Synthetic Data Generation Using Variable-Depth Random Decision Trees

OPM Pay Disparity Models

Tom Balmat, Duke University Sythenthetic Data Project

March 23, 2019

Using random decision tree (RDT) construction methods as proposed by Jagannathan et al., modified to continue branch splitting while observation count is above a specified minimum threshold (as observed in joint categories of the DIBBS synthetic data), with predictor attributes age, eduction, agency (first two positions), occupation (first two positions), and basic pay (rounded to various multiples of $1,000), ensembles of RDTs were used to generate synthetic sex and race labels for each authentic OPM observation in years 1988, 1996, 2003, and 2011. Table 1 describes RDT construction and data synthesis parameters. Two evaluations are made: one where sex alone is synthesized and one where race is synthesized after including synthetic sex as a predictor attribute. Once synthetic data sets are generated for each year, the following fixed effects pay disparity model is fit to each authentic and synthetic annual subset:

$$\ln(basic\ pay) = \beta_0 + \beta_{sex} + \beta_{race} + \beta_{age} \times age + \beta_{age^2} \times age^2 + \beta_{years_{ed}} \times years_{ed} + \beta_{agency} + \beta_{occupation} \qquad (1)$$

Comparison of β_{sex} and β_{race} estimates for corresponding sets of authentic and synthetic data are then made using year-longitudinal line graphs. β_{sex} and β_{race} estimates are labeled Sex(F), Race(A), Race(B), Race(C), and Race(D) for female, Native American, Asian, Hispanic, and Black respectively (sex male and race White are reference levels). Indicated values are an estimate of the effect of a parameter (sex or race) on $\ln(basic\ pay)$. Close agreement of annual values and overall trend through the study period indicates high utility of the synthetic data for the study of pay disparity by sex and race within the federal agencies and occupation categories included. A factorial-like experiment was conducted involving all combinations of various levels of RDT construction and data synthesis parameters. The following sections contain comparative plots that identify levels of synthesis parameters associated with high and low utility, along with parameters that do not have apparent effect on utility. It should be noted that significant interactions are present, in that a synthesis parameter may appear influential under a subset of other parameters and levels, but not all. Examples of such interactions are discussed.

Table 1: RDT construction and data synthesis parameters

Parameter	Function	Example Values
Attribute collapsing	Combine all labels of a given attribute with frequency below a specified threshold into a single label	pay (rounded to nearest \$2,000), agency (combine all with frequency<100), occupation (combine all with frequency<250))
Ensemble size	Number of RDTs to generate	3, 6
Branch growth threshold, n_B	RDT branch observation count below which to discontinue splitting (responsible for variable-depth trees)	25, 500
Node attribute selection weight	Weights used in sampling predictor attributes during RDT node splitting (if attributes x_1 and x_2 have weights w_1 and w_2, respectively, then for a given branch split, x_2 is w_2/w_1 as likely to be chosen than is x_1	1, 40
Global sensitivity	Maximum observable deviation in joint frequency after removal of a single observation	Fixed at 1 when, as in our case, all attributes are categorical
ϵ	Dispersion parameter of Laplace distribution used to to generate RDT leaf distributions (used in synthesis attribute label selection)	0.25, 0.75

1 Synthesis of Sex Using Authentic Age, Education, Agency, Occupation, Basic Pay, and Race

Subsections 1.1 through 1.5 contain panels of plots that assess the agreement of model (1) parameter estimates for compatible authentic and synthetic data subsets when levels of two synthesis parameters are simultaneously varied. Each panel contains two sets of plots (left and right), one for each level of a given synthesis parameter, each set contains one subset of graphs for each β_{sex} and β_{race} parameter, and each subset includes one dashed line for annual estimates derived from synthetic data sets generated for each level of a second synthesis parameter and a solid line for annual estimates derived from corresponding authentic data. High utility is indicated when points and lines corresponding to synthetic data are near, in proximity and trend, to those of corresponding authentic data. Divergence of synthetic lines within a single β_{sex} or β_{race} plot indicates variation in influence of the corresponding synthesis parameter.

1.1 Attribute Collapsing

During RDT construction, nodes are split at a given level by randomly selecting a predictor attribute, from those not previously selected in the branch, and creating a set of sub-nodes, one for each label of the selected attribute. High frequency attributes generate a large volume of sub-nodes, each to be subsequently split, which increases the time required to generate a tree. Combining attribute labels with observation frequency below a specified threshold into a single label reduces the number of branch splits required during RDT construction, reducing computation time. However, a reduction in utility may be expected, especially in attributes with low frequency labels, relative to specified collapsing thresholds. Figure 1 compares two collapsing levels for basic pay, agency, and occupation. Note that "collapsing" of basic pay actually results in rounding of pay to the nearest indicated value, in thousands ($2,500 or $5,000 for this comparison). It is seen that, in this case, neither ϵ nor collapsing have much effect on estimates. Figure 2 reveals negligible effect of attribute collapsing, but an apparent effect of ϵ. Since the sole difference in the two figures is the minimum branch growth threshold, n_B, an interaction effect involving ϵ and n_B is revealed, in that ϵ has effect for one level of n_B, but not the other. Due to a general lack of effect of collapsing agencies and occupations in all combinations studied, further analysis will limit collapsing to basic pay rounding.[1]

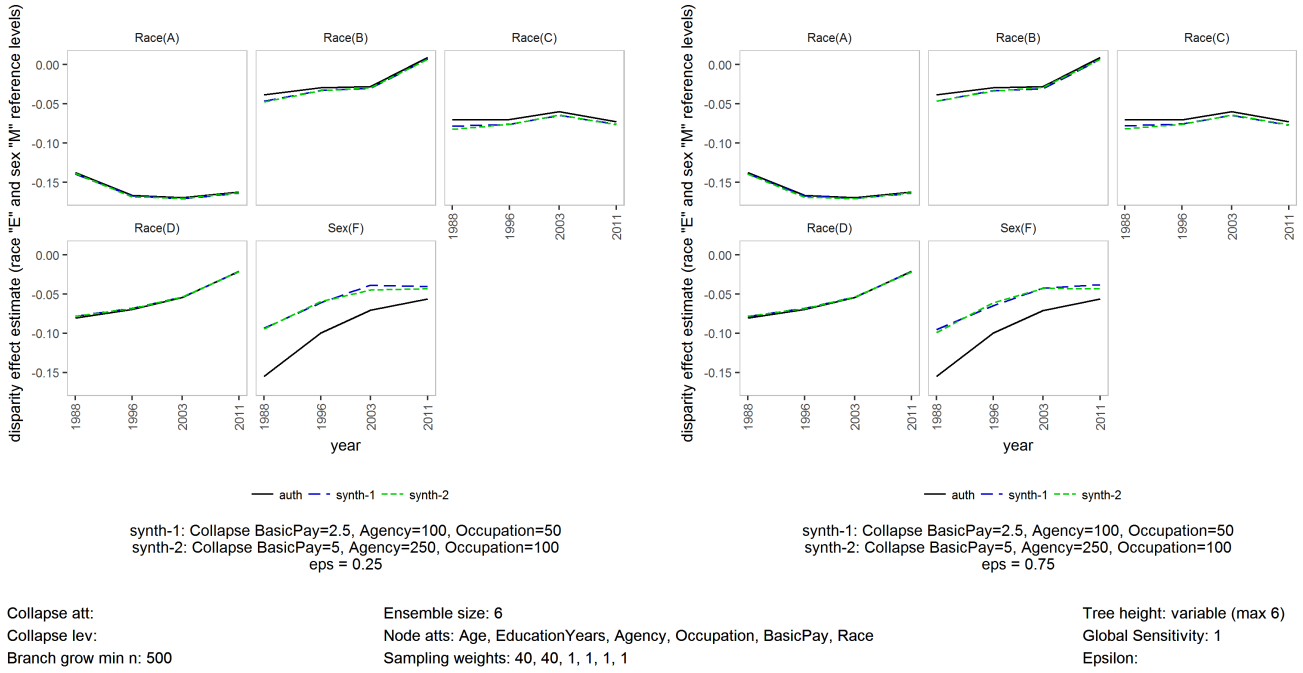


Figure 1: OPM pay disparity fixed effects model estimates. Results from authentic data (solid lines) and synthetic data (one dashed line for each level of attribute collapsing). Left and right panels compare results for $\epsilon \in \{0.25, 0.75\}$. $n_B = 500$. Levels of remaining synthesis parameters listed in table below plots.

[1] Due to existence of thousands of distinct pay values, some level of consolidation is required in order to make tree construction feasible.

