Differentially Private Synthetic Attribute Generation

Using Variable Depth Random Decision Trees
Branch Node Splitting Limited by Observation Frequencies in DIBBS Synthetic Data
Leaf Frequencies Based on Authentic Data
Fit Accuracy Assessed Using Publicly Available Buzzfeed Data

Duke University Synthetic Data Project

November 8, 2018

1 Objectives

- Generate differentially private synthetic attributes *Sex* and *Race* by reclassifying corresponding columns of authentic (private) data using ensembles of random decision trees (RDT) constructed with branches non-private columns from attribute labels in authentic data (*Agency, Occupation, Age, Education, Pay,* etc.) and leaf distributions from observed authentic frequencies with application of the Laplace mechanism to label counts prior to proportion computation.
- Study the effect of the following adjustable parameters on the utility and privacy of synthetic attributes
  - Choice of n_B, the minimum number of observations necessary in a branch in order to be extended, or split
  - Collapse of low frequency labels of high dimension attributes (many labels) into a common label
  - RDT ensemble size
  - Sampling weights used in selection of attributes during branch node splitting
  - Global sensitivity and ϵ used in the Laplace mechanism
- Assess the correlation of proportion observations in joint attribute, label categories between public and synthetic private attributes.
- Analyze efficiency of RDT, Laplace, and classification algorithms, develop improvement strategies.

2 Strategy for RDT Construction and Synthetic Attribute Generation

- RDT Construction
  - Limit to a single fiscal year
  - Retrieve public and private attributes from authentic data (AD)
  - Retrieve public attributes from DIBBS synthetic data (SD)
  - Collapse high dimension attributes (*Agency, Occupation*, and *Pay*)
  - Construct ensemble of RDTs using *Age, Education, Agency, Occupation, SupervisoryStatus,* and *Pay* as node attributes (these appear in AD, SD, and BD) and *Sex* as the dependent attribute (call this ensemble$_S$)
  - Synthesize *Sex* using constructed RDT and predictor observations in AD
  - Substitute AD *Sex* with corresponding (observation predicted) synthesized values
  - Construct ensemble$_R$ of RDTs using *Age, Education, Agency, Occupation, SupervisoryStatus, Pay*, and *Sex* as node attributes and *Race* as the dependent attribute
- Attribute synthesis
  - Retrieve public attributes (*Age, Education, Agency, Occupation, SupervisoryStatus,* and *Pay*) from Buzzfeed data (BD)

- Retrieve private attributes (*Sex* and *Race*) from AD by joining BD and AD longitudinal career identifiers (80% of all BD observations for the period 1988-2011 unequivocally match a single AD observation - in certain years, more than 95% have unequivocal matches)
- Omit BD observations with no corresponding AD observation (this is done to permit measurement of classification accuracy - where true private values must be known)
- Collapse *Agency, Occupation*, and *Pay*
- Use BD public attributes and ensemble$_S$ to synthesize *Sex*
- Use BD public attributes and ensemble$_R$ to synthesize *Race*

3 Random Decision Tree Properties

- A tree is assigned a classification (private) attribute for which label distributions are computed from frequencies observed in AD for associated node attributes and labels in the branch leading to a given leaf.
- Level one of a tree consists of a single node for each label of a randomly selected non-private attribute.
- For each node of a given level, each subsequent level consists of a single node for each label of a randomly selected attribute that has not already been selected for a previous level of the tree.
- Construction method is as described in Jagannathan et al., except that **SD observation counts are used to form variable height branches** as follows:
  - As nodes are split, SD observation counts for a branch (all parent and current node attribute) are maintained
  - A leaf is constructed when the (SD) branch observation count falls below the specified minimum, n_B, or when the maximum tree depth (the number of predictor attributes) is encountered
  - Leaf distributions are constructed from all AD observations with attributes and labels associated with the branch
- Use of SD to assess branch observation counts (in enforcing the requirement that variable depth tree branches have at least n_B associated observations up to their final split) eliminates the need to query AD - this reduces total *epsilon* used to construct RDTs.
- Node attributes: *Age, Education, Agency, Occupation, SupervisoryStatus,* and *Pay* (these appear in AD, SD, and BD).
- **Problem**: *Basic Pay* exists in AD and SD, while *Adjusted Basic* Pay exist in AD and BD
  - Presently SD joint frequencies involving basic pay are used to limit branch frequencies, but AD labels correspond to adjusted basic pay (collapsing should mitigate discrepancies due to this mismatch)
  - BD attributes correspond to those in AD and in RDT node labels, so that synthesized observation proportions correspond to those in AD
  - The sole discrepancy is in the decision to (or not) split a branch based on observation frequency when compared to n_B
  - Possible solutions to this problem include:
    - Adjust SD *Basic Pay* by a factor determined by the ratio of *Adjusted Pay* and *Basic Pay* for joint *Year, Agency, Occupation*, etc. in AD (does this violate or reduce DP?)
    - Construct RDTs using SD and AD *Basic Pay*, but maintain a separate set of nodes with corresponding leaves for *Adjusted Pay* (note that this would transform a very generic and adaptable RDT algorithm into one appropriate strictly for our data - it also involves additional AD queries)
    - Apply AD *Basic Pay* to *Adjusted Pay* ratio to BD *Adjusted Pay* prior to synthetic attribute generation (this requires knowledge of the ratio - something an attacker or data provider may not have)
- An *ensemble* consists of multiple RDTs, each with a randomly selected sequence of non-private attributes at each level.
- An outline for variable depth RDT construction follows (note the recursive nature of execution):

```
1    function constructNode(SD, AD, privateAttribute, nBranchGrowMin, maxTreeHeight, nodeLevel, nodeAttribute,
```

```
2                          attCandidates, attCandidateWeights, obsIndicesSD, obsIndicesAD)
3       if(observation_count[SD(obsIndicesSD)] >= nBranchGrowMin and nodeLevel < treeHt)
4         // Create a node for each label of a randomly selected attribute that has not been used in this node path
5         // Note that SD and AD observation indices are filtered for the current attribute and each label so that
6         // only observations related to the node path under construction remain when the corresponding leaf is
7         // composed
8         // Compose and return a leaf distribution when SD observation count falls below nBranchGrowMin or when
      node
9         // level reaches maxTreeHeight
10        for each label in nodeAttribute
11          o remove nodeAttribute from attCandidates
12          o currAttribute = randomly selected attribute form attCandidates sampled using attCandidateWeights
13          o identify SD and AD observation indices such that currentAttribute of data(obsIndices) = label
14          o return[constructNode(SD, AD, privateAttribute, nBranchGrowMin, maxTreeHeight, nodeLevel+1,
      currAttribute,
15                                 attCandidates-currentAttribute, attCandidateWeights, obsIndicesSD, obsIndicesAD)
      ]
16      else
17        // Create a leaf node (privateAttribute probability distribution) using AD observations subset using
      obsIndicesAD
18        if(count of SD observations subset using obsIndicesSD >= leafFreqMin)
19          o compute table of observation frequencies by privateAttribute label for AD(subset using obsIndicesAD)
20          o apply Laplace mechanism to label frequencies
21          o compute observation proportions by privateAttribute label
22        return(labels with corresponding proportions)
```

- RDT algorithm parameters that affect privacy and utility:
  - The number of non-private attributes from which to construct nodes. This regulates maximum tree depth since one attribute is chosen for each node at each level and attributes are not recycled. For the current exercise, nodes are constructed from attributes *Age, Education, Agency, Occupation, SupervisoryStatus,* and *Pay*, resulting in trees with a maximum depth of six nodes.
  - Minimum observation counts under which labels of high dimension attributes are combined (collapsed) into a common label.
  - RDT ensemble size (number of independent RDTs generated per synthesized attribute).
  - Sampling weights for selecting non-private attributes during node construction.
  - n_B: the minimum number of SD observations in a branch, under which node splitting discontinues and a leaf distribution is constructed.
  - $\Delta(f)$, global sensitivity: defined as $\max_{(\mathrm{AD}_1,\mathrm{AD}_2)\in\mathrm{AD}_0} \parallel f(\mathrm{AD}_1) - f(\mathrm{AD}_2) \parallel_1$, where AD_1 and AD_2 are subsets of AD differing by a single record (one of AD_1 or AD_2 have one record that does not appear the other, while all remaining records are identical). Note that each $f(AD)$ has a unique $\Delta(f)$. When RDT leaf distributions are computed from label frequencies $\Delta(f) = 1$, since omission of a single observation subtracts 1 from a single label frequency.
  - λ, the Laplace mechanism parameter: $\lambda = \frac{\epsilon}{\Delta(f)}$, where ϵ is selected by the data steward. Generally, lower values of ϵ reduce reduce the risk of disclosing private data features, but also reduce synthetic data utility (agreement between SD and AD attributes).
  - Queries of AD during RDT leaf construction are executed in parallel, retrieving disjoint subsets of observations, one for each node path - total accumulated ϵ should be ϵ.

