as the lowest MSE. Meaning that, possibly, accouting for more multilevel structure in the model improves it in terms of relative bias and MSE. These results agree with Dorie et al. (2022), who found that the `stan4bart` algorithm performed better in terms of coverage of the population values and RMSE compared to single-level BART models, BART models including a random intercept, Bayesian Causal Forests (BCF), and parametric MLMs.

However, this report has a few limitations. Since I only simulated 6 data sets per scenario to reduce computational time, the results might not be fully representative of the performance of the models. Furthermore, I only compared single- and multilevel BART models. Future research could Bayesian Causal Forests (BCF) and parametric MLMs as well. Lastly, I only considered the relative bias and MSE of the predictions, but did not consider the relative bias and MSE of the estimated parameters to visualize where the bias is present in the models, which could be interesting in evaluating the performance of the models.

Building on these results, I will implement the `stan4bart` as an imputation method within the package `MICE` (Buuren and Groothuis-Oudshoorn, 2011) in my thesis. I will compare the performance of the `stan4bart` model to other imputation methods: `2l.pmm, 2l.lmer, 2l.pan, 2l.jomo, rf` and single-level `pmm` and complete case analysis in the R-package `MICE` (Buuren and Groothuis-Oudshoorn, 2011). They will be evaluated in terms of relative bias, modeled variance, and the 95% confidence interval coverage of the estimates (Oberman and Vink, 2023). The simulation design will be extended to include more parameters: the missing data mechanism and amount of missingness. The simulation design is shown in table 2. For each combination of paramteres, a 1000 replicated datasets will be generated.

Table 2: Simulation design for the thesis

Parameter	**Values**
Number of clusters (j)	30, 50
Within-cluster sample size ($\mathrm{n_j}$)	5, 15, 35, 50
Intraclass Correlation (ICC)	0, .05, .3, .5
Missing data mechanism	MAR, MCAR
Amount of missingness	0%, 25%, 50%
Within-group effect size (γ)	.2, .5, .8

References

V. Audigier, I. R. White, S. Jolani, T. P. A. Debray, M. Quartagno, J. Carpenter, S. Van Buuren, and M. Resche-Rigon. Multiple Imputation for Multilevel Data with Continuous and Binary Variables. *Statistical Science*, 33(2), May 2018. ISSN 0883-4237. doi: 10.1214/18-STS646.

P. C. Austin, I. R. White, D. S. Lee, and S. Van Buuren. Missing Data in Clinical Research: A Tutorial on Multiple Imputation. *Canadian Journal of Cardiology*, 37(9):1322–1331, Sept. 2021. ISSN 0828282X. doi: 10.1016/j.cjca.2020.11.010.

J. W. Bartlett, S. R. Seaman, I. R. White, J. R. Carpenter, and for the Alzheimer's Disease Neuroimaging Initiative*. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, 24(4):462–487, Aug. 2015. ISSN 0962-2802, 1477-0334. doi: 10.1177/0962280214521348.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification And Regression Trees*. Routledge, 1 edition, 1984. ISBN 978-1-315-13947-0. doi: 10.1201/9781315139470.

L. F. Burgette and J. P. Reiter. Multiple Imputation for Missing Data via Sequential Regression Trees. *American Journal of Epidemiology*, 172(9):1070–1076, Nov. 2010. ISSN 0002-9262, 1476-6256. doi: 10.1093/aje/kwq260.

S. V. Buuren and K. Groothuis-Oudshoorn. **Mice** : Multivariate Imputation by Chained Equations in *R*. *Journal of Statistical Software*, 45(3), 2011. ISSN 1548-7660. doi: 10.18637/jss.v045.i03.

J. R. Carpenter and M. G. Kenward. *Multiple Imputation and Its Application*. Wiley, 1 edition, Jan. 2013. ISBN 978-0-470-74052-1 978-1-119-94228-3. doi: 10.1002/9781119942283.

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

H. A. Chipman, E. I. George, and R. E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), Mar. 2010. ISSN 1932-6157. doi: 10.1214/09-AOAS285.

M. Dong and A. Mitani. Multiple imputation methods for missing multilevel ordinal outcomes. *BMC Medical Research Methodology*, 23(1):112, May 2023. ISSN 1471-2288. doi: 10.1186/s12874-023-01909-5.

V. Dorie. Dbarts: Discrete bayesian additive regression trees sampler, 2023a.

V. Dorie. *Stan4bart: Bayesian Additive Regression Trees with Stan-Sampled Parametric Extensions*, 2023b.

V. Dorie, G. Perrett, J. L. Hill, and B. Goodrich. Stan and BART for Causal Inference: Estimating Heterogeneous Treatment Effects Using the Power of Stan and the Flexibility of Machine Learning. *Entropy*, 24(12):1782, Dec. 2022. ISSN 1099-4300. doi: 10.3390/e24121782.

