**Bayesian Born-Again Trees for Multilevel Data**

*Improving the Accuracy of Generalized Linear Mixed Model Trees Using Multilevel Bayesian Additive Regression Trees*

**Master Thesis Research Methods & Statistics**

**Faculty Social Sciences**

**By: David Eleveld 1741225**

**Thesis Supervisor: Dr. M. Fokkema**

**18 May 2023**

**Abstract**

To improve the accuracy and interpretability of the Generalized Linear Mixed Model (GLMM) Tree model, a novel version of the Born-Again (BA) method was tested using 6 real datasets and one artificial dataset. This method uses generated data (of size *Ngen*) based on the training data, and the Posterior Predictive Distribution (PPD) of a highly accurate, poorly interpretable Bayesian Additive Regression Tree (BART) model to train a new PPD-based BA GLMM tree. When generated sample size *Ngen* is 1 time the size of the original dataset, BA trees were more accurate or similarly accurate as GLMM trees. *Ngen* can be increased for more interpretability at the cost of a small decrease in accuracy. When generating data for traditional BA trees, setting permutation parameter *palt = 0* provides best results in terms of accuracy and interpretability. PPD-based BA trees were more accurate than traditional BA trees on five out of seven datasets, was outperformed on one, and was equivalent on one, suggesting that PPD-based BA trees generally provide more accurate results than traditional BA trees.

**Introduction**

In the field of predictive modeling, the relationship between accuracy and interpretability is often seen as a trade-off. Black box models like for example, a tree ensemble, often provides highly accurate results but the final predictive model is too complex for human understanding (James et al., 2013). These prediction models can be used to find relations in complex data structures to make predictions on new, unseen data. They could for example be used to process a large number of chemical compounds in a patient’s blood and predict if a disease is present, analyze a company’s tax data to predict if it’s committing fraud, or predict what movie you would most likely want to see next based on your previously watched movies. While the answers provided by these models are often accurate, it is hard to interpret how these answers came to be. A high-accuracy, low-interpretability model like this will from now on be referred to as a *black box* model. They

As the usage of black box models has become more commonplace it is increasingly important to understand how these models make their predictions (Bodó, & Janssen, 2021; Varshney & Alemzadeh, 2017). As pointed out by Rudin (2019): “the lack of transparency and accountability of predictive models can have (and has already had) severe consequences”. In the USA, for example, people have been incorrectly denied parole, or incorrectly released on bail (Wexler, 2017). Recently, the Dutch government came under scrutiny in an event called *Dutch childcare benefits scandal* (Dutch: *Toeslagenaffaire*), where, between 2013 and 2019, an estimated 26,000 parents were wrongly accused of making fraudulent benefit claims, (van Bruxvoort & van Keulen, 2021; Huisman, 2020). The European Digital Rights (EDRi) organization, a network of more than 40 European civil organizations, advocate for more regulation and transparency on artificial intelligence (AI) decision-making systems in the European Union, especially in the area of law enforcement and criminal justice. With the recent developments in language models like chatGPT by OpenAI (2022), it seems that the growth of the importance of algorithms in society is not slowing down (Future of Life Institute, n.d.). While highly accurate black box models are very useful, the call for models with higher interpretability is valid and cannot be ignored. For prediction models, the road to interpretability has two main directions: post-hoc explanation tools, and inherently interpretable glass box models.

In practice, post-hoc explanation techniques that attempt to interpret black box models are used often. Local Interpretable Model-agnostic Explanations (LIME; Ribeiro, et al., 2016) and Shapley additive explanations (SHAP; Lundberg & Lee, 2017) are popular examples. Kaur et al. (2020) however, point out that while these tools are useful in uncovering issues with datasets or models, they are often over-trusted and misused for interpretation. Rudin (2019) argues that instead of trying to explain black box models, the way forward is to create glass box models that are not only accurate, but also inherently interpretable. She disregards the trade-off between accuracy and interpretability as a myth. She argues that in numerous domains, highly interpretable glass box models exist that have accuracy close to, equal, or higher than black box models. Furthermore, she notes that a small increase in accuracy is often far less important than the ability to interpret results, as interpretability will lead to better data processing in next iteration, leading to better overall accuracy in the long-term.

Nevertheless, researchers should still aim to improve the accuracy of interpretable glass box models. In the current thesis, I will attempt to use the predictions generated by an accurate black box model (Bayesian tree ensemble, described below) to improve the accuracy of a less accurate glass box model, namely, the Generalized Linear Mixed-Model (GLMM) tree model. As explained below, this glass box model is easily interpretable, as it results in a single decision tree. It is a multilevel tree model, which has an advantage over a traditional single tree model, in that it is able to process multilevel data structures, leading to higher accuracy on multilevel data structures (Fokkema et al., 2018; Hajjem et al., 2017; Sela & Simonoff, 2012). GLMM trees however, fall short compared to the accuracy of black box models. By using the *born-again* approach described below, I will attempt to improve the accuracy of the GLMM tree glass box model, while maintaining high interpretability.

**Born-Again approach**

Breiman and Shang (1996) introduced a method to use a highly accurate black box model to improve the accuracy of an interpretable glass box model: the Born-Again (BA) tree algorithm. The BA approach aims to create a single tree model, which has accuracy close to a black box model, but is easier to interpret and apply by human decision makers.

