

IMPUTATION OF MISSING DATA WITH BAYESIAN ADDITIVE REGRESSION TREES

Todd Burrows

Supervised by Emmanuel Ogundimu,
Department of Mathematics, Durham University

11th February 2026

CONTENTS

1	Introduction to the Problem of Missing Data	2
2	Tree Based Regression	3
3	Bayesian Additive Regression Trees	4
4	The BART Process	5
5	missBART Imputation Method	7
6	Discussion and Research	16
7	Appendix	21

INTRODUCTION TO THE PROBLEM OF MISSING DATA

- ▶ Historically, **missingness** was commonly ignored.¹
- ▶ **Donald Rubin** formalised the analysis of missing data.
- ▶ **Multiple Imputation** proposed to capture the uncertainty caused by the missing data.²

¹ **Rubin, 1976**

² **Rubin, 1978**

TREE BASED REGRESSION

CLASSIFICATION AND REGRESSION TREES ³

- ▶ Statistical **classification** or **regression** that partitions the predictor space into subgroups.
- ▶ **Non-parametric** and can model high-level **interactions** and **non-linearities**.

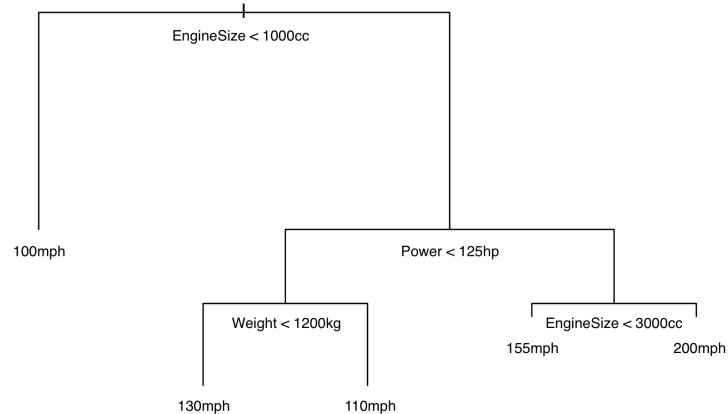


Figure. A simple regression tree to predict the top speed of a vehicle, given the engine size, power and weight.

³Formalised by Breiman et al., 1984

BAYESIAN ADDITIVE REGRESSION TREES

AN INTRODUCTION TO BART ⁴

- ▶ BART utilises a **sum-of-trees** structure, $Y = \sum_{j=1}^m g(\mathbf{X}; T_j, M_j) + \epsilon$, where for tree j
 - ▶ T_j is the **tree structure**.
 - ▶ $M_j = \{\mu_{1j}, \mu_{2j}, \dots, \mu_{b_jj}\}$ is the set of b_j **terminal parameters**.
- ▶ Trees built via a **Bayesian backfitting** algorithm.
- ▶ Can form the basis of an **imputation** model.

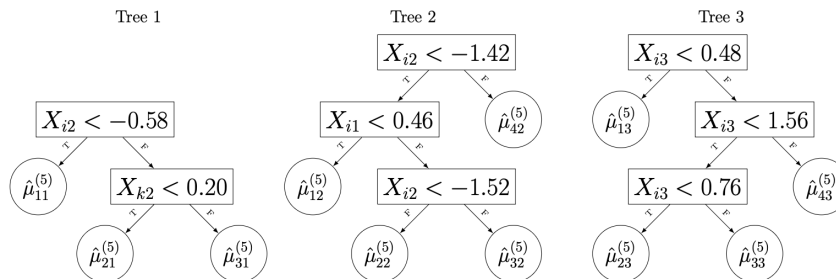


Figure. A sum-of-trees structure. Extracted from **Tan and Roy, 2019**.

THE BART PROCESS

THE PRIOR

- ▶ A strongly influential **regularising** prior guides the tree building process, tailoring the fit.

- ▶ **The Tree Prior** – $p(T_j)$.
- ▶ **The Terminal Parameter Prior** – $p(\mu_{ij} \mid T_j)$.
- ▶ **The Error Variance Prior** – $p(\sigma)$.