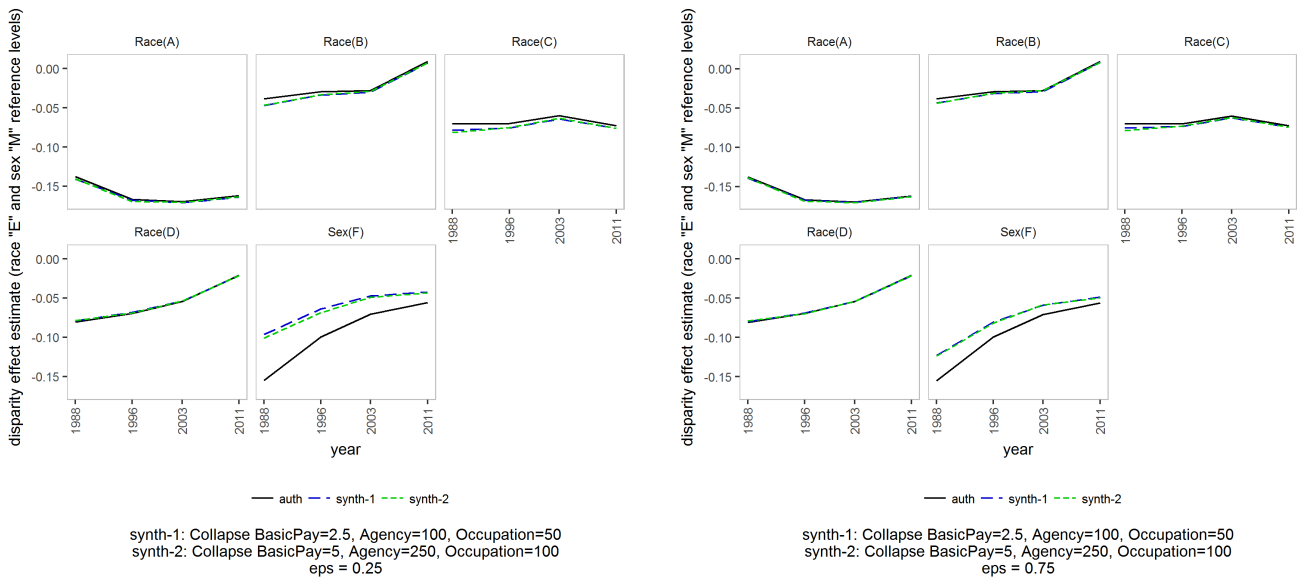
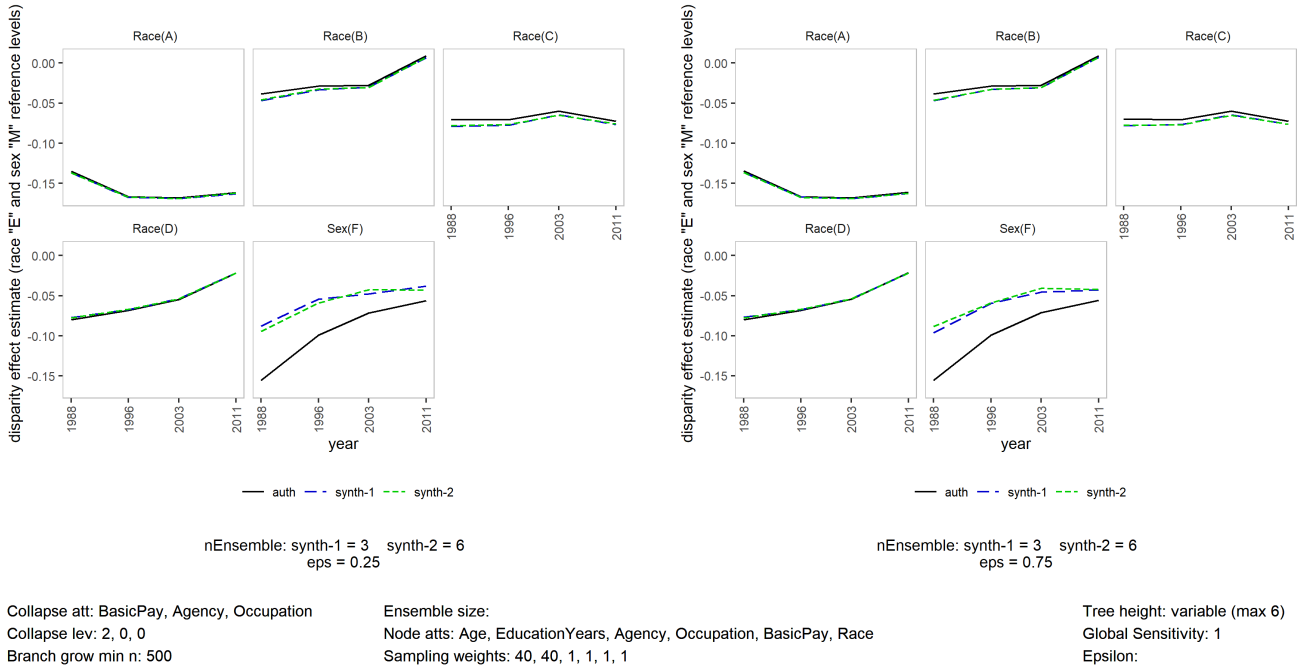


Figure 2: OPM pay disparity fixed effects model estimates. Results from authentic data (solid lines) and synthetic data (one dashed line for each level of attribute collapsing). Left and right panels compare results for $\epsilon \in \{0.25, 0.75\}$. $n_B = 25$. Levels of remaining synthesis parameters listed in table below plots.

1.2 Ensemble Size

Jagannathan et al. suggest that increasing the number of trees in an ensemble improves classification accuracy and generally recommend ensembles of size ten or more. However, with our authentic data and the OPM pay models studied, differences in utility of synthetic data generated with ensembles of three and six trees are negligible. Figures 3 and 4 show typical differences in model parameter estimates for ensembles of size three and six. A change in ϵ, from 0.25 to 0.75, appears to have an effect in these plots, but only when combined with a change in n_B, from 500 to 25. This is another interaction, but one involving ϵ and n_B, not ensemble size. Note that distinct branches within a single tree represent disjoint subsets of data and subsets for distinct years are also disjoint. Since all authentic observations in an annual subset are used in construction of each tree, total ϵ usage is $\epsilon \times n_{ensemble}$. From results presented so far, it appears that an ensemble size as small as three may give reasonable control over utility for the model entertained, in that specifying an ensemble size greater than three does not appear to improve utility.

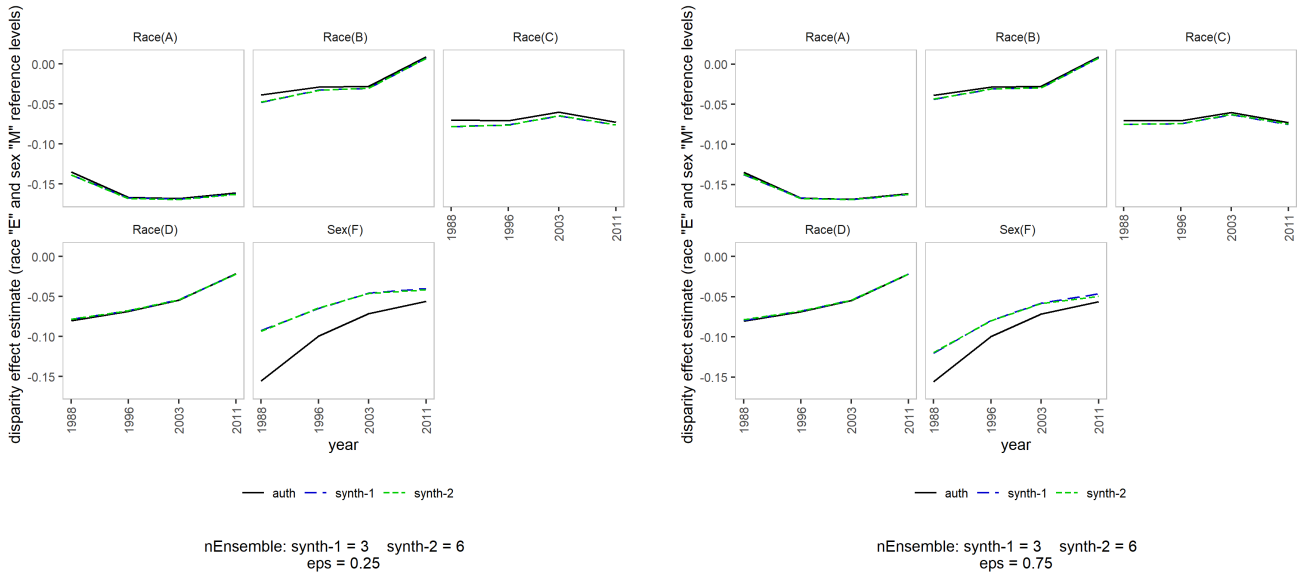


Collapse att: BasicPay, Agency, Occupation
Collapse lev: 2, 0, 0
Branch grow min n: 500

Ensemble size:
Node atts: Age, EducationYears, Agency, Occupation, BasicPay, Race
Sampling weights: 40, 40, 1, 1, 1, 1

Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon:

Figure 3: OPM pay disparity fixed effects model estimates. Results from authentic data (solid lines) and synthetic data (one dashed line for each $n_{ensemble}$). Left and right panels compare results for $\epsilon \in \{0.25, 0.75\}$. $n_B = 500$. Levels of remaining synthesis parameters listed in table below plots.



Collapse att: BasicPay, Agency, Occupation
Collapse lev: 2, 0, 0
Branch grow min n: 25

Ensemble size:
Node atts: Age, EducationYears, Agency, Occupation, BasicPay, Race
Sampling weights: 40, 40, 1, 1, 1, 1

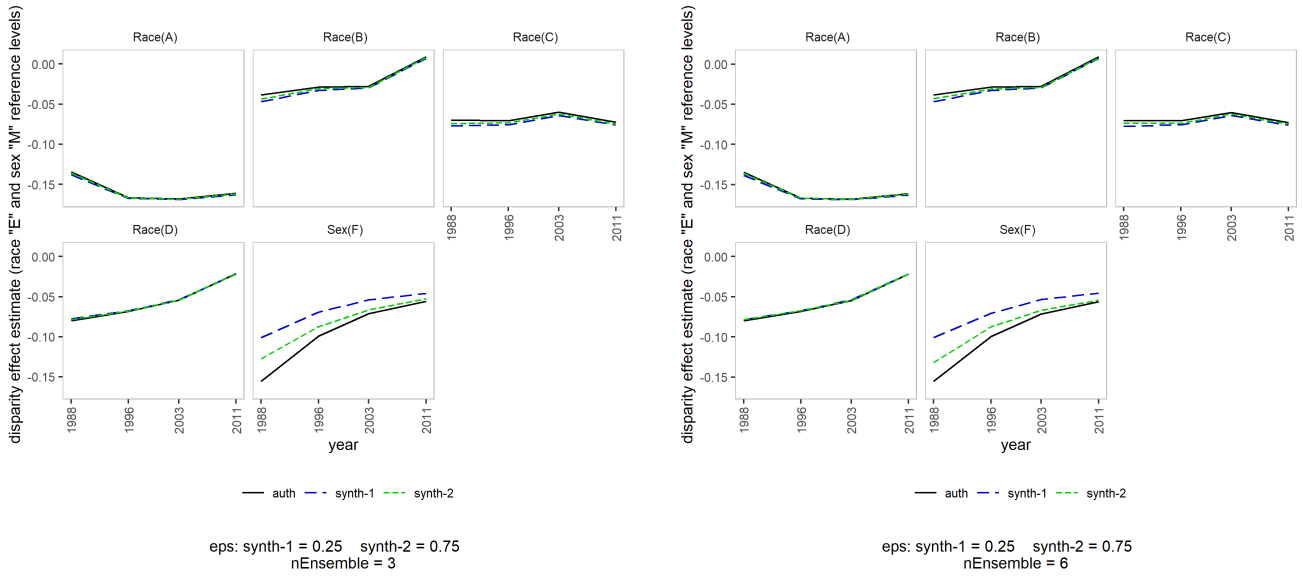
Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon:

Figure 4: OPM pay disparity fixed effects model estimates. Results from authentic data (solid lines) and synthetic data (one dashed line for each $n_{ensemble}$). Left and right panels compare results for $\epsilon \in \{0.25, 0.75\}$. $n_B = 25$. Levels of remaining synthesis parameters listed in table below plots.