Breiman and Shang’s (1996) BA tree algorithm follows the following four steps:

1. A black box model is fitted on the original predictor variables (X) to predict original outcome variable (Y).
2. Based on X, a large number of observations are artificially generated by resampling rows from X followed by columnwise permutation to create X­gen.
3. The black box model is applied to X­gen to obtain predictions Ygen for the observations generated in step 2.
4. A single tree is fitted on X­gen to predict Ygen resulting in a BA tree.

Whether predictive accuracy of the BA tree fitted in step 4) is improved, while its interpretability is retained, largely depends on three factors that are potentially controllable by the user:

* The accuracy of the black box model (steps 1 and 3)
* Whether the data generation approach yields a realistic set of s artificial observations Xgen (step 2)
* The interpretability and accuracy of the glass box model fitted on the artificial dataset (step 4)

Breiman and Shang (1996) found that while the accuracy of their glass box model increased, the interpretability worsened significantly. Below, I describe how the above factors are studied in the current thesis.

***The glass box model: GLMM tree***

Breiman and Shang (1996) used the CART algorithm to create the BA Tree. While a single CART is useful for decision making, as it functions as a flow-chart, the accuracy of the tree tends to be lower than that of black box models (Fernández-Delgado et al., 2014; Gacto et al., 2019; Zhang et al., 2017). Another limitation of the CART algorithm is that is does not account for multilevel data structures, which are common in psychological research.

When observations are nested in levels (i.e., groups), the data has a multilevel structure. These structures require multilevel modelling to account for the correlated nature of observations within higher-level units. Examples include multi-centre clinical trials with patients nested within treatment centres, longitudinal studies with repeated measurements nested within patients, and educational studies with students nested within classrooms. Multilevel models estimate random effects to account for variance attributable to grouping of observations, improving accuracy of predictions and standard errors, and reducing type-I and -II errors.

Tree models can be modified to account for multilevel data structures which have been shown to yield more accurate, as well as less complex trees than gained from the CART algorithm (Hajjem et al., 2017; Sela & Simonoff, 2012). One such multilevel tree model is the GLMM tree (Fokkema et al., 2018). The GLMM tree model builds on the Generalized Linear Model (GLM) tree, where every node in the tree is associated with a single GLM model (Zeileis et al., 2008) given by:

Where *i* indicates an individual observation, is represented as the GLM model, and is an error term, assumed to follow a normal distribution with mean 0 and variance . Errors are assumed identically and independently distributed, but this assumption is often violated when individual observations are nested within groups, for example, in longitudinal or multilevel data structures. This can be accounted for by inclusion of a random intercept:

Where reflects the intercept of the group of which individual observation *i* is part, and is assumed to follow a normal distribution with mean 0 and variance . Through the born-again approach, the current study aims to improve the accuracy of GLMM tree using the predictions of a black box method. Due to the multilevel nature of the data, the black box method needs to account for multilevel data structures as well.

***The black box model: M-BART***

Previous studies on BA trees, found accurate results with boosted tree ensembles (Breiman & Shang, 1996) or neural networks (Craven & Shavlik, 1995) as the black box. When multiple trees are combined to form tree ensembles, as in the BART, Random-Forest, Bagging, or Boosting algorithms, the predictive accuracy is improved and the risk of over-fitting is reduced, at the cost of dramatically decreasing interpretability. Random forests, for example, use bootstrap aggregating (bagging) to build many decision trees and then average their predictions (Breiman, 2001) and have been shown to be effective in a wide range of applications, including image classification and bioinformatics (Caruana & Niculescu-Mizil, 2006).

In the current study, I will focus on Bayesian Additive Regression Trees (BART) ensembles. The basic BART model uses a Bayesian framework to learn the ensemble, using prior distributions on the tree structure and the model parameters. Additionally, BART models provide predictions in the form of a posterior probability distribution (PPD), rather than deterministic point estimates (Chipman, et al., 2010). BART based ensembles have been shown to outperform Boosting, Neural Networks, Random Forests, and a host of other machine learning algorithms in terms of predictive accuracy (Chipman et al., 2010; Dorie et al., 2019). The BART model is given by:

Where *i* indicates an individual observation, is represented as a sum of regression trees, and is an error term, assumed to follow a normal distribution with mean = 0 and variance = . Though the BART model assumes independence among observations, this assumption is often violated when individual observations are nested within groups. The BART model can be altered to account for multilevel data structures (Sparapani et al., 2021; Tan, 2018; Wundervald et al., 2022) to fit Multilevel BART (M-BART) ensembles by inclusion of a random intercept:

Where reflects the intercept of the group of which individual observation *i* is part, and is assumed to follow a normal distribution with mean = 0 and variance = . A more detailed description of the estimation of (M-)BART models is given in the Method section. In the current thesis I will be using this M-BART model as the black box model, allowing the BA tree to be a multilevel tree as well. The PPD resulting from the M-BART models will be used in the data generation method.

***Data Generation Method***

*Smearing*

In order for the BA approach to improve accuracy, it requires a large sample size to train the tree (Breiman & Shang 1996). Breiman and Shang therefore suggest manufacturing data to increase the sample size. They sample from the training data X to create a new dataset X­genwith size *Ngen* times the original dataset. They also apply column-wise permutation (i.e. smearing) when generating X­gen. Smearing requires the user to a-priori specify the probability *palt* that a value in the data matrix will be permuted. Breiman and Shang’s (1996) method will be called the smearing-based BA tree approach in the current thesis. When using an M-BART model as the black box, the data generation of this method -as earlier described in step 2 and 3 in the BA tree algorithm- repeats the following steps:

2a. Sampling: A row *i* (*i* = 1, … , *N)* is randomly selected from X.