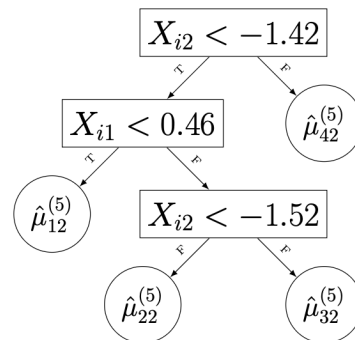


Figure. A single tree. Extracted from **Tan and Roy, 2019**.

THE BART PROCESS

TREE BUILDING

- ▶ The **backfitting** algorithm draws each tree (T_j, M_j) conditional on $[(T_{(j)}, M_{(j)}), \sigma]$.
- ▶ The conditional distribution of the j^{th} tree depends only on the other trees and the training data via,

$$R_j \equiv y - \sum_{k \neq j} g(\mathbf{x}; T_k, M_k).$$

- ▶ Tree design proposed iteratively by a **stochastic search** procedure that repeatedly **perturbs** the trees in one of four ways.

▶ **Grow**

▶ **Prune**

▶ **Change**

▶ **Swap**

MISSBART IMPUTATION METHOD

- Based on the popular **missForest** method of **Stekhoven and Bühlmann, 2012**.

missBART Algorithm

Require: \mathbf{X} an $n \times p$ matrix, stopping criterion γ

- 1: Initialise the missing values with a simple mean imputation
- 2: $\mathbf{k} \leftarrow$ vector of the indices of the columns in \mathbf{X} in order of increasing missingness
- 3: **while** not γ **do**
- 4: $\mathbf{X}_{\text{old}}^{\text{imp}} \leftarrow$ current imputed matrix
- 5: **for** each column, c in \mathbf{k} **do**
- 6: Fit a **BART** model: $\mathbf{y}_{\text{obs}}^{(c)} \sim \mathbf{x}_{\text{obs}}^{(c)}$
- 7: Predict $\mathbf{y}_{\text{mis}}^{(c)}$ using $\mathbf{x}_{\text{mis}}^{(c)}$
- 8: $\mathbf{X}_{\text{new}}^{\text{imp}} \leftarrow$ new imputed matrix, with updated $\mathbf{y}_{\text{mis}}^{(c)}$
- 9: **end for**
- 10: Update stopping criterion γ
- 11: **end while**
- 12: **return** the imputed matrix \mathbf{X}^{imp}

Table. Data based on the Theoph dataset [R Core Team, 2024].

Y	X ₁	X ₂	X ₃	X ₄
70.5	4.53	7.07	—	—
86.4	3.10	—	4.24	—
—	—	6.98	5.25	74
70.5	4.53	7.15	4.73	85
—	4.40	12.00	3.01	—
65.0	4.92	—	5.87	116
70.5	4.53	0.98	7.31	81
79.6	4.02	3.82	—	—
—	4.40	9.02	5.33	42
64.6	4.95	0.25	0.85	68

MISSBART IMPUTATION METHOD

- Based on the popular **missForest** method of **Stekhoven and Bühlmann, 2012**.

missBART Algorithm

Require: \mathbf{X} an $n \times p$ matrix, stopping criterion γ

- 1: Initialise the missing values with a simple mean imputation
- 2: $\mathbf{k} \leftarrow$ vector of the indices of the columns in \mathbf{X} in order of increasing missingness
- 3: **while** not γ **do**
- 4: $\mathbf{X}_{\text{old}}^{\text{imp}} \leftarrow$ current imputed matrix
- 5: **for** each column, c in \mathbf{k} **do**
- 6: Fit a **BART** model: $\mathbf{y}_{\text{obs}}^{(c)} \sim \mathbf{x}_{\text{obs}}^{(c)}$
- 7: Predict $\mathbf{y}_{\text{mis}}^{(c)}$ using $\mathbf{x}_{\text{mis}}^{(c)}$
- 8: $\mathbf{X}_{\text{new}}^{\text{imp}} \leftarrow$ new imputed matrix, with updated $\mathbf{y}_{\text{mis}}^{(c)}$
- 9: **end for**
- 10: Update stopping criterion γ
- 11: **end while**
- 12: **return** the imputed matrix \mathbf{X}^{imp}

Table. Data based on the Theoph dataset [R Core Team, 2024].

