# Variable Depth Random Decision Trees
# Growing Branches

Duke University Synthetic Data Project

November 14, 2018

## Background

Variable depth random decision trees are constructed by terminating or continuing branch splits based on a rule other than simply discontinuing at a fixed depth. A rule was implemented that terminates node splits when the observation count associated with a branch, as measured in the DIBBS synthetic data, diminishes to a specified threshold ($n_L$). This results in generation of variable depth branches with a guaranteed minimum observation count, which is beneficial in computing proportion synthetic label confidence intervals using the Laplace-binomial probability distributions appearing in terminating leaves. Each label of the attribute selected to split a given node generates a child node, forming a new branch, and it is desirable, from a data utility perspective, for high frequency branches to be extended as far as possible. Using the minimum leaf frequency rule, a terminating leaf must be appended to a branch prior to a split that would cause branch frequency to fall below $n_L$. But, this is at the same level as nodes with sufficient frequency to be split, causing an ambiguity. Options for treating this, along with associated deficiencies, are:

- Combine all observations for all low frequency labels into a single node. Deficiencies: combined frequency may not exceed the minimum $n_L$. Also, combined labels are removed from the leaf pmf support, so that synthetic labels will not be generated with them.
- Discontinue splitting of all nodes when the frequency of any one label of the attribute chosen for splitting would fall below $n_L$. Deficiency: a limit to branch depth is imposed by minimum frequency, and possibly least representative, labels.
- Construct a leaf from the subset of observations corresponding to the parent node whenever splitting a node would cause branch frequency to fall below $n_L$. Deficiency: this violates the requirement imposed by differential privacy that probability distributions used to synthesize data (from leaves here) represent and be derived from disjoint subsets of data.

## The $n_B$ Branch Growth Rule

Due to the deficiencies mentioned, an alternative to the $n_L$ rule was considered, such that a branch is extended beyond a given level provided that branch frequency at that level, as measured in the DIBBS synthetic data, is equal to or above a specified threshold ($n_B$). The new rule remedies $n_L$ rule deficiencies, but does not guarantee a minimum leaf frequency, making definite confidence statements regarding proportion label synthesized problematic. While constructing trees with the $n_L$ rule, it was observed that placing a predictor attribute early in the order of selection (depth or level at which an attribute appears in a branch) tends to improve agreement of proportion synthetic and authentic observations relative to labels of that attribute. Figures 1 through 3 demonstrate, for $n_B = 1,000$, improved utility through use of variable, instead of fixed, depth branches while controlling the order of attribute selection. Observations:

- Figure 1 compares pairwise slopes and correlation of proportion observations for all synthesized attributes and labels with both Sex labels (female and male). Fixed depth results are on the left, those for variable depth are on the right. Each joint attribute/label combination is represented by a point. Points with coordinates near the reference line of slope 1.0 indicate close agreement in corresponding

proportion observations. Of note is an improvement in agreement for nearly every attribute and label generated by the variable depth ensemble. Note that attribute selection at each node level was uniformly random (all sampling weights = 1), making observed improvement due solely to use of variable depth branches.

- Figure 2 Compares proportion slopes and correlation with prioritized sampling of Agency (50 times the sampling weight of remaining attributes). A significant improvement in agreement of Agency proportions is observed in both plots, with an associated reduction in agreement for remaining attributes. Proportion agreement is improved in the variable depth panel.

- Figure 3 confirms the benefit of increased attribute sampling weight, with Agency returning to standard weight and Age and Supervisory Status advancing to 50 times that of remaining attributes. A reduction in Agency proportion agreement, along with improvement for both Age and Supervisor Status are observed.

## Branch Depth, Leaf Frequency, and Proportion Label Synthesized Confidence Intervals

Each additional level, or depth, of a tree further subsets observations, such that increased branch depth results in reduced observation frequency in leaves. As previously stated, the $n_B$ rule does not guarantee leaf frequency, which raises the question of exactly how leaf frequency is influenced by choice of $n_B$. Figure 4 plots, for $n_B = 1,000$ and equally weighted attribute sampling, by individual tree in an ensemble of five, the distribution of branch depth (top), the distribution of leaf frequency (center), and individual branch leaf frequency given branch depth (bottom). Observations:

- Approximately 85% of branches are of depth four, five, or six - depths that are greater than the value of three that is suggested for a fixed depth tree constructed from six predictor attributes.[1] This suggests that, as constructed here, a typical variable depth branch is one to three levels deeper than its corresponding fixed depth branch. Since depth is equal to the number of attributes describing associated data, a longer branch is expected to better predict leaf proportions. However, due to increased specificity, extending branches to include (nearly) all possible attributes could be expected to reduce generality and cause overfitting of trees to the training data (authentic OPM in our case).

- Approximately 85% of leaves have frequency less than 100, which is significantly less than $n_B = 1,000$, indicating that the final split of a branch creates a relatively small observation subset. Although tempting to consider $n_B$ as providing reliable protection against low frequency leaves, it is clear that confidence intervals of proportion synthetic observation generated by label, computed using $\frac{n_B}{10}$ (our situation), are significantly larger than those computed using $n_B$ (further discussion synthetic label proportion confidence intervals follows).

- As expected, and as observed in the leaf frequency by branch depth plot, branch depths greater than $n_B$ occur only with branches of maximum depth (six). For branch depths between two and four, the shift in high density leaf frequency (heavy blue regions) with respect to branch depth is a rather small proportion of $n_B$, indicating relative independence of leaf frequency and these branch depths.