4 Private Attribute Synthesis Properties

- Non-private attributes in BD are used to span an RDT beginning at the root and continuing through branches corresponding to node attributes and labels until a leaf distribution is encountered.
- Since AD and BD have a common set of non-private attributes and labels and RDT nodes at each level correspond to a comprehensive, disjoint set of labels for one attribute, each BD observation maps to a single branch and leaf in an RDT constructed from the AD.
- Maximum tree depth and n_B limit the number of attributes appearing in a given RDT branch and, therefore, the number of attributes jointly used in composing leaf distributions. Reduced attribute representation in leaf distributions causes greater confounding of attributes (attributes appearing in observations are not represented in leaves and do not affect synthesized label sampling).
- Due to attribute confounding, **using leaf distributions to generate synthetic records can produce combinations of attributes and labels that do not exist in the AD**. For example, given

an RDT that classifies *Race*, where *Age* does not appear in the branch that a given observation maps to, then *Race* is synthesized for the observation without regard for *Age* (mass for all *Age* labels is uniformly distributed across all *Race* labels), so that synthesis may produce a *Race, Age* combination of *A, 17* when, in fact, no such AD observations exist.

5 Specification of Branch and Leaf Observation Frequency, n_B

Given an RDT branch and leaf distribution, the number X_i of n synthesized labels generated with label i has a binomial(n, p_i) distribution, where p_i Laplace(λ, p_{i0}) and p_{i0} is the proportion of observations in the AD subset corresponding to the branch and leaf that are coded with label i. Note that the Laplace mechanism is applied during leaf construction and sampling is done later during data synthesis. This makes synthesis of label i conditional on probability p_i, giving

$$P(X_i = k|p_i) = \frac{P(X_i = k, p_i)}{P(p_i|\lambda)} \Rightarrow P(X_i = k, p_i) = P(X_i = k|p_i)P(p_i|\lambda) \tag{1}$$

Using the right side of equation 1 as the joint (Laplace-binomial) probability of X_1 and p_i, 1-α confidence bounds on $p_i = x_i/n$ can be derived for given $x_i,\ n,\ \lambda$, and α. Since $X_i \in \{0, \ldots, n\}, \lambda$ is constrained by privacy constraints, and $1 - \alpha$ is a desired confidence level, the remaining control parameter is $n = n_B$, the minimum branch splitting frequency. For various values of $n_B L$, given x_i, probability mass can be accumulated (using eq 1) for intervals $[x_i - 0,\ x_i + 0], [x_i - 1,\ x_i + 1],\ \ldots,\ [x_i - \delta_0,\ x_i + \delta_1]$, until a total mass of $1 - \alpha$ is achieved, where $\delta_0,\ \delta_1\ \in \{0,\ \ldots\ n_B\}$. n_B can be increased or decreased to identify the minimum value that satisfies confidence bound requirements. The resulting $1 - \alpha$ bound on p_i is $[\frac{x_i - \delta_0}{n_B},\ \frac{x_i + \delta_1}{n_B}]$. Since δ_0 and δ_1 are integers, total mass will likely be greater than $1 - \alpha$. However, δ endpoints can be interpolated to yield an exact $1 - \alpha$ interval. Figure 1 shows, for various values of p_{i0} (labeled "specified p" in panel titles), λ (0.0 to 1.0 in 1/10 increments), and n_B (on the x-axis), 0.90 confidence intervals on p_i. Given specified values of $p_{i0},\ \lambda$, and maximum width of a 0.90 confidence interval, the minimum n_B that should be used in RDT construction can be located on the x-axis of the corresponding plot. For instance, given $p_{i0} = 0.25$ and $\lambda = 0.4$, n_B should be at least 1,000 to restrict the 0.90 confidence interval on p_i to within [0.22, 0.27]. Alternatively, the privacy cost (λ) of constructing trees with a given n_B can be determined, for each p_{i0}, by locating (on the corresponding plot) the widest error bar that is less than the desired interval, and reading its associated value of λ. A few key observations are:

- Consistent with the binomial distribution, due to minimum variance at $p_{i0} = 0$ increasing to maximum variance at $p_{i0} = 0.5$, the width of confidence intervals is minimized at $p_{i0} = 0$ and maximized at $p_{i0} = 0.5$
- For all values of p_{i0} and n_B less than approximately 1,000, confidence interval widths rapidly decrease as λ increases; when $n_B \geq 1,000$, the effect of λ on confidence interval width is negligible

It must be emphasized that n_B affects the number of observations required in a branch in order to be extended, or split. There may be fewer than n_B observations in a given leaf. Values of n appearing in figure 1 represent leaf frequencies.

To verify theoretical 0.90 confidence intervals, simulated p_i values were drawn from Laplace-binomial distributions using $\lambda = 0.3, n_B = 1,000$, and $p_{i0} \in \{0, 0.01, 0.05, 0.10, 0.25, 0.50\}$. 10,000 random p_i values were computed for each p_{i0}. Histograms along with empirical 0.9 confidence intervals are shown in figure 2. It is seen that indicated confidence intervals correspond closely to theoretical ones represented in figure 1.

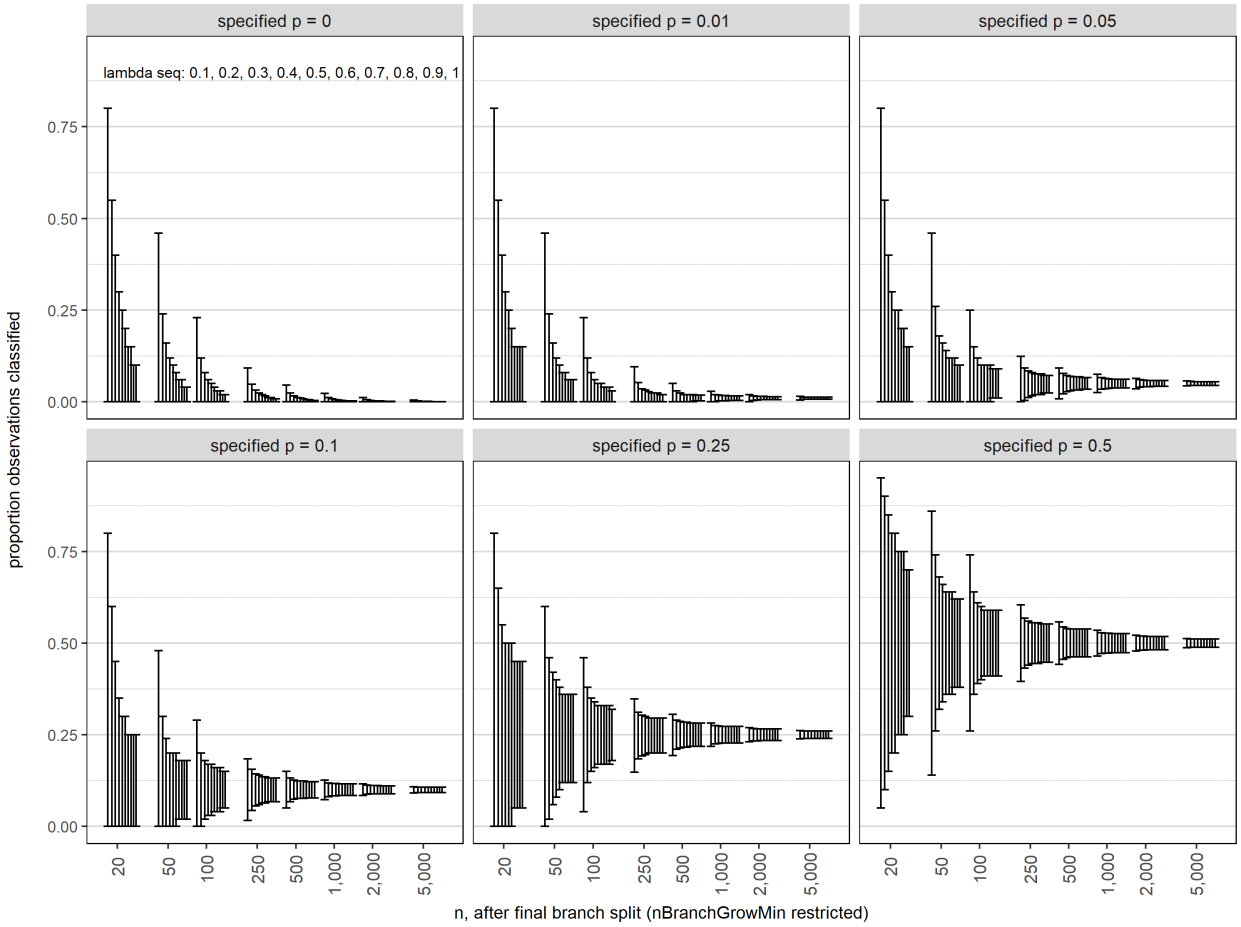


Figure 1: Theoretical joint Laplace-binomial 0.90 confidence intervals on p_i, given specified values of p_{i0}. λ from 0.0 to 1.0 in 1/10 increments. Minimum value of n_B to achieve CI on x-axis.