Y	X ₁	X ₂	X ₃	X ₄
70.5	4.53	7.07	4.57	77.7
86.4	3.10	5.91	4.24	77.7
72.4	4.38	6.98	5.25	74
70.5	4.53	7.15	4.73	85
72.4	4.40	12.00	3.01	77.7
65.0	4.92	5.91	5.87	116
70.5	4.53	0.98	7.31	81
79.6	4.02	3.82	4.57	77.7
72.4	4.40	9.02	5.33	42
64.6	4.95	0.25	0.85	68

MISSBART IMPUTATION METHOD

- Based on the popular **missForest** method of **Stekhoven and Bühlmann, 2012**.

missBART Algorithm

Require: \mathbf{X} an $n \times p$ matrix, stopping criterion γ

- 1: Initialise the missing values with a simple mean imputation
- 2: $\mathbf{k} \leftarrow$ vector of the indices of the columns in \mathbf{X} in order of increasing missingness
- 3: **while** not γ **do**
- 4: $\mathbf{X}_{\text{old}}^{\text{imp}} \leftarrow$ current imputed matrix
- 5: **for** each column, c in \mathbf{k} **do**
- 6: Fit a **BART** model: $\mathbf{y}_{\text{obs}}^{(c)} \sim \mathbf{x}_{\text{obs}}^{(c)}$
- 7: Predict $\mathbf{y}_{\text{mis}}^{(c)}$ using $\mathbf{x}_{\text{mis}}^{(c)}$
- 8: $\mathbf{X}_{\text{new}}^{\text{imp}} \leftarrow$ new imputed matrix, with updated $\mathbf{y}_{\text{mis}}^{(c)}$
- 9: **end for**
- 10: Update stopping criterion γ
- 11: **end while**
- 12: **return** the imputed matrix \mathbf{X}^{imp}

Table. Data based on the Theoph dataset [R Core Team, 2024].

Y	X_1	X_2	X_3	X_4
70.5	4.53	7.07	4.57	77.7
86.4	3.10	5.91	4.24	77.7
72.4	4.38	6.98	5.25	74
70.5	4.53	7.15	4.73	85
72.4	4.40	12.00	3.01	77.7
65.0	4.92	5.91	5.87	116
70.5	4.53	0.98	7.31	81
79.6	4.02	3.82	4.57	77.7
72.4	4.40	9.02	5.33	42
64.6	4.95	0.25	0.85	68

MISSBART IMPUTATION METHOD

- Based on the popular **missForest** method of **Stekhoven and Bühlmann, 2012**.

missBART Algorithm

Require: \mathbf{X} an $n \times p$ matrix, stopping criterion γ

- 1: Initialise the missing values with a simple mean imputation
- 2: $\mathbf{k} \leftarrow$ vector of the indices of the columns in \mathbf{X} in order of increasing missingness
- 3: **while** not γ **do**
- 4: $\mathbf{X}_{\text{old}}^{\text{imp}} \leftarrow$ current imputed matrix
- 5: **for** each column, c in \mathbf{k} **do**
- 6: Fit a **BART** model: $\mathbf{y}_{\text{obs}}^{(c)} \sim \mathbf{x}_{\text{obs}}^{(c)}$
- 7: Predict $\mathbf{y}_{\text{mis}}^{(c)}$ using $\mathbf{x}_{\text{mis}}^{(c)}$
- 8: $\mathbf{X}_{\text{new}}^{\text{imp}} \leftarrow$ new imputed matrix, with updated $\mathbf{y}_{\text{mis}}^{(c)}$
- 9: **end for**
- 10: Update stopping criterion γ
- 11: **end while**
- 12: **return** the imputed matrix \mathbf{X}^{imp}

Table. Data based on the Theoph dataset [R Core Team, 2024].