1.3 ϵ

The dispersion parameter of the Laplace distribution is a function of ϵ, such that smaller values of ϵ result in increased dispersion and reduced utility of synthetic data, while larger values result in less dispersion and higher utility. To minimize total ϵ used during synthesis, RDTs should be generated using the smallest possible value of ϵ that produces data with adequate utility. Figures 5 through 7 compare estimates from models fit to authentic data and synthetic data generated using ϵ values of 0.25 and 0.75, ensemble sizes of three and six, and various levels of remaining synthesis parameters. Figure 5 reveals an apparent effect of ϵ that does not change with ensemble size (this result is similar to that observed in figure 4 with, however, a different set of node attribute sampling weights). Figure 6 shows a difference in effect of ϵ for minimum branch growth values $n_B = 25$ and $n_B = 500$. Figure 7 also shows a change in effect of ϵ for different values of n_B, but in this case from relatively low utility ($n_B = 25$) to apparent high utility ($n_B = 500$). Recognition of synthesis parameter interaction, combined with observation of apparent high utility combinations of parameter values, reinforces the need for careful study and selection of particular values used to generate synthetic data for actual use. However, the question of how much ϵ is consumed in making such inquiries must be raised.

An important consideration is that ensemble size and ϵ are the sole synthesis parameters that contribute to ϵ consumption, in a differentially private sense, and we have observed cases in which an increase in either does not improve the utility of resulting synthetic data. This presents an opportunity to identify combinations of synthesis parameter values that include ensemble size and ϵ that lead to an absolute minimum ϵ consumption.

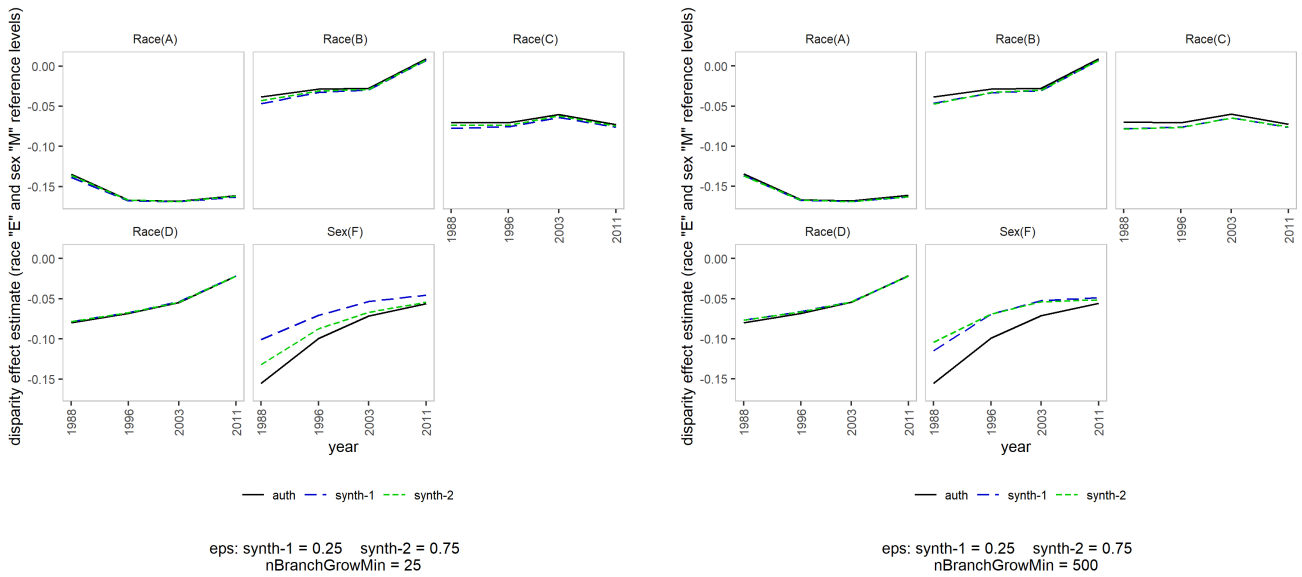


Collapse att: BasicPay, Agency, Occupation
Collapse lev: 2, 0, 0
Branch grow min n: 25

Ensemble size:
Node atts: Age, EducationYears, Agency, Occupation, BasicPay, Race
Sampling weights: 1, 1, 40, 40, 1, 1

Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon:

Figure 5: OPM pay disparity fixed effects model estimates. Results from authentic data (solid lines) and synthetic data (one dashed line for each ϵ value). Left and right panels compare results from ensembles of size three and six. Levels of remaining synthesis parameters listed in table below plots.



eps: synth-1 = 0.25 synth-2 = 0.75
nBranchGrowMin = 25

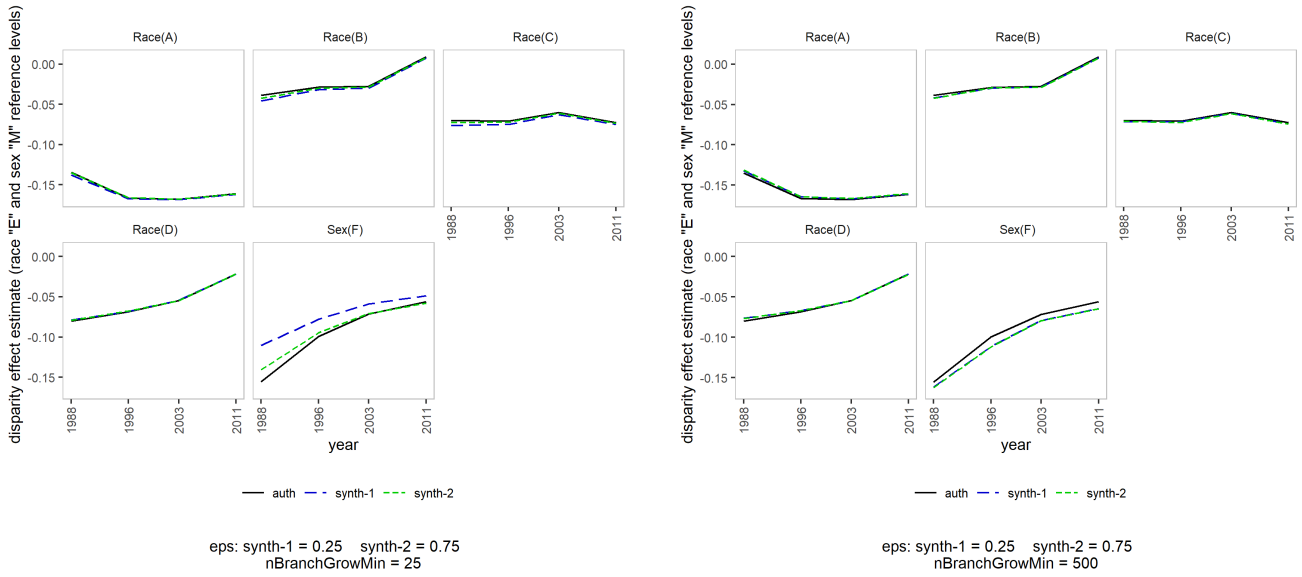
eps: synth-1 = 0.25 synth-2 = 0.75
nBranchGrowMin = 500

Collapse att: BasicPay, Agency, Occupation
Collapse lev: 2, 0, 0
Branch grow min n:

Ensemble size: 6
Node atts: Age, EducationYears, Agency, Occupation, BasicPay, Race
Sampling weights: 1, 1, 40, 40, 1, 1

Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon:

Figure 6: OPM pay disparity fixed effects model estimates. Results from authentic data (solid lines) and synthetic data (one dashed line for each ϵ value). Left and right panels compare results from $n_B = 25$ and $n_B = 500$. Levels of remaining synthesis parameters listed in table below plots.



eps: synth-1 = 0.25 synth-2 = 0.75
nBranchGrowMin = 25

eps: synth-1 = 0.25 synth-2 = 0.75
nBranchGrowMin = 500

Collapse att: BasicPay, Agency, Occupation
Collapse lev: 2, 0, 0
Branch grow min n:

Ensemble size: 6
Node atts: Age, EducationYears, Agency, Occupation, BasicPay, Race
Sampling weights: 1, 1, 1, 1, 40, 1

Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon:

Figure 7: OPM pay disparity fixed effects model estimates. Results from authentic data (solid lines) and synthetic data (one dashed line for each ϵ value). Left and right panels compare results from $n_B = 25$ and $n_B = 500$. Levels of remaining synthesis parameters listed in table below plots.

1.4 Minimum Branch Growth Threshold, n_B

During RDT construction each branch is split at nodes corresponding to labels of randomly chosen predictor attributes and branch depth is the number of times a branch is split. With each split a new branch is created that represents a subset of observations of diminishing count. Each complete branch terminates with a leaf containing an attribute classification probability distribution derived from the total set of observations corresponding to all node attributes and labels appearing in the branch. Confidence intervals of proportions of labels sampled from the leaf distributions are a function of branch observation count and ϵ. To minimize confidence interval widths, branch splitting is terminated when observation frequency diminishes to below a threshold n_B. This leads to branches of different depths and a competition between attribute label sampling precision and overall data utility, in that high values of n_B result in shallow trees with low representation of covariate relationships, but high classification precision due to relatively narrow sampling confidence intervals, while low vales of n_B yield deep branches with good representation of covariate relationships, but also relatively wide label sampling confidence intervals. Figure 8 plots 0.90 confidence intervals for label sampling proportions from a Laplace perturbed leaf distribution, given actual (specified) proportions of labels and ϵ values (ϵ is indicated as lambda in the plot and CI bars are arranged left to right within each n_B group from $\epsilon = 0.1$ to $\epsilon = 1.0$). For typical values of $n_B \in \{25, 500\}$ and $\epsilon \in \{0.25, 0.75\}$ used in the study, a significant reduction of confidence interval width is observed within the $n_B = 25$ group (first set of CIs in each plot) from $\epsilon = 0.25$ to $\epsilon = 0.75$ (third to eighth intervals), but the reduction for $n_B = 500$ (fifth set of intervals) does not appear significant. This raises the expectation that, for a given set of synthesis parameter values, one combination of tested ϵ and n_B values yields optimal utility.