2b. Permutation: for every predictor *j* (*j* = 1, … , *P)*, a draw is taken from a Bernoulli distribution with *p* = *palt* and *n* = 1, which determines whether the current value *xij* is retained, or replaced by a random draw from x***j***.

2c. Step 1. and step 2. are repeated *Ngen* \* *N* times to create Xgen.

3a. The black box model fitted on the original training data (X, Y) is used to create a PPD for each row in Xgen.

3b. The mean of the PPD for each row in Xgen is taken in order to generate Ygen.

Where *N* is the number of participants, *P* is the number of predictors, *palt* is a tuning parameter between 0 and 1, representing the probability that a given value is replaced by a random draw, and *Ngen* is the number of times the generated data is larger than the original data. Note that for *palt* = 1, we are performing complete permutation and when *palt* = 0, we are sampling random participants as is. *palt* can be optimized for more accurate results. Breiman and Shang (1996) use *palt* = 0.25 and *palt* = 0.50 in their study, as they found that those values yielded most accurate results. They do not go in to further detail about the exact relation between *palt* on and the accuracy or the interpretability. In the current thesis, I will test different values of *palt* (*palt* = 0, *palt* = 0.25, *palt* = 0.50) to get a better understanding of its effect. Furthermore, I will introduce a novel data generation method: PPD sampling.The values of *Ngen*and *palt* influence the accuracy and interpretability of the final glass box model and will be studied in the current thesis.

*PPD sampling*

The use of permutation may lead to unrealistic data patterns and has been criticized (Hooker et al., 2021). For this reason, PPD sampling can be utilized to create Xgen­ and Ygen­that more closely resemble the original data. This method will be called the PPD-based BA tree approach in the current thesis. The PPD sampling algorithm -as earlier described in step 2 and 3 in the BA tree algorithm- repeats the following steps:

2. Sampling: A row *i* (*i* = 1, … , *N)* is randomly selected from X. No permutation is applied (this is equivalent to setting *palt* = 0).

3a. A random prediction is sampled from the PPD of the corresponding row from step 2a.

3b. Step 2 and step 3a. are repeated for *Ngen* number of times the size of X to create Xgen and Ygen.

Data generated in this manner should better retain the intercorrelation between predictor variables present original data and thus lead to a more accurate BA tree. As stated before, this is possible when using a M-BART or BART model as a black box.

***Research Questions and Hypotheses***

The main goal of the current paper is to study how BA trees perform compared to “regular” GLMM trees. The first two research questions are thus:

**RQ1:** Does a PPD-based BA approach improve the predictive accuracy of GLMM trees?

**RQ2:** Does a PPD-based BA approach improve the interpretability of GLMM trees?

I will also compare PPD-based BA tree approach to smearing-based BA tree approach, as proposed by Breiman and Shang (1996). This leads us to the third research question:

**RQ3:** In terms of predictive accuracy and interpretability, does a PPD-based BA approach outperform a smearing-based approach?

Finally, I will study the effect of *N­gen* and *palt* on interpretability and accuracy. The final research question is thus:

**RQ4:** In terms of predictive accuracy and interpretability, what are the effects of *Ngen* and *palt* on the performance of BA trees?

**Methods**

***Datasets***

The research questions were addressed on four different benchmark datasets described below. Only the Safety dataset was artificially generated. The datasets all have a multilevel structure.

*AIDS Clinical Trials Group Study 175 (ACT)*

This dataset created by Hammer et al. (1996), measures human immunodeficiency virus type 1 (HIV-1) infected patients and measures the amount of CD4 T cells present in the blood as outcome measure. A number of medical drugs, demographics and confounding variables, like prior antiviral therapy and CD4 T cell count 30 days before treatment, are present to predict CD4 T cell count. Multiple measurements were taken over time within patients, creating a multilevel structure. There are 6417 observations from 2139 patients 24 variables. A sample of *N* = 1000 was used for the current study, containing repeated observations on 377 patients.

*Safety*

Hox et al. (2017) released an artificially generated dataset comprising of *N* = 1000 observations containing information from a survey conducted on 100 streets. The survey randomly sampled 10 individuals on each street to report the frequency of feeling unsafe while walking along that street. The responses were categorized as never, sometimes, and often. The dataset includes predictor variables such as age, gender, an economic index (standardized Z-score), and a rating of street crowdedness (7-point scale) to predict how frequently people feel unsafe. The data is organized in a multilevel structure with individuals nested within streets.

*Marriage*

This is a simplified version of a dataset created by Lax and Phillips (2009) where support for gay marriage in the USA is predicted based on age, education, gender, Christianity and political vote. 6525 Participants are nested within the 49 states they live in. A sample of *N* = 1000 containing all 49 states was taken.

*Early Childhood Longitudinal Study Kindergarten Class of 2010-11 (ECLS-K*)

Mulligan et al. (2016) collected data about kindergarteners over the course of five years. The data includes information on selected child and family characteristics, such as poverty status, parental education, family type, and primary home language. This is used to predict the children's knowledge and skills in math, reading and science, in separate datasets. As multiple measurements are taken within the children, we get a multilevel structure. For the math and reading datasets, samples of *N* = 1000 containing 200 children were taken. For the science dataset a sample of *N* = 999 containing 333 children were taken, as the science test was only measured at three ages.

