Tracing Trees

Visualizing Random Forest Tree Variability with Trace Plots

Katherine Goode (5573)

kjgoode@sandia.gov

July 11, 2022

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND 2022-XXXX.

Introduction to Katherine

Research and development statistician in 5573

Education

- BA in mathematics from Lawrence University
- MS in statistics from University of Wisconsin - Madison
- PhD in statistics from Iowa State University

Sandia Journey

- Dec 2019: Intern (mentored by Daniel Ries, 5574)
- Sep 2021: Post-doc (mentored by J. Derek Tucker, 5573)
- Dec 2021: FTE (mentored by J. Derek Tucker, 5573)

Research Interests

- Explainable machine learning
- Data visualization
- Model assessment

Personal website: goodekat.github.io
Personal github: github.com/goodekat

Overview

- Background: Trace Plots
- Methods: Extending Trace Plots
  - TreeTracer: Implementation and Structural Augmentations in R
  - Tree Summaries: Identifying Representative Trees
- Music Example: Application with "larger" random forest
- Conclusions: Pros, Cons, and Possible Research Directions

Credits: Joint work with Heike Hofmann (Professor at Iowa State University)

Background: Trace Plots

Common Tree Visualization

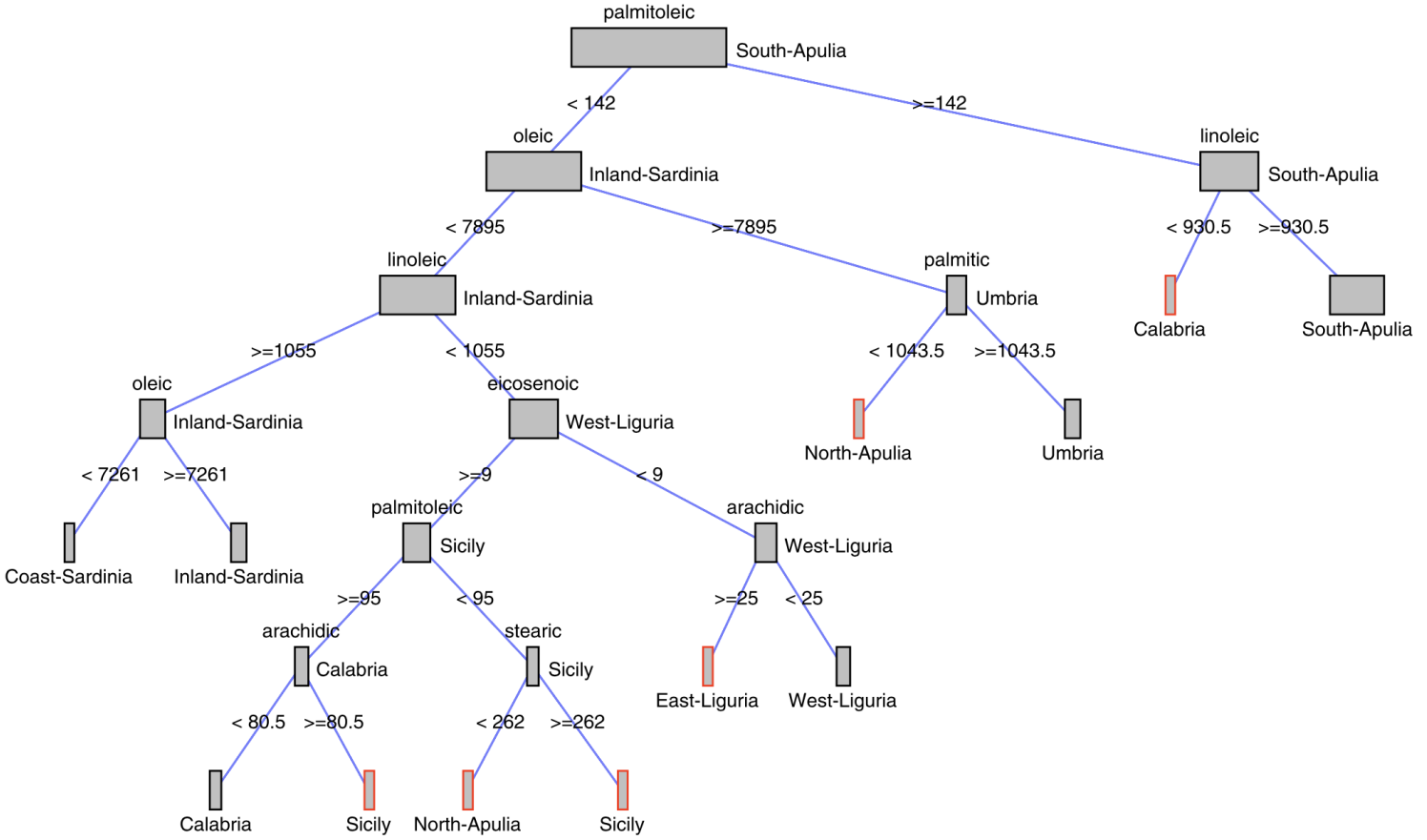
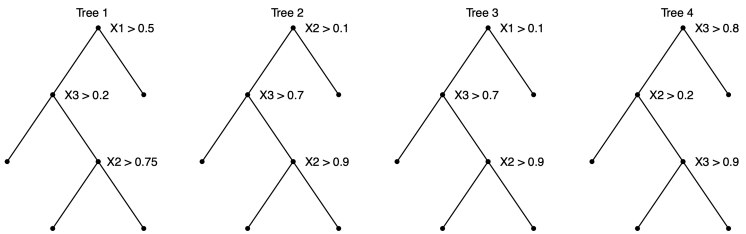


Image source: Urbanek (2008)

Visual Comparisons of Multiple Trees

Issues with "traditional" visuals:

- Difficult direct visual comparison
- Non-efficient use of space
- Identifying patterns is cognitively difficult (figure classification French, Ekstrom, and Price (1963))



(A) Random Forest with 100 trees

(B) Random Forest with 625 trees

Image source: Kuznetsova (2014)

Trace Plots (one tree) Urbanek (2008)

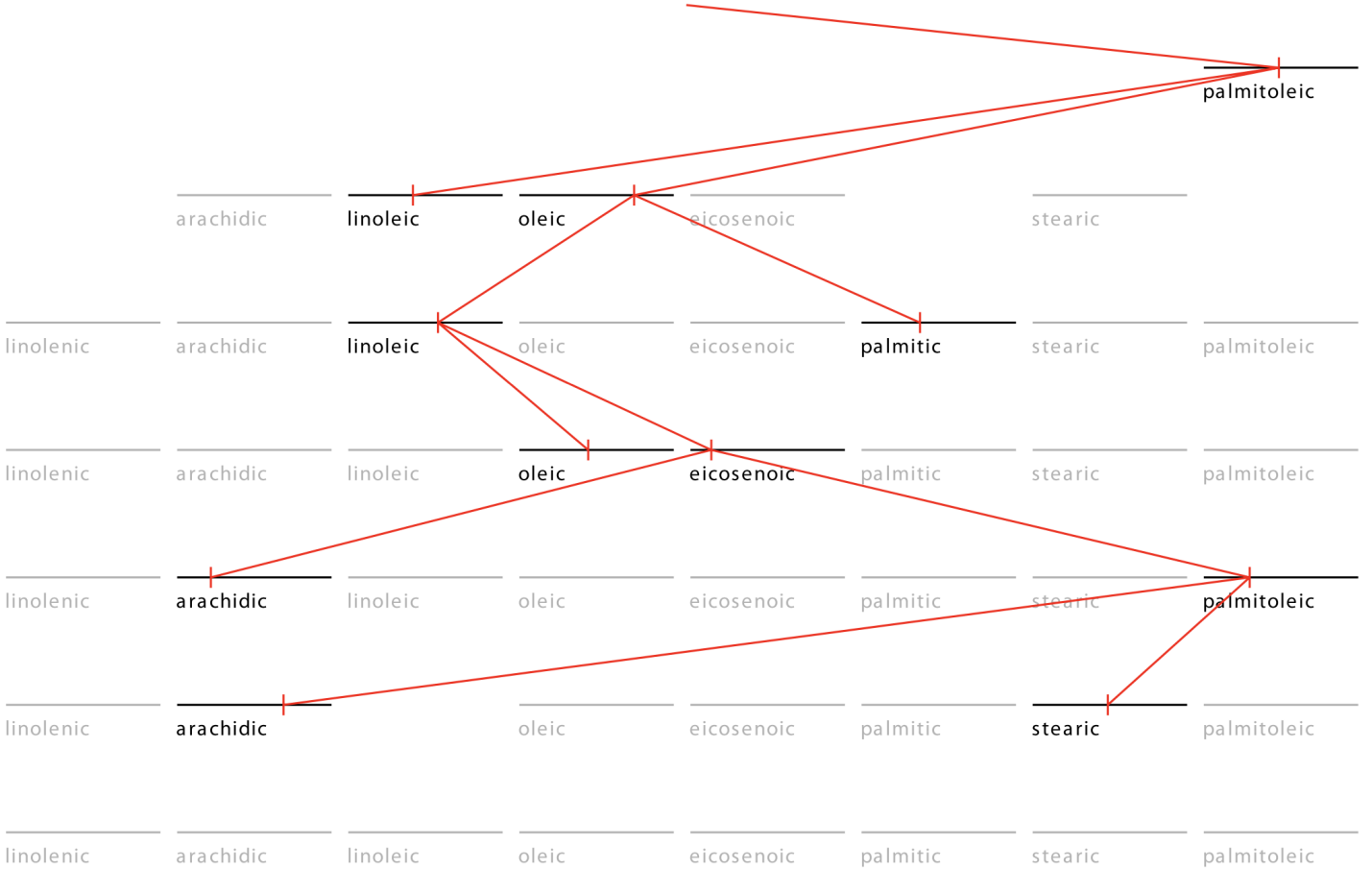


Figure 10.13. A classification tree and its *trace plot*

Image source: Urbanek (2008)

Trace Plots (esemble of trees) Urbanek (2008)

Designed to compare (1) variables used for splitting, (2) location of split points, and (3) hierarchical structure

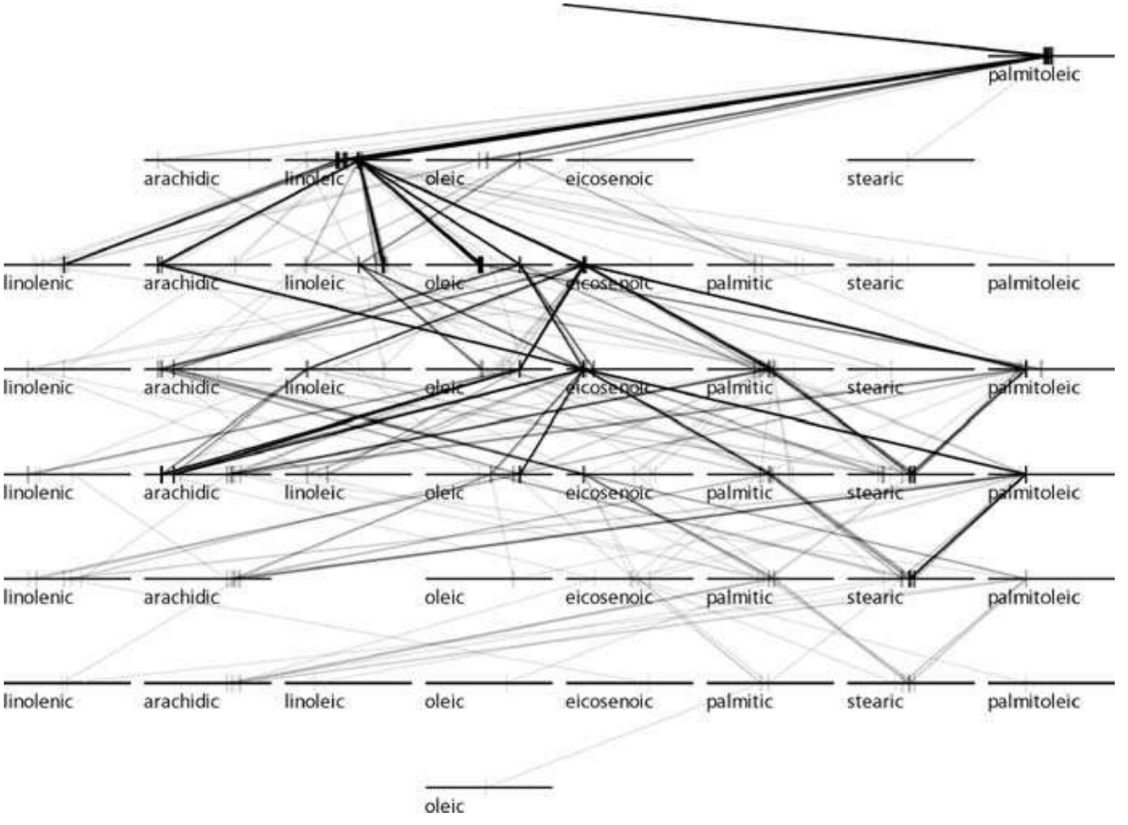


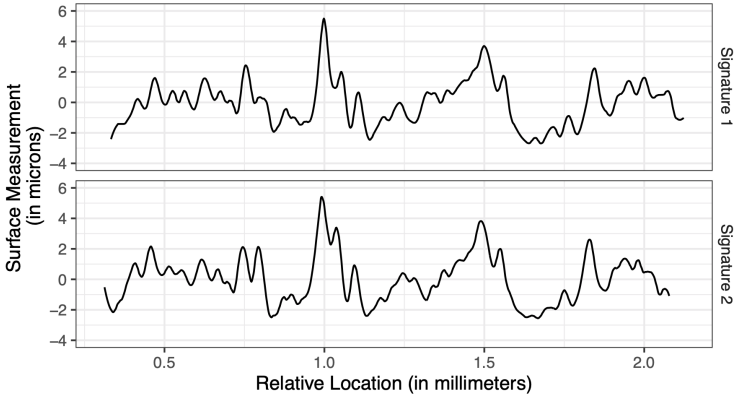
Figure 10.14. Trace plot of 100 bootstrapped trees

