



# Bayesian Additive Regression Trees with Model Trees

Estevão B. Prado, Rafael A. Moral, Andrew C. Parnell

Hamilton Institute, Maynooth University, Maynooth, Ireland  
Corresponding author. Email: [estevao.prado@mu.ie](mailto:estevao.prado@mu.ie)

## 1 Introduction

- Bayesian Additive Regression Trees (BART) is a statistical method proposed by [1] that has become popular in recent years due to its competitive performance on regression and classification problems, when compared to other supervised machine learning methods, such as Random Forests (RF) and Gradient Boosting (GB).
- In this work, we introduce the algorithm MOTR-BART, which combines Model Trees [2] with BART to deal with local linearity at node levels. In MOTR-BART, rather than estimating a constant as the predicted value as BART does, for each terminal node a linear predictor is estimated, including only the covariates that have been used as a split in the corresponding tree. With this approach, we aim to capture linear associations between the response and covariates and then improve the final prediction.

## 2 BART

Introduced by [1], BART is a tree-based machine learning method that estimates a univariate response by a sum-of-trees as

$$y_i = \sum_{t=1}^m g(\mathbf{x}_i; T_t, M_t) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

where  $g(\mathbf{x}_i; T_t, M_t)$  is a function that assigns a predicted value  $\mu_{t\ell}$  based on the covariates  $\mathbf{x}_i$ ,  $T_t$  is the set of splitting rules that defines the  $t$ -th tree and  $M_t = (\mu_{t1}, \dots, \mu_{tb_t})$  is the set of predicted values for all nodes in the tree  $t$ .