We also transformed the dataset to include school wide statistics. We calculated the percentage of males, the mean Social Economic Score, the mean race, and mean certification per school. This new data is called ECLSK-School was used to predict the reading score of the individual children, nested in schools, and is not longitudinal. From this dataset a sample of *N* = 999 containing 146 schools was taken.

***BART model***

BART models are created by first creating K trees (by default, K = 200) with a single root node. The mean of this node is sampled from a prior, which is a normal distribution where 95% of the means lie between y­min and ymax (when using default shrinkage parameter *k* = 2). The model then calculates the residuals and goes into the second iteration, where the trees (splits and means) are randomly permuted based on the specified priors (or tuning parameters). A tree can grow extra nodes or prune off nodes based on a beta distribution prior. By default, = 0.95 and = 2 is set so that most trees end up having size 2 or 3 (55% and 28% respectively). The number of iterations is set to 1000 by default. Every iteration, trees that improve the fit are favored. The end result is not a single *ŷi* per observation, but a distribution of predicted values for every iteration. This distribution is called the Posterior Probability Distribution (PPD). The first 200 iterations are usually very inaccurate and are thus removed; these are called the burn-in samples. A more detailed explanation of the BART model and estimation is provided by Chipman et al. (2010).

All parameters can be optimized by cross-validation, but this is computationally very demanding. As the default parameters are usually effective, it is recommended to use them instead (Chipman et al., 2010; Sparapani et al., 2021). Using the default parameters is also a more data-centric approach that is less likely to lead to overfitting (Carnegie, 2020). For these reasons I will be using the default parameters to create BART models in the current thesis.

***Procedures***

First, samples of *N* = 1000 (or 999, see *Datasets* section above) were taken from each dataset to reduce computational burden. The data was then randomly split in ten equal folds to use in 10-fold Cross Validation (10-CV). When a fold was assigned as test set, it was checked whether all factor levels in the test set were also present in the training set. If not, the data was moved to the training set. Next, three algorithms were applied to the data:

1. GLMM tree.
2. M-BART.
3. BA tree.

To create the BA tree, artificial datasets were created according to the BA-tree algorithm described in the Introduction. That is, a set of artificial observations Xgen was generated using resampling and permutation. The response Ygen was sampled from the PPD of the M-BART ensemble (in the case of PPD sampling), or the mean was taken from the PPD of the M-BART model (in the case of smearing). The GLMM tree algorithm was then applied to this artificial dataset (Xgen, Ygen). The following parameters used for resampling and permutation (*Ngen* and *palt*) were varied:

* *Ngen*: The sample size of the artificial dataset, with *Ngen* =1, 5 or 10 times the size of the original dataset.
* *palt*: A parameter of the smearing data-generation method, with *palt* = 0.00, 0.25, 0.50.

This gives us three BA tree models: MPPD: *N* = 1, MPPD: *N* = 5, MPPD: *N* = 10, and different smearing-based GLMM tree models: Msmear: *N* = 1, palt = 0, Msmear: *N* = 1, palt = 0.25, …, Msmear: *N* = 10, palt = 0.50.

*Ngen*

Breiman and Shang (1996) state that a larger size of *Ngen*could lead to more accurate trees. In their own preliminary experiments, they generated data that was a factor 10 to 20 times the original training set (i.e., *Ngen ­*= 10 to *Ngen ­*= 20). They found trees that were more accurate, but so large that they were not interpretable. For their main experiment, they used *Ngen ­*= 1. They further stated that their results were not sensitive to the value of *Ngen.* In the current thesis,I will attempt to replicate these results by varying *Ngen* between 1, 5, and 10, and studying the result on interpretability and accuracy. The generated observations were weighted to compensate for artificially inflated sample size. Namely, weights were set equal to 1/ *Ngen*.

***Assessment of performance***

The models were compared in terms of interpretability and accuracy. These were measured as tree size, and MSE respectively:

*Tree size:*

Tree size is defined as the number of splits in the tree. Tree size will not be computed for the BART ensembles as they contain a very large number of trees.

*MSE*

To compare the predictive accuracy of the models, the mean squared error (MSE) will be computed on test observations. This is divided by the variance of the dependent variable Y to obtain a coefficient of determination *R2* (Consonni et. al, 2010): *R2* measures the proportion of variance of Y explained by the predictions of the model. Note that *R2* evaluated in this manner on test data, can reach below 0 when the MSE is higher than the test sample variance of Y.

*Cross validation*

The 10-CV was repeated 10 times, resulting in 10 × 10 = 100 different values of MSE and tree size for every model on every sample. For time-series data, random persons were sampled instead of random observations, so that every level had enough data.

*Comparisons*

To account for the correlated nature of the *R2* of the models within folds within every fold, repeated-measure ANOVAs were used to detect differences between *R2*sin every dataset. For post-hoc testing, Bonferroni corrected paired-samples t-tests were used. When the *R2*s and tree sizes of every dataset were combined to test the effects of *p­alt* and *Ngen*, Friedman’s non-parametric repeated-measures ANOVAs were used to allow for the analysis of unrelated data between datasets. Effect size is calculated to further interpret the results, using Landis & Koch’s (1977) rule of thumb: 0.00 ≤ < 0.20 indicating a slight effect, 0.20 ≤ < 0.40 a fair effect, 0.40 ≤ < 0.60 a moderate effect, 0.60 ≤ < 0.80 a substantial effect and ≥ 0.80 an almost perfect effect.