Image source: Urbanek (2008)

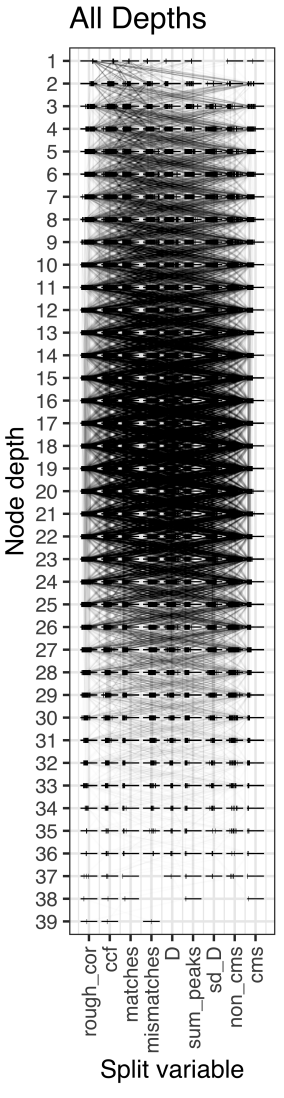
Limitations of Trace Plots

Example:

- Objective: Are two bullets fired from same gun?
- Model: Random forest (300 trees) Hare, Hofmann, and Carriquiry (2017)
- Response variable: Same gun?
- Predictor variables: 9 characteristics comparing two signatures such as cross correlation function (CCF)



Limitations of Trace Plots



Info gained

- Deep trees (max node depth of 39)
- Certain variables more commonly used for first split
- All variables commonly used between node depths of 3 and 30

Difficult to extract patterns when...

- Many trees in a forest
- Deep trees
- Large number of predictors

Methods: Extending Trace Plots

Overview

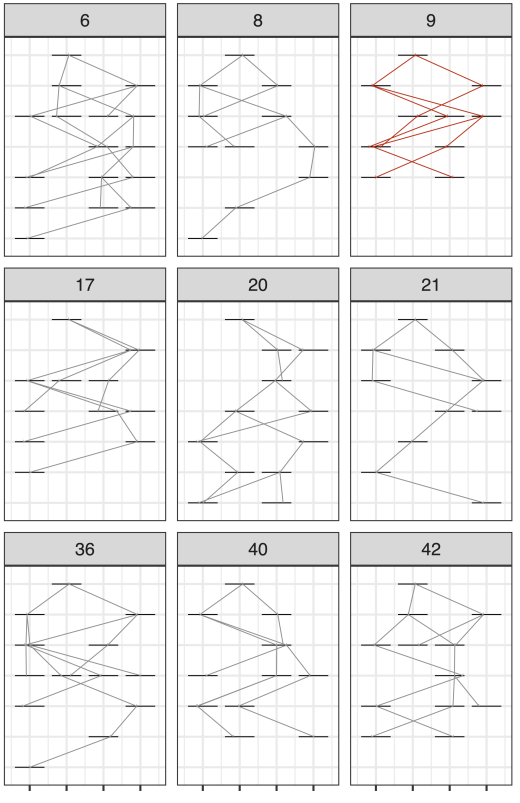
Objective: Extend trace plots to improve ability to find patterns in random forest architecture

	Intentions
Who	Data analysts
What	- Visualization of random forest architecture - One tool in toolbox for explaining random forests
When/Where	- After model training - Model assessment - Model "explanation"
Why	- Help understand how variables are used - Compare variability in split locations at different node depths - Identify patterns to explore further
How	Using `TreeTracer` R package

Approaches

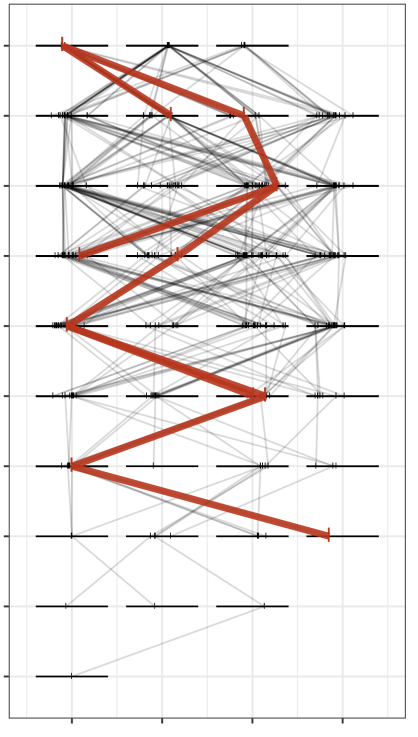
Structural Augmentations

- Highlight patterns
- Lessen cognitive load



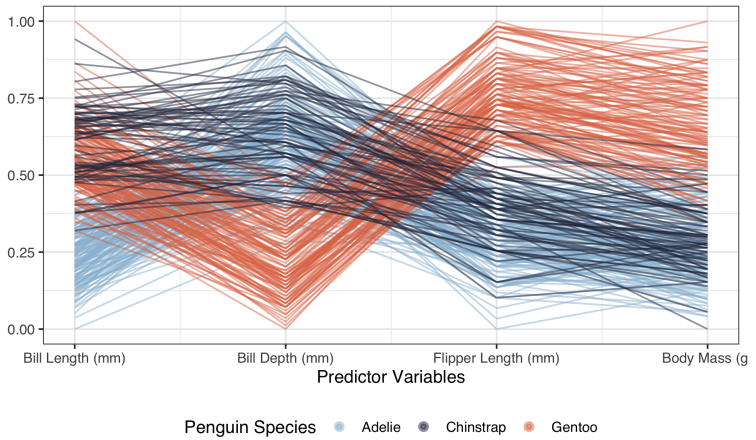
Tree Summaries

- Identify summary trees
- Re-purpose trace plots for highlighting summary trees

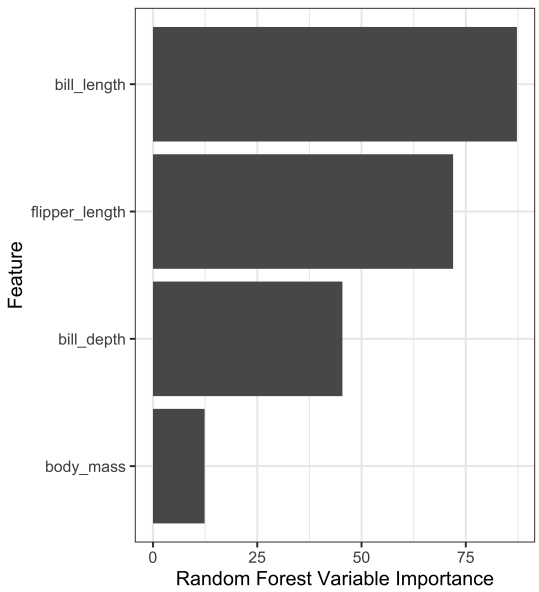


Example: Palmer Penguins

- Data: 342 penguins from Palmer Archipelago in Antarctica
- Three species: Adelie, Chinstrap, and Gentoo
- Four body measurements: Bill length, bill depth, flipper length, body mass
- Random Forest: Predict species using 50 trees



	Adelie	Chinstrap	Gentoo	Class Error
Adelie	146	4	1	0.03
Chinstrap	4	64	0	0.06
Gentoo	0	1	122	0.01



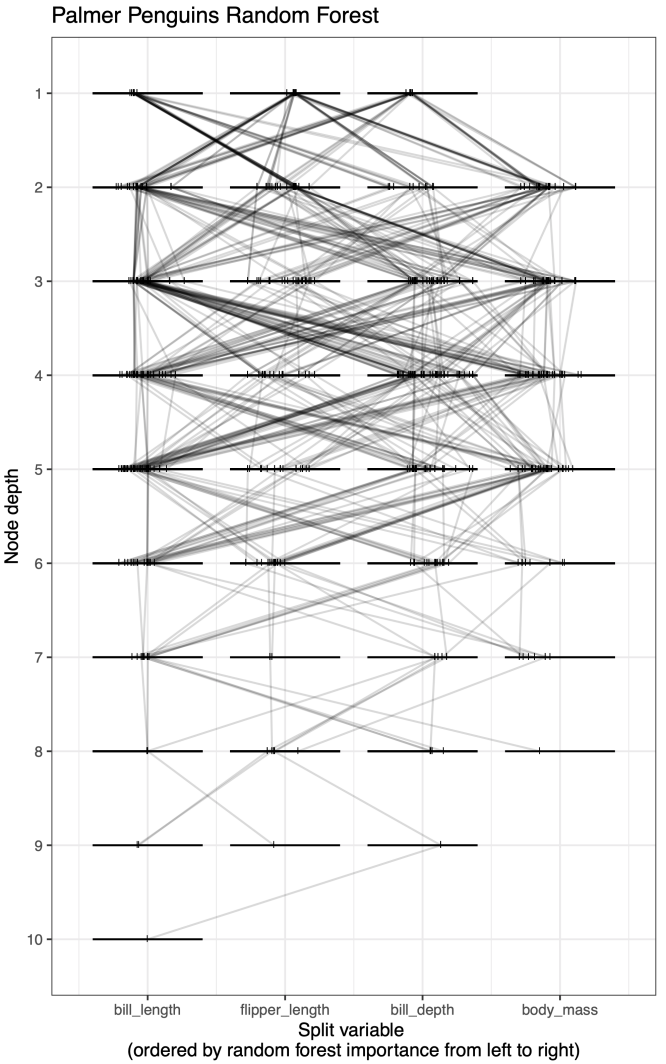
Implementation of trace plots (and extensions)

Overview

- R package `TreeTracer`
- First readily available implementation in R
- GitHub repo: https://github.com/goodekat/TreeTracer