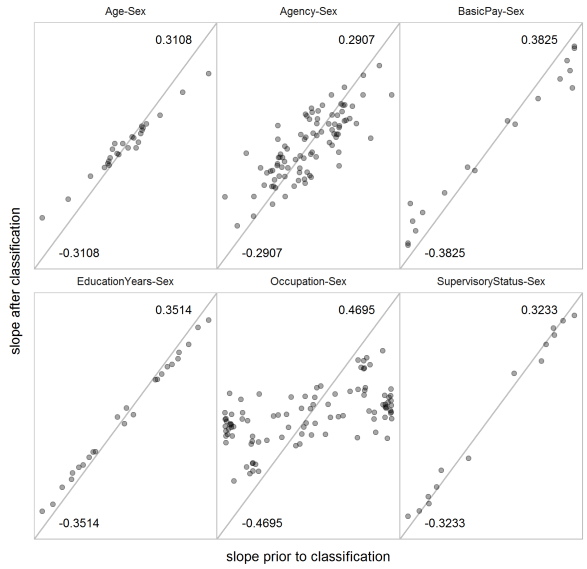
From the proportion synthetic label generated confidence interval plot (figure 9), with leaf frequency = 1,000 and $\epsilon = 0.3$, proportion label generated confidence intervals range from approximately 0.01 to 0.05, depending on expected proportion generated. It is also seen in figure 9 that confidence intervals significantly expand as actual leaf frequency decreases. Concern over prediction accuracy arises when a high proportion of low frequency leaves is observed, as in those generated by ensembles constructed with $n_B = 1,000$. In our study, $n_B$ is the sole tree construction parameter available for adjustment that is expected to affect branch depth, an increase in $n_B$ generally expected to decrease depth. Given this, an ensemble of trees was constructed using node attributes, uniform sampling weights, and $\epsilon$ as before, but modifying the minimum branch frequency required to split from $n_B = 1,000$ to $n_B = 5,000$. Figure 5 confirms a more centralized distribution of branch depths along with a significant reduction in proportion low frequency leaves. Although beneficial

[1] Jagannathan, et al, suggest half the number of predictor attributes.

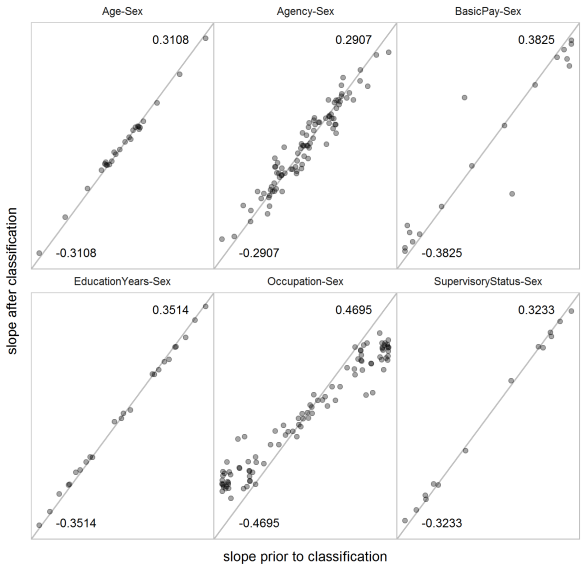
from a proportion label synthesized confidence interval perspective, a reduction in utility also results, as observed in figure 6 (when compared to figure 1 which was generated with $n_B = 1,000$). To test the other extreme, an ensemble of trees was constructed with $n_B = 250$. Figure 7 confirms an expected shift in branch depth distribution to higher depths along with a significant increase in proportion low frequency leaves. A modest increase in utility is observed with data from this ensemble, as seen in figure 8 (compared to figure 1, $n_B = 1,000$), but with a significant reduction in proportion synthesized prediction accuracy, due to an approximate four-fold increase in very low frequency leaves and associated widening of confidence intervals.

## Conclusion

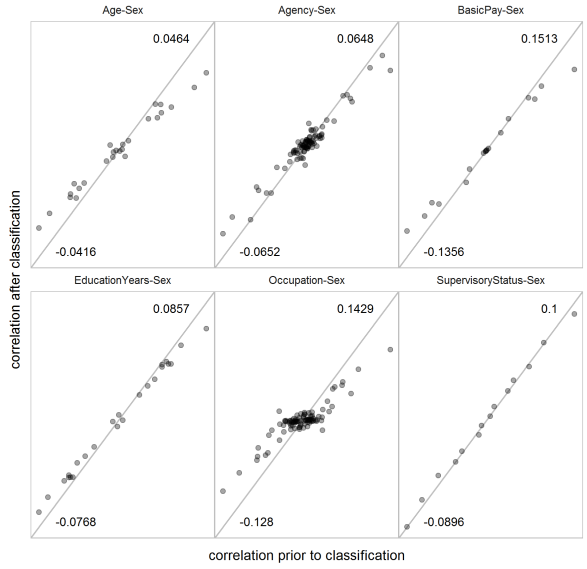
- A significant improvement in utility (agreement in joint category proportion observations between authentic and synthesized labels, Sex in this study) is observed when labels are synthesized using ensembles of variable depth trees as compared to those generated using a similar ensemble of fixed depth trees.
- Requiring a minimum leaf frequency (the $n_L$ rule) enables estimation of definite confidence intervals on proportion observations synthesized by label, but implementing the rule leads to ambiguous situations regarding leaf sampling weights and violation of differential privacy.
- Requiring a minimum branch frequency to continue splitting ($n_B$ rule) resolves $n_L$ rule deficiencies, but fails to provide definite leaf frequencies that are necessary for computing synthesized label proportion confidence intervals.
- Once an ensemble is constructed, distributions of branch depth and leaf frequency can be composed to study the relationship of $n_B$ to depth, leaf frequency, and expected proportion label confidence intervals. The ensembles used in this study were composed from both authentic and synthetic data and queries of authentic data consume $\epsilon$ (in a differentially privacy sense). An alternative is to study branch depth and leaf frequency distribution resulting from an ensemble constructed strictly using synthetic data. Assuming unlimited ability to query synthetic data, this would permit extensive experimentation and characterization of $n_B$, branch, and leaf relationships.
- Care must be exercised in assessing actual proportion label confidence intervals, given an ensemble of variable depth trees generated with $n_B$. Regardless of the value of $n_B$, actual resulting leaf frequency must be considered in estimating realistic confidence intervals.