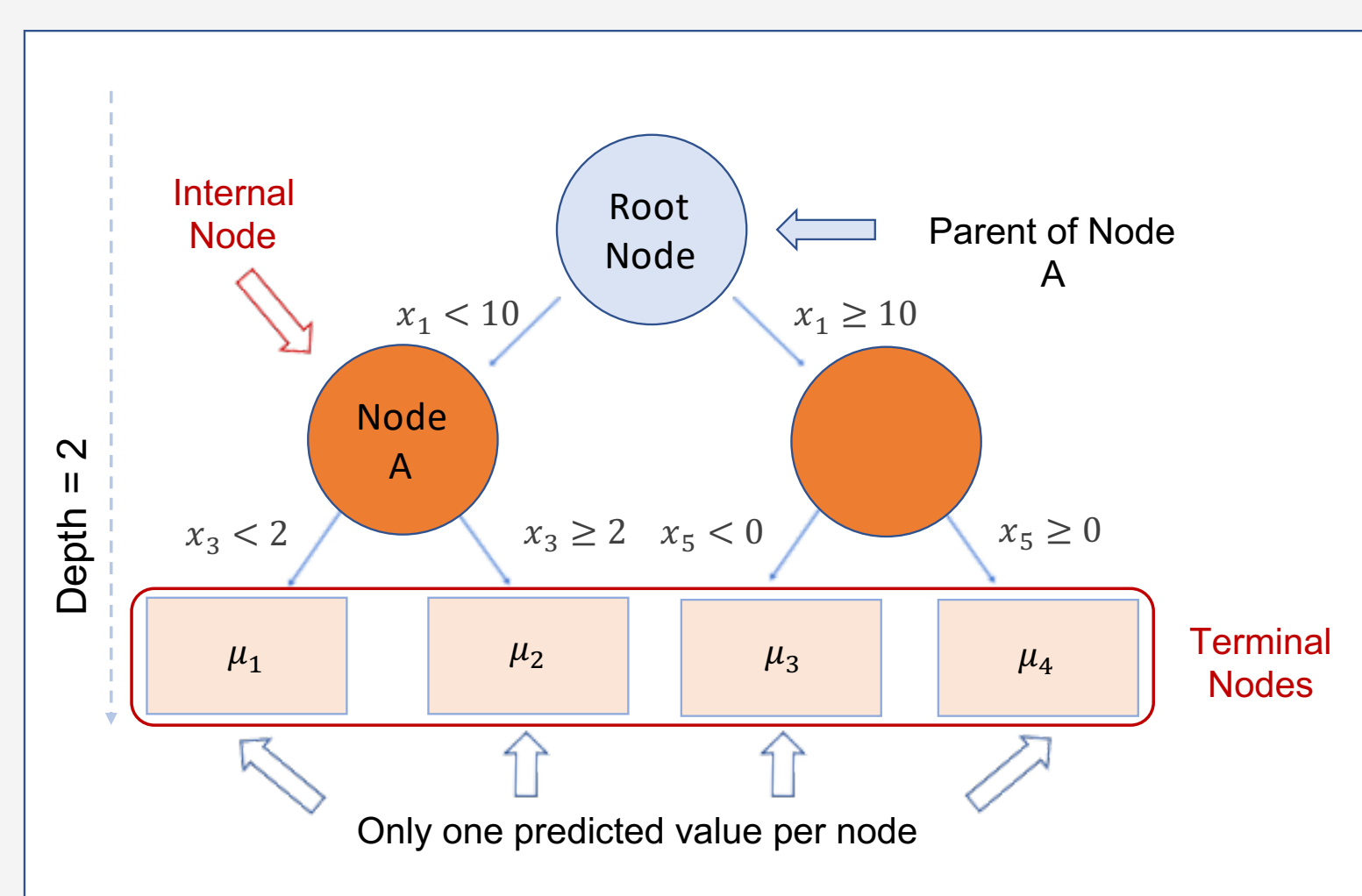


Figure 1: An example of a single tree generated by BART. In practice, BART generates multiple trees for which the predictions are added together.

## 3 BART with Model Trees

In MOTR-BART, we consider that the response variable is a sum of trees in the form of

$$\mathbf{y} = \sum_{t=1}^m g(\mathbf{X}; T_t, B_t) + \epsilon,$$

where  $B_t$  is the set of parameters of all linear predictors for the tree  $t$ . In terms of partial residuals, MOTR-BART can be represented as

$$r_i | \mathbf{x}_i, \boldsymbol{\beta}_{t\ell}, \sigma^2 \sim N(\mathbf{x}_i \boldsymbol{\beta}_{t\ell}, \sigma^2),$$

where  $r_i = y_i - \sum_{j \neq t} g(\mathbf{x}_i; T_j, B_j)$  and  $\boldsymbol{\beta}_{t\ell}$  is the parameter vector associated to the terminal node  $\ell$  of the tree  $t$ . In this sense, all observations  $i \in \mathcal{P}_{t\ell}$  will have predicted values based on  $\boldsymbol{\beta}_{t\ell}$  and the values of their covariates  $\mathbf{X}_{t\ell}$ . Thus, each observation  $i \in \mathcal{P}_{t\ell}$  may have different predicted values.