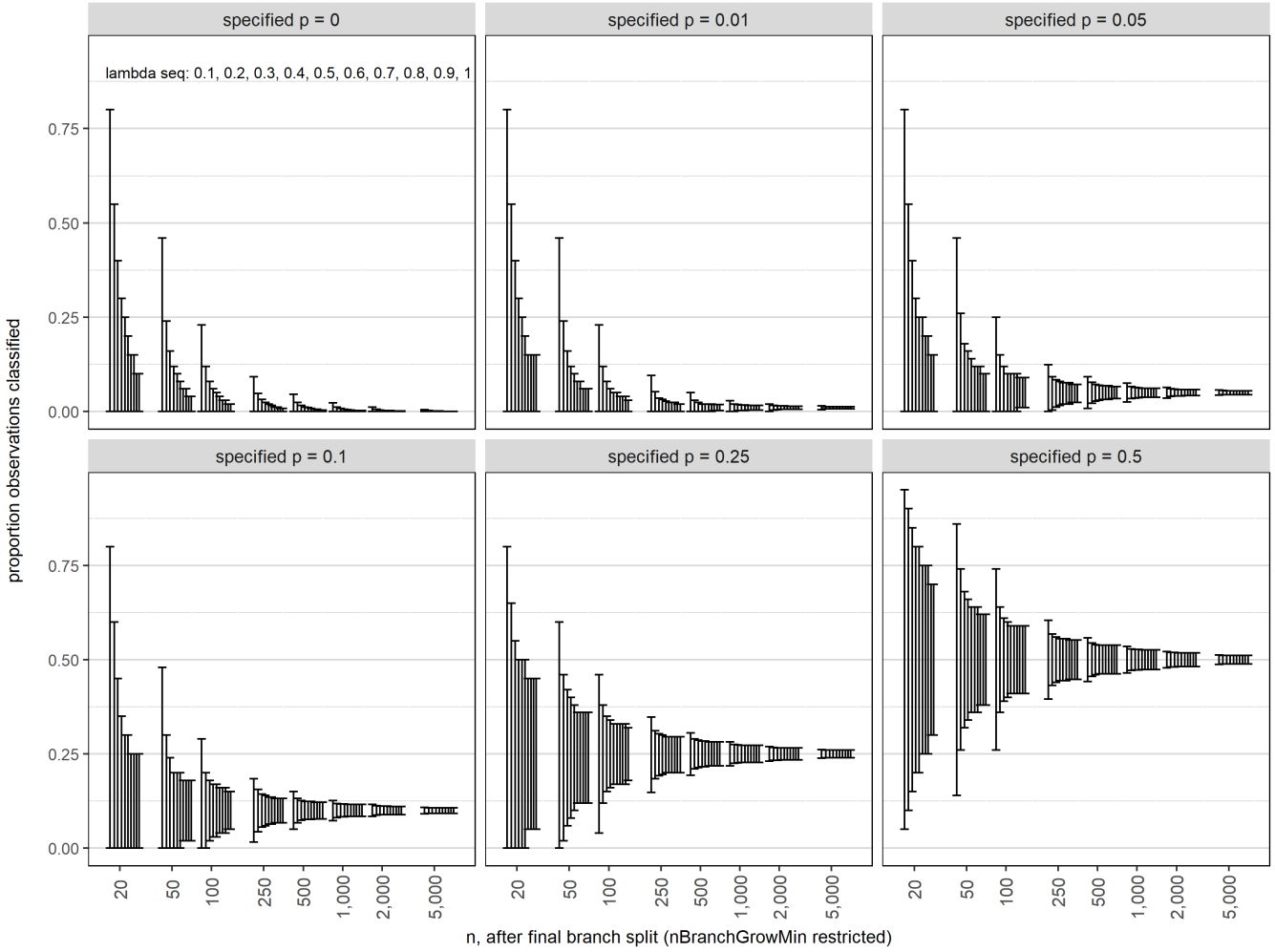
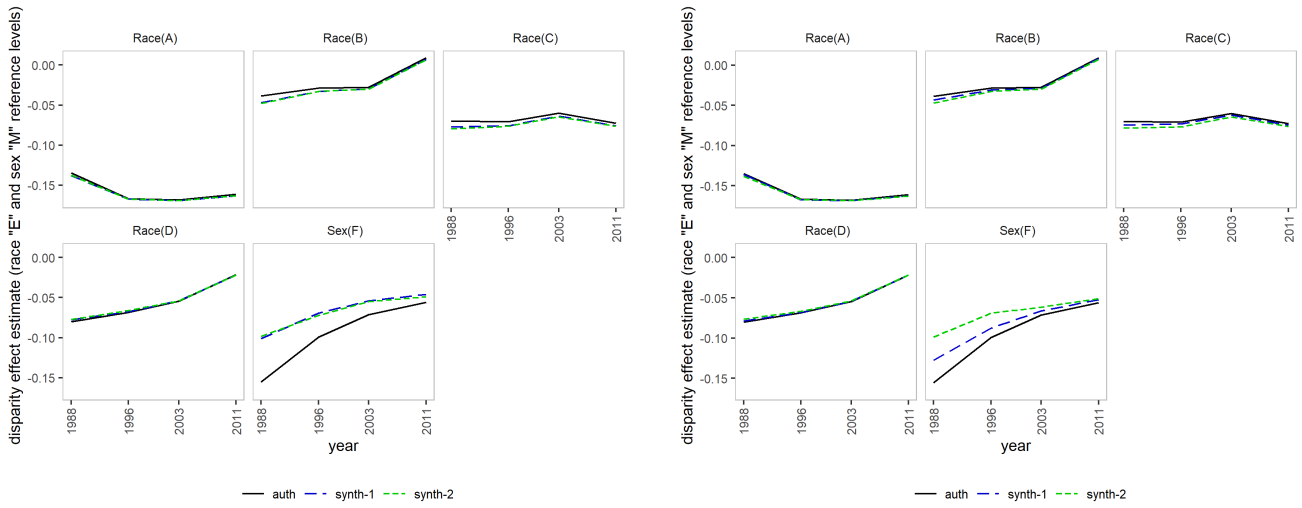


Figure 8: Theoretical joint Laplace-binomial 0.90 confidence intervals on label sampling proportions, given specified theoretical (specified) label proportions. $\lambda = \epsilon$ from 0.1 to 1.0 in 1/10 increments. Minimum value of n_B to achieve CI on x-axis.

Figures 9 and 10 show n_B interacting with several other synthesis parameters at specific levels. In figure 9 we see n_B and ϵ interact with $n_B = 500$ having different effects for the two ϵ values indicated, and in figure 10 we see n_B and ϵ interact with $n_B = 25$ having different effects. Note the difference in node attribute sampling weights between the two figures. This reveals further interaction involving attribute sampling.

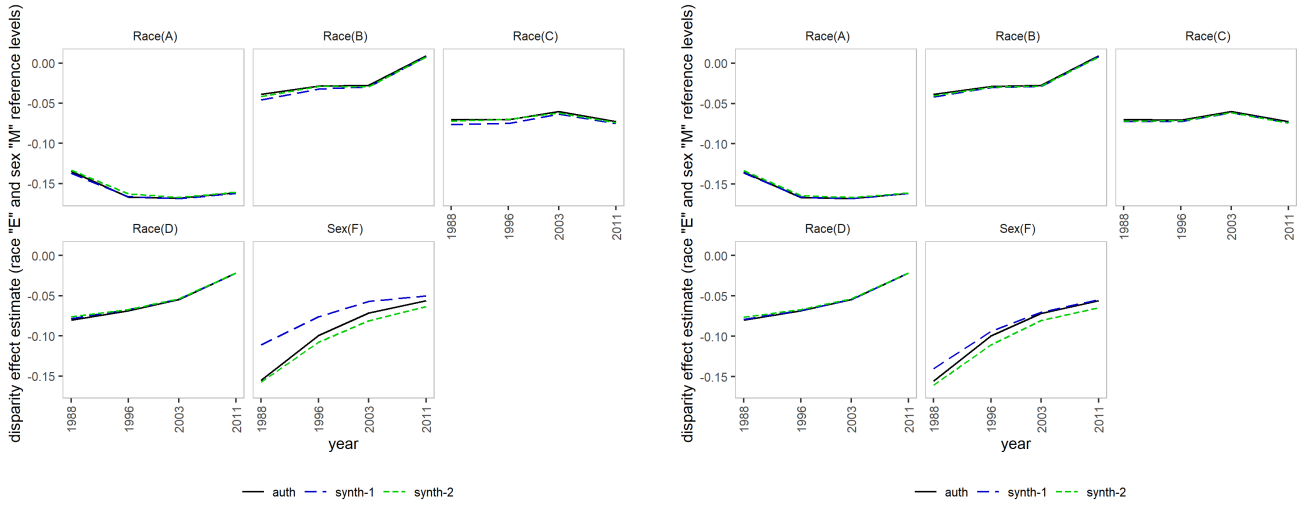


Collapse att: BasicPay, Agency, Occupation
Collapse lev: 2, 0, 0
Branch grow min n:

Ensemble size: 3
Node atts: Age, EducationYears, Agency, Occupation, BasicPay, Race
Sampling weights: 1, 1, 40, 40, 1, 1

Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon:

Figure 9: OPM pay disparity fixed effects model estimates. Results from authentic data (solid lines) and synthetic data (one dashed line for each n_B value). Left and right panels compare results from $\epsilon = 0.25$ and $\epsilon = 0.5$. Levels of remaining synthesis parameters listed in table below plots.