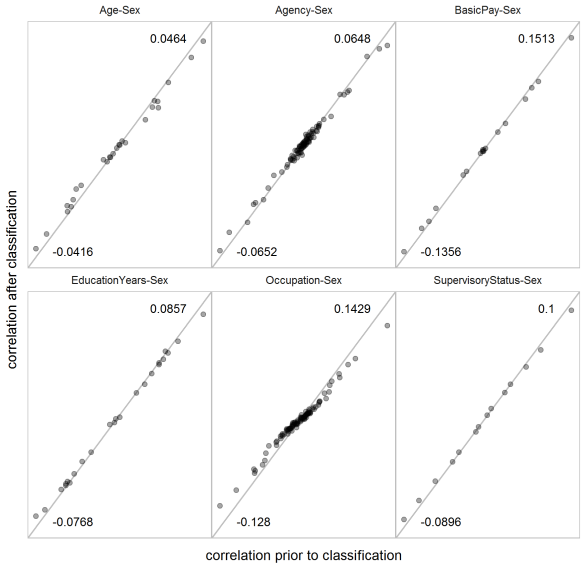

(c1) Classified vs. Unclassified p-Slopes, nBrGrow=0




(c2) Classified vs. Unclassified p-Slopes, nBrGrow=1000




(d1) Classified vs. Unclassified p-Correlation, nBrGrow=0



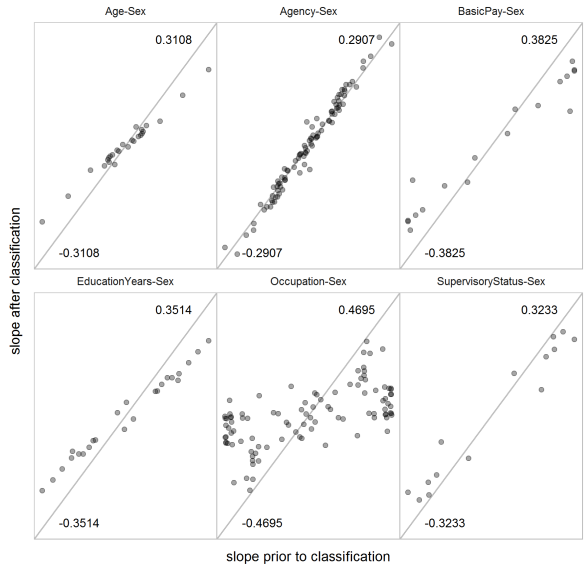

(d2) Classified vs. Unclassified p-Correlation, nBrGrow=1000

Year: 1988
Collapse att: BasicPay, Agency, Occupation
Collapse lev: 25, 500, 200
nUnsynthesized: 0

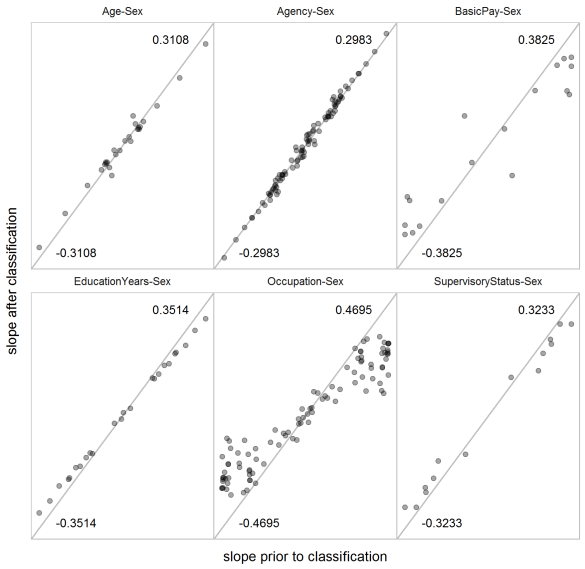
Ensemble size: 5
Node atts: Age, EducationYears, Agency, Occupation, SupervisoryStatus, BasicPay
Sampling weights: 1, 1, 1, 1, 1, 1
pTrueClassified: 0.6746

Tree height: 3, var
Global Sensitivity: 1
Epsilon: 0.3
Branch grow min n:

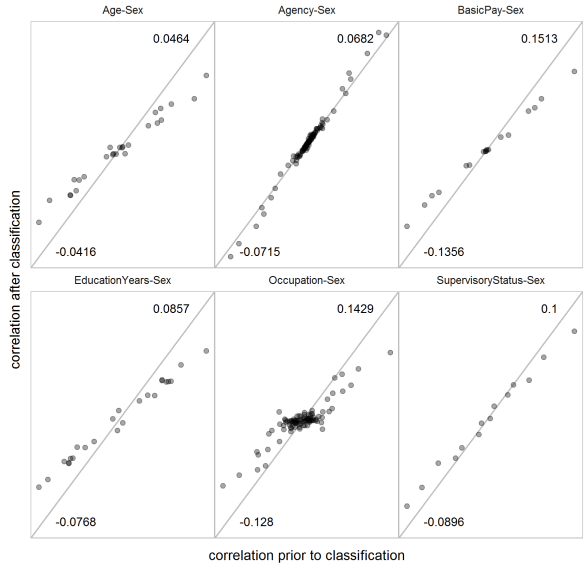
Figure 1: Comparison of $n_B = 0$ and $n_B = 1,000$. Attributes equally weighted when sampled for node splitting. Ensemble size and $\epsilon$ as indicated.



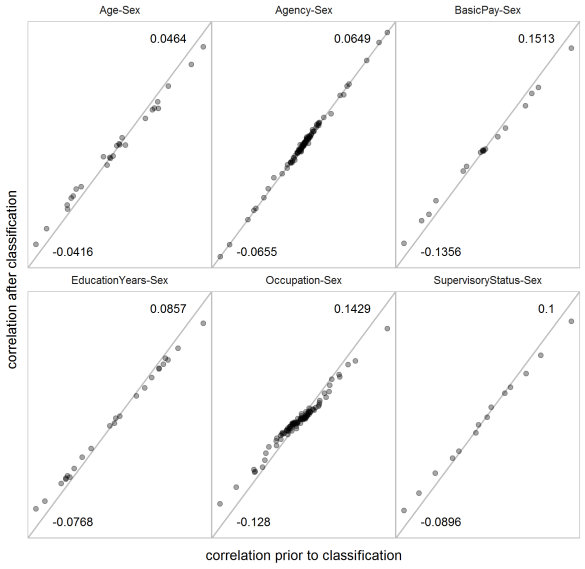

(c1) Classified vs. Unclassified p-Slopes, nBrGrow=0




(c2) Classified vs. Unclassified p-Slopes, nBrGrow=1000




(d1) Classified vs. Unclassified p-Correlation, nBrGrow=0



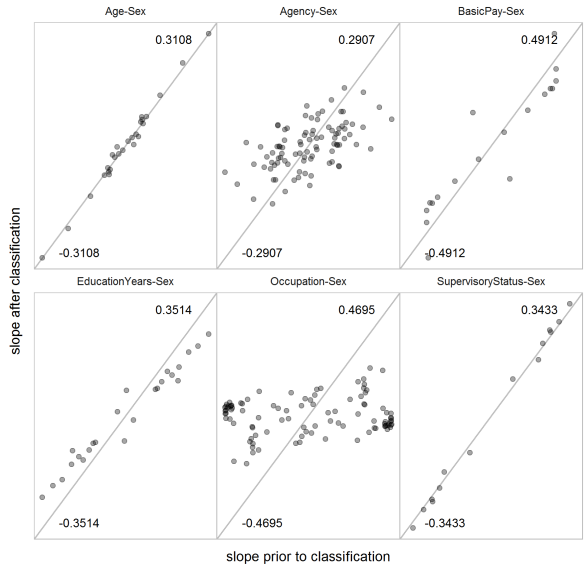

(d2) Classified vs. Unclassified p-Correlation, nBrGrow=1000

Year: 1988
Collapse att: BasicPay, Agency, Occupation
Collapse lev: 25, 500, 200
nUnsynthesized: 0