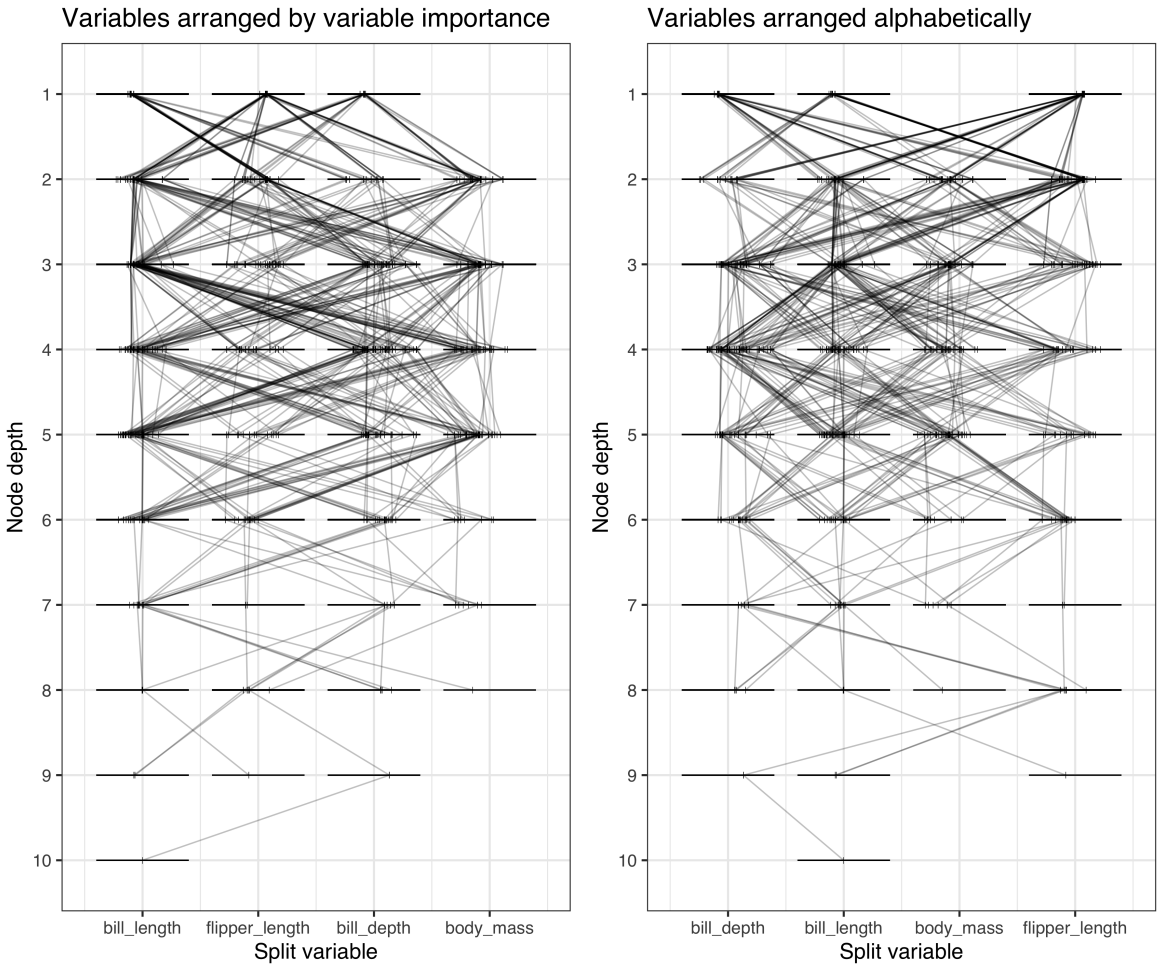
Functions

- Create trace plots from `randomForest` R package
- Structural augmentations
- Compute distances between trees