C. K. Enders. Multiple imputation as a flexible tool for missing data handling in clinical research. *Behaviour Research and Therapy*, 98:4–18, Nov. 2017. ISSN 00057967. doi: 10.1016/j.brat.2016.11.008.

C. K. Enders, S. A. Mistler, and B. T. Keller. Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods*, 21(2):222–240, June 2016. ISSN 1939-1463, 1082-989X. doi: 10.1037/met0000063.

C. K. Enders, T. Hayes, and H. Du. 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, Sept. 2018a. ISSN 0027-3171, 1532-7906. doi: 10.1080/00273171.2018.1477040.

C. K. Enders, B. T. Keller, and R. Levy. A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods*, 23(2):298–317, June 2018b. ISSN 1939-1463, 1082-989X. doi: 10.1037/met0000148.

C. K. Enders, H. Du, and B. T. Keller. A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. *Psychological Methods*, 25(1): 88–112, Feb. 2020. ISSN 1939-1463, 1082-989X. doi: 10.1037/met0000228.

S. Grund, O. Lüdtke, and A. Robitzsch. Multiple imputation of missing covariate values in multilevel models with random slopes: A cautionary note. *Behavior Research Methods*, 48(2):640–649, June 2016. ISSN 1554-3528. doi: 10.3758/s13428-015-0590-3.

S. Grund, O. Lüdtke, and A. Robitzsch. Multiple Imputation of Missing Data for Multilevel Models: Simulations and Recommendations. *Organizational Research Methods*, 21(1):111–149, Jan. 2018a. ISSN 1094-4281, 1552-7425. doi: 10.1177/1094428117703686.

S. Grund, O. Lüdtke, and A. Robitzsch. 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, June 2018b. ISSN 1076-9986, 1935-1054. doi: 10.3102/1076998617738087.

S. Grund, O. Lüdtke, and A. Robitzsch. Multiple imputation of missing data in multilevel models with the R package mdmb: A flexible sequential modeling approach. *Behavior Research Methods*, 53(6): 2631–2649, May 2021. ISSN 1554-3528. doi: 10.3758/s13428-020-01530-0.

M. Gulliford, G. Adams, O. Ukoumunne, R. Latinovic, S. Chinn, and M. Campbell. Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. *Journal of Clinical Epidemiology*, 58(3):246–251, Mar. 2005. ISSN 08954356. doi: 10.1016/j.jclinepi.2004.08.012.

M. C. Gulliford, O. C. Ukoumunne, and S. Chinn. 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, May 1999. ISSN 0002-9262, 1476-6256. doi: 10.1093/oxfordjournals.aje.a009904.

T. J. Hastie, editor. *Statistical Models in S*. Routledge, 1st edition, 2017. ISBN 978-1-351-41422-7.

J. Hill, A. Linero, and J. Murray. Bayesian Additive Regression Trees: A Review and Look Forward. *Annual Review of Statistics and Its Application*, 7(1):251–278, Mar. 2020. ISSN 2326-8298, 2326-831X. doi: 10.1146/annurev-statistics-031219-041110.

J. Hox and J. K. Roberts, editors. *Handbook of Advanced Multilevel Analysis*. Routledge, 0 edition, Jan. 2011. ISBN 978-1-136-95127-5. doi: 10.4324/9780203848852.

J. J. Hox, M. Moerbeek, and R. Van De Schoot. *Multilevel Analysis: Techniques and Applications*. Routledge, Third edition. — New York, NY : Routledge, 2017. —, 3 edition, Sept. 2017. ISBN 978-1-315-65098-2. doi: 10.4324/9781315650982.

R. A. Hughes, I. R. White, S. R. Seaman, J. R. Carpenter, K. Tilling, and J. A. Sterne. Joint modelling rationale for chained equations. *BMC Medical Research Methodology*, 14(1):28, Dec. 2014. ISSN 1471-2288. doi: 10.1186/1471-2288-14-28.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics. Springer US, New York, NY, 2021. ISBN 978-1-07-161417-4 978-1-07-161418-1. doi: 10.1007/978-1-0716-1418-1.

H. Kang. The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 64(5):402, 2013. ISSN 2005-6419, 2005-7563. doi: 10.4097/kjae.2013.64.5.402.

D. Lee, B. Carpenter, P. Li, M. Morris, M. Betancourt, Maverickg, M. Brubaker, R. Trangucci, M. Inacio, A. Kucukelbir, S. Buildbot, Bgoodri, Seantalts, J. Arnold, D. Tran, M. Hoffman, C. Margossian, M. Modrák, A. Adler, K. Sakrejda, A. Stukalov, M. Lawrence, R. J. Goedman, K. S. Van Horn, A. Vehtari, J. Gabry, J. S. Casallas, and B. Bales. Stan-dev/stan: V2.17.1. Zenodo, Dec. 2017.