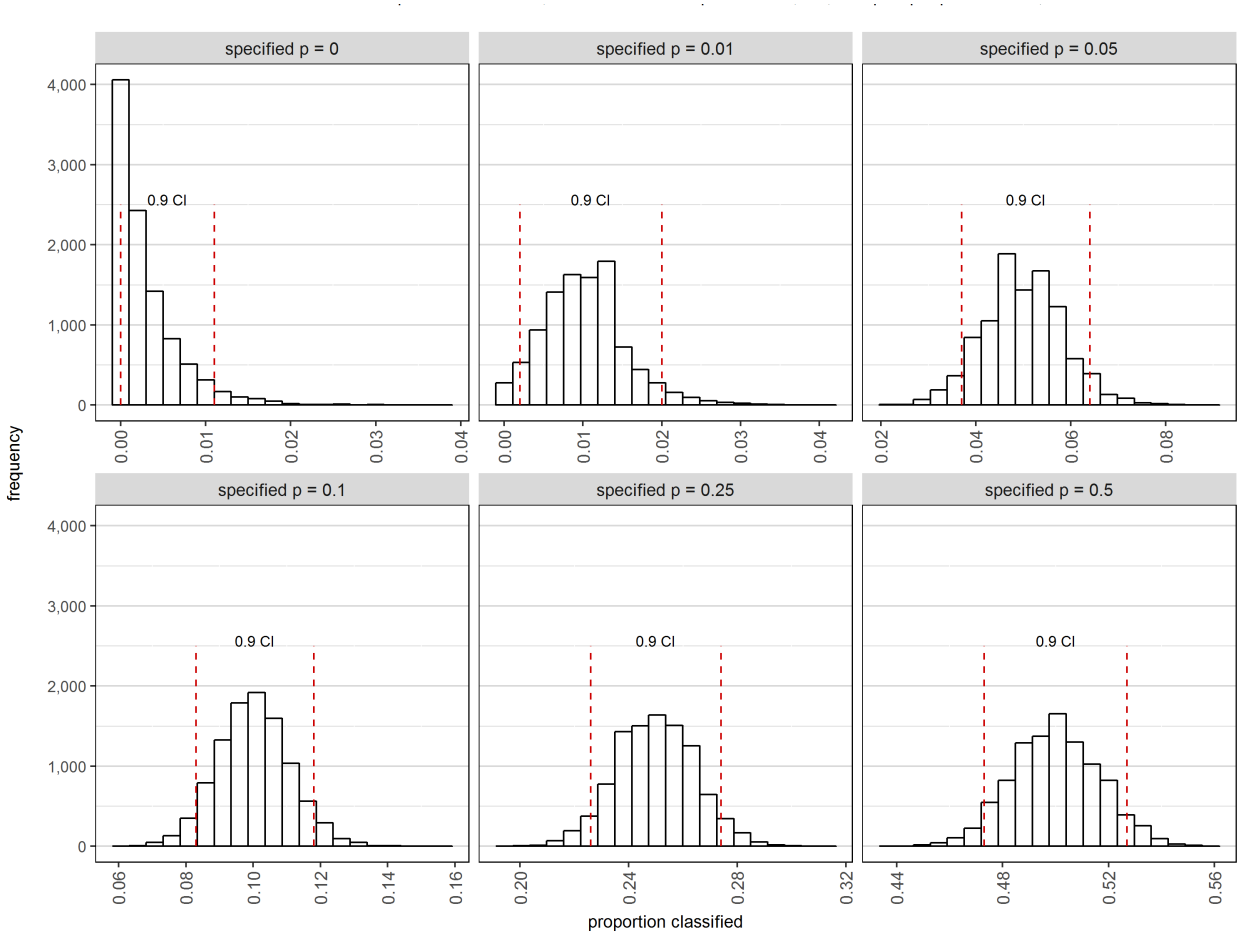


Figure 2: Simulated Laplace-binomial p_i values with 0.90 confidence intervals. $\lambda = 0.3, n_B = 1,000$, p_{i0} as indicated ("specified p").

6 Utility of Variable Depth RDT Generated Synthetic Data

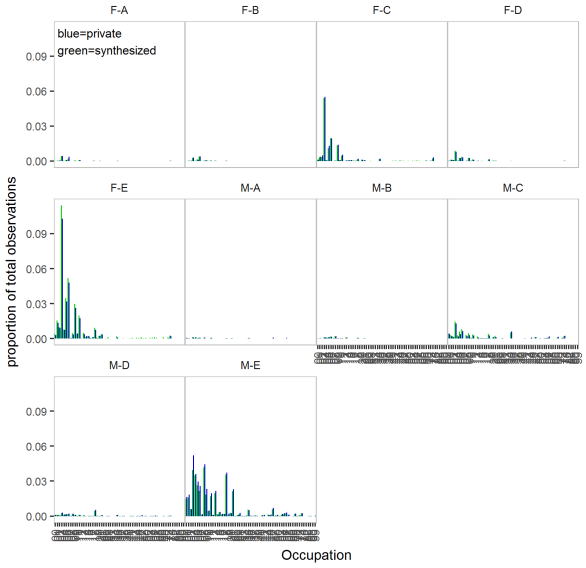
Attributes and observations used to generate RDTs are limited to a discrete variables and data from a single year. To assess utility of synthetic attributes, proportion of observations in joint attribute/label categories within Ad and SD are compared. Figures 3 through 9 each contain four comparative plots as follows:

- (a) A histogram for joint categories of *Sex, Race*, and a third attribute of interest, with blue bars representing AD proportions and green bars representing those for SD. Similar height blue and green bars for common x-axis labels indicate equal proportion of observations between data sets.
- (b) A plot of proportion observations for joint categories of *Sex, Race*, and six node attributes. SD on y-axis, AD on x-axis. Points near reference line of slope 1 indicate strong agreement. Correlation involving all categories as indicated.
- (c) A plot of the least squares slope between attributes and levels within SD (y-axis) and the slope of corresponding attributes and levels within AD (x-axis). Note that, since individual levels of attributes form dichotomous indicator vectors (position i=1 if observation i is coded for the level and is 0 for all other levels), the slope of attribute/level y on attribute/level x is the difference in the proportions y=1 (given x=1) and y=1 (given x=0). Points near the reference line of slope 1 indicate agreement. Note the relationship of SD-AD agreement to collapsing values and attribute sampling weights.
- (d) A Plot of correlation between attributes and levels within SD (y-axis) and the correlation of corresponding attributes and levels within AD (x-axis). These values are related to those in plot (c) since the correlation of two vectors x and y is a function of their slope: $r = \beta_1 \frac{\sigma_x}{\sigma_y} = \beta_1 \sqrt{\frac{P(X=1)[1-P(X=1)]}{P(Y=1)[1-P(Y=1)]}}$.

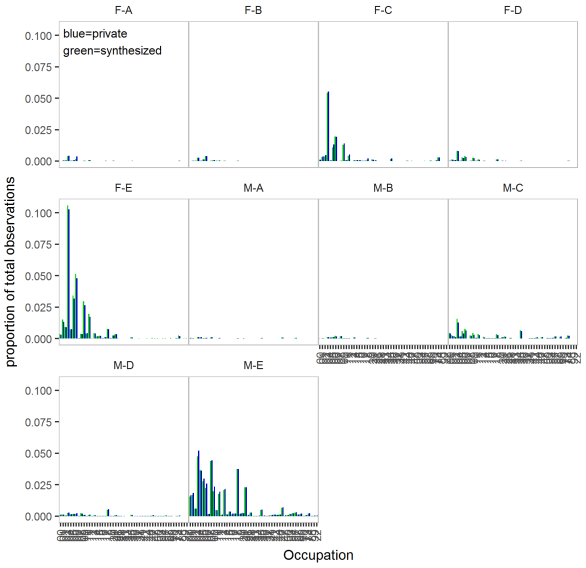
Effect of Collapsing

- Method 1: *Pay* converted to units of \$10,000, no other attributes adjusted
- Method 2: *Pay* converted to units of \$25,000, *Agencies* with fewer than 500 observations (in the target year) combined, *Occupations* with fewer than 200 observations (in the target year) combined

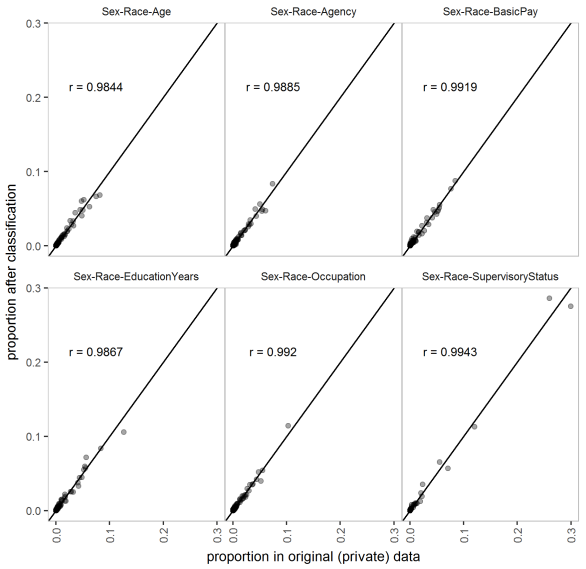
Figure 3 compares methods 1 and 2 with histograms and joint category proportion plots. Figure 4 compares collapsing methods with slope and correlation plots. Improvements in slope are observed for *Agency* and *Occupation*. Note that *Occupation* has the highest sampling weight.