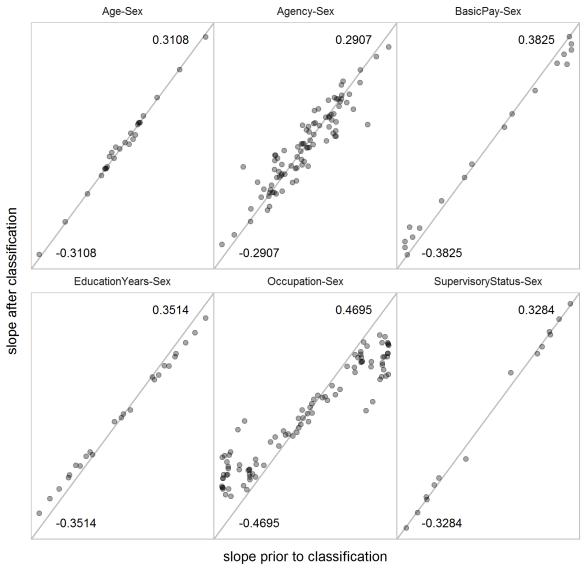
Ensemble size: 5
Node atts: Age, EducationYears, Agency, Occupation, SupervisoryStatus, BasicPay
Sampling weights: 1, 1, 50, 1, 1, 1
pTrueClassified: 0.6662

Tree height: 3, var
Global Sensitivity: 1
Epsilon: 0.3
Branch grow min n:

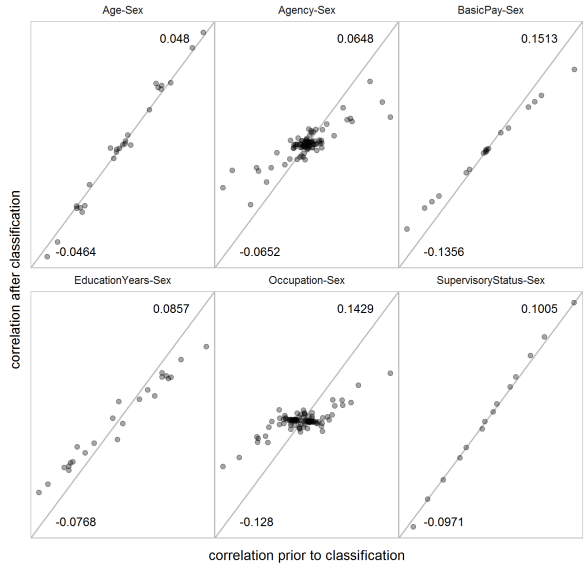
Figure 2: Comparison of $n_B = 0$ and $n_B = 1,000$. Agency given selection preference in sampling for node splitting. Ensemble size and $\epsilon$ as indicated.



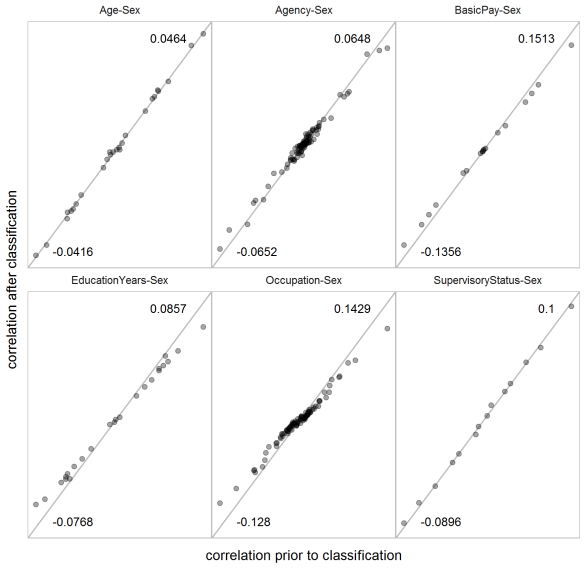

(c1) Classified vs. Unclassified p-Slopes, nBrGrow=0




(c2) Classified vs. Unclassified p-Slopes, nBrGrow=1000




(d1) Classified vs. Unclassified p-Correlation, nBrGrow=0




(d2) Classified vs. Unclassified p-Correlation, nBrGrow=1000

Year: 1988
Collapse att: BasicPay, Agency, Occupation
Collapse lev: 25, 500, 200
nUnsynthesized: 0

Ensemble size: 5
Node atts: Age, EducationYears, Agency, Occupation, SupervisoryStatus, BasicPay
Sampling weights: 50, 1, 1, 1, 50, 1
pTrueClassified: 0.6666

Tree height: 3, var
Global Sensitivity: 1
Epsilon: 0.3
Branch grow min n:

Figure 3: Comparison of $n_B = 0$ and $n_B = 1,000$. Age and Supervisory Status given selection preference in sampling for node splitting. Ensemble size and $\epsilon$ as indicated.