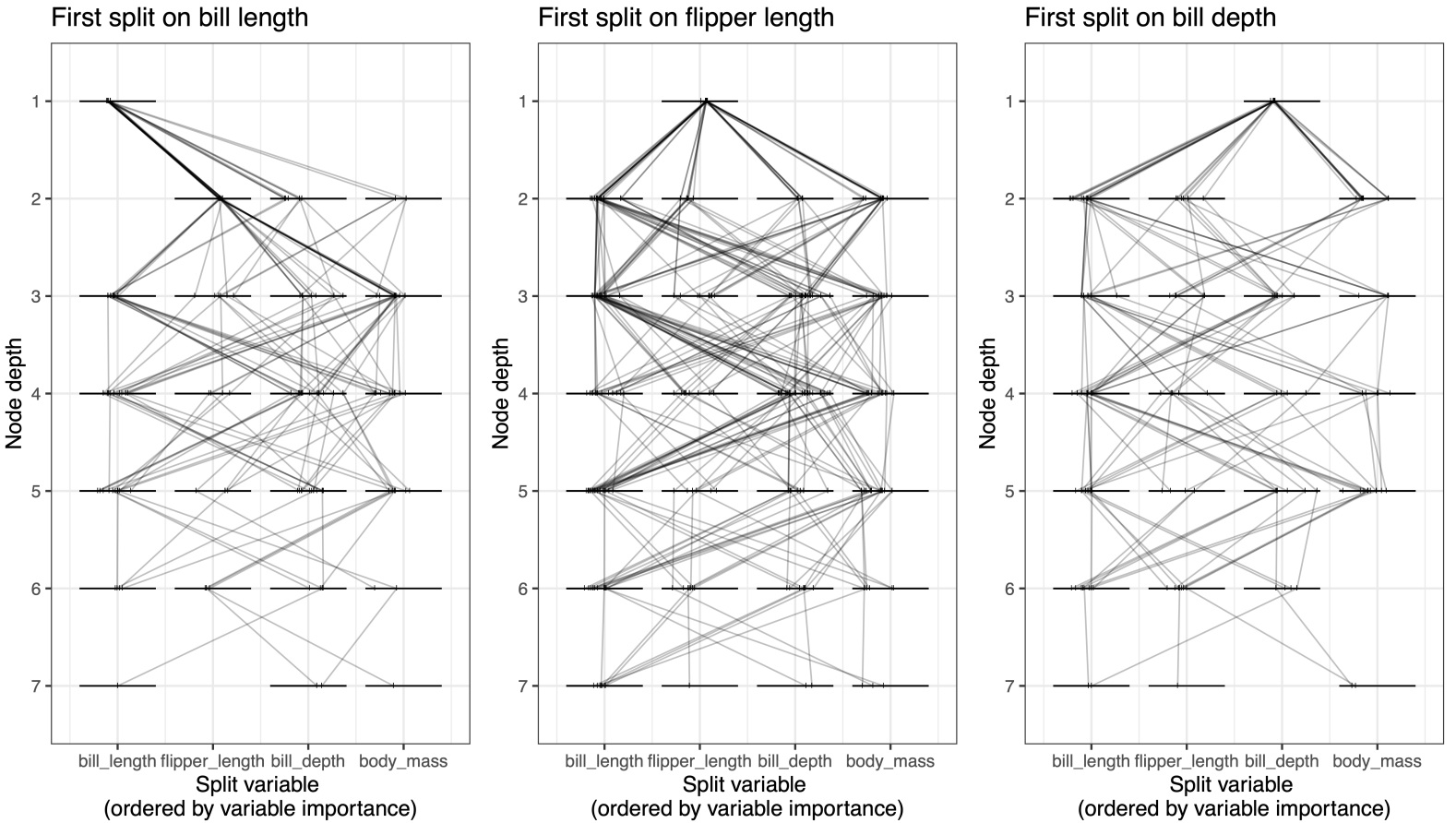
Extensions: Structural Augmentations

Ordering of split variables: Provides different perspectives



Extensions: Structural Augmentations

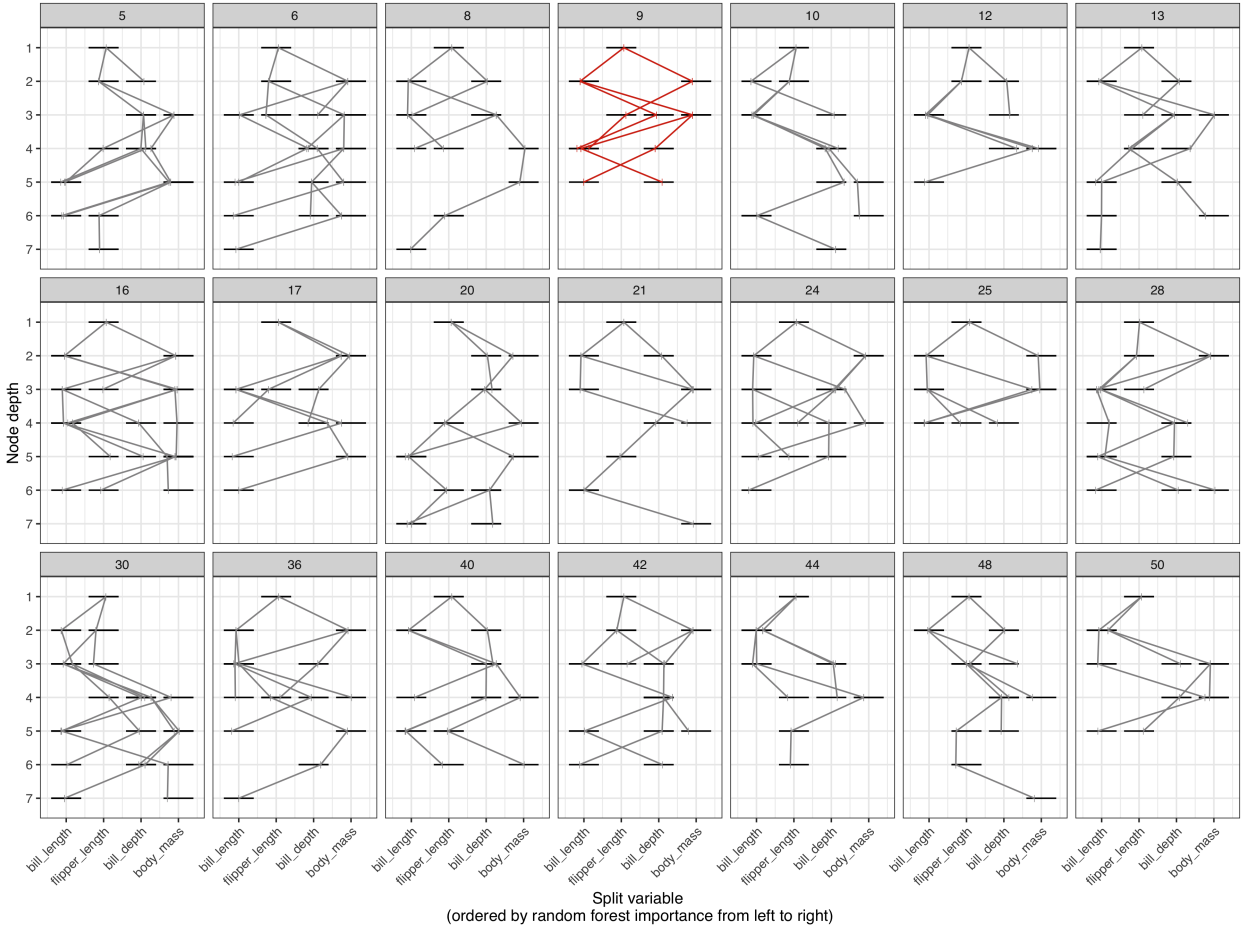
Subsets of trees: Lessen cognitive load



Extensions: Structural Augmentations

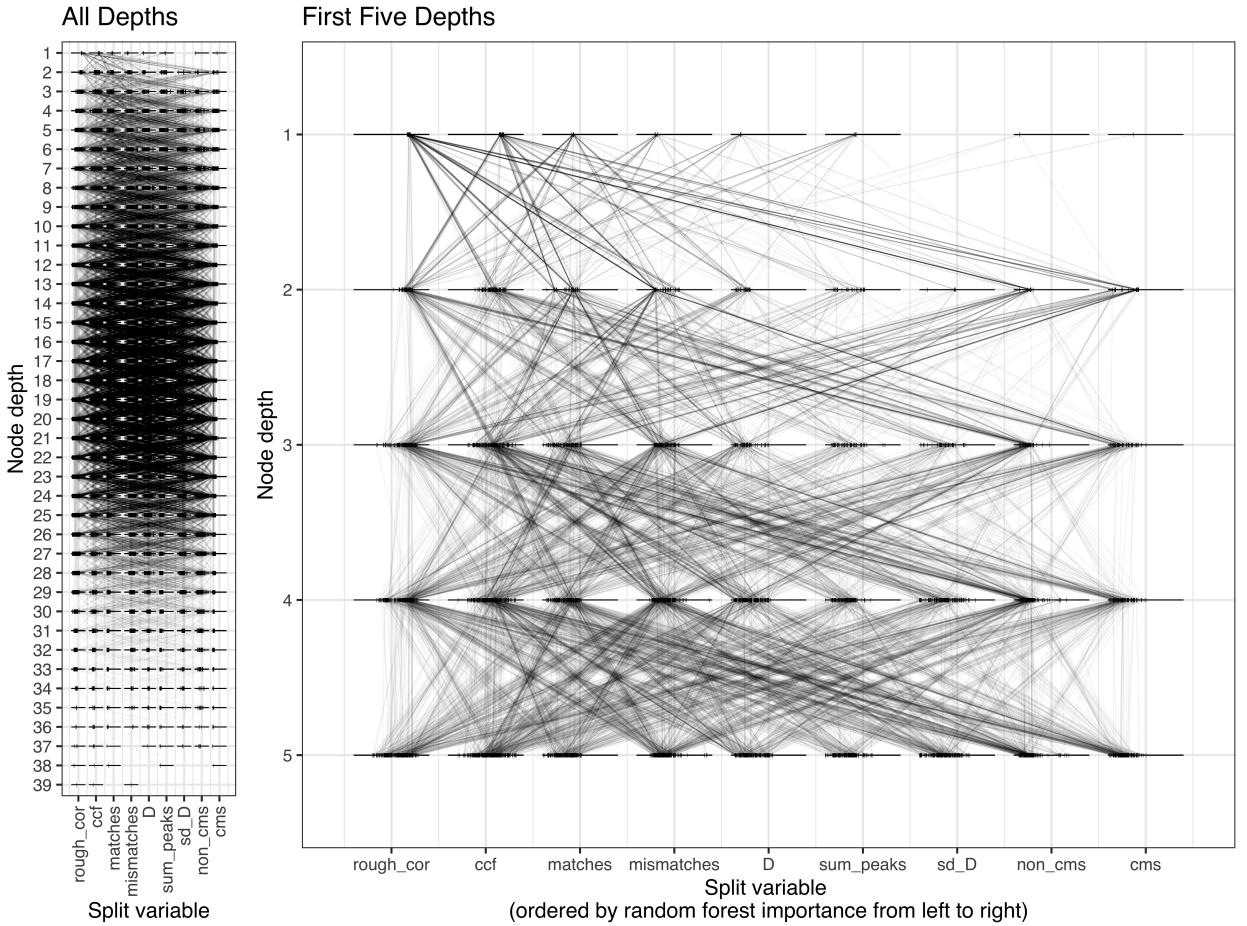
Facets: Separate trees using facets

Use of color and line size: Highlight individual or groups of trees



Extensions: Structural Augmentations

Maximum node depth: Focus on upper node depths where global structures may exist (e.g., considering the "canopy")

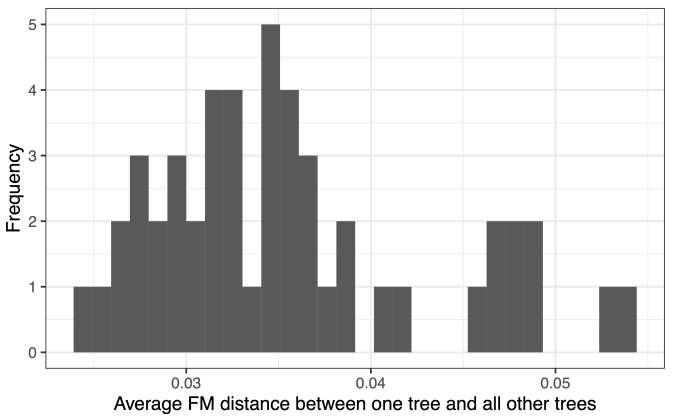


Extensions: Tree Summaries

Background (summarizing tree ensembles)

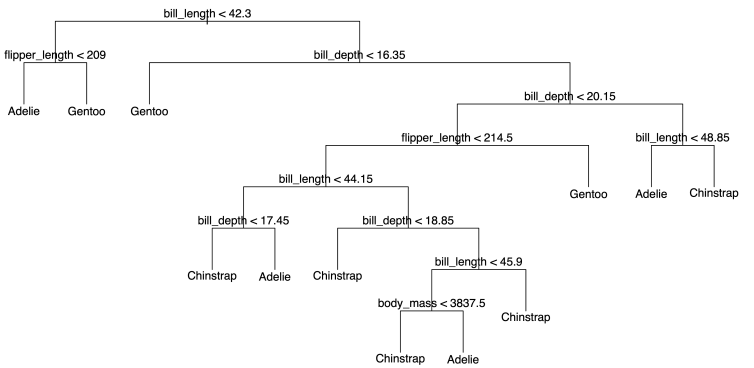
Representative tree (Shannon and Banks, 1999; Banerjee, Ding, and Noone, 2012; Weinberg and Last, 2019)

- Identify a tree that is representative of the forest
- One approach: Find tree that has smallest average distance to all other trees



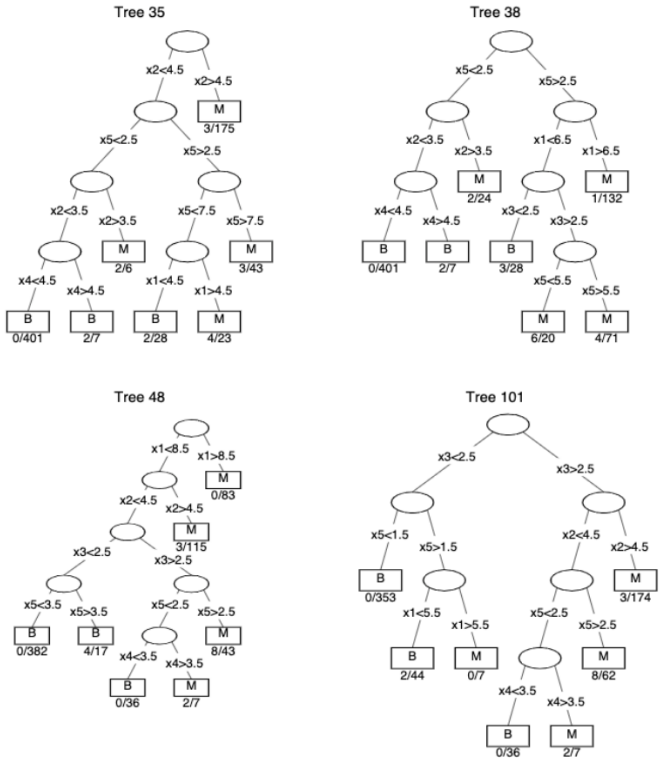
Clusters of trees (Chipman, George, and McCulloch, 1998; Sies and Mechelen, 2020)

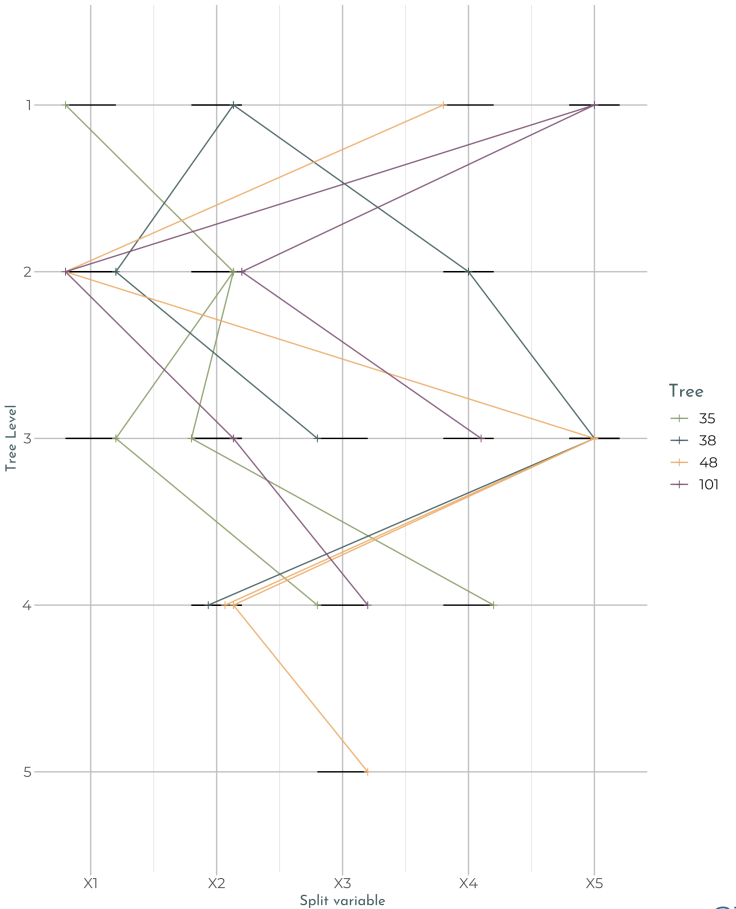
- Compute distances between trees
- Identify clusters via MDS, K-means, etc.



Extensions: Tree Summaries

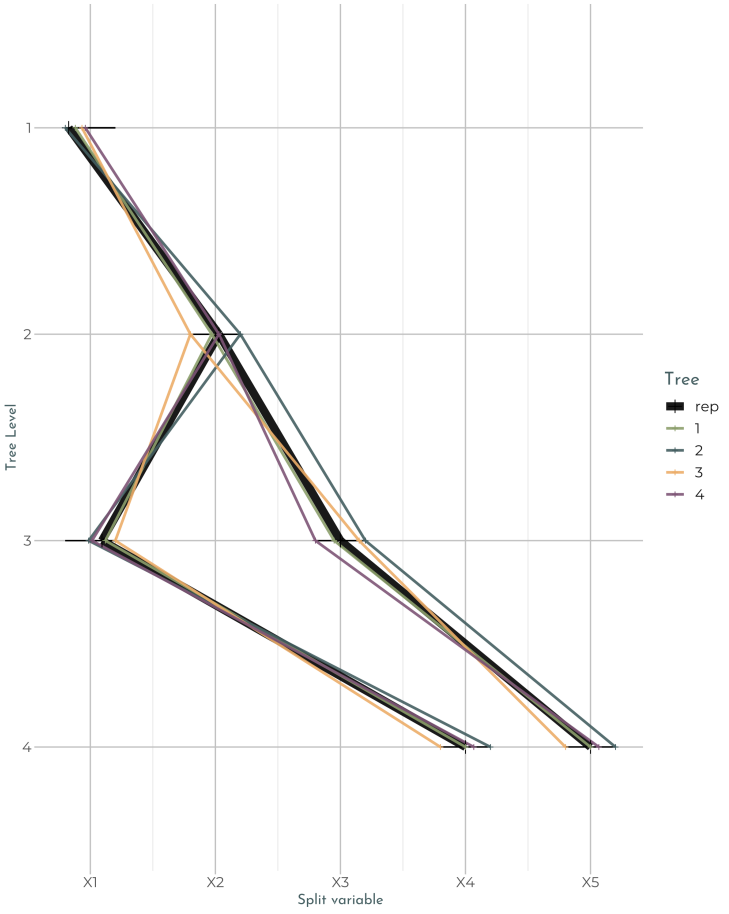
Benefits of trace plots: Example of representative trees from clusters within a tree ensemble (Chipman, George, and McCulloch, 1998; Sies and Mechelen, 2020)

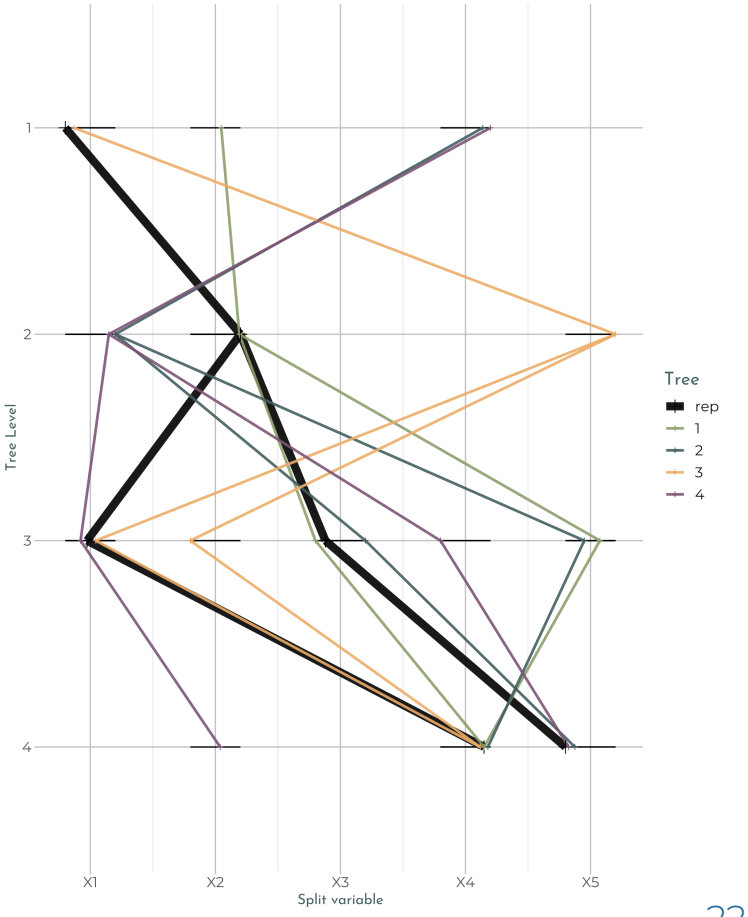




Extensions: Tree Summaries

Benefits of trace plots: Two scenarios of visualizing representative trees with variability