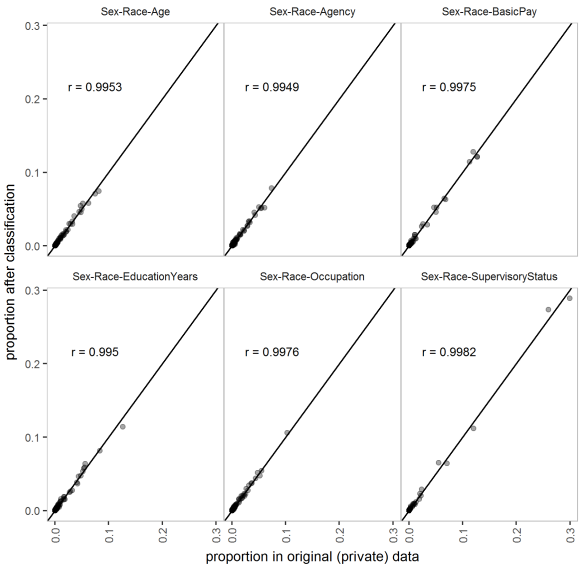
(a1) Distribution of Occupation by Sex and Race, Collapse lev: 10, 0, 0



(a2) Distribution of Occupation by Sex and Race, Collapse lev: 25, 500, 200



(b1) Correlation of Joint Observation Proportions, Collapse lev: 10, 0, 0



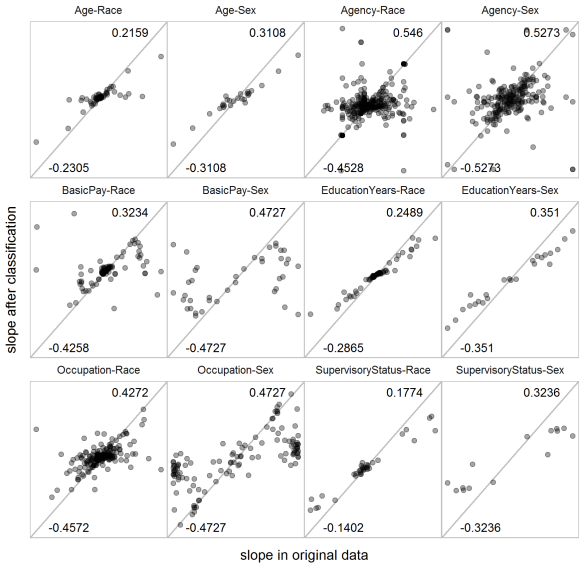
(b2) Correlation of Joint Observation Proportions, Collapse lev: 25, 500, 200

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

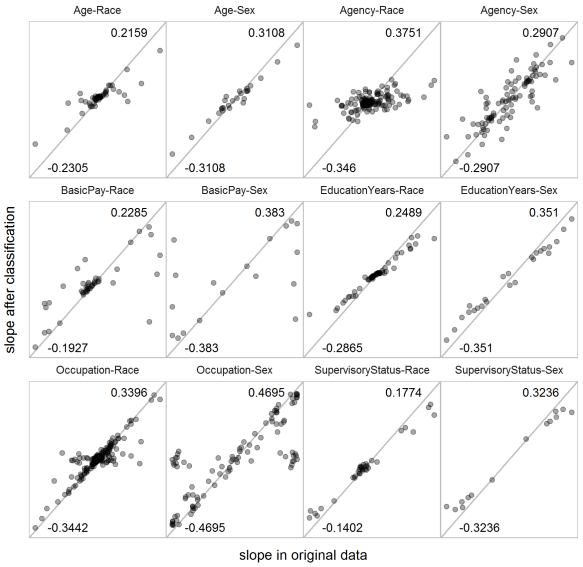
Ensemble size: 10
Node atts: Age, EducationYears, Agency, Occupation, SupervisoryStatus, BasicPay
Sampling weights: 1, 1, 1, 20, 1, 10
pTrueClassified: 0.6779, 0.5813

Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon: 0.4
Leaf min n:1000

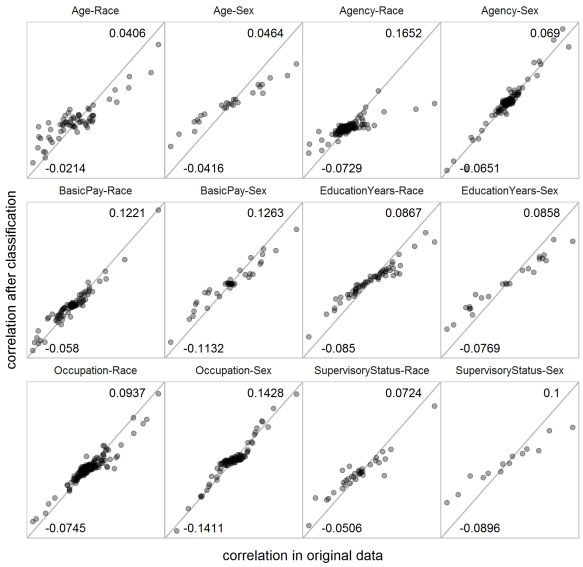
Figure 3: AD/SD attribute proportion observation comparison panel. Year, attribute collapsing values, node attribute sampling weights, ensemble size, n_B, and $\lambda = \epsilon/(\text{global sensitivity})$ indicated in table below images.



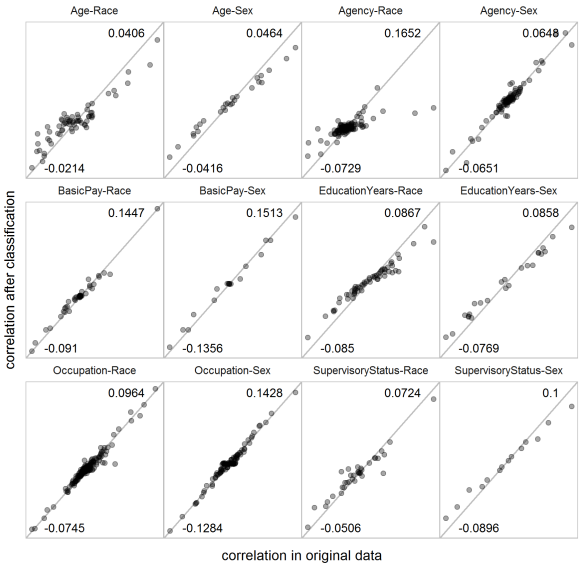
(c1) Private vs. Synth p-Slopes, Collapse lev: 10, 0, 0



(c2) Private vs. Synth p-Slopes, Collapse lev: 25, 500, 200



(d1) Private vs. Synth p-Correlation, Collapse lev: 10, 0, 0



(d2) Private vs. Synth p-Correlation, Collapse lev: 25, 500, 200

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

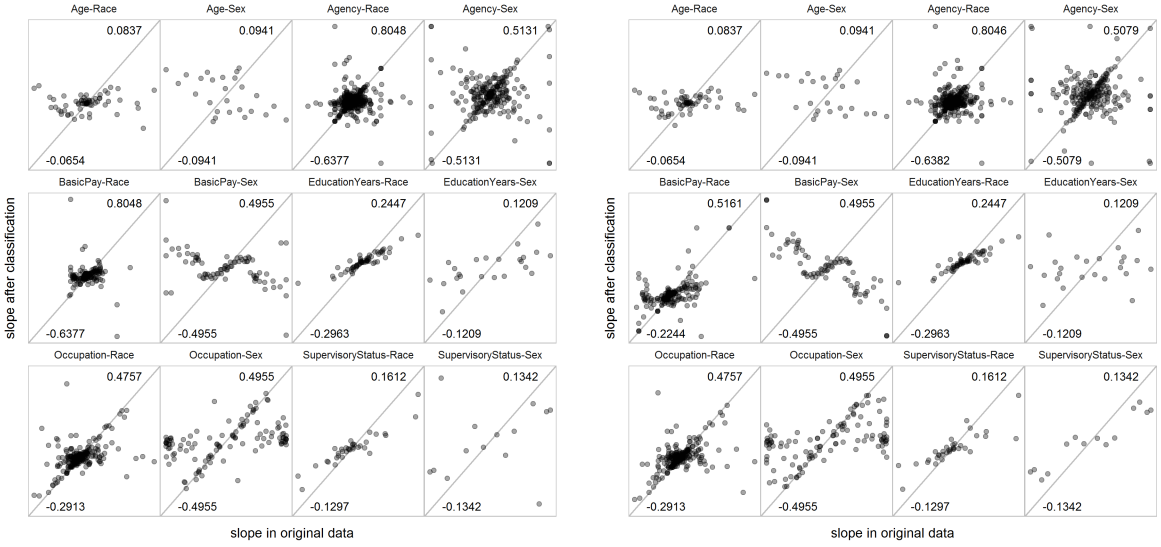
Ensemble size: 10
Node atts: Age, EducationYears, Agency, Occupation, SupervisoryStatus, BasicPay
Sampling weights: 1, 1, 1, 20, 1, 10
pTrueClassified: 0.6779, 0.5813

Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon: 0.4
Leaf min n:1000

Figure 4: AD/SD attribute proportion observation comparison panel. Year, attribute collapsing values, node attribute sampling weights, ensemble size, n_B, and $\lambda = \epsilon/(\text{global sensitivity})$ indicated in table below images.