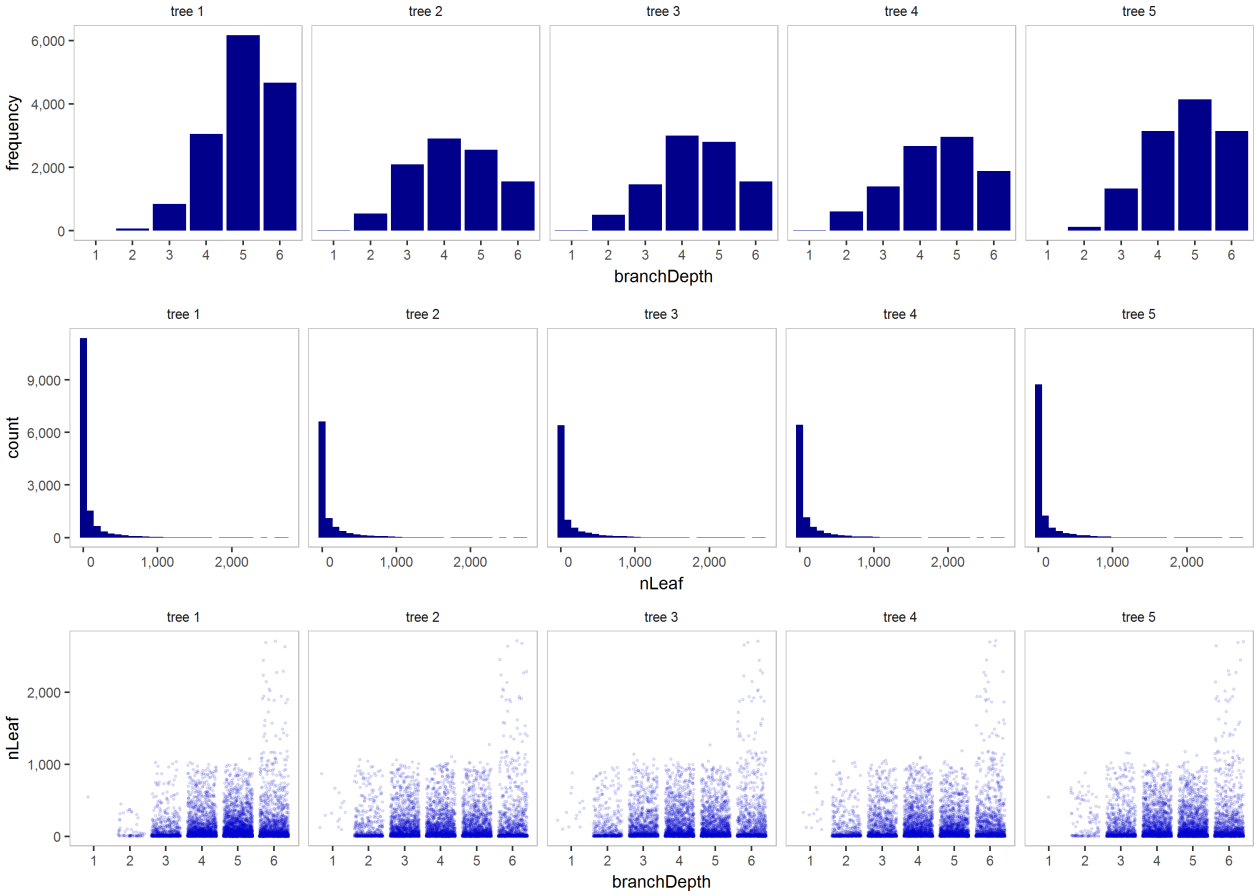



Figure 4: Distribution of branch depth and leaf frequency. $n_B = 1,000$. Attributes equally weighted when sampled for node splitting. Ensemble size and $\epsilon$ as indicated.

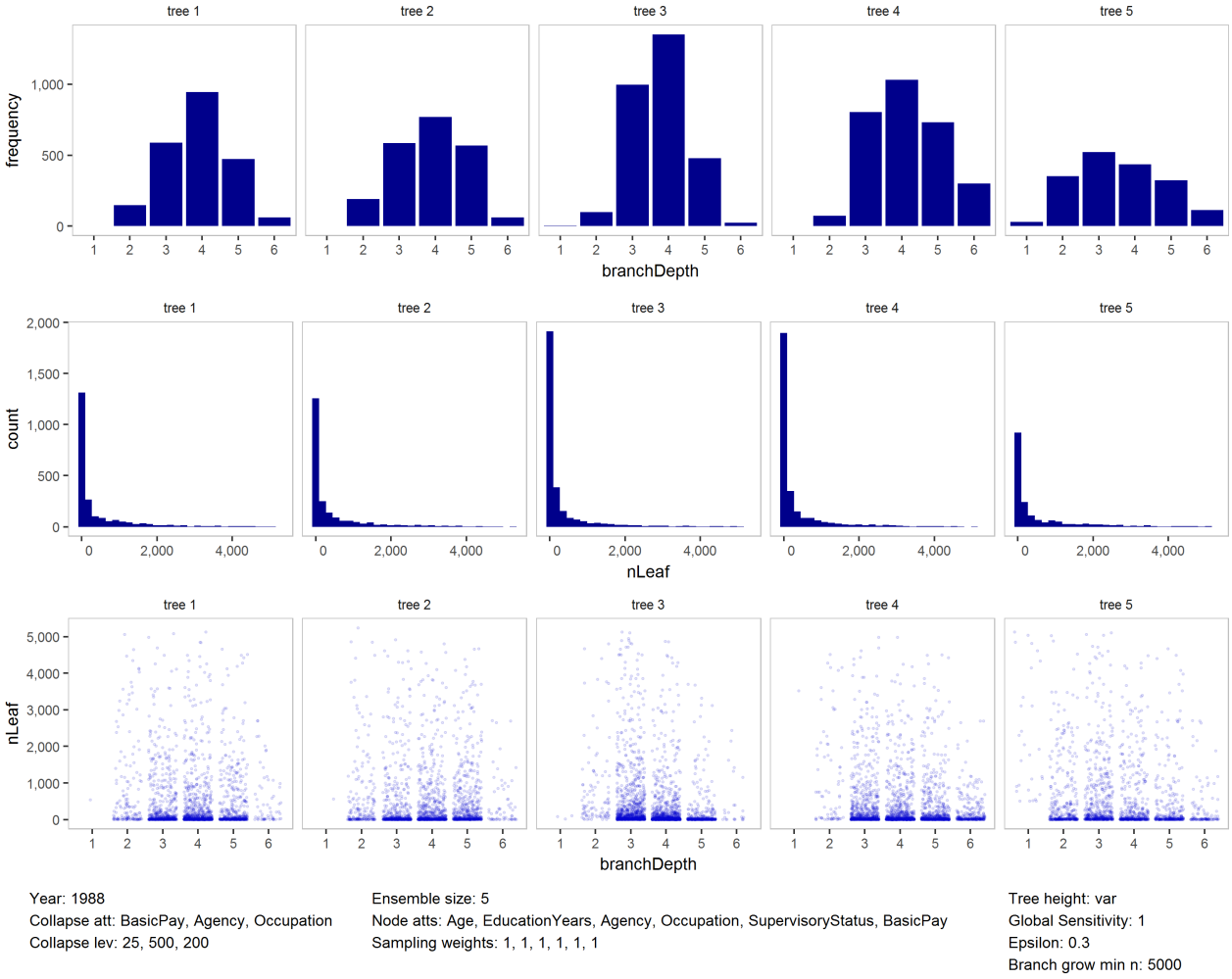
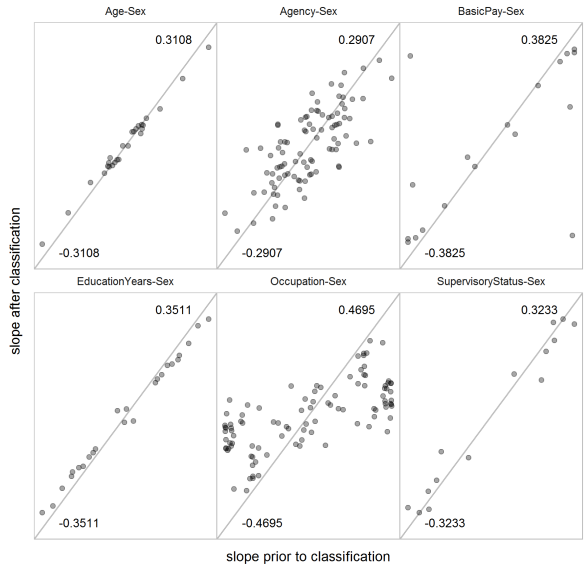
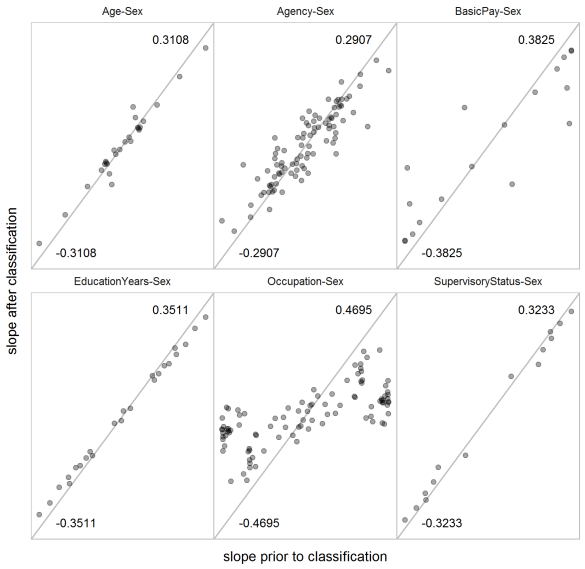