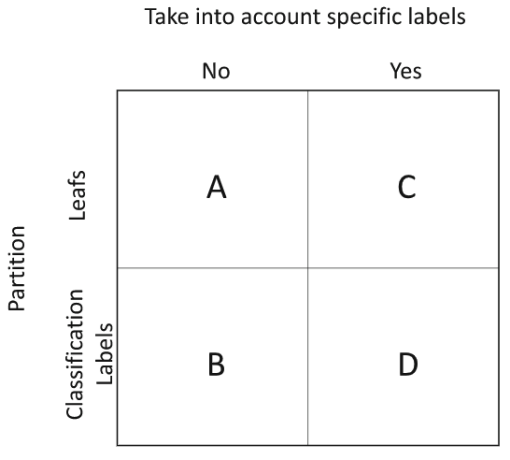




Extensions: Tree Summaries

Background (distances between trees): Various metrics proposed (Chipman, George, and McCulloch, 1998; Shannon and Banks, 1999; Miglio and Soffritti, 2004; Banerjee, Ding, and Noone, 2012; Sies and Mechelen, 2020)

Comparing Predictions



Comparing Topology

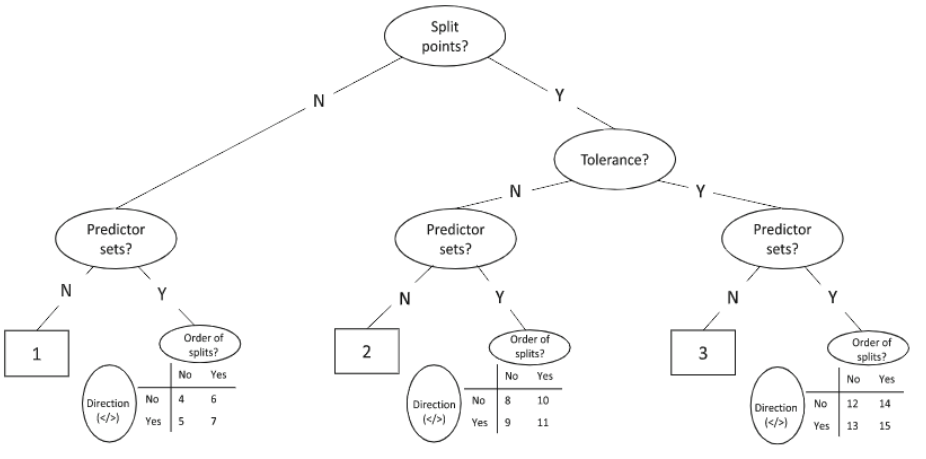


Image source: Sies and Mechelen (2020)

Extensions: Tree Summaries

Example Distance Metrics

Covariate metric: Compares split variables from two trees (Banerjee, Ding, and Noone, 2012)

$$d_{CM}(T_1, T_2) = \frac{\text{Number of covariate mismatches between } T_1 \text{ and } T_2}{k}.$$

Fit metric: Compares predictions from two trees (Chipman, George, and McCulloch, 1998)

$$d_{FM}\left(T_1, T_2\right) = \frac{1}{n}\sum_{i=1}^{n} m\left(\hat{y}_{i1}, \hat{y}_{i2}\right)$$

Partition metric: Compares how observations are divided between leaves (Chipman, George, and McCulloch, 1998)

$$d_{PM}\left(T_1, T_2\right) = \frac{\sum_{i>j} \left|I_1(i,j) - I_2(i,j)\right|}{\binom{n}{2}}$$

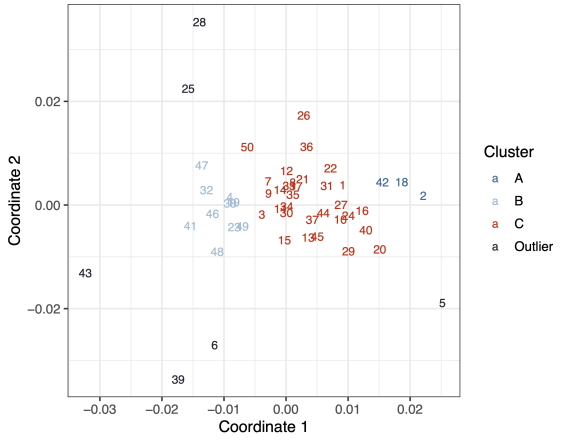
$$I_t(i,j) = \begin{cases} 1 & \text{if } T_t \text{ places observations } i \text{ and } j \text{ in the same terminal node} \\ 0 & \text{o.w.} \end{cases}$$

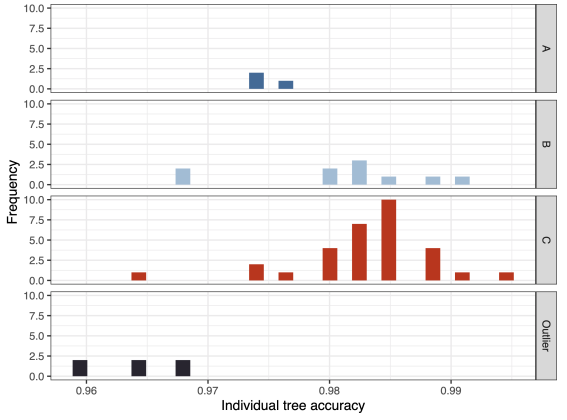
Details:

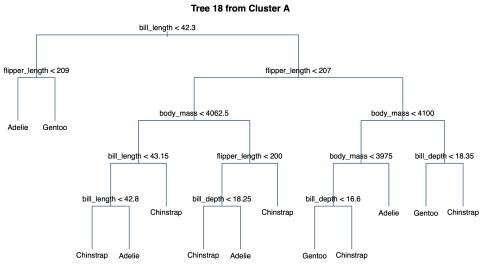
- Observation: i with $i \in \{1, \ldots, n\}$ or j with $j \in \{1, \ldots, n\}$
- Response: y_i
- Predictor variables: $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})$
- Fitted value: $\hat{y}_{it}$
- Trees: T_t with $t \in \{1, 2\}$
- Metric: m
  - Regression: $m\left(\hat{y}_{i1}, \hat{y}_{i2}\right) = \left(\hat{y}_{i1} - \hat{y}_{i2}\right)^2$
  - Classification: $m\left(\hat{y}_{i1}, \hat{y}_{i2}\right) = \begin{cases} 1 & \text{if } \hat{y}_{i1} \neq \hat{y}_{i2} \\ 0 & \text{o.w.} \end{cases}$

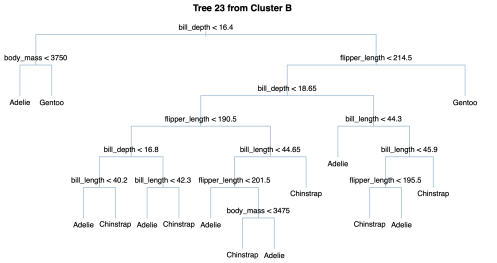
Extensions: Tree Summaries

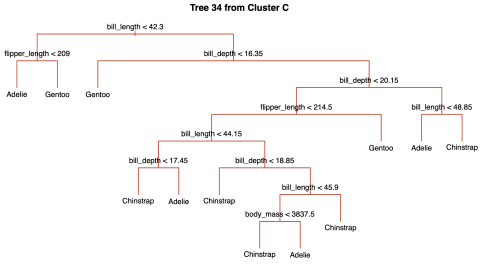
Penguins Example: Clusters identified using *multidimensional scaling* with fit metric and representative trees from clusters based on smallest average fit metric distance to all other trees in cluster