Y	X ₁	X ₂	X ₃	X ₄
70.5	4.53	7.07	4.57	77.7
86.4	3.10	5.91	4.24	77.7
72.4	4.38	6.98	5.25	74
70.5	4.53	7.15	4.73	85
72.4	4.40	12.00	3.01	77.7
65.0	4.92	5.91	5.87	116
70.5	4.53	0.98	7.31	81
79.6	4.02	3.82	4.57	77.7
72.4	4.40	9.02	5.33	42
64.6	4.95	0.25	0.85	68

MISSBART IMPUTATION METHOD

- Based on the popular **missForest** method of **Stekhoven and Bühlmann, 2012**.

missBART Algorithm

Require: \mathbf{X} an $n \times p$ matrix, stopping criterion γ

- 1: Initialise the missing values with a simple mean imputation
- 2: $\mathbf{k} \leftarrow$ vector of the indices of the columns in \mathbf{X} in order of increasing missingness
- 3: **while** not γ **do**
- 4: $\mathbf{X}_{\text{old}}^{\text{imp}} \leftarrow$ current imputed matrix
- 5: **for** each column, c in \mathbf{k} **do**
- 6: Fit a **BART** model: $\mathbf{y}_{\text{obs}}^{(c)} \sim \mathbf{x}_{\text{obs}}^{(c)}$
- 7: Predict $\mathbf{y}_{\text{mis}}^{(c)}$ using $\mathbf{x}_{\text{mis}}^{(c)}$
- 8: $\mathbf{X}_{\text{new}}^{\text{imp}} \leftarrow$ new imputed matrix, with updated $\mathbf{y}_{\text{mis}}^{(c)}$
- 9: **end for**
- 10: Update stopping criterion γ
- 11: **end while**
- 12: **return** the imputed matrix \mathbf{X}^{imp}

Table. Data based on the Theoph dataset [R Core Team, 2024].

Y	X ₁	X ₂	X ₃	X ₄
70.5	4.53	7.07	4.57	77.7
86.4	3.10	5.91	4.24	77.7
72.4	4.92	6.98	5.25	74
70.5	4.53	7.15	4.73	85
72.4	4.40	12.00	3.01	77.7
65.0	4.92	5.91	5.87	116
70.5	4.53	0.98	7.31	81
79.6	4.02	3.82	4.57	77.7
72.4	4.40	9.02	5.33	42
64.6	4.95	0.25	0.85	68

MISSBART IMPUTATION METHOD

- Based on the popular **missForest** method of **Stekhoven and Bühlmann, 2012**.

missBART Algorithm

Require: \mathbf{X} an $n \times p$ matrix, stopping criterion γ

- 1: Initialise the missing values with a simple mean imputation
- 2: $\mathbf{k} \leftarrow$ vector of the indices of the columns in \mathbf{X} in order of increasing missingness
- 3: **while** not γ **do**
- 4: $\mathbf{X}_{\text{old}}^{\text{imp}} \leftarrow$ current imputed matrix
- 5: **for** each column, c in \mathbf{k} **do**
- 6: Fit a **BART** model: $\mathbf{y}_{\text{obs}}^{(c)} \sim \mathbf{x}_{\text{obs}}^{(c)}$
- 7: Predict $\mathbf{y}_{\text{mis}}^{(c)}$ using $\mathbf{x}_{\text{mis}}^{(c)}$
- 8: $\mathbf{X}_{\text{new}}^{\text{imp}} \leftarrow$ new imputed matrix, with updated $\mathbf{y}_{\text{mis}}^{(c)}$
- 9: **end for**
- 10: Update stopping criterion γ
- 11: **end while**
- 12: **return** the imputed matrix \mathbf{X}^{imp}

Table. Data based on the Theoph dataset [R Core Team, 2024].

Y	X ₁	X ₂	X ₃	X ₄
70.5	4.53	7.07	4.57	77.7
86.4	3.10	5.91	4.24	77.7
72.4	4.92	6.98	5.25	74
70.5	4.53	7.15	4.73	85
72.4	4.40	12.00	3.01	77.7
65.0	4.92	5.91	5.87	116
70.5	4.53	0.98	7.31	81
79.6	4.02	3.82	4.57	77.7
72.4	4.40	9.02	5.33	42
64.6	4.95	0.25	0.85	68

MISSBART IMPUTATION METHOD

- Based on the popular **missForest** method of **Stekhoven and Bühlmann, 2012**.