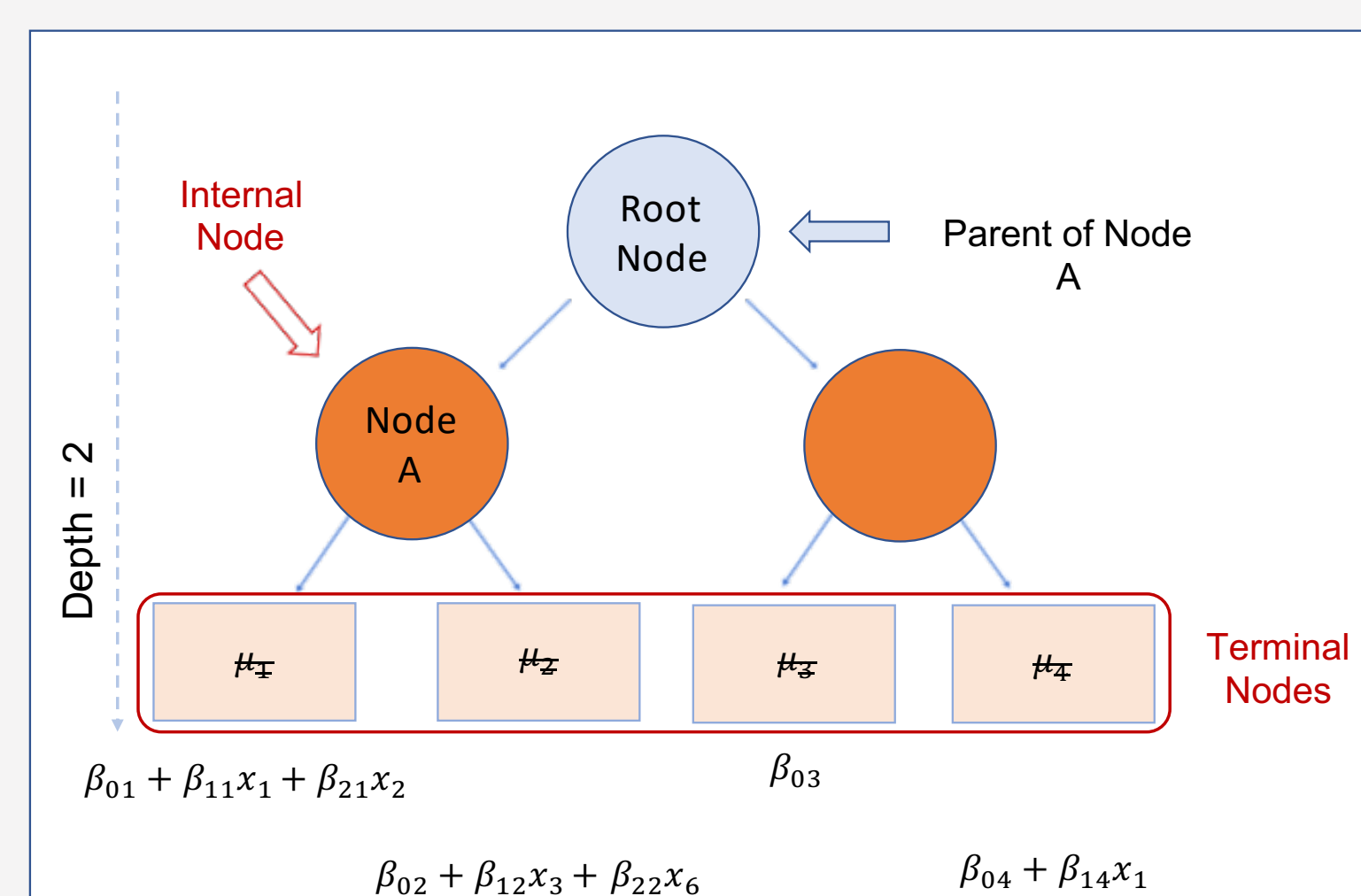


Figure 2: An example of a single tree generated by MOTR-BART.

## 4 Implementation

MOTR-BART implementation is available in the format of an R package at <https://github.com/ebprado/MOTR-BART>.

## 5 Results for real data

We compare the predictive performance of MOTR-BART to RF, GB, BART, soft BART, and Local Linear Forests (LLF), in terms of RMSE, on four real data sets. Below, we show the results for two data sets.