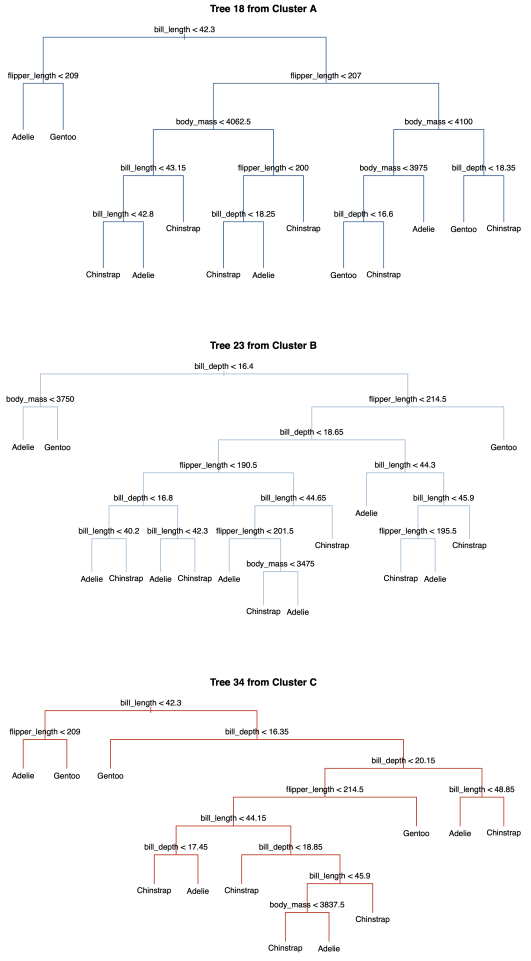


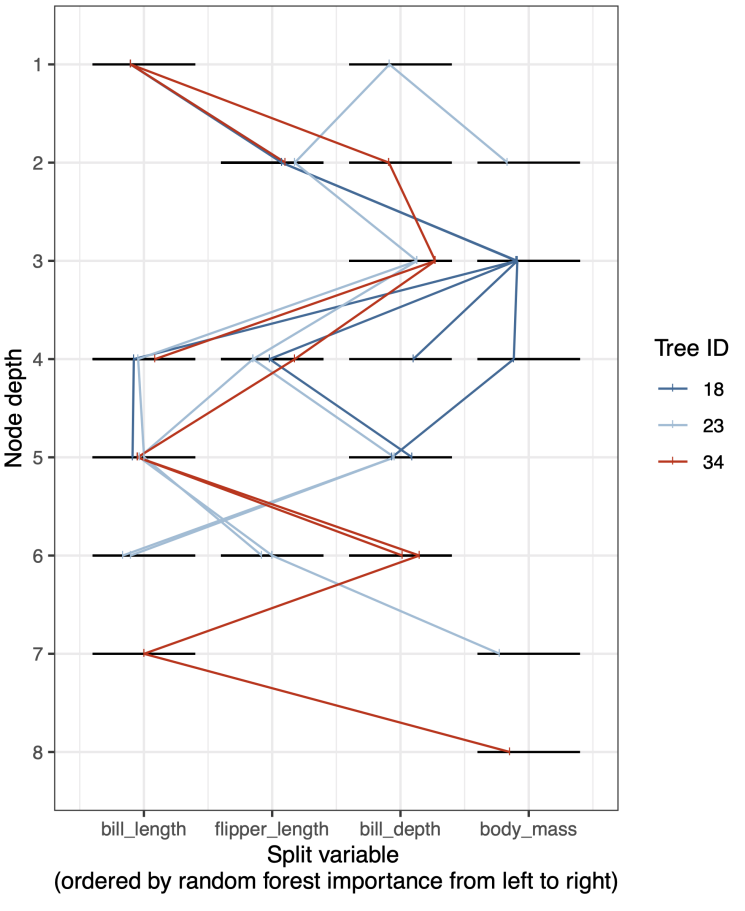




Extensions: Tree Summaries

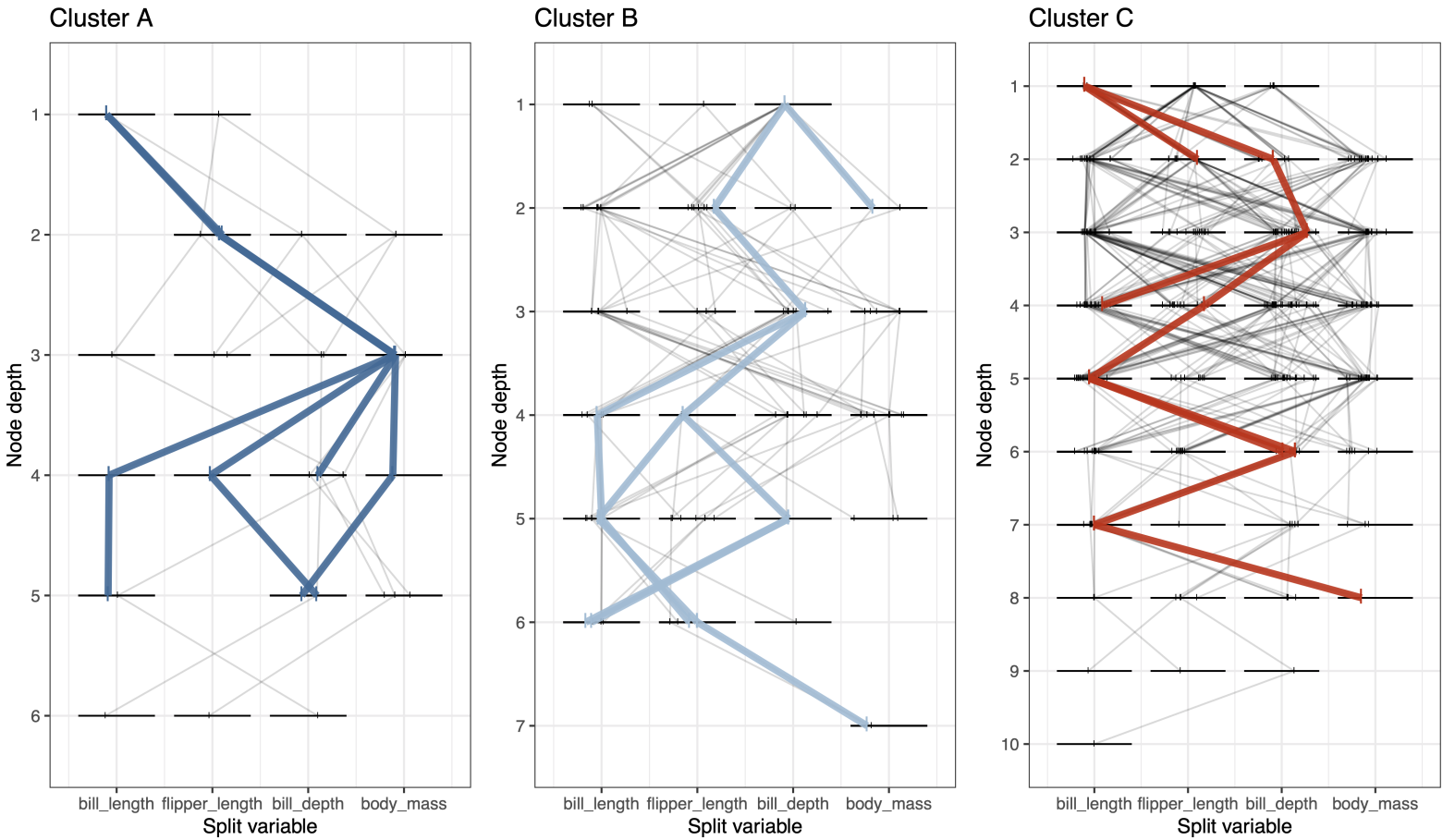
Example 1: Visualizing representative trees with a trace plot





Extensions: Tree Summaries

Example 2: Incorporating variability within a cluster



Music Example: Application with "larger" random forest

Music Example

Objective/Response:

- Predict song genre of 40 songs

Features

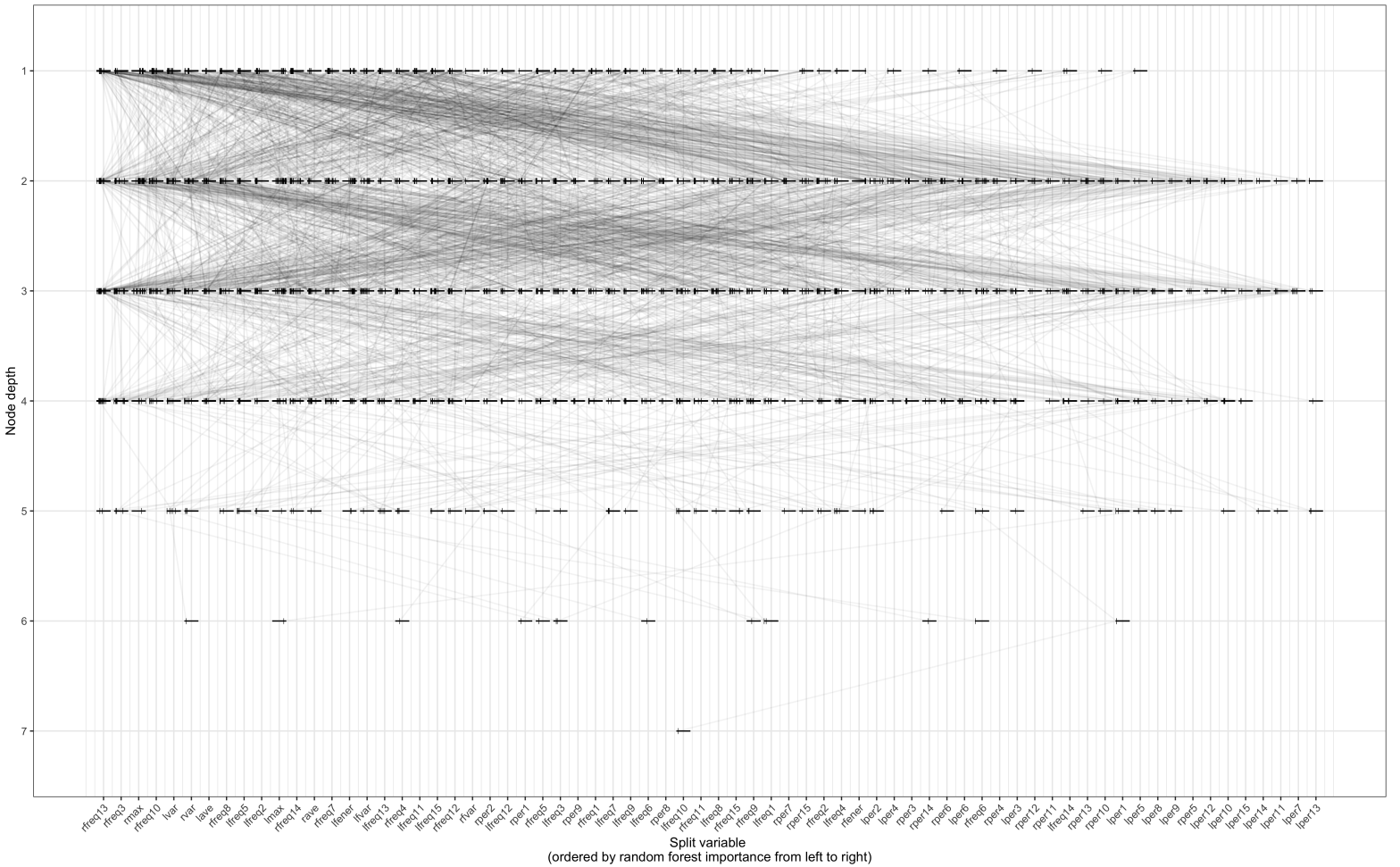
- 70 numeric variables
- Extracted from WAV files (Cook and Swayne, 2007)
- Ex: left and right channel frequencies

Model

- Random forest (`randomForest` R package)
- Default tuning parameters (e.g., 500 trees)
- Out-of-bag class errors:
  - Classical = 0.15
  - New wave = 0.67
  - Rock = 0.24

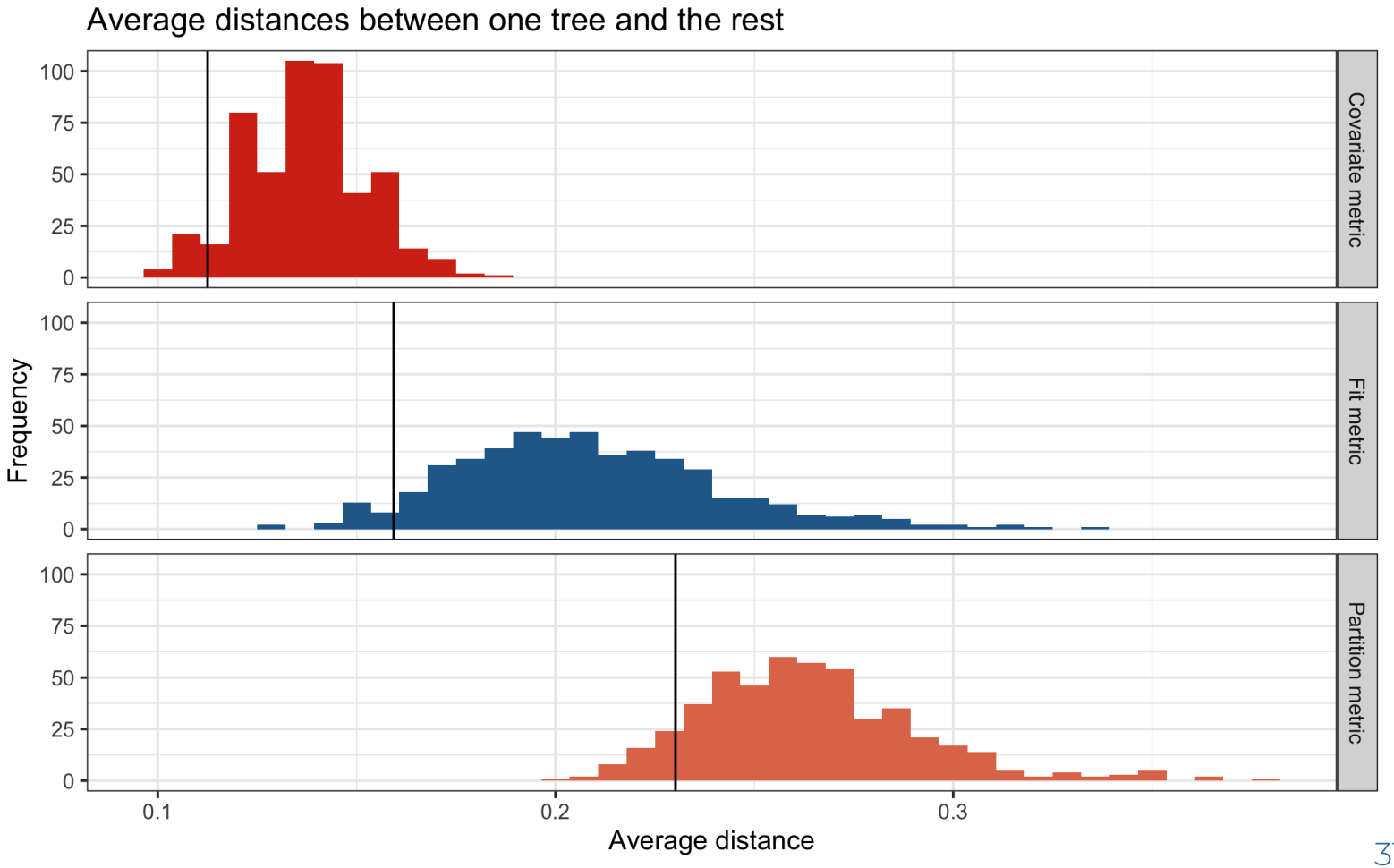
Genre	Artist	Number of Songs
Classical	Beethoven	6
Classical	Mozart	5
Classical	Vivaldi	9
New wave	Enya	3
Rock	Abba	6
Rock	Beatles	6
Rock	Eels	5