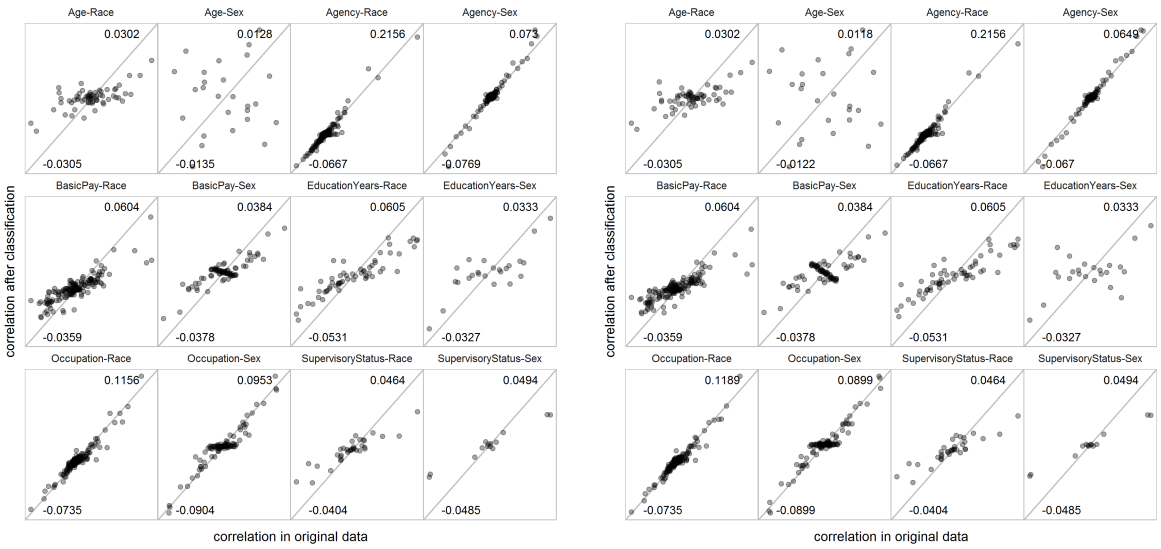
Effect of Ensemble Size

Figure 5 compares ensembles of five and ten RDTs. Differences do not appear significant. Although a specific example, difference in slopes and correlation have generally not been observed between five and ten tree ensembles, regardless of other values of attribute collapsing, attribute selection weighting, n_B, and ϵ. An important result is that, with our data, it appears adding trees beyond five (with additional ϵ consumption) is not helpful in improving synthetic data utility.



(c1) Private vs. Synth p-Slopes, n-Ensemble = 5

(c2) Private vs. Synth p-Slopes, n-Ensemble = 10



(d1) Private vs. Synth p-Correlation, n-Ensemble = 5

(d2) Private vs. Synth p-Correlation, n-Ensemble = 10

Year: 2011
Collapse att: BasicPay, Agency, Occupation
Collapse lev: 10, 0, 0
nUnsynthesized: 0, 0

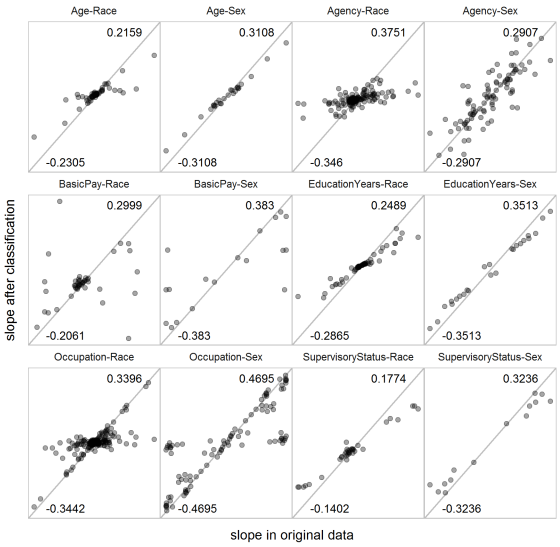
Ensemble size:
Node atts: Age, EducationYears, Agency, Occupation, SupervisoryStatus, BasicPay
Sampling weights: 1, 1, 60, 20, 1, 1
pTrueClassified: 0.5849, 0.4907

Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon: 0.4
Leaf min n:1000

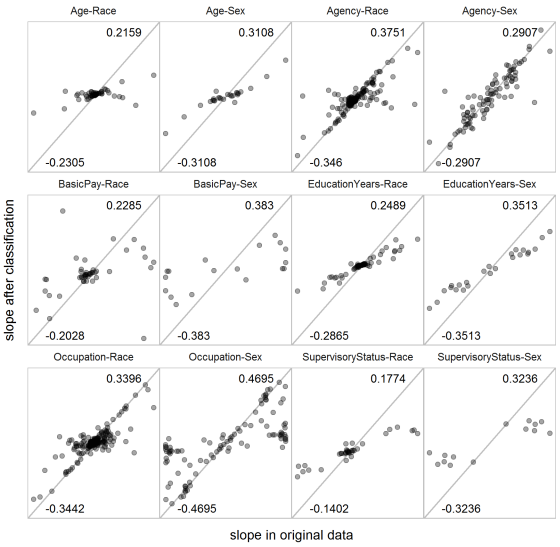
Figure 5: AD/SD attribute proportion observation comparison panel. Year, attribute collapsing values, node attribute sampling weights, ensemble size, n_B, and $\lambda = \epsilon/(\text{global sensitivity})$ indicated in table below images.

Effect of Node Attribute Sampling Weight

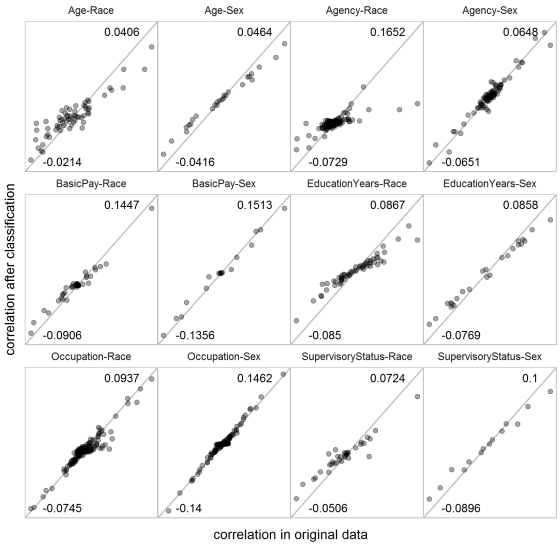
At each level of tree construction, attributes are randomly selected for node splitting. Altering sampling weights can give gives substantively important attributes preference. Figures 6 and 6 compare low to high sampling weight for *Agency* (1988). A significant difference is observed in the fit of slopes and correlation with attributes collapsed (figure 6), but not when *Pay* is the sole collapsed attribute (figure 7).



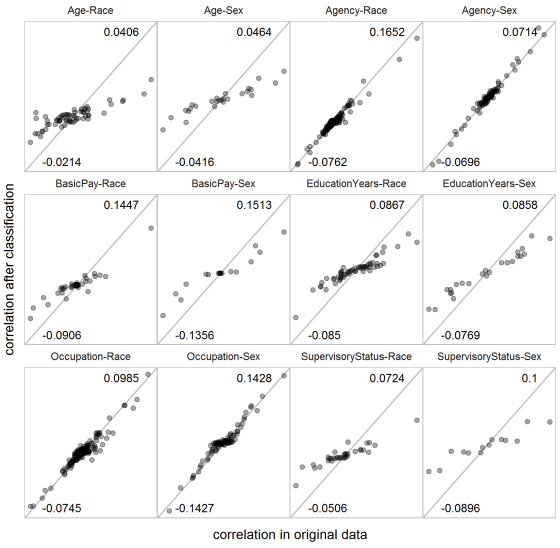
(c1) Private vs. Synth p-Slopes, Att-Wt = 1, 1, 1, 20, 1, 10



(c2) Private vs. Synth p-Slopes, Att-Wt = 1, 1, 60, 20, 1, 1



(d1) Private vs. Synth p-Correlation, Att-Wt = 1, 1, 1, 20, 1, 10



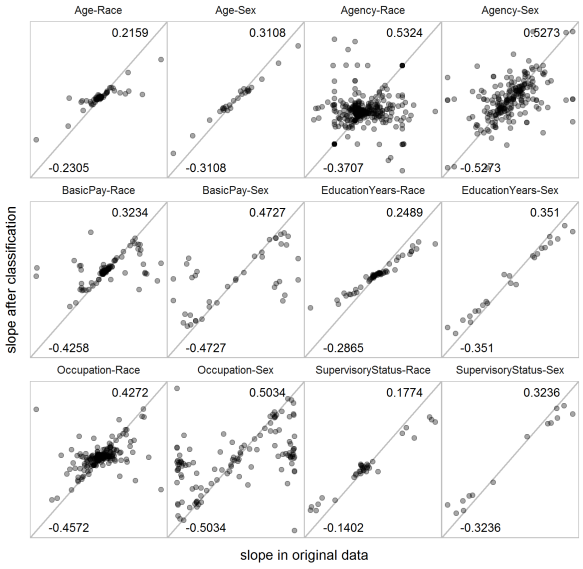
(d2) Private vs. Synth p-Correlation, Att-Wt = 1, 1, 60, 20, 1, 1