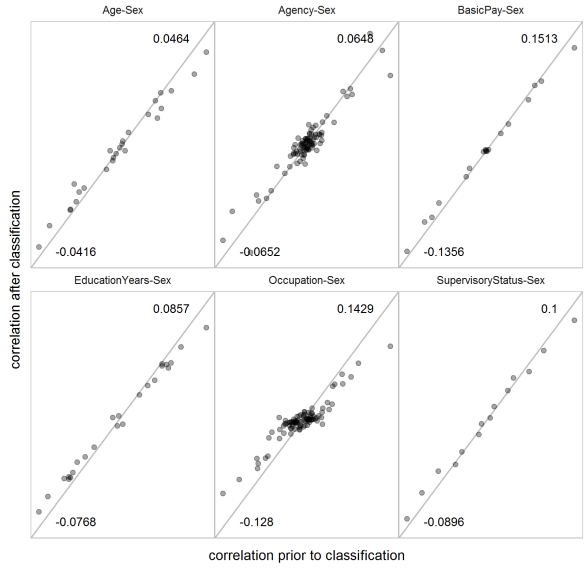

Figure 5: Distribution of branch depth and leaf frequency. $n_B = 5,000$. Attributes equally weighted when sampled for node splitting. Ensemble size and $\epsilon$ as indicated.