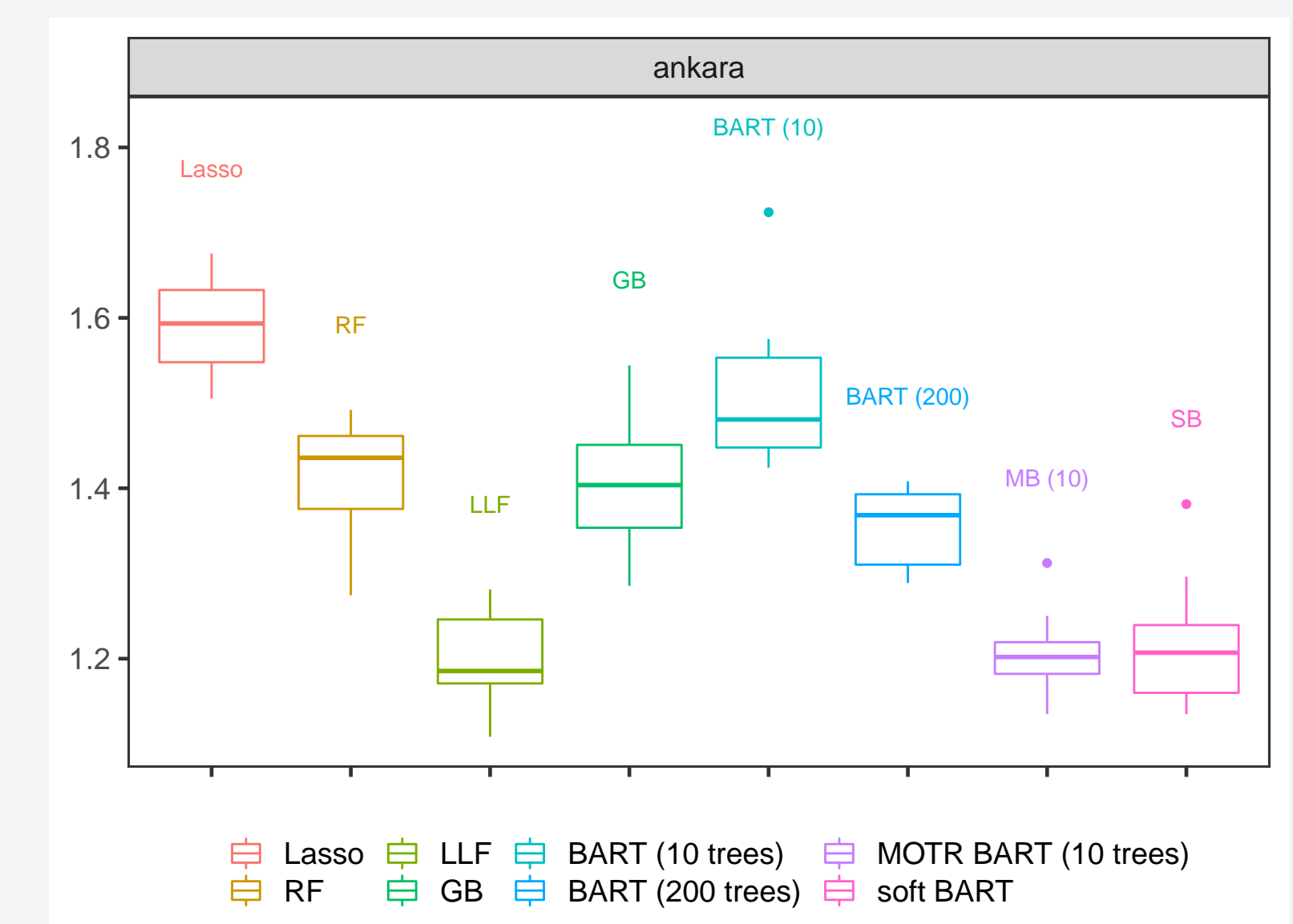


Figure 3: Comparison of RMSE for Ankara data set on 10 different test data.