missBART Algorithm

Require: \mathbf{X} an $n \times p$ matrix, stopping criterion γ

- 1: Initialise the missing values with a simple mean imputation
- 2: $\mathbf{k} \leftarrow$ vector of the indices of the columns in \mathbf{X} in order of increasing missingness
- 3: **while** not γ **do**
- 4: $\mathbf{X}_{\text{old}}^{\text{imp}} \leftarrow$ current imputed matrix
- 5: **for** each column, c in \mathbf{k} **do**
- 6: Fit a **BART** model: $\mathbf{y}_{\text{obs}}^{(c)} \sim \mathbf{x}_{\text{obs}}^{(c)}$
- 7: Predict $\mathbf{y}_{\text{mis}}^{(c)}$ using $\mathbf{x}_{\text{mis}}^{(c)}$
- 8: $\mathbf{X}_{\text{new}}^{\text{imp}} \leftarrow$ new imputed matrix, with updated $\mathbf{y}_{\text{mis}}^{(c)}$
- 9: **end for**
- 10: Update stopping criterion γ
- 11: **end while**
- 12: **return** the imputed matrix \mathbf{X}^{imp}

Table. Data based on the Theoph dataset [R Core Team, 2024].

Y	X ₁	X ₂	X ₃	X ₄
70.5	4.53	7.07	4.57	77.7
86.4	3.10	5.91	4.24	77.7
72.4	4.92	6.98	5.25	74
70.5	4.53	7.15	4.73	85
72.4	4.40	12.00	3.01	77.7
65.0	4.92	5.91	5.87	116
70.5	4.53	0.98	7.31	81
79.6	4.02	3.82	4.57	77.7
72.4	4.40	9.02	5.33	42
64.6	4.95	0.25	0.85	68

MISSBART IMPUTATION METHOD

- Based on the popular **missForest** method of **Stekhoven and Bühlmann, 2012**.

missBART Algorithm

Require: \mathbf{X} an $n \times p$ matrix, stopping criterion γ

- 1: Initialise the missing values with a simple mean imputation
- 2: $\mathbf{k} \leftarrow$ vector of the indices of the columns in \mathbf{X} in order of increasing missingness
- 3: **while** not γ **do**
- 4: $\mathbf{X}_{\text{old}}^{\text{imp}} \leftarrow$ current imputed matrix
- 5: **for** each column, c in \mathbf{k} **do**
- 6: Fit a **BART** model: $\mathbf{y}_{\text{obs}}^{(c)} \sim \mathbf{x}_{\text{obs}}^{(c)}$
- 7: Predict $\mathbf{y}_{\text{mis}}^{(c)}$ using $\mathbf{x}_{\text{mis}}^{(c)}$
- 8: $\mathbf{X}_{\text{new}}^{\text{imp}} \leftarrow$ new imputed matrix, with updated $\mathbf{y}_{\text{mis}}^{(c)}$
- 9: **end for**
- 10: Update stopping criterion γ
- 11: **end while**
- 12: **return** the imputed matrix \mathbf{X}^{imp}

Table. Data based on the Theoph dataset [R Core Team, 2024].

Y	X ₁	X ₂	X ₃	X ₄
70.5	4.53	7.07	4.57	77.7
86.4	3.10	6.98	4.24	77.7
72.4	4.92	6.98	5.25	74
70.5	4.53	7.15	4.73	85
72.4	4.40	12.00	3.01	77.7
65.0	4.92	4.10	5.87	116
70.5	4.53	0.98	7.31	81
79.6	4.02	3.82	4.57	77.7
72.4	4.40	9.02	5.33	42
64.6	4.95	0.25	0.85	68

MISSBART IMPUTATION METHOD

- Based on the popular **missForest** method of **Stekhoven and Bühlmann, 2012**.

missBART Algorithm

Require: \mathbf{X} an $n \times p$ matrix, stopping criterion γ

- 1: Initialise the missing values with a simple mean imputation
- 2: $\mathbf{k} \leftarrow$ vector of the indices of the columns in \mathbf{X} in order of increasing missingness
- 3: **while** not γ **do**
- 4: $\mathbf{X}_{\text{old}}^{\text{imp}} \leftarrow$ current imputed matrix
- 5: **for** each column, c in \mathbf{k} **do**
- 6: Fit a **BART** model: $\mathbf{y}_{\text{obs}}^{(c)} \sim \mathbf{x}_{\text{obs}}^{(c)}$
- 7: Predict $\mathbf{y}_{\text{mis}}^{(c)}$ using $\mathbf{x}_{\text{mis}}^{(c)}$
- 8: $\mathbf{X}_{\text{new}}^{\text{imp}} \leftarrow$ new imputed matrix, with updated $\mathbf{y}_{\text{mis}}^{(c)}$
- 9: **end for**
- 10: Update stopping criterion γ
- 11: **end while**
- 12: **return** the imputed matrix \mathbf{X}^{imp}

Table. Data based on the Theoph dataset [R Core Team, 2024].