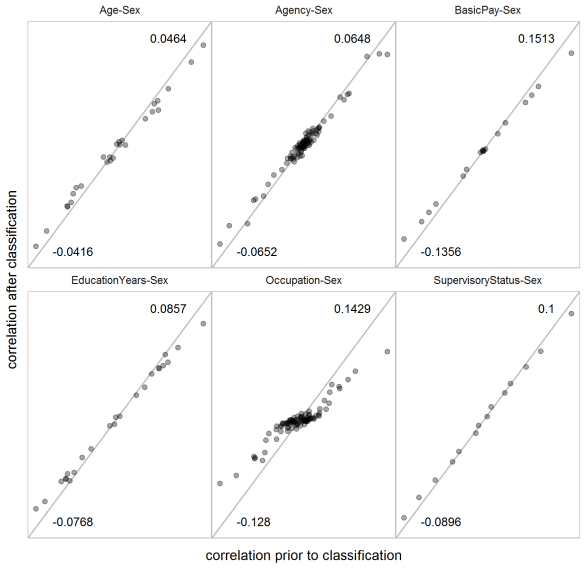

(c1) Classified vs. Unclassified p-Slopes, nBrGrow=0




(c2) Classified vs. Unclassified p-Slopes, nBrGrow=5000



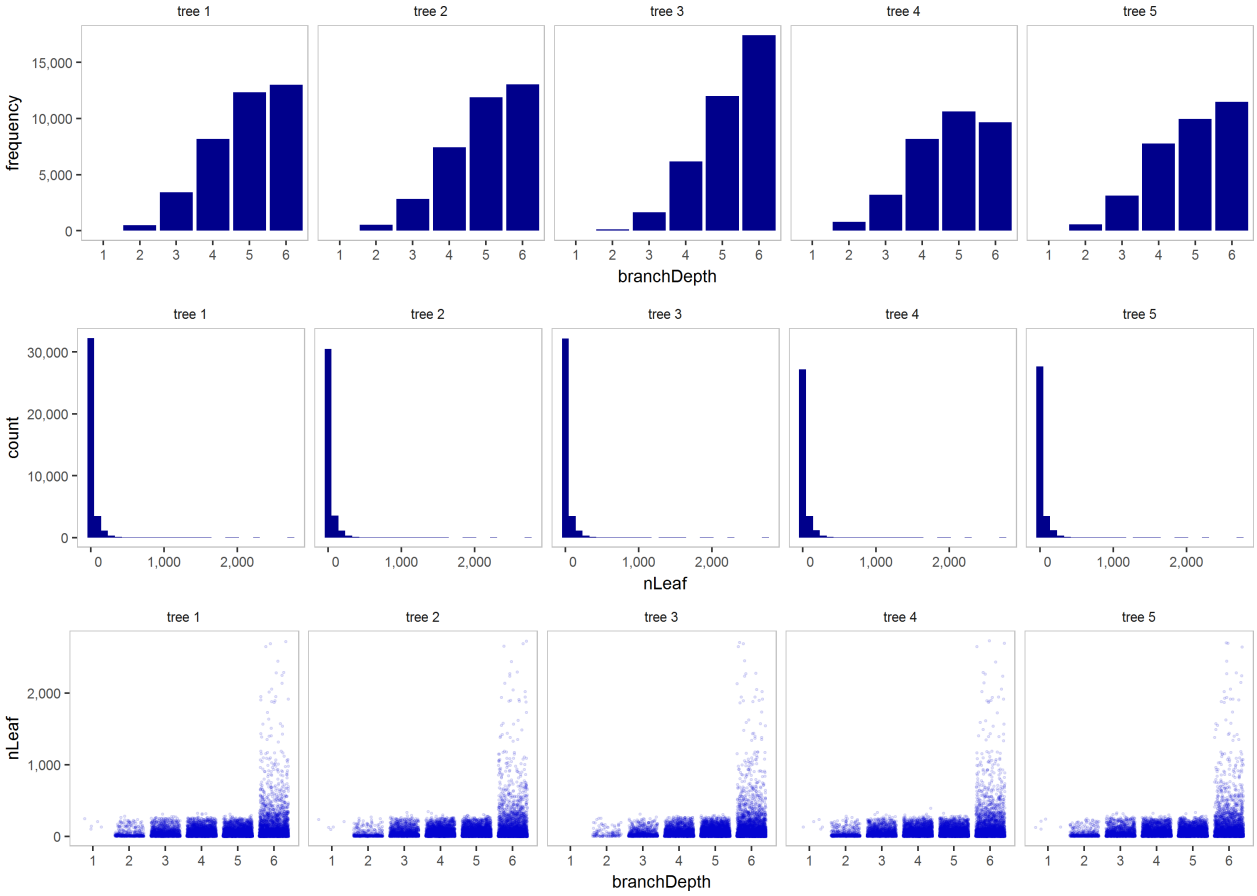

(d1) Classified vs. Unclassified p-Correlation, nBrGrow=0




(d2) Classified vs. Unclassified p-Correlation, nBrGrow=5000

| | | |
|---|---|---|
| Year: 1988 | Ensemble size: 5 | Tree height: 3, var |
| Collapse att: BasicPay, Agency, Occupation | Node atts: Age, EducationYears, Agency, Occupation, SupervisoryStatus, BasicPay | Global Sensitivity: 1 |
| Collapse lev: 25, 500, 200 | Sampling weights: 1, 1, 1, 1, 1, 1 | Epsilon: 0.3 |
| nUnsynthesized: 0 | pTrueClassified: 0.6294 | Branch grow min n: |

Figure 6: Comparison of $n_B = 0$ and $n_B = 5,000$. Attributes equally weighted when sampled for node splitting. Ensemble size and $\epsilon$ as indicated.

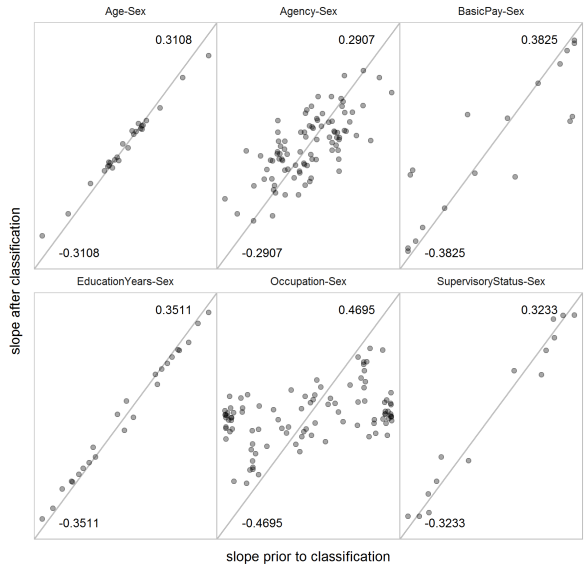



Year: 1988
Collapse att: BasicPay, Agency, Occupation
Collapse lev: 25, 500, 200

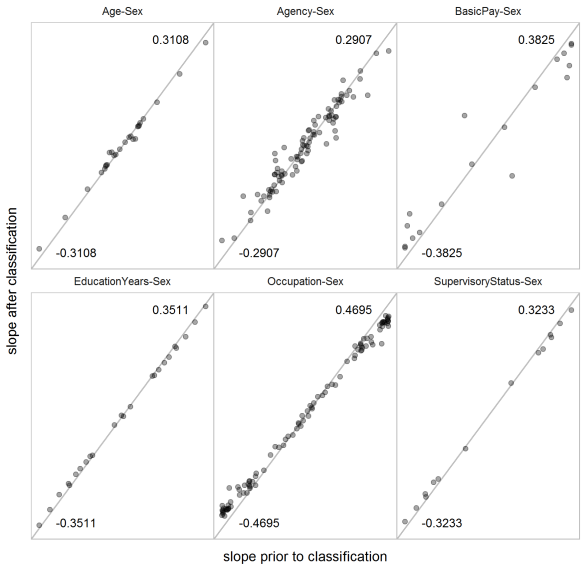
Ensemble size: 5
Node atts: Age, EducationYears, Agency, Occupation, SupervisoryStatus, BasicPay
Sampling weights: 1, 1, 1, 1, 1, 1

Tree height: var
Global Sensitivity: 1
Epsilon: 0.3
Branch grow min n: 250

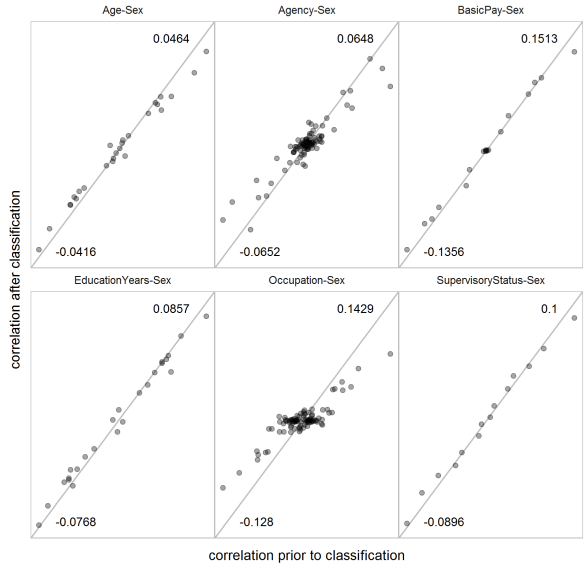
Figure 7: Distribution of branch depth and leaf frequency. $n_B = 250$. Attributes equally weighted when sampled for node splitting. Ensemble size and $\epsilon$ as indicated.