Collapse att: BasicPay, Agency, Occupation
Collapse lev: 2, 0, 0
Branch grow min n:

Ensemble size: 3
Node atts: Age, EducationYears, Agency, Occupation, BasicPay, Race
Sampling weights: 1, 1, 1, 1, 40, 1

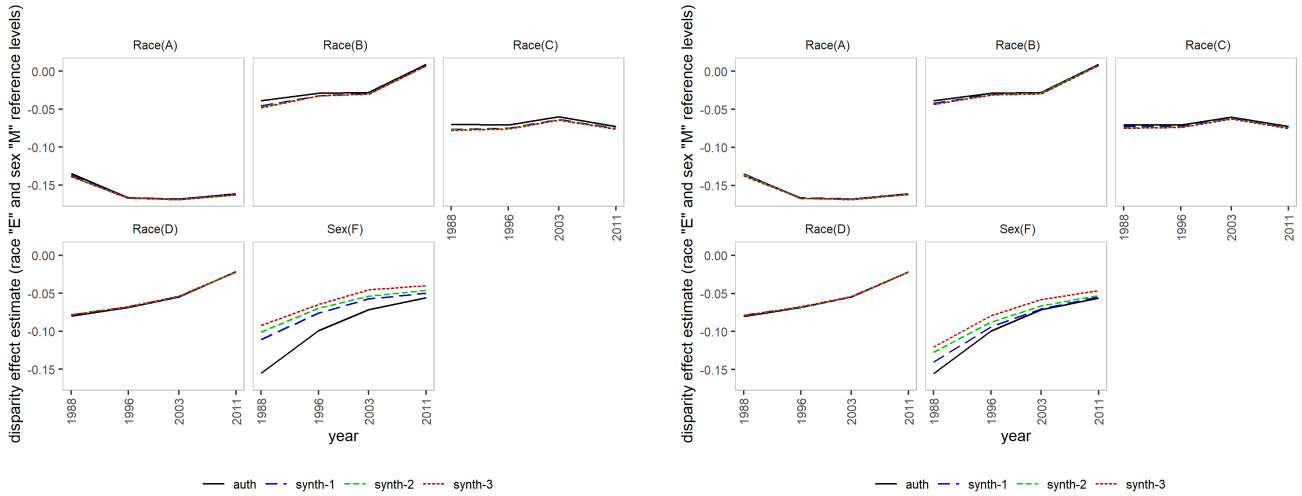
Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon:

Figure 10: OPM pay disparity fixed effects model estimates. Results from authentic data (solid lines) and synthetic data (one dashed line for each n_B value). Left and right panels compare results from $\epsilon = 0.25$ and $\epsilon = 0.5$. Levels of remaining synthesis parameters listed in table below plots.

1.5 Node Attribute Sampling Weight

Of the synthesis parameter value combinations reviewed so far, only two generate synthetic data with apparent utility (in the context of model 1): those used to generate the right hand panels of figures 7 and 10. In both of these sets the RDT node attribute sampling weight of basic pay is significantly greater than that of the remaining parameters. This increases the probability of basic pay appearing as a predictor attribute in RDT branches and, since the disparity model has pay as its response, it seems reasonable for data generated by trees that capture covariate relationships involving pay to yield better fitting models than do data from trees that lack such relationships. "Better" implies models with parameter estimates near those derived from equivalent models fit to the authentic data. However, as the left hand panels of figures 7 and 10 reveal, increased sampling weight of basic pay is not sufficient to guarantee utility. In reviewing results of all tested synthesis parameter value combinations, it was observed that all combinations with apparent utility involved increased basic pay sampling weight, but not all combinations with increased basic pay sampling weight exhibited utility. Figures 11 through 16 compare results from data generated using three node attribute sampling vectors. It is seen that there exist several combinations of synthesis parameter values that yield data with apparent utility. Each involves increased basic pay sampling priority and we are especially interested those that specify low values of ϵ and ensemble size.

Figures 11 and 12 compare results for $\epsilon \in \{0.25, 0.75\}$ and $n_B \in \{25, 500\}$. The effect of ϵ appears negligible while, limited to increased basic pay sampling weight, an improvement in utility is observed comparing $n_B = 500$ to $n_B = 25$.



synth-1: Node-att-wt Age=1, Ed=1, Agency=1, Occ=1, Pay=40, Race=1
synth-2: Node-att-wt Age=1, Ed=1, Agency=40, Occ=40, Pay=1, Race=1
synth-3: Node-att-wt Age=40, Ed=40, Agency=1, Occ=1, Pay=1, Race=1
eps = 0.25

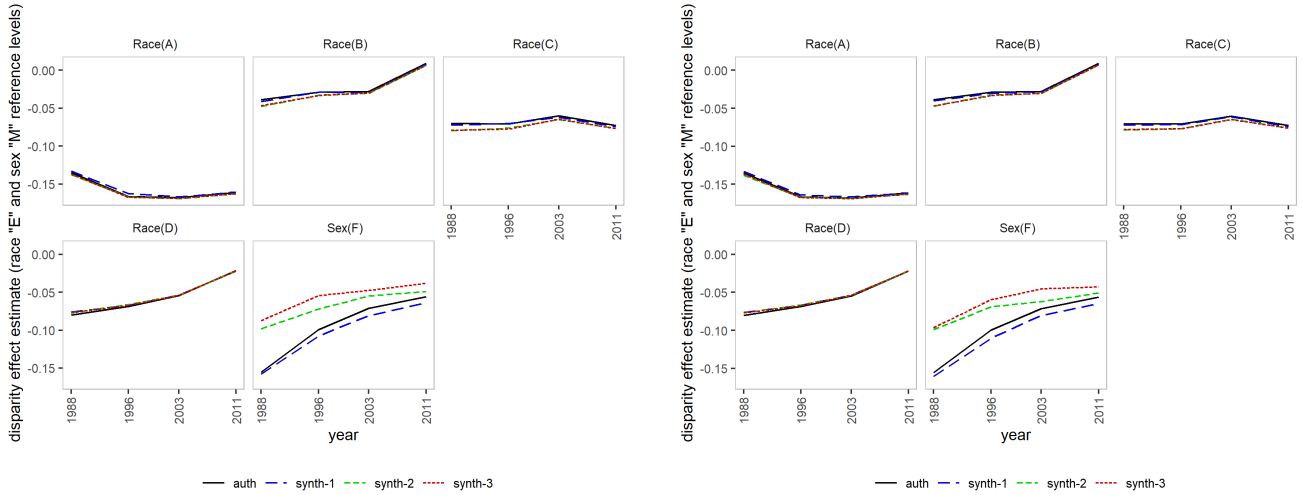
synth-1: Node-att-wt Age=1, Ed=1, Agency=1, Occ=1, Pay=40, Race=1
synth-2: Node-att-wt Age=1, Ed=1, Agency=40, Occ=40, Pay=1, Race=1
synth-3: Node-att-wt Age=40, Ed=40, Agency=1, Occ=1, Pay=1, Race=1
eps = 0.75

Collapse att: BasicPay, Agency, Occupation
Collapse lev: 2, 0, 0
Branch grow min n: 25

Ensemble size: 3
Node atts: Age, EducationYears, Agency, Occupation, BasicPay, Race
Sampling weights:

Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon:

Figure 11: OPM pay disparity fixed effects model estimates. Results from authentic data (solid lines) and synthetic data (one dashed line for each attribute weighting vector). Left and right panels compare results for $\epsilon = 0.25$ and $\epsilon = 0.75$. $n_B = 25$. Ensemble size=three.



synth-1: Node-att-wt Age=1, Ed=1, Agency=1, Occ=1, Pay=40, Race=1
synth-2: Node-att-wt Age=1, Ed=1, Agency=40, Occ=40, Pay=1, Race=1
synth-3: Node-att-wt Age=40, Ed=40, Agency=1, Occ=1, Pay=1, Race=1
eps = 0.25

synth-1: Node-att-wt Age=1, Ed=1, Agency=1, Occ=1, Pay=40, Race=1
synth-2: Node-att-wt Age=1, Ed=1, Agency=40, Occ=40, Pay=1, Race=1
synth-3: Node-att-wt Age=40, Ed=40, Agency=1, Occ=1, Pay=1, Race=1
eps = 0.75

Collapse att: BasicPay, Agency, Occupation
Collapse lev: 2, 0, 0
Branch grow min n: 500

Ensemble size: 3
Node atts: Age, EducationYears, Agency, Occupation, BasicPay, Race
Sampling weights:

Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon:

Figure 12: OPM pay disparity fixed effects model estimates. Results from authentic data (solid lines) and synthetic data (one dashed line for each attribute weighting vector). Left and right panels compare results for $\epsilon = 0.25$ and $\epsilon = 0.75$. $n_B = 500$. Ensemble size=three.

Figures 13 and 14 compare results for ensembles of three and six trees and $n_B \in \{25, 500\}$. The effect of ensemble size appears negligible, while, limited to increased basic pay sampling weight, an improvement in utility is observed comparing $n_B = 500$ to $n_B = 25$.

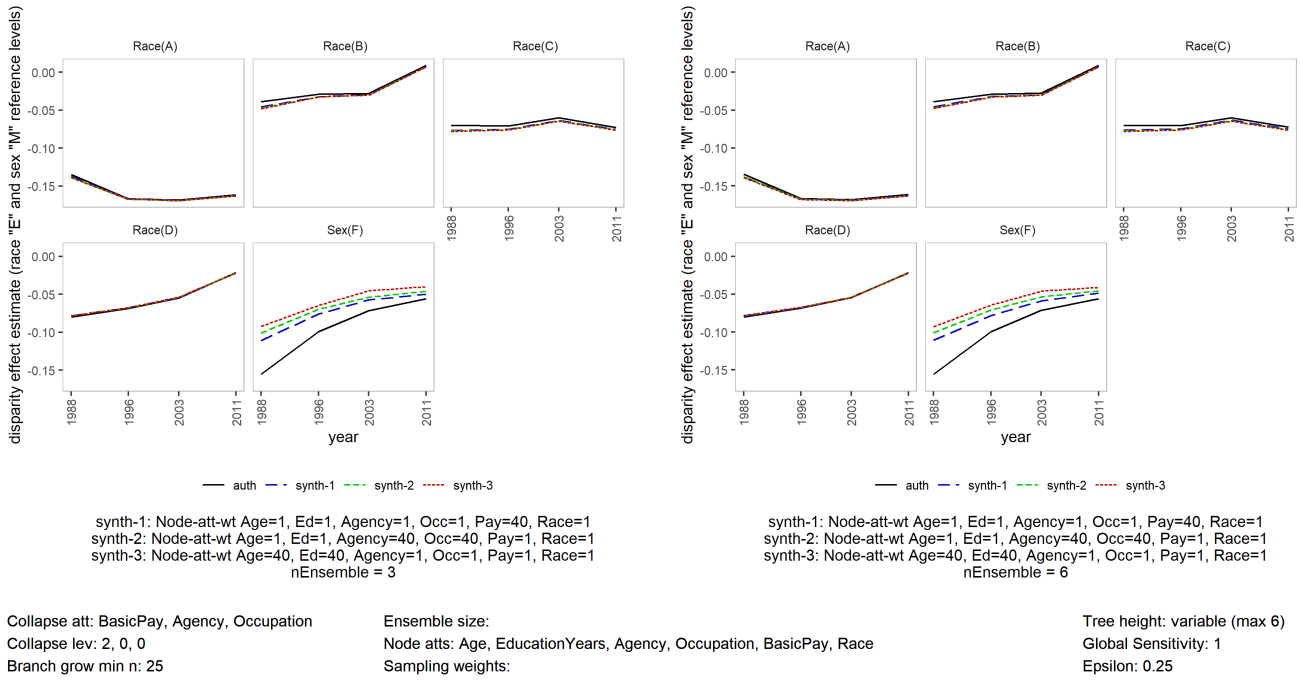


Figure 13: OPM pay disparity fixed effects model estimates. Results from authentic data (solid lines) and synthetic data (one dashed line for each attribute weighting vector). Left and right panels compare results for ensemble sizes of three and six. $n_B = 25$. $\epsilon = 0.25$.