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

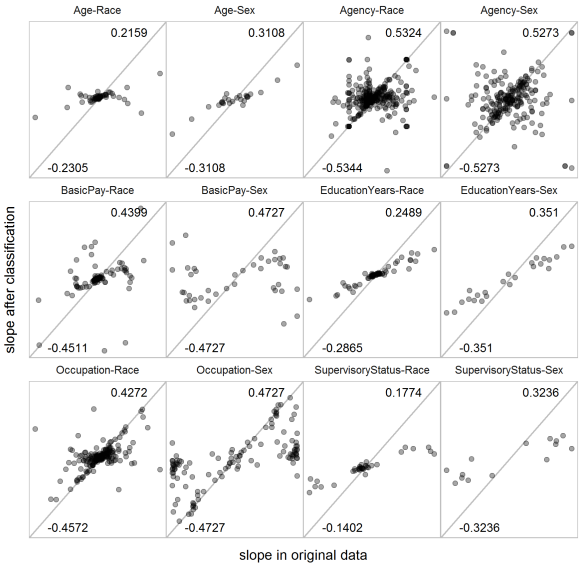
Ensemble size: 5
Node atts: Age, EducationYears, Agency, Occupation, SupervisoryStatus, BasicPay
Sampling weights:
pTrueClassified: 0.6391, 0.577

Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon: 0.1
Leaf min n:1000

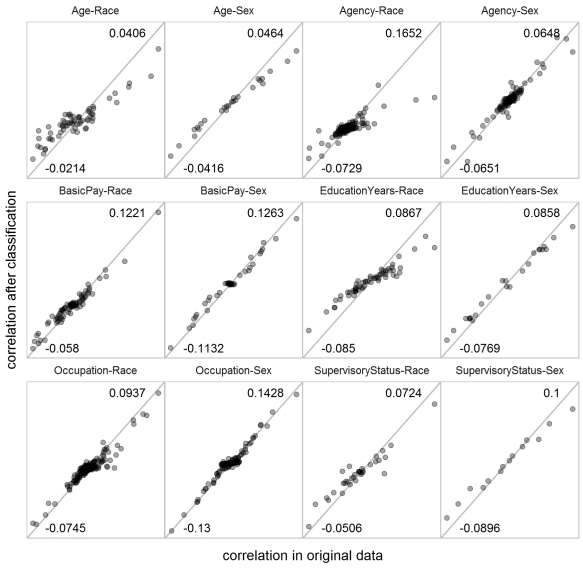
Figure 6: AD/SD attribute proportion observation comparison panel. Year, attribute collapsing values, node attribute sampling weights, ensemble size, n_B, and $\lambda = \epsilon/(\text{global sensitivity})$ indicated in table below images.



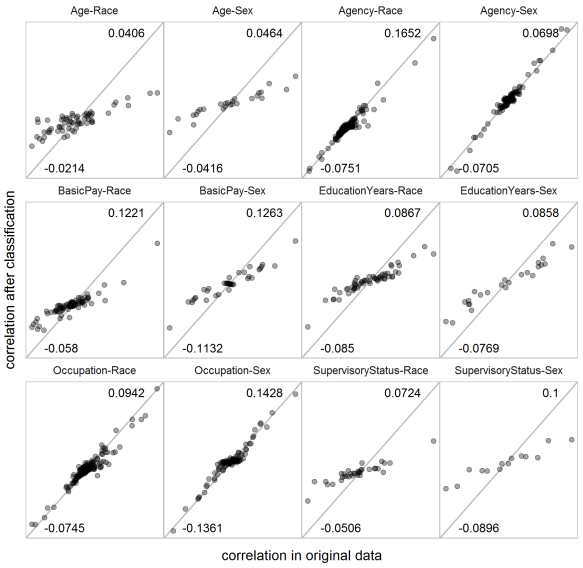
(c1) Private vs. Synth p-Slopes, Att-Wt = 1, 1, 1, 20, 1, 10



(c2) Private vs. Synth p-Slopes, Att-Wt = 1, 1, 60, 20, 1, 1



(d1) Private vs. Synth p-Correlation, Att-Wt = 1, 1, 1, 20, 1, 10



(d2) Private vs. Synth p-Correlation, Att-Wt = 1, 1, 60, 20, 1, 1

Year: 1988
Collapse att: BasicPay, Agency, Occupation
Collapse lev: 10, 0, 0
nUnsynthesized: 0, 0

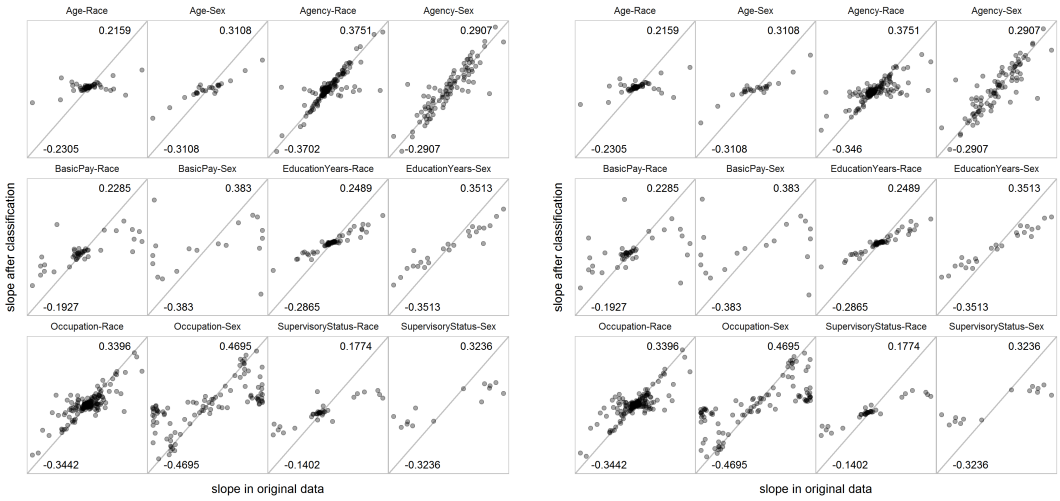
Ensemble size: 5
Node atts: Age, EducationYears, Agency, Occupation, SupervisoryStatus, BasicPay
Sampling weights:
pTrueClassified: 0.6365, 0.5743

Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon: 0.1
Leaf min n:1000

Figure 7: AD/SD attribute proportion observation comparison panel. Year, attribute collapsing values, node attribute sampling weights, ensemble size, n_B, and $\lambda = \epsilon/(\text{global sensitivity})$ indicated in table below images.

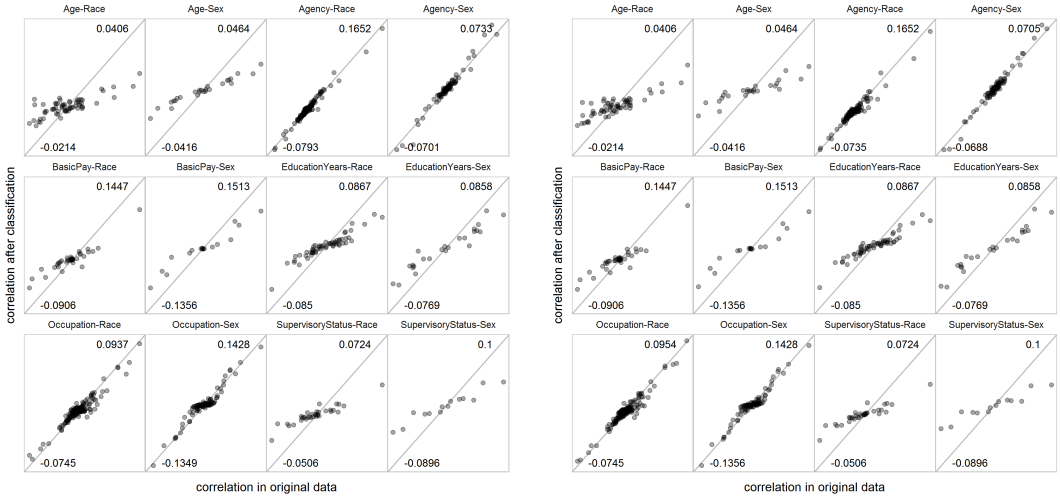
Effect of ϵ

Figure 8 compares $\epsilon = 0.1$ and $\epsilon = 0.4$. With $n_B = 1,000$, there does not appear to be a significant difference in distribution. Perhaps this is consistent the with p_i confidence intervals of figure 1 (p_i confidence intervals with respect to λ and ϵ). ϵ comparison plots generally do not show a significant difference between AD and BD observation distribution. An interesting feature of figure 8 is an apparent improved agreement between AD and BD *Agency* slopes for $\epsilon = 0.1$ vs. $\epsilon = 0.4$. Although counterintuitive, this may result from variation in attribute and label sampling during RDT construction and attribute synthesis.