***Software***

All analyses were performed using R Statistical Software (v4.1.1; R Core Team, 2021). The BART models were calculated using the rbart\_vi function in the dbarts package (v0.9-20; Dorie, 2021). The GLMM trees were made using the lmertree function in the glmertree package (v0.2-0; Fokkema et al., 2018).

**Results**

***Predictive Accuracy***

To statistically test whether it is possible for BA trees to improve the predictive accuracy of GLMM trees, the mean *R2* of the GLMM tree was compared to the mean *R2* of the most accurate PPD-based BA tree and the mean *R2* of the most accurate smearing-based BA tree for every dataset. These ANOVAs revealed significant differences in *R2* on every dataset (*p* > .001 for every ANOVA). Post-hoc planned paired-samples t-tests revealed that the PPD-based BA trees had significantly higher mean *R2* than GLMM trees on the ACT (*p* < .001), ECLSK math (*p* < .001), ECLSK reading (*p* < .001), ECLSK science (*p* < .001), ECLSK school (*p* < .001), and Safety (*p* < .001) datasets. There was no significant difference on the Marriage dataset. Smearing-based BA trees had significantly higher mean *R2* than GLMM trees on the ACT (*p* < .001), ECLSK math (*p* = .014), ECLSK science (*p* < .001), and Safety (*p* < .001) datasets, but not on the ECLSK reading, ECLSK school and Marriage datasets.

Using paired-samples t-tests, we tested whether PPD-based BA trees improve the predictive accuracy of smearing-based BA trees. This was the case on the ECLSK math (*p* < .001), ECLSK reading (*p* = .004), and the ECLSK science (*p* < .001), ECLSK school (*p* < .001), and Safety (*p* < .001) datasets. Smearing-based BA trees outperformed PPD-based BA trees on the Marriage dataset (*p* < .001), but no significant difference was found on the ACT dataset.

|  |
| --- |
| **Figure 1** |
| *Violin plots of R2 for every model on ACT and ECLSK science datasets* |
|  |
| *Note: R2* over tenrepeats of 10-CV (100 folds). Black dots indicate mean *R2.* Red violins indicates GLMM tree, blue violins indicates M-BART, yellow violins indicates the most accurate PPD-based BA tree and green violins indicate the most accurate smearing-based BA tree. |

*R2*s and tree sizes for the ACT and ECLSK science datasets are shown in Table 1 and plotted in Figure 1, as the results of these datasets are representative of the results in the other datasets. The results on the other datasets can be found in Appendix A and Appendix B. The conclusions drawn from the results depend on whether or not an accurate M-BART model was found. When the predictive accuracy of M-BART is very low (*R2* is close to, or equal to 0), as is the case for the Marriage and ATC datasets, the glass box models (e.g., GLMM tree, BA tree) all have similarly low predictive accuracy. The predictive accuracy in *R2* for the ACT dataset is displayed in Figure 1 (left). When the predictive accuracy of the black box model is high(er), the glass box models also have higher predictive accuracy. In Figure 1 (right), the predictive accuracy of every model trained on the ECLSK science dataset is displayed.

|  |
| --- |
| **Table 1** |
| *R2 and tree size for ACT and ECLSK science datasets* |
|  |
| *Note.* For PPD- and Smearing-based BA trees, only the model with the highest R2 is displayed here. For the complete results see Appendix A and Appendix B. |

***Tree size***

As seen in Table 1, the tree size of the BA trees increases drastically compared to the GLMM tree (except for PPD: *Ngen* = 5). There is a direct reverse relationship between tree size and *Ngen*. That is, the higher *Ngen*, the higher the tree size. As seen in Appendix B, when choosing *Ngen* = 1 for BA trees, the tree size is always higher than for GLMM trees. When we choose *Ngen* = 10, as seen in Table 2, the tree size is significantly lower than for BA trees than for GLMM trees. The *R2* on the other hand, only increases a little. In Figure 2 we can see this relation more clearly.

|  |
| --- |
| **Table 2** |
| *R2 and tree size for ACT and ECLSK science datasets where generated data is Ngen = 10 times the size of the original dataset* |
|  |
| *Note.* For Smearing-based BA GLMM trees, only the model with the highest R2 is displayed here. For the complete results see Appendix A and Appendix B. |
| **Figure 2** |
| *Effect of N on R2 and tree size for every dataset* |
|  |
| *Note:* R2 (left) and tree size (right) over ten times 10-CV (100 folds) for all BA GLMM models on all seven datasets. Blue dotted line marked by “GLMM” marks the median R2 or tree size of GLMM models. Red dotted line marked by “BART” marks the median R2 of BART model. Red dots connected by red solid line indicate median R2 or tree size. |