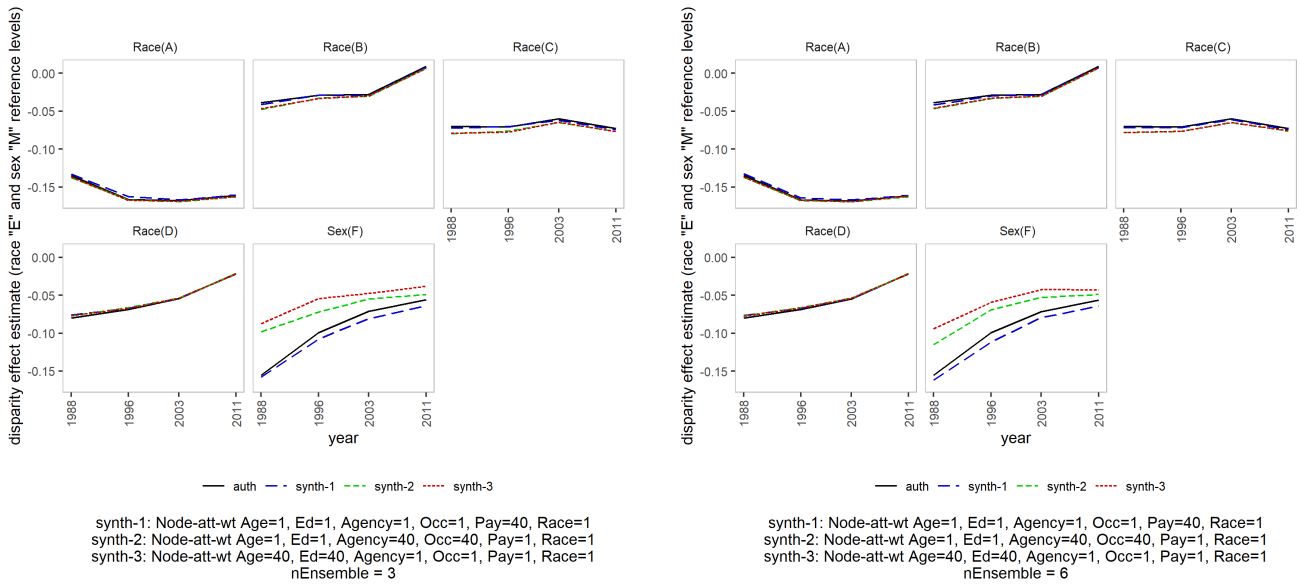
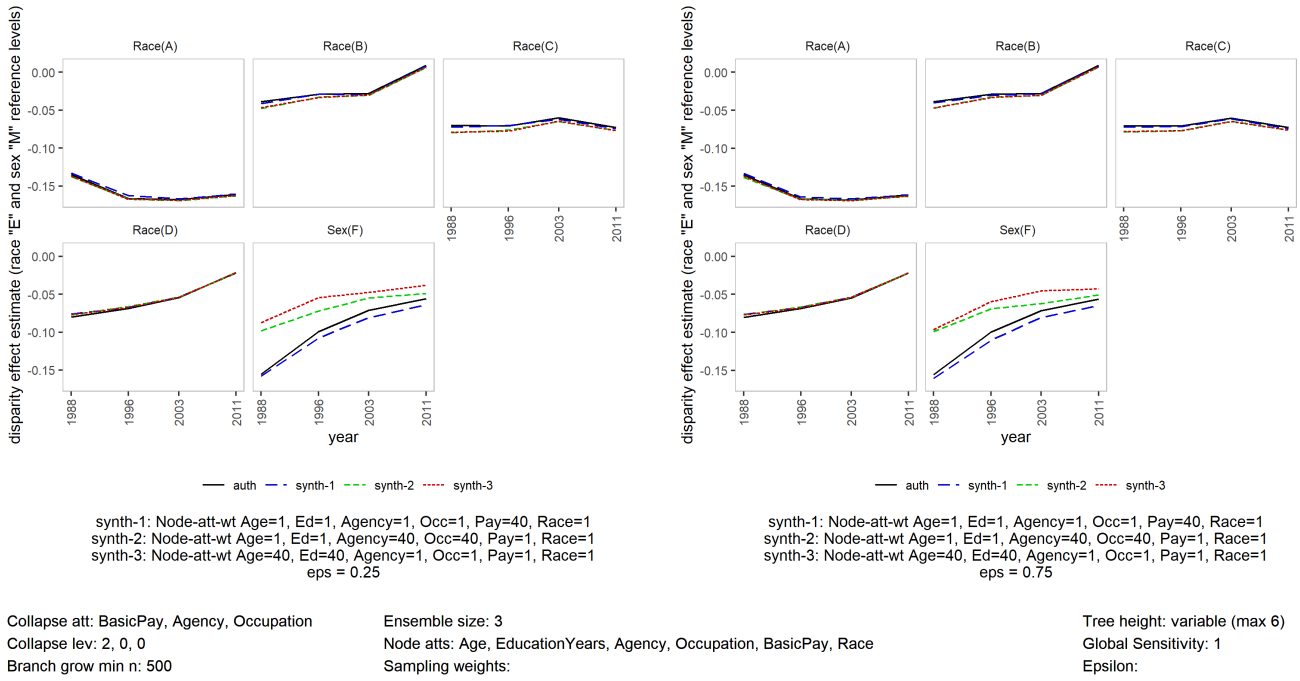


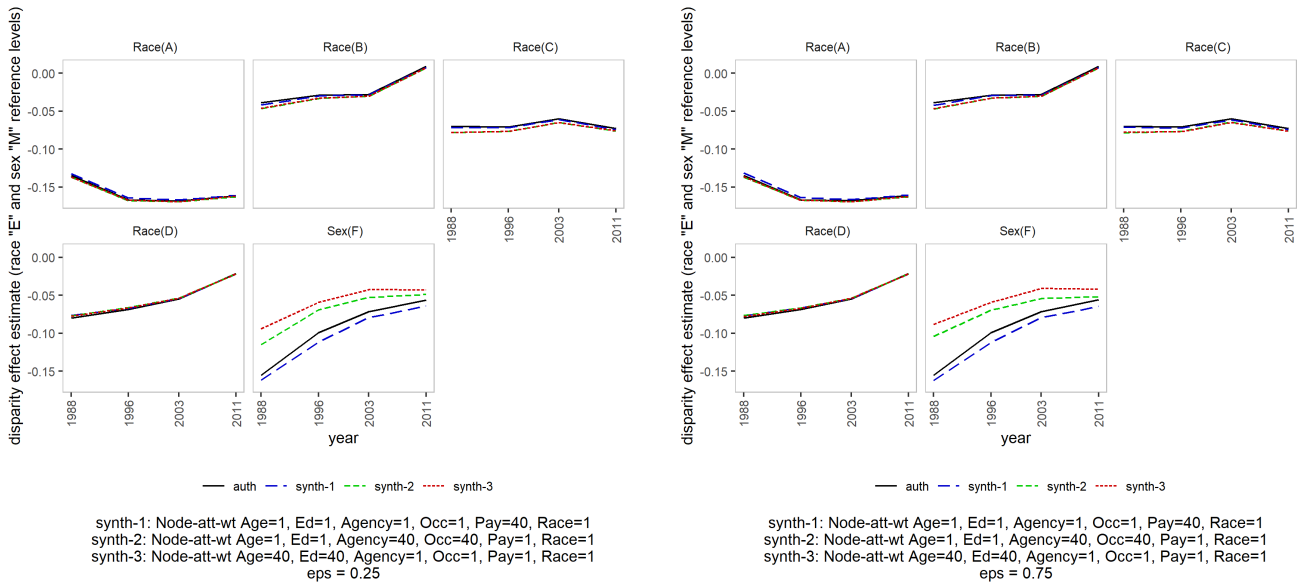
Figure 14: OPM pay disparity fixed effects model estimates. Results from authentic data (solid lines) and synthetic data (one dashed line for each attribute weighting vector). Left and right panels compare results for ensemble sizes of three and six. $n_B = 500$. $\epsilon = 0.25$.

Figures 15 and 16 compare results for $\epsilon \in \{0.25, 0.75\}$ and ensemble sizes of three and six. The effects of both parameters appear negligible, indicating an opportunity to minimize overall ϵ usage by employing small values of ϵ and ensemble size.



Collapse att: BasicPay, Agency, Occupation
Collapse lev: 2, 0, 0
Branch grow min n: 500
Ensemble size: 3
Node atts: Age, EducationYears, Agency, Occupation, BasicPay, Race
Sampling weights:
Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon:

Figure 15: OPM pay disparity fixed effects model estimates. Results from authentic data (solid lines) and synthetic data (one dashed line for each attribute weighting vector). Left and right panels compare results for $\epsilon \in \{0.25, 0.75\}$. Ensemble size=three. $n_B = 500$.



Collapse att: BasicPay, Agency, Occupation
Collapse lev: 2, 0, 0
Branch grow min n: 500
Ensemble size: 6
Node atts: Age, EducationYears, Agency, Occupation, BasicPay, Race
Sampling weights:
Tree height: variable (max 6)
Global Sensitivity: 1
Epsilon:

Figure 16: OPM pay disparity fixed effects model estimates. Results from authentic data (solid lines) and synthetic data (one dashed line for each attribute weighting vector). Left and right panels compare results for $\epsilon \in \{0.25, 0.75\}$. Ensemble size=six. $n_B = 500$.

2 Synthesis of Race and Sex

Suppose two attributes, A_1 and A_2, are to be synthesized from an initial set of RDT predictor variables R, creating new vectors of attribute labels A_1^* and A_2^*. Successful generation of A_1^* implies that covariate relationships observed between R and A_1 are also observed between R and A_1^*. In generating A_2^*, joint covariate relationships observed between R, A_1^*, and A_2 must be maintained in R, A_1^*, and A_2^*. To accomplish this, synthesis is conducted in two phases:[2]

1. Construct an ensemble of RDTs using R and A_2 as predictors and A_1 label frequencies as leaf distributions. Generate A_1^*.
2. Construct an ensemble of RDTs using R and A_1^* as predictors and A_2 label frequencies as leaf distributions. Generate A_2^*.