(c1) Private vs. Synth p-Slopes, Epsilon = 0.1

(c2) Private vs. Synth p-Slopes, Epsilon = 0.4



(d1) Private vs. Synth p-Correlation, Epsilon = 0.1

(d2) Private vs. Synth p-Correlation, Epsilon = 0.4

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

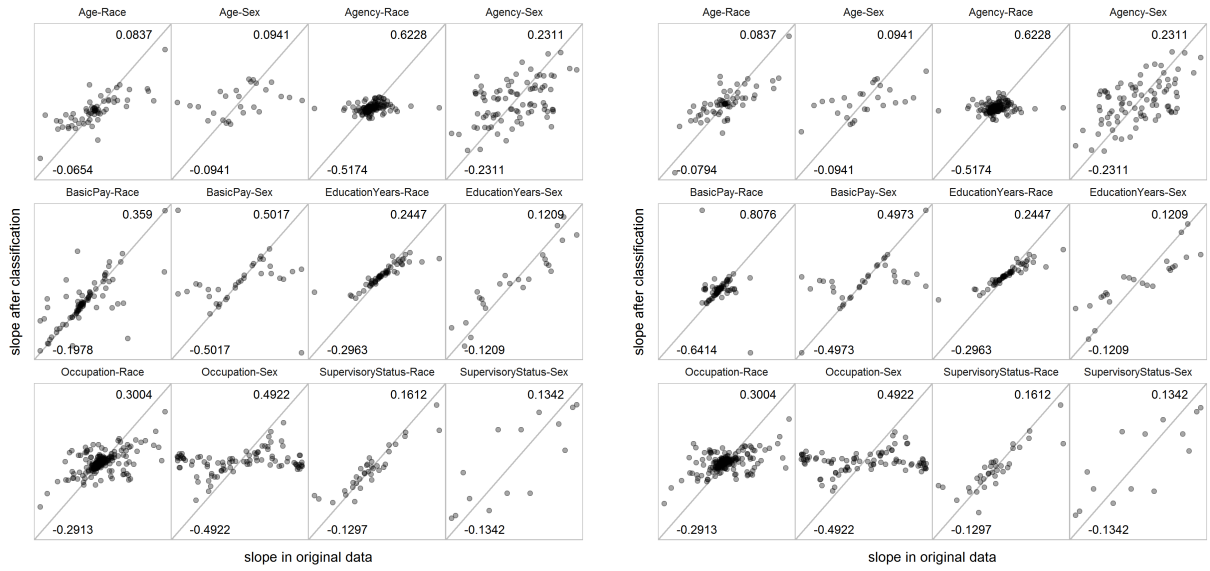
Ensemble size: 5
Node atts: Age, EducationYears, Agency, Occupation, SupervisoryStatus, BasicPay
Sampling weights: 1, 1, 60, 20, 1, 1
pTrueClassified: 0.6366, 0.5772

Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon:
Leaf min n:1000

Figure 8: AD/SD attribute proportion observation comparison panel. Year, attribute collapsing values, node attribute sampling weights, ensemble size, n_B, and $\lambda = \epsilon/(\text{global sensitivity})$ indicated in table below images.

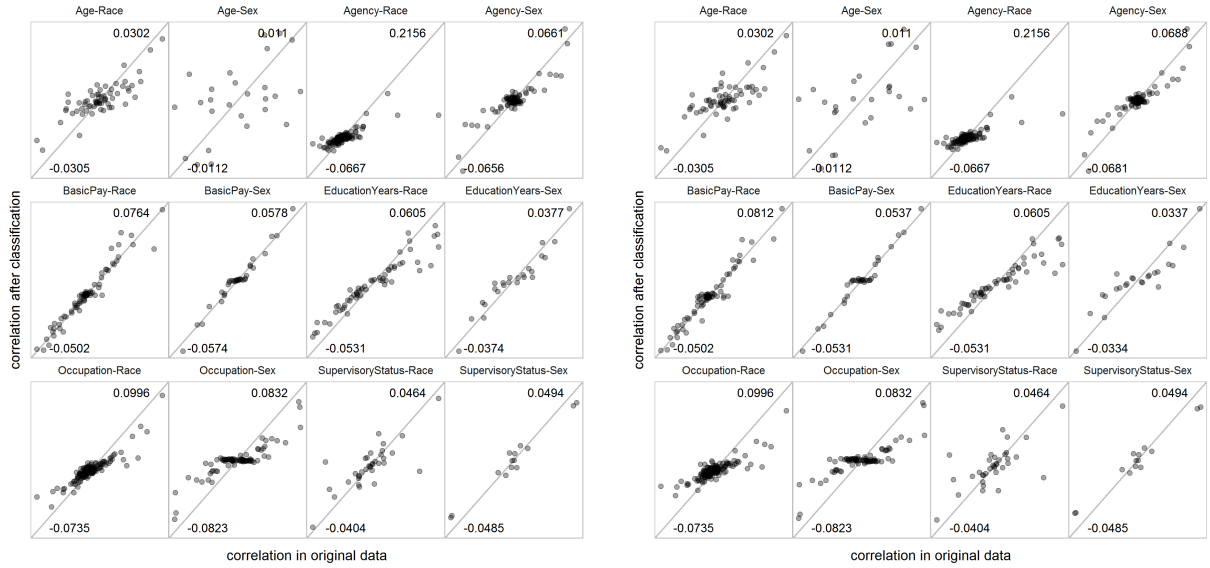
Effect of n_B

From the p_i confidence interval plot (figure **??**), we do not expect a significant difference in utility between $n_B = 1,000$ and $n_B = 2,000$. This is confirmed in figure 9 for a given year, level of collapsing, attribute sampling weight, and ϵ. Lack of influence of n_B is generally observed in remaining comparison plots.



(c1) Private vs. Synth p-Slopes, n-Leaf-Min = 1000

(c2) Private vs. Synth p-Slopes, n-Leaf-Min = 2000



(d1) Private vs. Synth p-Correlation, n-Leaf-Min = 1000

(d2) Private vs. Synth p-Correlation, n-Leaf-Min = 2000

Year: 2011
Collapse att: BasicPay, Agency, Occupation
Collapse lev: 25, 500, 200
nUnsynthesized: 0, 0

Ensemble size: 5
Node atts: Age, EducationYears, Agency, Occupation, SupervisoryStatus, BasicPay
Sampling weights: 1, 1, 1, 1, 1, 20
pTrueClassified: 0.5541, 0.4794

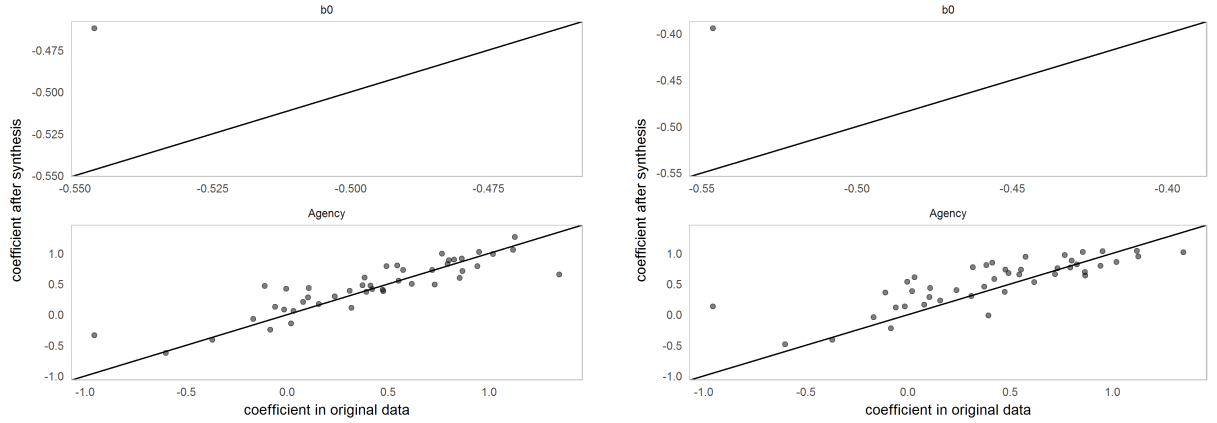
Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon: 0.1
Leaf min n:

Figure 9: AD/SD attribute proportion observation comparison panel. Year, attribute collapsing values, node attribute sampling weights, ensemble size, n_B, and $\lambda = \epsilon/(\text{global sensitivity})$ indicated in table below images.

Logistic Regression

Figures 10 through 14 compare coefficients from logistic regression models fit from AD and BD with proportion female as the response. Predictor attributes are as annotated.