As seen in Figure 2 (right), Friedman’s non-parametric repeated measures ANOVA revealed that the tree size consistently got significantly lower when *Ngen* was increased (*X2*(2)= 3070.53, *p* < .001). The effect size (W­kendall­ = .55) indicates a moderate effect. Figure 2 (left) shows that *R2* also consistently decreases when *Ngen* increases (*X2*(2)= 920.88, *p* < .001). The effect size however (W­kendall­ =.16), indicates a slight effect. This suggests that increasing *Ngen* could be useful when interpretability is important, while mostly maintaining the accuracy.

|  |
| --- |
| **Figure 3** |
| *Effect of palt on R2 and tree size for every dataset* |
|  |
| *Note:* R2 (left) and tree size (right) over ten times 10-CV (100 folds) for all BA GLMM models on all seven datasets. Blue dotted line marked by “GLMM” marks the median R2 or tree size of GLMM models. Red dotted line marked by “BART” marks the median R2 of BART model. Red dots connected by red solid line indicate median R2 or tree size. |

***palt***

For smearing-based BA trees, varying palt has a clear effect. Figure 3 (left) shows that *R2* consistently decreases when *Ngen* increases (*X2*(2)= 1159.18, *p* < .001). The effect size (*W­kendall­* = .28) indicates a fair (slight/medium) effect Figure 3 (right) reveals that *palt* = 0 consistently results in the lowest tree size (*X2*(2)= 857.65, *p* < .001) with (*W­kendall­* = .20) indicating a fair effect (Landis & Koch, 1977). Although it may appear that setting *palt* to 0 consistently yields the best results in terms of accuracy and predictability, this is not the case when *Ngen* is also set to 1. While this configuration results in the highest accuracy, it comes at the expense of reduced interpretability (see Appendix C).

**Discussion**

The results have provided evidence that BA trees improve predictive performance compared to GLMM trees (measured in *R2*). Paired-samples t-tests revealed that PPD-based BA trees obtained higher *R2* than GLMM trees on six out of seven datasets. On the Marriage dataset, there was no significant difference between the *R2* of PPD-based BA trees and GLMM trees. This is likely because the BA trees are modelled based on the predictions of the M-BART model, and the M-BART model did not find a more accurate fit than the GLMM tree on the Marriage dataset. When the black box model does not provide accuracy substantially higher than the glass-box model accurate predictions, the BA approach is less likely to yields improved accuracy. It is thus good to keep in mind that the potential for BA trees to improve accuracy increases with the accuracy of the black-box model. Further paired-samples t-tests revealed that smearing-based BA trees obtained higher *R2* than GLMM trees on four out of seven datasets: on the ACT, ECLSK math, ECLSK science, and Safety datasets, but not on the ECLSK reading, ECLSK school and Marriage datasets.

The results also provided evidence that PPD-based BA trees have higher accuracy than smearing-based BA trees on the same datasets. Paired-samples t-tests revealed that PPD-based GLMM trees obtained higher *R2* than smearing-based BA trees on five out of seven datasets. Smearing-based BA trees only outperformed PPD-based trees on the Marriage dataset. There was no significant difference on the ACT dataset. This is likely due to the fact that the Marriage and ACT datasets produced inaccurate models. This shows that PPD-based BA trees outperform smearing-based BA trees on datasets with accurate black box models.

The results showed that the tree size for BA trees increased to uninterpretable levels (often reaching tree size > 40 nodes) when using the most accurate model with *Ngen* = 1. When increasing *Ngen*, however the tree size decreases and can get lower than the GLMM tree. This increase in interpretability comes at the cost of lowering the accuracy, but this decrease is a small one. Researchers using BA tree models should thus take into account their goals and vary *Ngen* based on the importance of interpretability in their study. As stated in the introduction I believe that there is an inherent value to interpretability that can outweigh a small decrease in accuracy. Using a high value of *Ngen* could lead to a very informative tree, even if the accuracy is slightly lower.

When setting *palt* = 0, the smearing-based BA trees consistently had the highest interpretability and highest accuracy (accuracy was not statistically different on the marriage dataset). This contrasts the findings by Breiman and Shang (1996) that found *palt* = 0.50 to lead to the most accurate results. The current study thus leads to the conclusion that smearing, or permuting the values of predictor variables, should not be considered as a data-generated method for BA trees. With *palt* = 0 we are not permuting the predictor variable, leaving existing intercorrelations between predictors intact and merely resampling observations from the original data with replacement. Setting palt > 0 can yield unrealistic data patterns, leading to more inaccurate results.

**Limitations**

***Non-accurate models***

As mentioned before, the BA trees had the most accurate and interpretable results when the M-BART model was substantially more accurate than the GLMM tree model. There were no models that were able to have a better than chance fit (*R2* > 0) on the ACT and Marriage datasets. The ECLSK school and Safety datasets had models with better but still mediocre fits (*R2* < 0.25). This means that only three out of seven datasets had had M-BART models with good fits (*R2* > 0.50). Although it is interesting to see the BA trees work on real-life data which is often unpredictable, stronger conclusions could perhaps have been drawn about the mechanisms of BA trees if more datasets with better fitting M-BART models had been used.

***Observation weights***

The BA trees were trained on generated data that was weighted according to the generated sample size *Ngen*. When for example *Ngen* = 5, each observation was weighted by a factor of . This was done to prevent overfitting with large trees, as bigger sample sizes could have increased the power too much when statistically testing whether or not a new node should be made. In the current study, I did not independently vary *Ngen* and the weights to test the effects on accuracy and interpretability. It can be expected that the tree size decreases for PPD-based BA trees when *Ngen* increases, as the sampled outcome variables will more closely resemble the PPD and thus become less noisy. It is however interesting that this also happened for smearing-based BA trees. Future research could study the individual effects of varying the factors weights and *N­gen* to gain further insights on their effects.