Subsections 2.1 through 2.3 contain panels of plots that compare model (1) parameter estimates when race and sex are sequentially synthesized in order. Each panel contains two sets of plots, one for each level of a given synthesis parameter, each set contains one subset of graphs for each β_{sex} and β_{race} parameter, and each subset includes one (dashed) line plotting synthetic annual subset estimates by year for each level of a second synthesis parameter, along with a (solid) line plotting annual subset authentic model estimates by year. High utility is indicated when points and lines corresponding to synthetic data are near, in proximity and trend, to those of corresponding authentic data. Divergence of synthetic lines within a single plot indicates variation in influence of the corresponding synthesis parameter. Analysis is limited to data synthesized with basic pay collapsed (rounded) to the nearest $2,000.

[2] An alternative is to construct RDTs using R as predictors and joint A_1, A_2 probability distributions in leaves. Since, with this method, only one ensemble is generated, total ϵ usage is half that of the sequential method described. However, the proposed, sequential method can be used, unmodified, to synthesize any number of attributes by simply introducing newly synthethized attributes as predictors, followed by further RDT regeneration.

2.1 Node Attribute Sampling Weight

Figure 17 is a typical RDT node attribute sampling weight comparison plot. As in section 1, when sex was synthesized strictly using authentic predictor data, node attribute sampling weight has apparent significant effect on utility when attributes are sequentially synthesized, since increased sampling weight for basic pay is associated with improved proximity of synthetic and authentic β_{sex} and β_{race} estimates. Further analysis is limited to results involving increased basic pay node sampling weight.

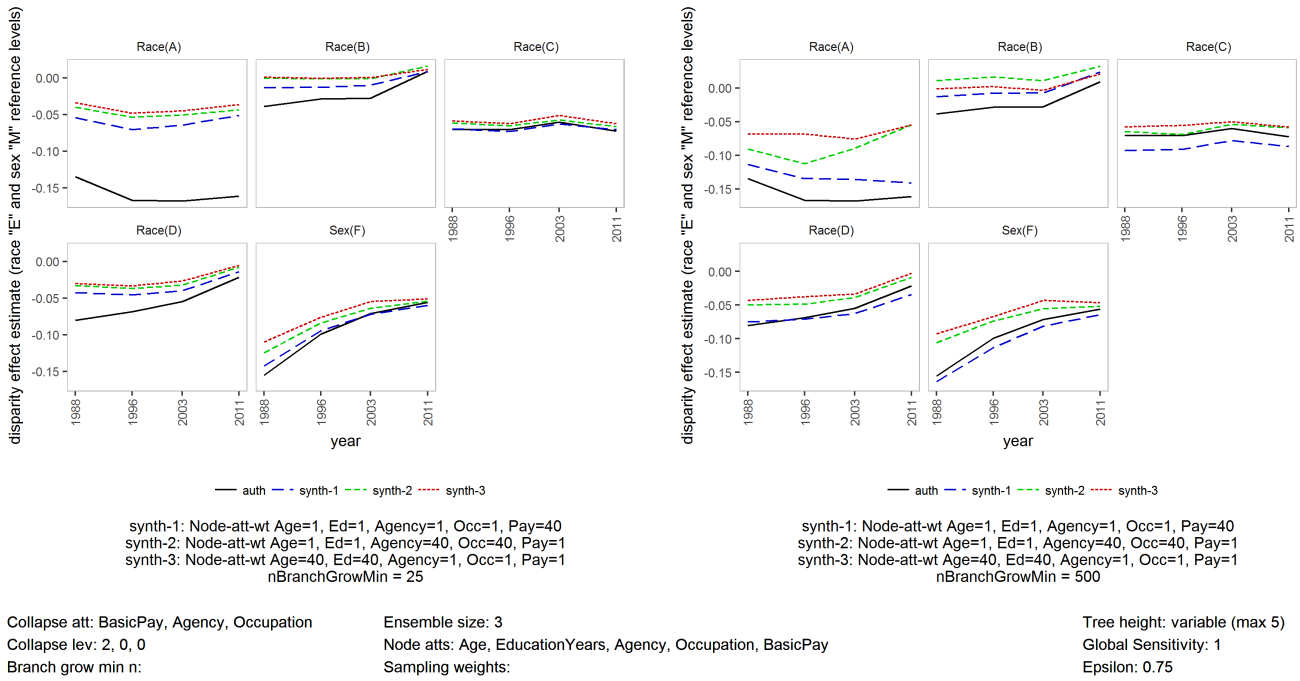
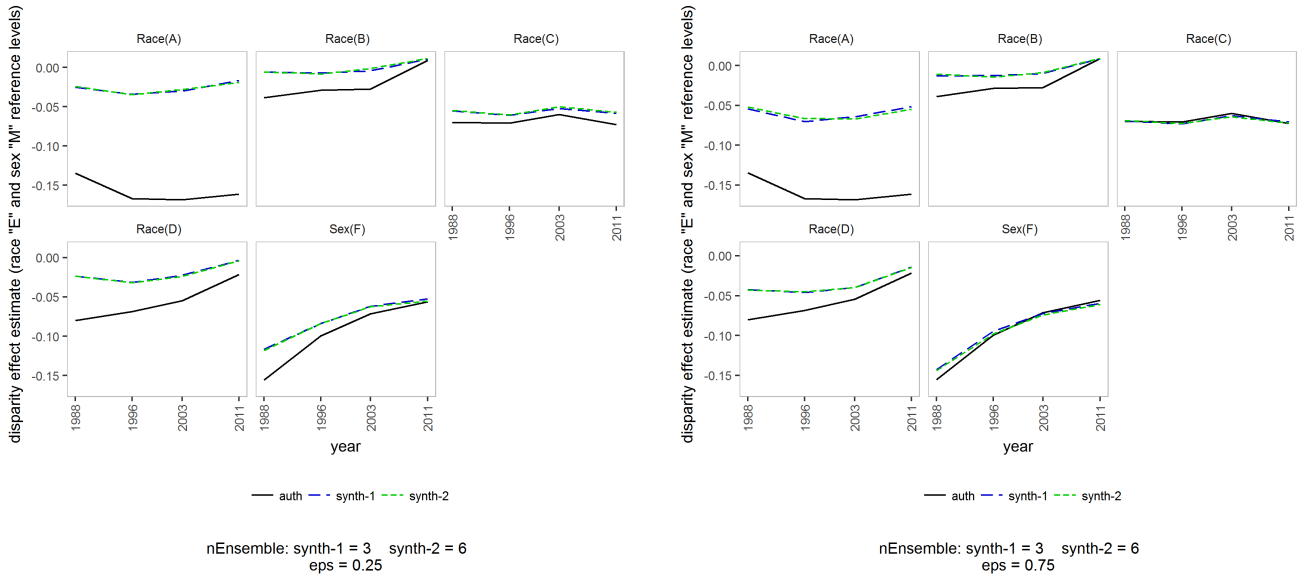


Figure 17: OPM pay disparity fixed effects model estimates. Sequential race and sex synthesis. Results from authentic data (solid lines) and synthetic data (one dashed line for each vector of node sampling weights). Left and right panels compare results for $n_B \in \{25, 500\}$.

2.2 Ensemble Size and ϵ

Figures 18 and 19 are typical joint ensemble size, ϵ comparison plots. Ensemble size has negligible effect on utility, since synthetic (dashed) lines for both levels have near identical coordinates. ϵ appears to have effect for $n_B = 25$ only. Note that figure 19 reveals a combination of synthesis parameter values such that an increase in ensemble size or ϵ does not yield improved utility. This combination may be in a region of optimal ϵ usage.

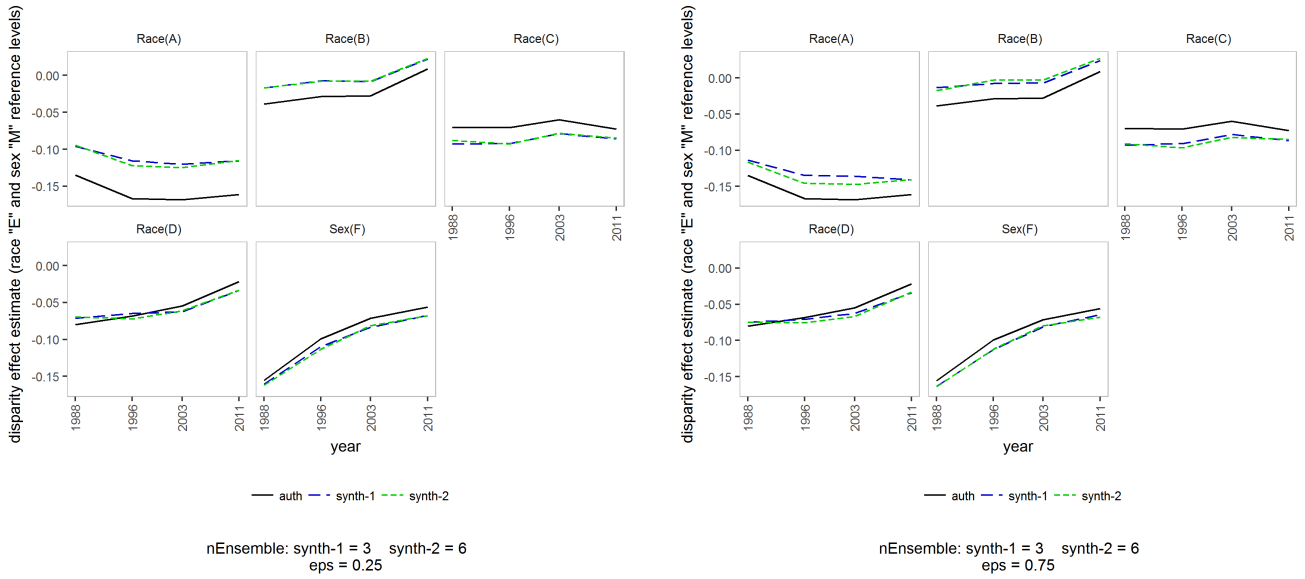


Collapse att: BasicPay, Agency, Occupation
Collapse lev: 2, 0, 0
Branch grow min n: 25

Ensemble size:
Node atts: Age, EducationYears, Agency, Occupation, BasicPay
Sampling weights: 1, 1, 1, 1, 40

Tree height: variable (max 5)
Global Sensitivity: 1
Epsilon:

Figure 18: OPM pay disparity fixed effects model estimates. Sequential race and sex synthesis. Authentic estimate lines solid, synthetic lines dashed (ensemble size in $\in \{3, 6\}$). Left, right panels compare $\epsilon \in \{0.25, 0.75\}$. $n_B = 25$.