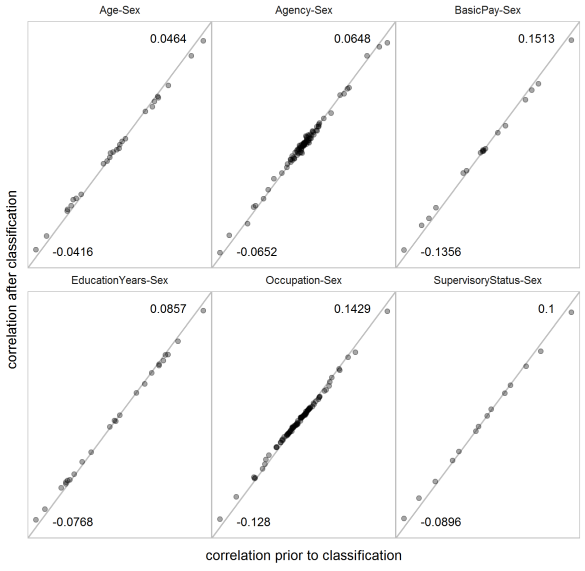

(c1) Classified vs. Unclassified p-Slopes, nBrGrow=0




(c2) Classified vs. Unclassified p-Slopes, nBrGrow=250




(d1) Classified vs. Unclassified p-Correlation, nBrGrow=0




(d2) Classified vs. Unclassified p-Correlation, nBrGrow=250

| | | |
|---|---|---|
| Year: 1988 | Ensemble size: 5 | Tree height: 3, var |
| Collapse att: BasicPay, Agency, Occupation | Node atts: Age, EducationYears, Agency, Occupation, SupervisoryStatus, BasicPay | Global Sensitivity: 1 |
| Collapse lev: 25, 500, 200 | Sampling weights: 1, 1, 1, 1, 1, 1 | Epsilon: 0.3 |
| nUnsynthesized: 0 | pTrueClassified: 0.6959 | Branch grow min n: |

Figure 8: Comparison of $n_B = 0$ and $n_B = 250$. Attributes equally weighted when sampled for node splitting. Ensemble size and $\epsilon$ as indicated.

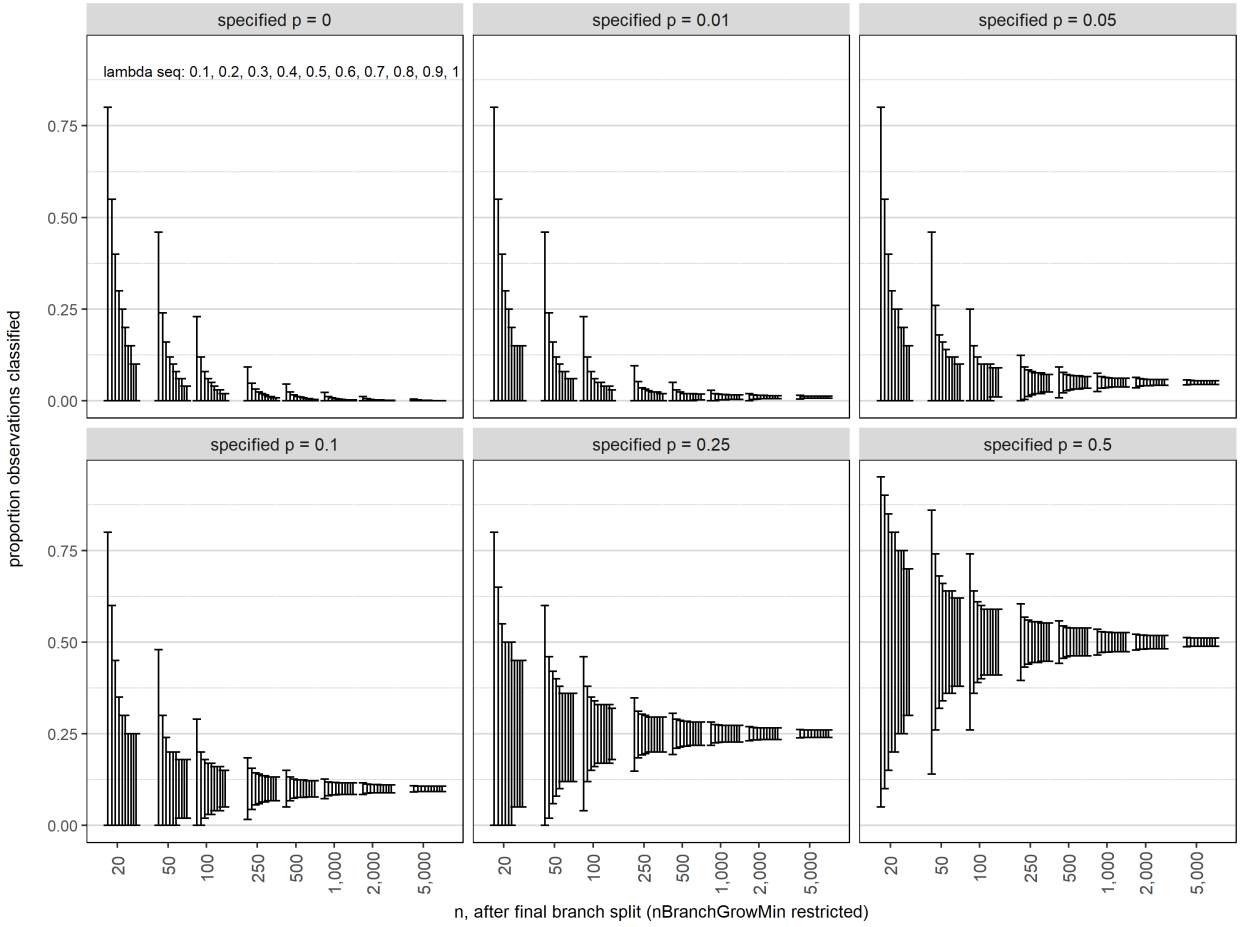



Figure 9: Theoretical joint Laplace-binomial 0.90 confidence intervals on proportion synthesized observations, given specified hypothetical proportions. Laplace $\lambda$ parameter values from 0.0 to 1.0 in 1/10 increments. Actual leaf frequency on x-axis.