Y	X ₁	X ₂	X ₃	X ₄
70.5	4.53	7.07	5.16	40.2
86.4	3.10	6.98	4.24	88.4
65.2	4.92	6.98	5.25	74
70.5	4.53	7.15	4.73	85
72.7	4.40	12.00	3.01	11.7
65.0	4.92	4.10	5.87	116
70.5	4.53	0.98	7.31	81
79.6	4.02	3.82	8.33	12.2
71.1	4.40	9.02	5.33	42
64.6	4.95	0.25	0.85	68

DISCUSSION AND RESEARCH

IMPUTATION MODELS

- ▶ **missForest/BART** produces a single imputed dataset.
- ▶ Is this a disadvantage?

DISCUSSION AND RESEARCH

IMPUTATION MODELS

- ▶ **missForest/BART** produces a single imputed dataset.
- ▶ Is this a disadvantage?
- ▶ Yes – not multiple imputation! ⁵

⁵van Buuren, 2018

DISCUSSION AND RESEARCH

IMPUTATION MODELS

- ▶ **missForest/BART** produces a single imputed dataset.
- ▶ Is this a disadvantage?
- ▶ Yes – not multiple imputation! ⁶
- ▶ Bootstrap the data and produce multiple imputed datasets.
- ▶ Compare missForest vs missBART vs both bootstrapped.
- ▶ Utilise BART's intrinsic **Bayesian probability model** to generate multiple imputations from **posterior predictive** draws and combine via **Rubin's rules**.

⁶van Buuren, 2018

DISCUSSION AND RESEARCH









IMPUTATION METRICS

- ▶ Evaluating accuracy alone **fails** to detect **variability** issues.
 - ▶ Calculate the **empirical bias**.
 - ▶ Calculate the **coverage**.
-
- ▶ **Imputation is not prediction.**⁷

⁷van Buuren, 2018

REFERENCES AND QUESTIONS

THANK YOU FOR LISTENING, ANY QUESTIONS? ⁸

-  Breiman, L., Friedman, J., Olshen, R., & Stone, C. J. (1984). **Classification and regression trees (1st ed.)**. Chapman; Hall/CRC. <https://doi.org/10.1201/9781315139470>
-  Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). **Bart: Bayesian additive regression trees**. *The Annals of Applied Statistics*, 4(1). <https://doi.org/10.1214/09-aos285>
-  R Core Team. (2024). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing. Vienna, Austria. <https://www.R-project.org/>
-  Rubin, D. B. (1976). **Inference and missing data**. *Biometrika*, 63(3), 581–592. <https://doi.org/10.1093/biomet/63.3.581>
-  Rubin, D. B. (1978). **Multiple imputations in sample surveys—a phenomenological bayesian approach to nonresponse**. *Proceedings of the Survey Research Methods Section of the American Statistical Association*. <https://api.semanticscholar.org/CorpusID:197861764>
-  Stekhoven, D. J., & Bühlmann, P. (2012). **Missforest—non-parametric missing value imputation for mixed-type data**. *Bioinformatics*, 28(1), 112–118. <https://doi.org/10.1093/bioinformatics/btr597>
-  Tan, Y. V., & Roy, J. (2019). **Bayesian additive regression trees and the general bart model**. *Statistics in Medicine*, 38(25), 5048–5069. <https://doi.org/https://doi.org/10.1002/sim.8347>
-  van Buuren, S. (2018). **Flexible imputation of missing data, second edition (2nd ed.)**. Chapman; Hall/CRC. <https://doi.org/10.1201/9780429492259>