Collapse att: BasicPay, Agency, Occupation
Collapse lev: 2, 0, 0
Branch grow min n: 500

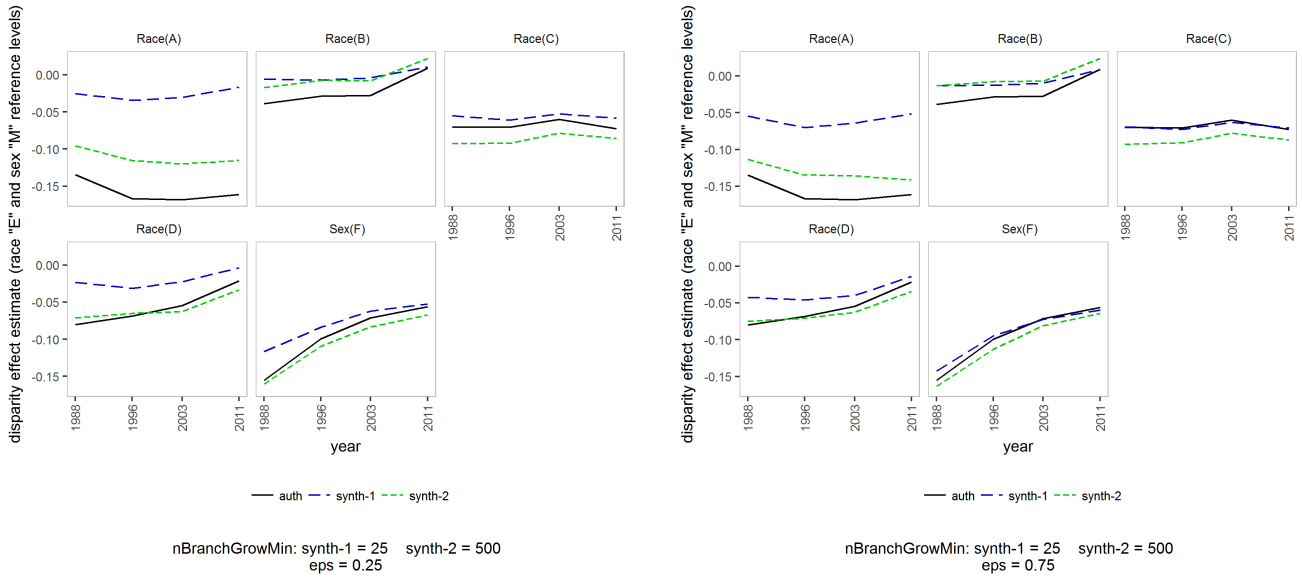
Ensemble size:
Node atts: Age, EducationYears, Agency, Occupation, BasicPay
Sampling weights: 1, 1, 1, 1, 40

Tree height: variable (max 5)
Global Sensitivity: 1
Epsilon:

Figure 19: OPM pay disparity fixed effects model estimates. Sequential race and sex synthesis. Authentic estimate lines solid, synthetic lines dashed (ensemble size in $\in \{3, 6\}$). Left, right panels compare $\epsilon \in \{0.25, 0.75\}$. $n_B = 500$.

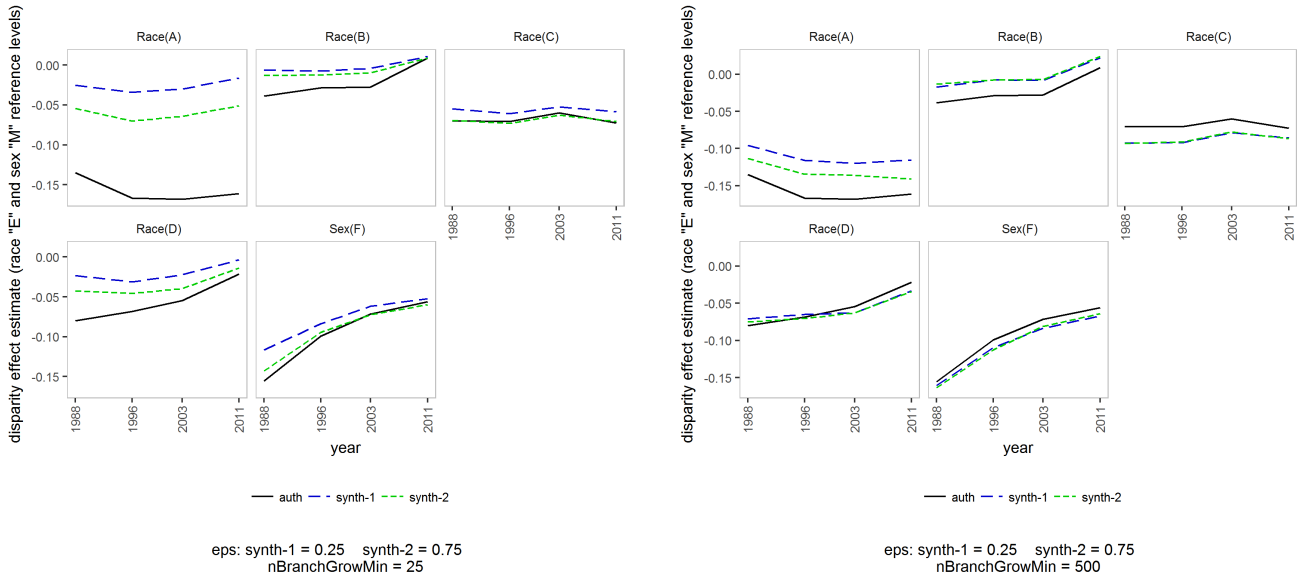
2.3 n_B and ϵ

Figures 20 and 21 jointly compare $n_B \in \{25, 500\}$ and $\epsilon \in \{0.25, 0.75\}$. Although both figures plot the same results, figure 21 indicates greater influence of n_B than ϵ on utility, with synthetic estimates for all races, except A, converging to identical, near-authentic, coordinates in the $n_B = 500$ plot. Authentic annual trends are represented in synthetic estimates, but bias also appears, with a systematic separation of authentic and synthetic estimates.



Collapse att: BasicPay, Agency, Occupation
Collapse lev: 2, 0, 0
Branch grow min n:
Ensemble size: 3
Node atts: Age, EducationYears, Agency, Occupation, BasicPay
Sampling weights: 1, 1, 1, 1, 40
Tree height: variable (max 5)
Global Sensitivity: 1
Epsilon:

Figure 20: OPM pay disparity fixed effects model estimates. Sequential race and sex synthesis. Authentic estimate lines solid, synthetic lines dashed ($n_B \in \{3, 6\}$). Left, right panels compare $\epsilon \in \{0.25, 0.75\}$. Ensemble of three trees.



Collapse att: BasicPay, Agency, Occupation
Collapse lev: 2, 0, 0
Branch grow min n:
Ensemble size: 3
Node atts: Age, EducationYears, Agency, Occupation, BasicPay
Sampling weights: 1, 1, 1, 1, 40
Tree height: variable (max 5)
Global Sensitivity: 1
Epsilon:

Figure 21: OPM pay disparity fixed effects model estimates. Sequential race and sex synthesis. Authentic estimate lines solid, synthetic lines dashed ($\epsilon \in \{0.25, 0.75\}$). Left, right panels compare $n_B \in \{3, 6\}$. Ensemble of three trees.

3 Conclusion

- Factorial-experiment-like combinations of synthesis parameter values were used to generate ensembles of random decision trees (RDT methods as proposed by Jagannathan et al., modified by generating variable depth branches, controlled by observation counts taken from DIBBS synthetic data)
- RDT Ensembles were used to generate synthetic sex/race label vectors
- Federal employee pay disparity model (1) was fit to corresponding sets of authentic and synthetic data
- Resulting β_{sex} and β_{race} parameter estimates were graphically compared for agreement in proximity and trend
- Synthesis parameter values were studied in an attempt to identify combinations leading to maximum utility and minimum ϵ usage
- Many interactions in the synthesis parameters were observed, so that for instance, increasing ensemble size or epsilon improves utility in certain combinations, but not others
- An opportunity for low ϵ usage may exist since, within the synthesis parameter values studied, there are high utility combinations with $\epsilon = 0.25$ and ensemble size = 3 (for total ϵ consumption of 0.75)
- A question arises as to total ϵ consumed by exploring factorial-like combinations of synthesis parameter values (optimal values are chosen from n generated synthetic data sets, so is total usage $n \times \epsilon$?)
- Two attribute synthesis strategies were studied:

  1. Synthesize sex from authentic age, ed, agency, occ, pay, and race
  2. Synthesize race from authentic age, ed, agency, occ, and pay then synthesize sex by including synthetic race with the previous authentic attributes

  There is some deterioration in sex coefficient estimate utility for item 2, but not a significant amount
- The sequential synthesis method (previous bullet, item 2) generates two ensembles, so that total ϵ usage is twice that of generating a single ensemble
- Leaf distributions could be composed with multivariate distributions, requiring a single ensemble, but this would depart from the generalized solution presented (with the current method, introducing a new variable simply requires constructing a new ensemble where you decide which of the predictors are authentic and which are synthetic, then you run)
- Perhaps the most significant finding of the study is that, given the combinations employed, the most influential synthesis parameter is the order of predictor attribute selection during tree construction (recall that this is random)
- Because pay is the response of the disparity model studied, it must be selected, otherwise utility suffers (disappears) and, to "force selection of pay, its sampling weight is increased to a significant multiple of remaining parameters (as a result, the resulting trees may be considered semi-random)
- Another significant parameter is n_B, the number of branch observations required to split a node (recall that this is what gives variable depth branches)

  - There is an interesting interplay between predictor covariate relationships and leaf distribution confidence intervals in that short branches (large n_B) confound covariate relationships, but improve leaf CIs, while long branches (small n_B) do the opposite
  - There is a sweet n_B spot that is certainly dependent on the data and characterizing it would be an interesting phenomenon to model or simulate
- The method presented is sufficiently generalized to be publishable as a package [there are several tree packages available and some that generate synthetic data (synthpop, ROSE random over-sampling examples, OpenSDPsynthR), but none that appear to implement differential privacy]
- All functions developed and used in this study for tree generation, attribute synthesis, and analysis of results are parameterized and should adapt to a variety of data sets