S. Lin and W. Luo. A New Multilevel CART Algorithm for Multilevel Data with Binary Outcomes. *Multivariate Behavioral Research*, 54(4):578–592, July 2019. ISSN 0027-3171, 1532-7906. doi: 10.1080/00273171.2018.1552555.

O. Lüdtke, A. Robitzsch, and S. Grund. Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods*, 22(1):141–165, Mar. 2017. ISSN 1939-1463, 1082-989X. doi: 10.1037/met0000096.

X.-L. Meng. Multiple-imputation inferences with uncongenial sources of input. *Statistical science*, pages 538–558, 1994.

S. A. Mistler and C. K. Enders. A Comparison of Joint Model and Fully Conditional Specification Imputation for Multilevel Missing Data. *Journal of Educational and Behavioral Statistics*, 42(4): 432–466, Aug. 2017. ISSN 1076-9986, 1935-1054. doi: 10.3102/1076998617690869.

T. P. Morris, I. R. White, and M. J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102, May 2019. ISSN 0277-6715, 1097-0258. doi: 10.1002/sim.8086.

D. M. Murray and J. L. Blitstein. Methods To Reduce The Impact Of Intraclass Correlation In Group-Randomized Trials. *Evaluation Review*, 27(1):79–103, Feb. 2003. ISSN 0193-841X, 1552-3926. doi: 10.1177/0193841X02239019.

H. I. Oberman and G. Vink. Toward a standardized evaluation of imputation methodology. *Biometrical Journal*, page 2200107, Mar. 2023. ISSN 0323-3847, 1521-4036. doi: 10.1002/bimj.202200107.

M. Quartagno and J. R. Carpenter. Substantive model compatible multilevel multiple imputation: A joint modeling approach. *Statistics in Medicine*, 41(25):5000–5015, Nov. 2022. ISSN 0277-6715, 1097-0258. doi: 10.1002/sim.9549.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2023.

M. Resche-Rigon and I. R. White. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research*, 27(6):1634–1649, June 2018. ISSN 0962-2802, 1477-0334. doi: 10.1177/0962280216666564.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/63.3.581.

D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York, 1987. ISBN 978-0-470-31669-6.

M. Salditt, S. Humberg, and S. Nestler. Gradient Tree Boosting for Hierarchical Data. *Multivariate Behavioral Research*, pages 1–27, Jan. 2023. ISSN 0027-3171, 1532-7906. doi: 10.1080/00273171.2022.2146638.

G. Shieh. A comparison of two indices for the intraclass correlation coefficient. *Behavior Research Methods*, 44(4):1212–1223, Dec. 2012. ISSN 1554-3528. doi: 10.3758/s13428-012-0188-y.

G. C. Silva and R. Gutman. Multiple imputation procedures for estimating causal effects with multiple treatments with application to the comparison of healthcare providers. *Statistics in Medicine*, 41(1): 208–226, Jan. 2022. ISSN 0277-6715, 1097-0258. doi: 10.1002/sim.9231.

M. Taljaard, A. Donner, and N. Klar. Imputation Strategies for Missing Continuous Outcomes in Cluster Randomized Trials. *Biometrical Journal*, 50(3):329–345, June 2008. ISSN 0323-3847, 1521-4036. doi: 10.1002/bimj.200710423.

Y. V. Tan, C. A. C. Flannagan, and M. R. Elliott. Predicting human-driving behavior to help driverless vehicles drive: Random intercept Bayesian Additive Regression Trees. 2016. doi: 10.48550/ARXIV.1609.07464.

S. Van Buuren. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3):219–242, June 2007. ISSN 0962-2802, 1477-0334. doi: 10.1177/0962280206074463.

S. van Buuren. *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Interdisciplinary Statistics Series. CRC Press, Taylor & Francis Group, Boca Raton London New York, second edition edition, 2018. ISBN 978-1-138-58831-8.

J. Wagner, B. T. West, M. R. Elliott, and S. Coffey. 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, Dec. 2020. ISSN 2001-7367. doi: 10.2478/jos-2020-0043.

A. K. Waljee, A. Mukherjee, A. G. Singal, Y. Zhang, J. Warren, U. Balis, J. Marrero, J. Zhu, and P. D. Higgins. Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, 3(8): e002847, Aug. 2013. ISSN 2044-6055, 2044-6055. doi: 10.1136/bmjopen-2013-002847.

B. Wundervald, A. Parnell, and K. Domijan. Hierarchical Embedded Bayesian Additive Regression Trees. 2022. doi: 10.48550/ARXIV.2204.07207.

D. Xu, M. J. Daniels, and A. G. Winterstein. Sequential BART for imputation of missing covariates. *Biostatistics*, 17(3):589–602, July 2016. ISSN 1465-4644, 1468-4357. doi: 10.1093/biostatistics/kxw009.