**Conclusion**

In conclusion, it can be stated that Born-Again GLMM trees can outperform regular GLMM trees in terms of accuracy and interpretability, provided that the black box model is more accurate than the glass box model. When generated sample size *Ngen* is 1 , yielding the same sample size as for the original dataset, BA trees are more accurate or similarly accurate as GLMM trees. *Ngen* can be increased for more interpretability at the cost of a small decrease in accuracy. When using a smearing-based BA tree, setting *p­alt* = 0 provides best results in terms of accuracy and interpretability.

**References**

Bodó, B., & Janssen, H. (2021). Here Be Dragons–Maintaining Trust in the Technologized Public Sector. *Policy & Society 41*(3), 414–29

Breiman, L., & Shang, N. (1996). Born again trees. *University of California, Berkeley, Berkeley, CA, Technical Report*, *1*(2), 4.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

van Bruxvoort, X., & van Keulen, M. (2021). Framework for Assessing Ethical Aspects of Algorithms and Their Encompassing Socio-Technical System. *Applied Sciences*, *11*(23), 11187.

Carnegie, N. B., (2020, May 13). *Introduction to Bayesian Additive Regression Trees for Causal Inference* [Targeted Learning Webinar Series]. Putnam Data Sciences. <https://www.youtube.com/watch?v=9d5-3_7u5a4&t=2093s>

Caruana, R., & Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning* (pp. 161-168).

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266-298.

Consonni, V., Ballabio, D., & Todeschini, R. (2010). Evaluation of model predictive ability by external validation techniques. *Journal of chemometrics*, *24*(3‐4), 194-201.

Craven, M., & Shavlik, J. (1995). Extracting tree-structured representations of trained networks. *Advances in Neural Information Processing Systems*, *8, 24–30*.

Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, *34*(1), 43-68.

Fair Trials and European Digital Rights (EDRi), (2022, March 1). *Civil society calls on the EU to ban predictive AI systems in policing and criminal justice in the AI Act*. Retrieved from <https://edri.org/our-work/civil-society-calls-on-the-eu-to-ban-predictive-ai-systems-in-policing-and-criminal-justice-in-the-ai-act/>

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, *15*(1), 3133-3181.

Fokkema, M. (2020). Fitting Prediction Rule Ensembles with R Package pre. *Journal of Statistical Software*, *92*(12), 1–30. https://doi.org/10.18637/jss.v092.i12

Fokkema, M., Edbrooke-Childs, J., & Wolpert, M. (2021). Generalized linear mixed-model (GLMM) trees: A flexible decision-tree method for multilevel and longitudinal data. *Psychotherapy Research*, *31*(3), 329-341.

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior research methods*, *50*(5), 2016-2034.

Gacto, M. J., Soto-Hidalgo, J. M., Alcalá-Fdez, J., & Alcalá, R. (2019). Experimental study on 164 algorithms available in software tools for solving standard non-linear regression problems. *IEEE Access*, *7*, 108916-108939.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., ... & Merigan, T. C. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, *335*(15), 1081-1090.

Hooker, G., Mentch, L., & Zhou, S. (2021). Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, *31*(6), 1-16.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Huisman, P. (2020). Hoe de toeslagenaffaire kon gebeuren. *Management Kinderopvang*, *26*(2), 36-37.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020, April). Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-14).

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174. *33*(1), 159-174.

Lax, J. R., & Phillips, J. H. (2009). How should we estimate public opinion in the States?. *American Journal of Political Science*, *53*(1), 107-121.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020, April). Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-14).

Mulligan, G. M., McCarroll, J. C., Flanagan, K. D., & Potter, D. (2016). Findings from the Third-Grade Round of the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K: 2011): First Look. NCES 2016-094. *National Center for Education Statistics*.

OpenAI. (2022, November 30., *ChatGPT: Optimizing Language Models for Dialogue.* Retrieved January 25, 2023, <https://openai.com/blog/chatgpt/>

Future of Life Institute. (n.d.). *Artificial Intelligence.* Retrieved January 25, 2023, from <https://futureoflife.org/cause-area/artificial-intelligence/>

R Core Team (2021). R: A language and environment for statistical computing*. R Foundation for Statistical Computing,* Vienna, Austria. URL <https://www.R-project.org/>

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.

Sparapani, R., Spanbauer, C., & McCulloch, R. (2021). Nonparametric machine learning and efficient computation with bayesian additive regression trees: the BART R package. *Journal of Statistical Software*, *97*, 1-66.

Tan, Y. V. (2018). *Novel Applications and Extensions for Bayesian Additive Regression Trees (BART) in Prediction, Imputation, and Causal Inference* [Doctoral dissertation, The University of Michigan].

Varshney, K. R., & Alemzadeh, H. (2017). On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big data*, *5*(3), 246-255.