Trace Plot of Model

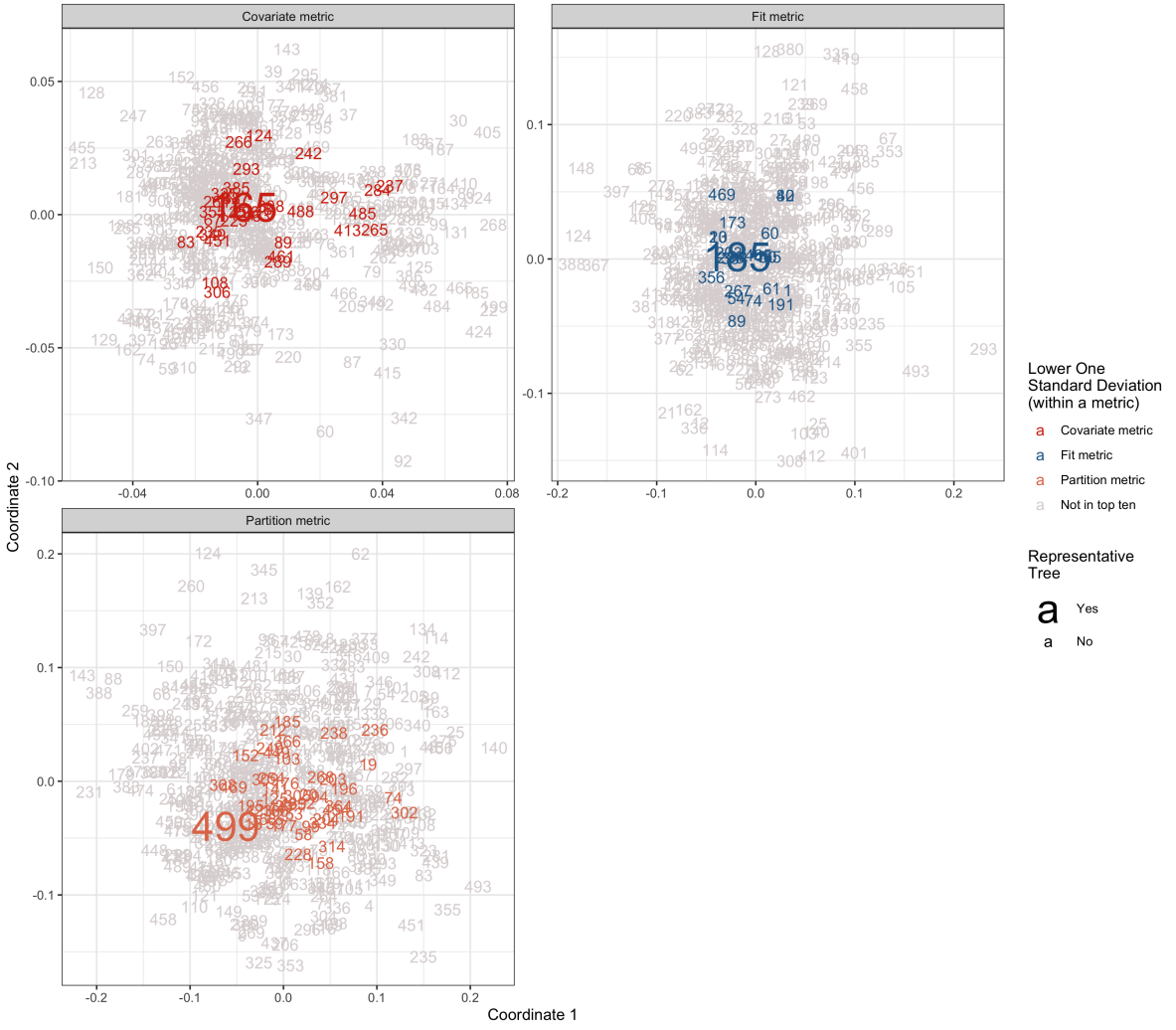


Average Distances

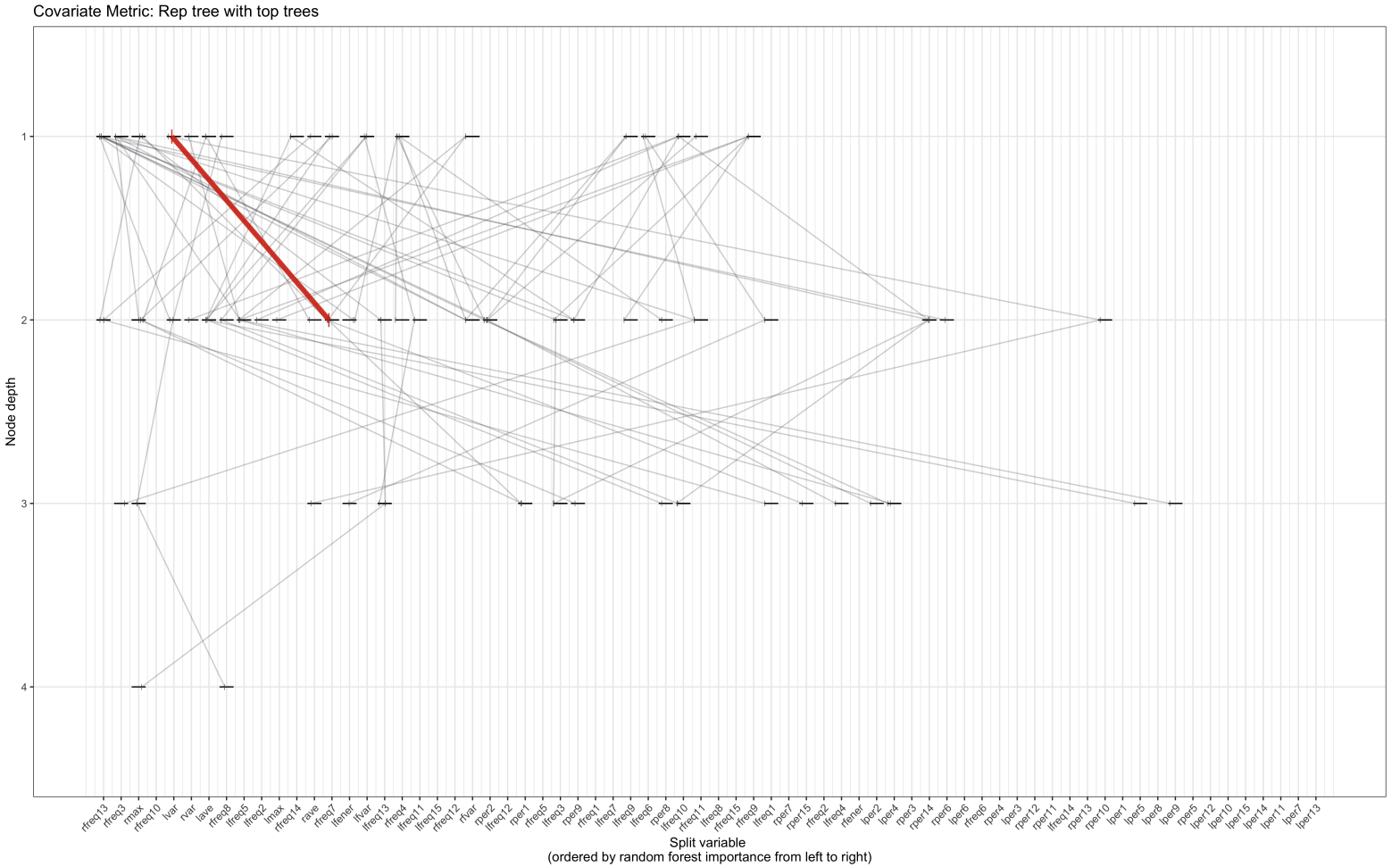
Vertical lines indicate location of smallest average distance plus one standard deviation of distances for a metric



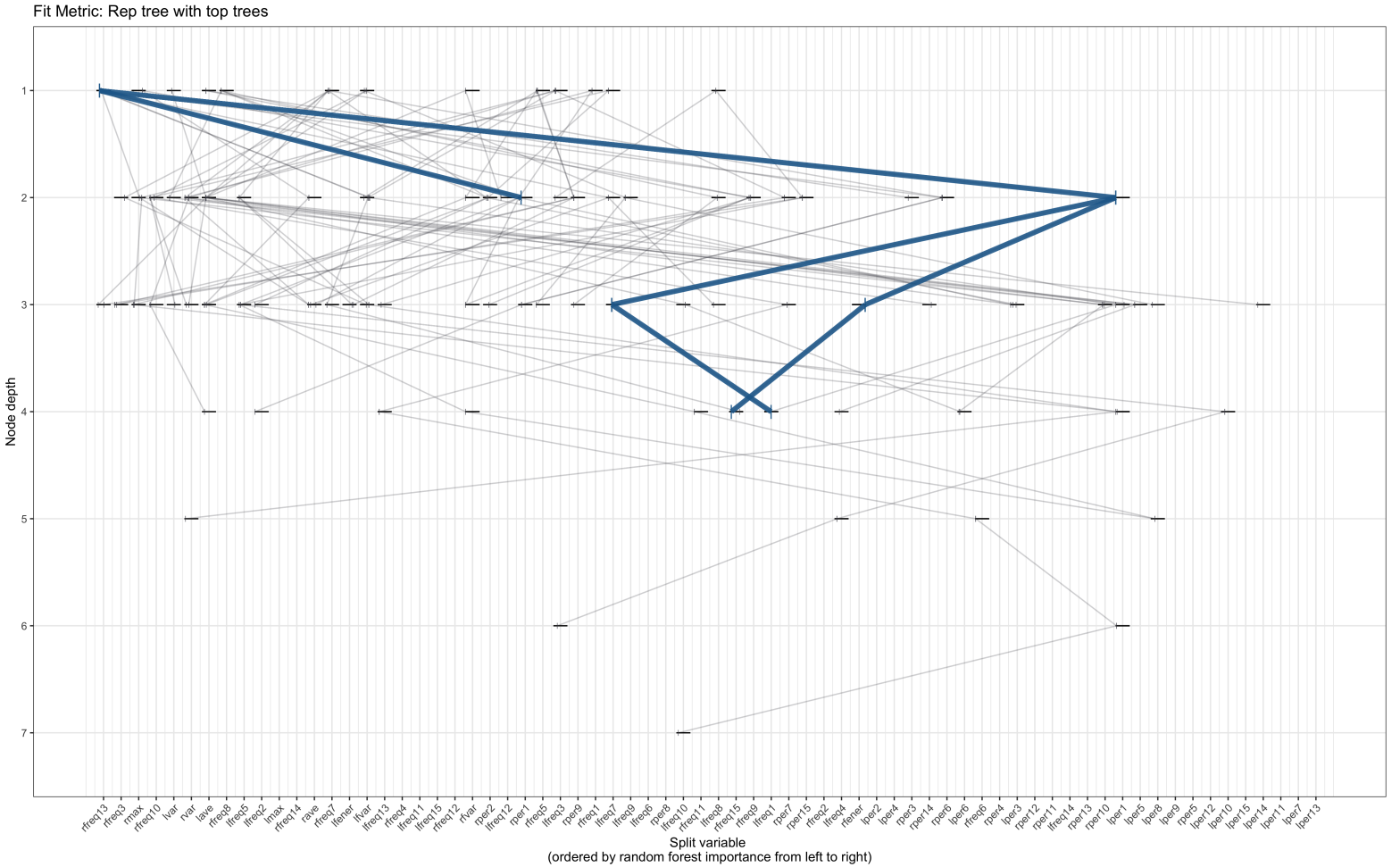
MDS Results


Covariate metric
Fit metric
Partition metric
Coordinate 2
Coordinate 1
Lower One Standard Deviation (within a metric)
a Covariate metric
a Fit metric
a Partition metric
a Not in top ten
Representative Tree
a Yes
a No