- Figures 10 and 11 involve a single predictor (*Agency*) and indicate little difference in the results using two values of ϵ (0.1 and 0.4)
- Figures 12 and 13 involve *Agency* and *Occupation* predictors (note the increased weighting of *Occupation* in figure 13 (60) along with significant improvement of fit for certain occupations - note also the difference in ϵ)
- Figure 14 confirms the close fit for certain occupations when *Occupation* is the sole predictor after high RDT sampling preference



(a) P(female) Logistic Regression, Private vs. Synthetic Coefficients
Attribute Weights: Agency=60

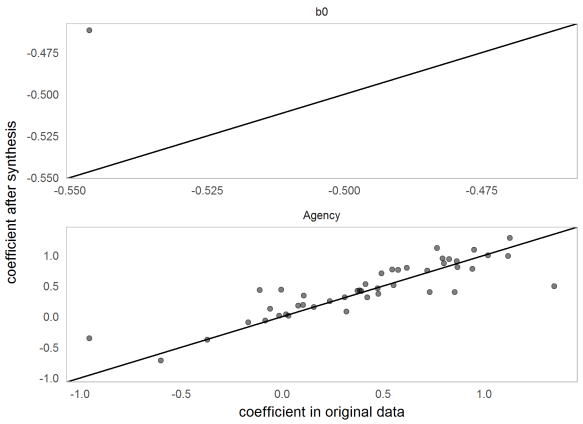
(b) P(female) Logistic Regression, Private vs. Synthetic Coefficients
Attribute Weights: Agency=1

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

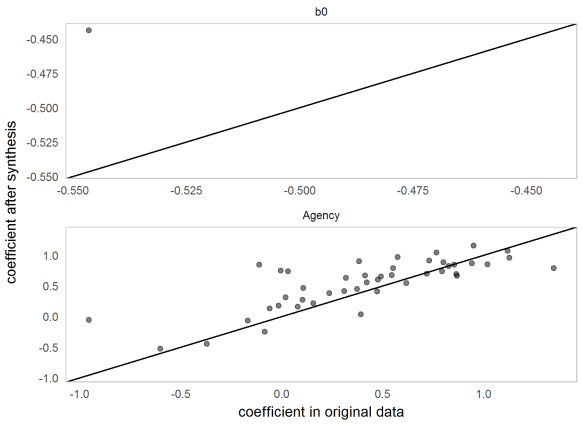
Ensemble size: 5
Node atts: Age, EducationYears, Agency, Occupation, SupervisoryStatus, BasicPay
Sampling weights:

Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon: 0.1
Leaf min n:1000

Figure 10: AD/SD proportion female logistic regression model coefficients. Year, attribute collapsing values, node attribute sampling weights, ensemble size, n_B, and $\lambda = \epsilon/(\text{global sensitivity})$ indicated in table below images.



(a) P(female) Logistic Regression, Private vs. Synthetic Coefficients
Attribute Weights: Agency=60



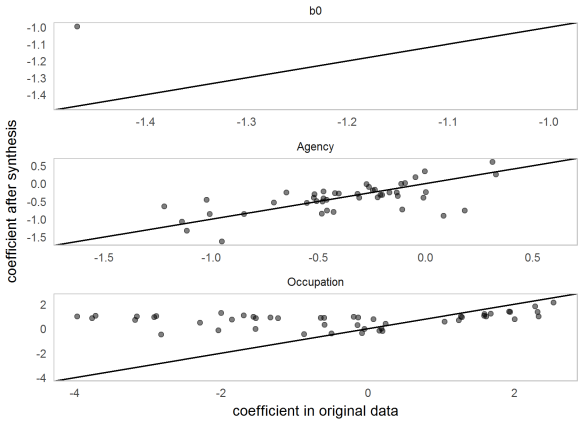
(b) P(female) Logistic Regression, Private vs. Synthetic Coefficients
Attribute Weights: Agency=1

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

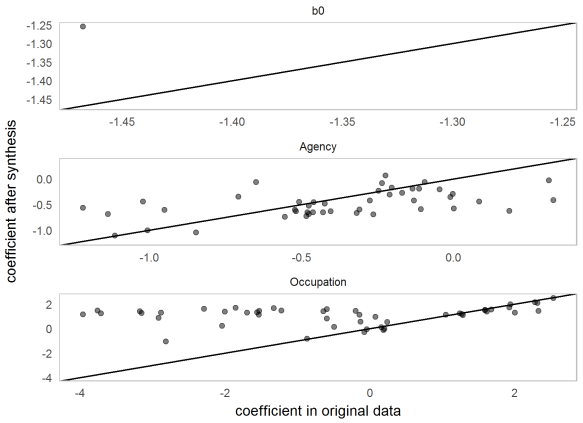
Ensemble size: 5
Node atts: Age, EducationYears, Agency, Occupation, SupervisoryStatus, BasicPay
Sampling weights:

Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon: 0.4
Leaf min n:1000

Figure 11: AD/SD proportion female logistic regression model coefficients. Year, attribute collapsing values, node attribute sampling weights, ensemble size, n_B, and $\lambda = \epsilon/(\text{global sensitivity})$ indicated in table below images.



(a) P(female) Logistic Regression, Private vs. Synthetic Coefficients
Attribute Weights: Agency=60, Occupation=20



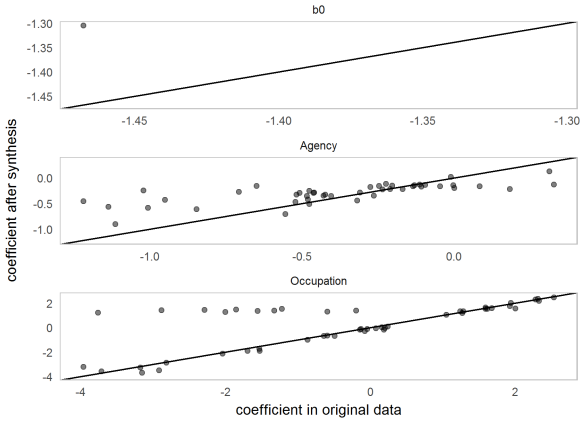
(b) P(female) Logistic Regression, Private vs. Synthetic Coefficients
Attribute Weights: Agency=1, Occupation=20

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

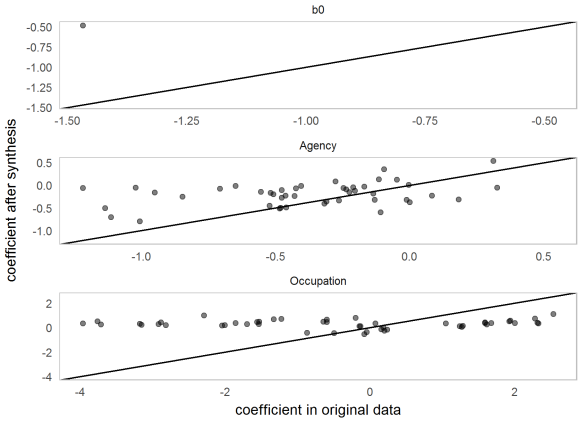
Ensemble size: 5
Node atts: Age, EducationYears, Agency, Occupation, SupervisoryStatus, BasicPay
Sampling weights:

Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon: 0.1
Leaf min n:1000

Figure 12: AD/SD proportion female logistic regression model coefficients. Year, attribute collapsing values, node attribute sampling weights, ensemble size, n_B, and $\lambda = \epsilon/(\text{global sensitivity})$ indicated in table below images.



(a) P(female) Logistic Regression, Private vs. Synthetic Coefficients
Attribute Weights: Agency=1, Occupation=60



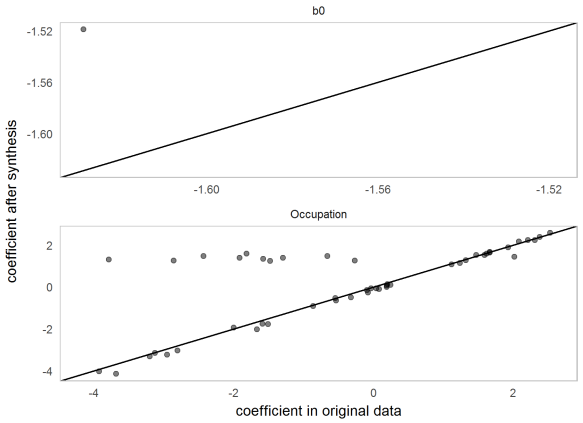
(b) P(female) Logistic Regression, Private vs. Synthetic Coefficients
Attribute Weights: Agency=1, Occupation=1

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

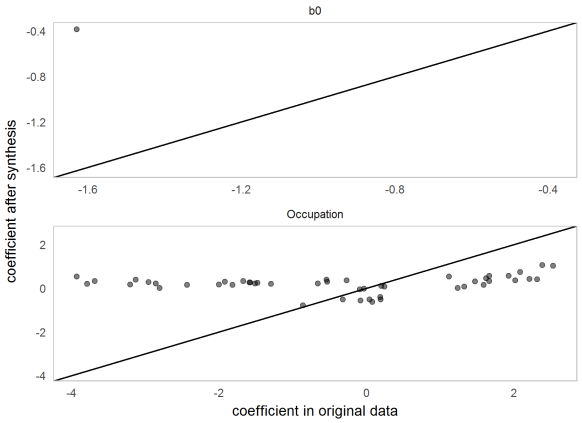
Ensemble size: 5
Node atts: Age, EducationYears, Agency, Occupation, SupervisoryStatus, BasicPay
Sampling weights:

Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon: 0.4
Leaf min n:1000

Figure 13: AD/SD proportion female logistic regression model coefficients. Year, attribute collapsing values, node attribute sampling weights, ensemble size, n_B, and $\lambda = \epsilon/(\text{global sensitivity})$ indicated in table below images.



(a) P(female) Logistic Regression, Private vs. Synthetic Coefficients
Attribute Weights: Occupation=60



(b) P(female) Logistic Regression, Private vs. Synthetic Coefficients
Attribute Weights: Occupation=1

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:

Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon: 0.4
Leaf min n:1000

Figure 14: AD/SD proportion female logistic regression model coefficients. Year, attribute collapsing values, node attribute sampling weights, ensemble size, n_B, and $\lambda = \epsilon/(\text{global sensitivity})$ indicated in table below images.