Dorie, V. (2021). dbarts: Discrete Bayesian Additive Regression Trees Sampler. R package version 0.9-20. URL <https://CRAN.R-project.org/package=dbarts>

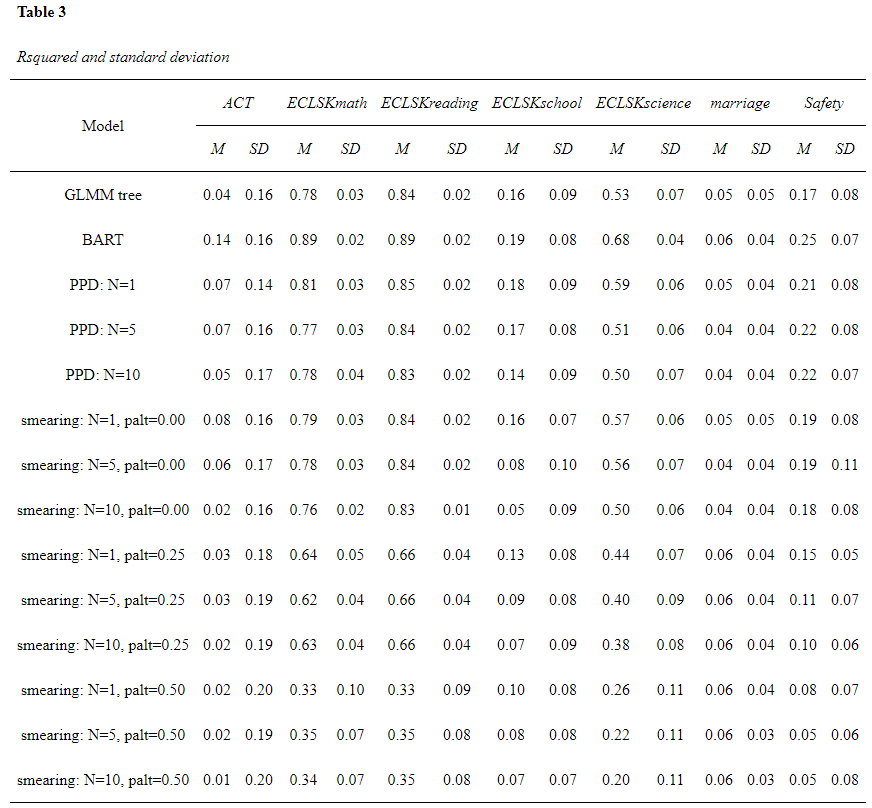
Wexler, R. (2017). When a computer program keeps you in jail. *New York Times*.

Wundervald, B., Parnell, A., & Domijan, K. (2022). Hierarchical Embedded Bayesian Additive Regression Trees. *arXiv preprint arXiv:2204.07207*.

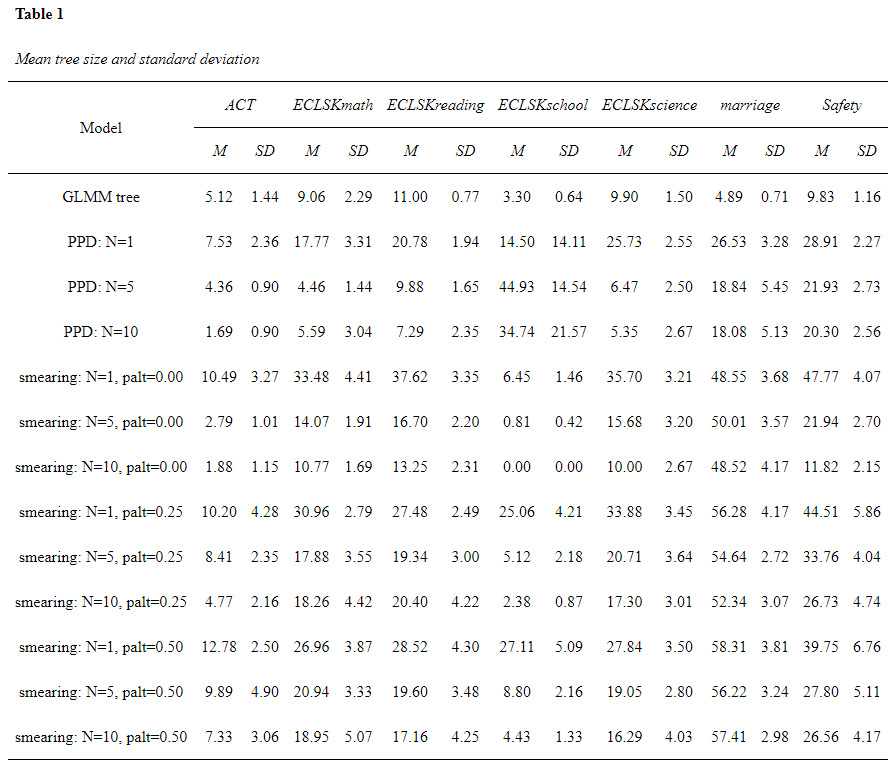
Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492-514.

**Appendix**

**Appendix A**

*Mean R2 and standard deviation for every model on every dataset* *Note: M* is mean *R2* over tentimes 10-CV (100 folds). *SD* is standard deviation.

**Appendix B**

*Mean tree size and standard deviation for every model on every dataset* *Note: M* is mean tree sizeover tentimes 10-CV (100 folds). *SD* is standard deviation.

|  |
| --- |
| **Appendix C** |
| *Effect of palt on tree size for every model grouped by N* |
|  |
| *Note: Blue dotted line marked by “GLMM” marks the median tree size of GLMM model. Red dots connected by red solid line indicate median tree size.* |