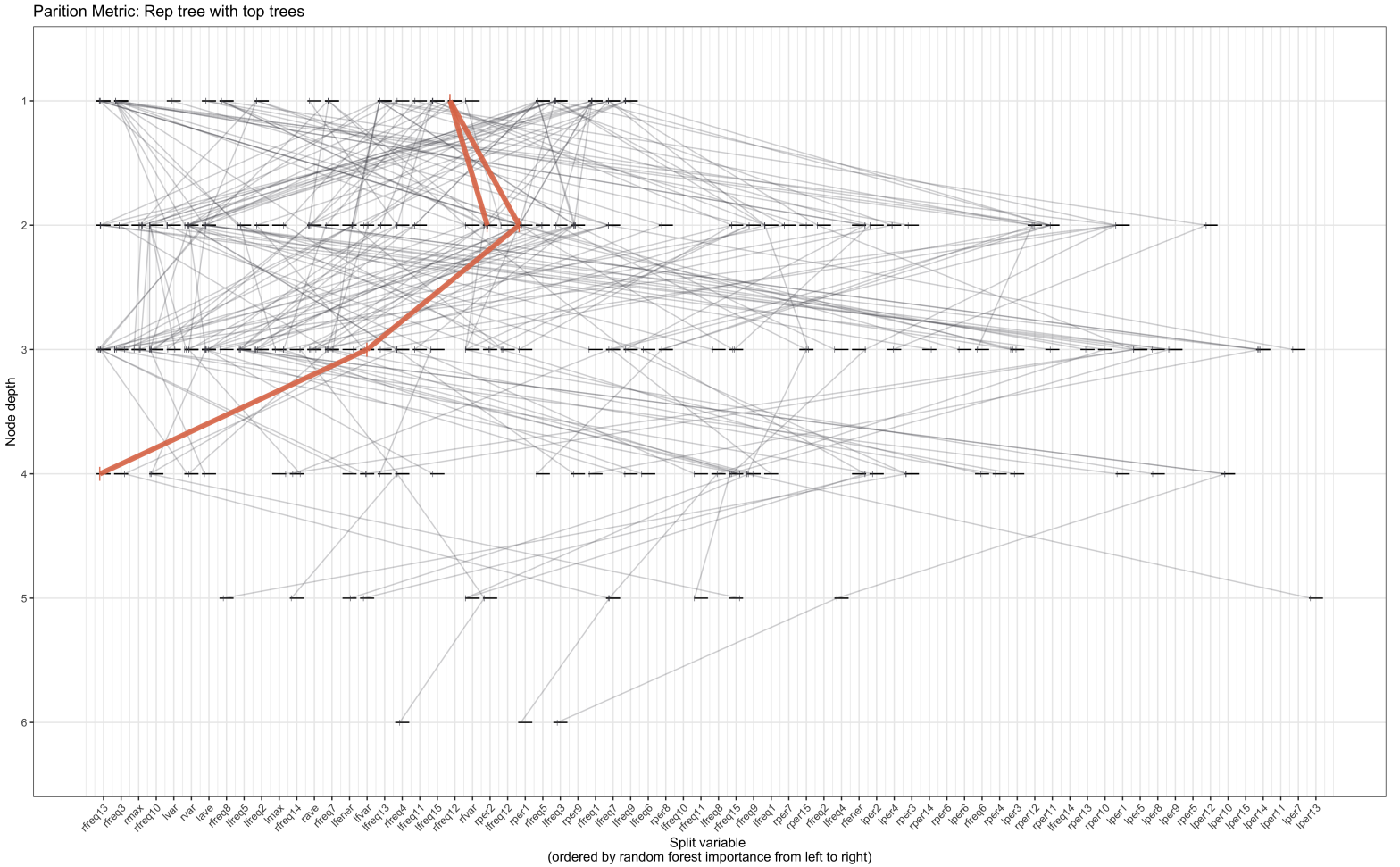
Covariate Metric



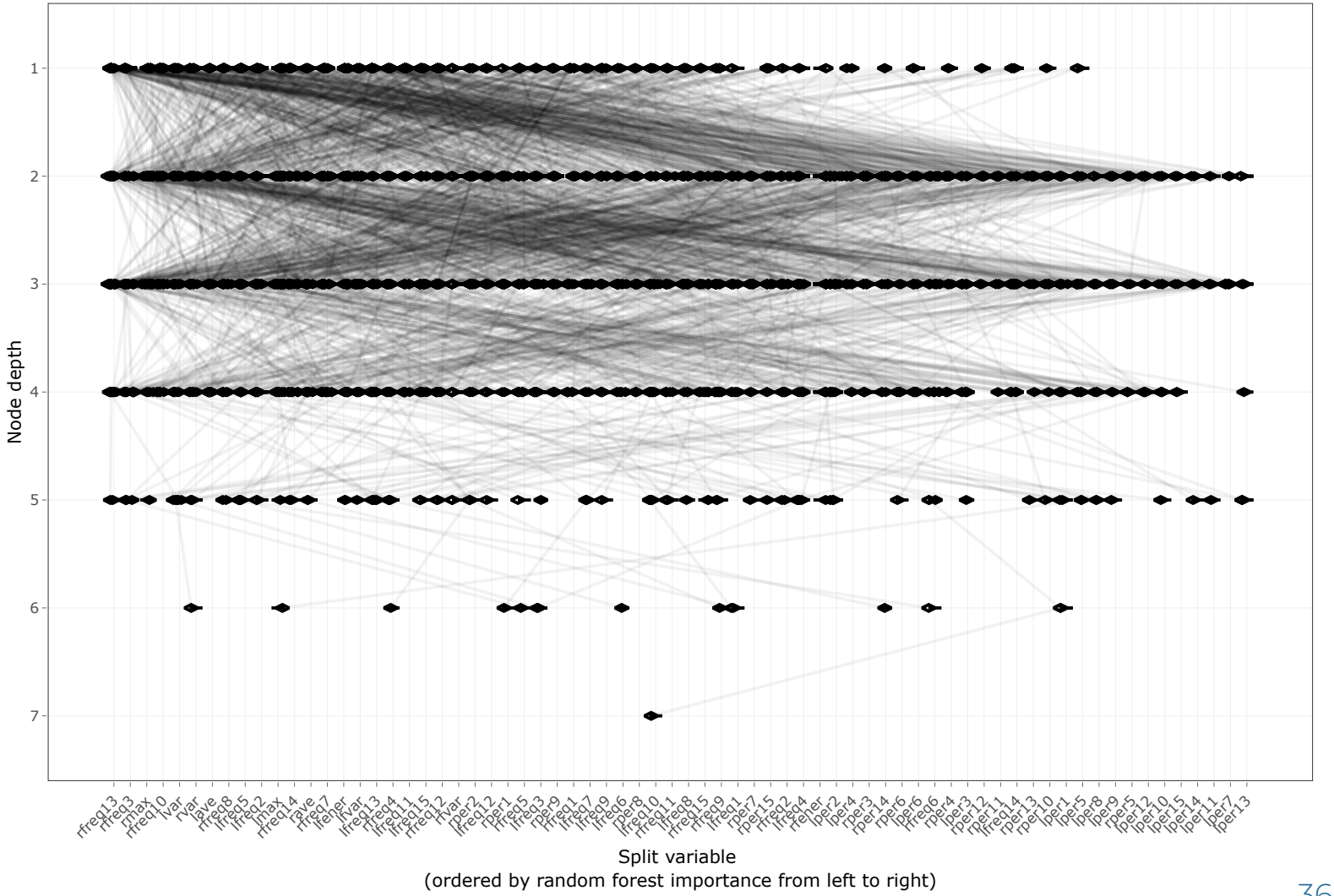
Fit Metric


Fit Metric: Rep tree with top trees
Node depth
1
2
3
4
5
6
7
rfreq13 rfreq3 rmax rfreq10 lvar rvar lave rfreq8 lfreq5 lfreq2 lmax rfreq14 rave rfreq7 lfener lfvar lfreq13 rfreq4 lfreq11 lfreq15 lfreq12 rfvar rper2 lfreq12 rper1 rfreq5 lfreq3 rper9 rfreq1 lfreq7 lfreq9 lfreq6 rper8 rfreq10 rfreq11 lfreq8 rfreq15 rfreq9 lfreq1 rper7 rper15 rfreq2 lfreq4 rfener lper2 lper4 rper3 rper14 rper6 lper6 rfreq6 rper4 lper3 rper12 rper11 lfreq14 rper13 rper10 lper1 lper5 lper8 lper9 rper5 lper12 lper10 lper15 lper14 lper11 lper7 lper13
Split variable
(ordered by random forest importance from left to right)

Partition Metric



Interactive Version



Conclusions: Pros, Cons, and Possible Research Directions

Summary

Proposed trace plot extensions

- Structural augmentations
- Repurpose trace plots for visualizing tree summaries

Implemented trace plots

- *TreeTracer* R package

Benefits of trace plot extensions

- Help extract patterns from random forest architectures
- Inspire new questions and hypotheses

Strenths and Weaknesses

Strengths

- Added organization of traces
- Reduction in the cognitive load
- Increased ability to visually compare trees

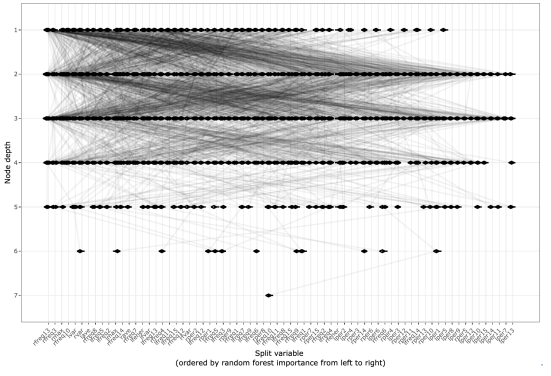
Weaknesses

- Simplification leads to loss of information
  - May be worthwhile to view signal among noise
  - May present a view that is not practically helpful
- Not simplified enough
  - Too much information to expose patterns
- Finding optimal balance
  - Can be challenging
  - Dependent on model

Future Work

Interactivity

- Link trace plot to visualizations focused on more nuanced aspects of random forests:
  - Click on intersection of node depth and split variable
  - Produces plot of split in data space
- Zoom in on large trace plots



Computation

- R package for management of tree data
- Create a geom for trace plots
- Implementation in Python

Other

- Color branches based on dominate class or average value of observations
- How to select maximum depth?
- Consider other metrics more focused on topology

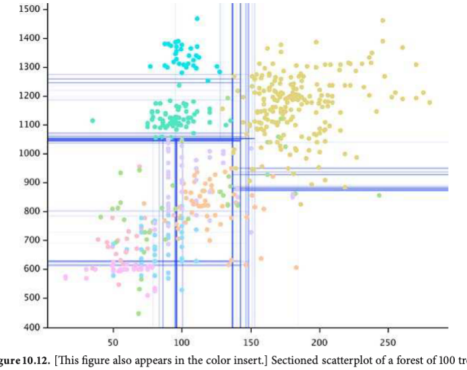


Figure 10.12. [This figure also appears in the color insert.] Sectioned scatterplot of a forest of 100 trees

Sectioned scatter plot image source: Urbanek (2008)

References

Banerjee, M., Y. Ding, and A. Noone (2012). "Identifying representative trees from ensembles". In: *Statistics in Medicine* 31.15, pp. 1601-1616. ISSN: 1097-0258. DOI: 10.1002/sim.4492.

Chipman, H. A., E. I. George, and R. E. McCulloch (1998). "Making sense of a forest of trees". In: *Proceedings of the 30th Symposium on the Interface.* , pp. 84-92. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.2598.

Cook, D. and D. F. Swayne (2007). *Interactive and Dynamic Graphics for Data Analysis, With R and Ggobi*. 1st ed. Springer-Verlag New York. ISBN: 9780387717616. DOI: 10.1007/978-0-387-71762-3.

French, J. W., R. B. Ekstrom, and L. A. Price (1963). *Kit of reference tests for cognitive factors*. Educational Testing Service. Princeton, NJ.

Hare, E., H. Hofmann, and A. Carriquiry (2017). "Automatic matching of bullet land impressions". In: *Annals of Applied Statistics* 11.4, pp. 2332-2356. DOI: 10.1214/17-AOAS1080.

Kuznetsova, N. (2014). "Random forest visualization". Supervised by Michel Westenberg. Eindhoven, Netherlands.

Miglio, R. and G. Soffritti (2004). "The comparison between classification trees through proximity measures". In: *Computational Statistics & Data Analysis* 45.3, pp. 577-593. ISSN: 0167-9473. DOI: 10.1016/s0167-9473(03)00063-x.

Shannon, W. D. and D. Banks (1999). "Combining classification trees using MLE". In: *Statistics in Medicine* 18.6, pp. 727-740. ISSN: 1097-0258. DOI: 10.1002/(sici)1097-0258(19990330)18:6<727::aid-sim61>3.0.co;2-2. URL: https://onlinelibrary.wiley.com/doi/epdf/10.1002/%28SICI%291097-0258%2819990330%2918%3A6%3C727%3A%3AAID-SIM61%3E3.0.CO%3B2-2.

Sies, A. and I. V. Mechelen (2020). "C443: a Methodology to See a Forest for the Trees". In: *Journal of Classification* 37.3, pp. 730-753. ISSN: 0176-4268. DOI: 10.1007/s00357-019-09350-4. URL: https://link.springer.com/article/10.1007/s00357-019-09350-4.

Urbanek, S. (2008). "Visualizing Trees and Forests". In: *Handbook of Data Visualization*. Ed. by C. Chen, W. Härdle and A. Unwin. Vol. 3. Berlin, Germany: Springer-Verlag, pp. 243-266. ISBN: 9783540330363. URL: https://haralick.org/DV/Handbook_of_Data_Visualization.pdf.

Weinberg, A. I. and M. Last (2019). "Selecting a representative decision tree from an ensemble of decision-tree models for fast big data classification". In: *Journal of Big Data* 6.1, p. 23. DOI: 10.1186/s40537-019-0186-3. URL: https://link.springer.com/article/10.1186/s40537-019-0186-3.

Thank you!