⁸<https://github.com/toddburrows>

APPENDIX

APPENDIX A: THE BART MCMC CYCLE

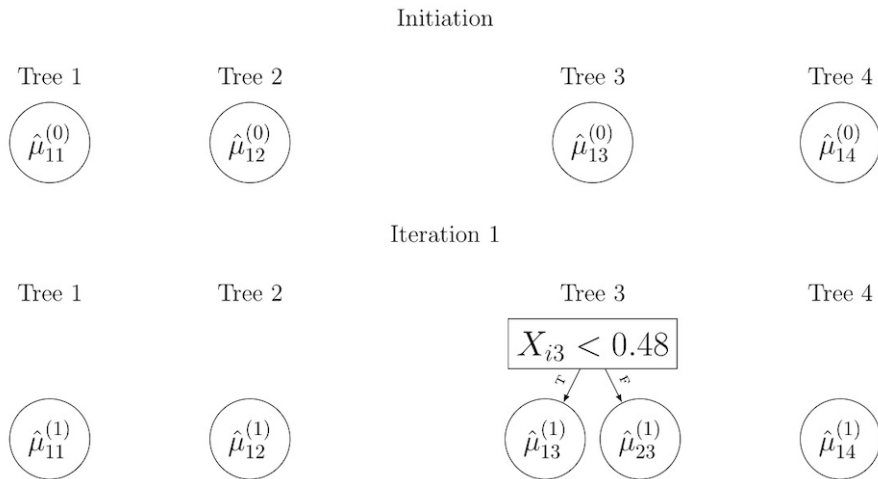


Figure. Initialisation and Iteration 1 of a BART model. Extracted from [Tan and Roy, 2019](#).

APPENDIX

APPENDIX A: THE BART MCMC CYCLE

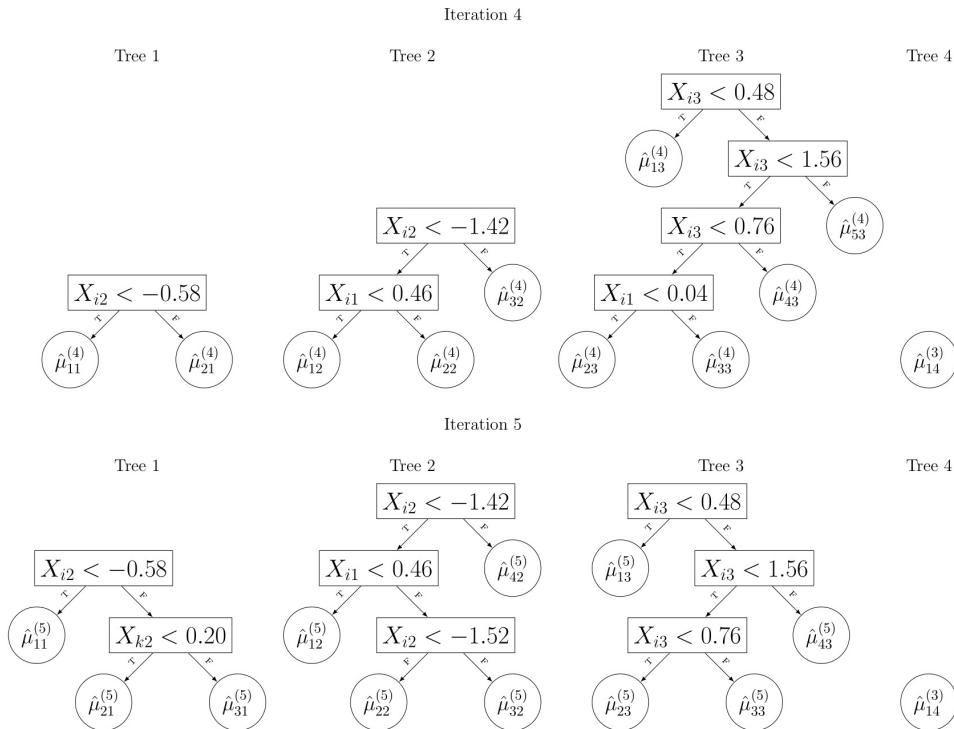


Figure. Iteration 4 and 5 of a BART model. Extracted from [Tan and Roy, 2019](#).