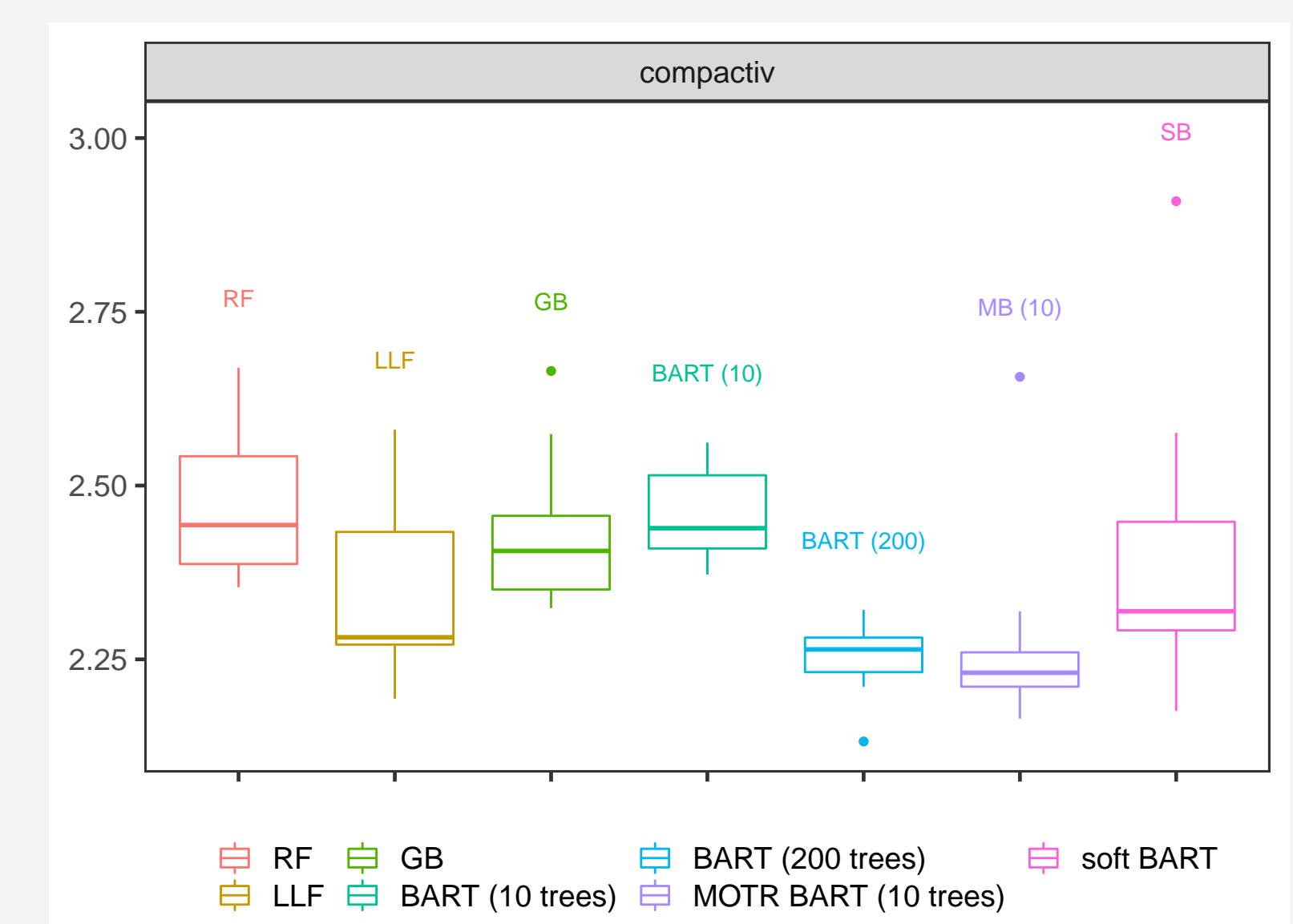


Figure 4: Comparison of RMSE for Compactiv data set on 10 different test data.

- MOTR-BART presents the **best or the second best** results.
- It requires **far fewer** trees to achieve equal or better performance than BART.

## 6 Discussion

- To evaluate variable importance or even to select the covariates in the linear predictors is not straightforward in MOTR-BART.
- Splines can be introduced to capture local non-linear behaviour.

## References

- Hugh A Chipman, Edward I George, Robert E McCulloch, et al. “BART: Bayesian additive regression trees”. In: *The Annals of Applied Statistics* 4.1 (2010), pp. 266–298.
- John R Quinlan. “Learning with continuous classes”. In: *5th Australian joint conference on artificial intelligence*. Vol. 92. World Scientific. 1992, pp. 343–348.