# Generation of Differentially Private Synthetic Data
# Utility of Random Decision Trees

Duke University Sythenthetic Data Project

September 4, 2018

## 1 Objective

- Generate a differentially private synthetic data (SD) from a given data set ($D_0$) by reclassifying given columns of data (attributes) using ensembles of random decision trees (RDT) fit to $D_0$ with application of the Laplace mechanism to leaf distributions. For this exercise, $D_0$ is generated using attributes named after those found in the OPM data, but with labels and covariate relationships absent in and unrelated to that data.
- Study the effect of
  - Global sensitivity and $\epsilon$ (on the Laplace mechanism)
  - Stage of applying Laplace mechanism (applied to frequencies prior to computing proportions compared to applying to computed proportions, etc.)
  - Order of attributes synthesized on distribution of synthesized labels
  - Tree height on distribution of synthesized attributes
- Assess agreement of covariate relationships and distribution of observation mass between $D_0$ and resulting SD
- Analyze efficiency of RDT, Laplace, and classification algorithms, develop improvement strategies

## 2 General Strategy

- Develop $D_0$ covariate relationships
- Develop RDT construction algorithm
- Develop algorithm to apply Laplace mechanism to RDT leaf distributions
- Develop leaf sampling attribute classification algorithm
- Develop sequential attribute classification algorithm, where a given attribute is classified using a hybrid RDT using previously synthesized attributes from SD along with remaining unsynthesized attributes from $D_0$
- Generate $D_0$
- Measure global sensitivity using a select group of attributes
- Sequentially execute RDT, Laplace (with parameter adjusted by global sensitivity), and sampling algorithms to generate complete set of synthetic columns
- Retain execution times of various computations along with relevant parameters for efficiency evaluation
- Plot various three-way joint distributions comparing those in $D_0$ to those in SD

## 3 Sample Data

$D_0$ consists of 1,000,000 observations generated using the attributes, labels, and covariate relationships presented in table 1.

Table 1: Sample Data Attributes

| Attribute | Labels | Covariate Relationships |
|---|---|---|
| FY | 1988-2011 | Independent, uniformly distributed |
| Sex | F, M | Proportion female uniformly increased from 0.45 in 1988 to 0.52 in 2011 |
| Race | A, B, C, D, E | Dependent on sex<br>Female weights (A, B, C, D, E): 0.03, 0.07, 0.35, 0.15, 0.40<br>Male weights (A, B, C, D, E): 0.05 0.10 0.30 0.10 0.45 |
| Age | 22, 27, 32, 37, 42, 47, 52, 57, 62 | Dependent on sex and race, weights in order of age are (note the rather extreme correlation, intended to create discernible distributions)<br>F, A: 0.11, 0.22, 0.33, 0.22, 0.11, 0.00, 0.00, 0.00, 0.00, 0.00<br>F, B: 0.00, 0.11, 0.22, 0.33, 0.22, 0.11, 0.00, 0.00, 0.00, 0.00<br>F, C: 0.00, 0.00, 0.11, 0.22, 0.33, 0.22, 0.11, 0.00, 0.00, 0.00<br>F, D: 0.00, 0.00, 0.00, 0.11, 0.22, 0.33, 0.22, 0.11, 0.00, 0.00<br>F, E: 0.00, 0.00, 0.00, 0.00, 0.11, 0.22, 0.33, 0.22, 0.11, 0.00<br>M, A: 0.06, 0.22, 0.56, 0.11, 0.06, 0.00, 0.00, 0.00, 0.00, 0.00<br>M, B: 0.00, 0.06, 0.22, 0.56, 0.11, 0.06, 0.00, 0.00, 0.00, 0.00<br>M, C: 0.00, 0.00, 0.06, 0.22, 0.56, 0.11, 0.06, 0.00, 0.00, 0.00<br>M, D: 0.00, 0.00, 0.00, 0.06, 0.22, 0.56, 0.11, 0.06, 0.00, 0.00<br>M, E: 0.00, 0.00, 0.00, 0.00, 0.06, 0.22, 0.56, 0.11, 0.06, 0.00 |
| Agency | AAAA, AABB, CCCC, DDDD, EEEE, FFFF, GGGG, HHHH | Agency and occupation are assigned jointly, with a dependency on sex |
| Occupation | 0001, 0002, 0003, 0004, 0005, 0006, 0007, 0008 | 10% random sample of observations, agencies AAAA and BBBB, occupations 0005, 0006, 0007, 0008 assigned with following weights by sex:<br>F: 0.2, 0.2, 0.4, 0.2<br>M: 0.2, 0.2, 0.2, 0.4<br>6.67% random sample of observations, agencies CCCC and DDDD, occupations 0005, 0006, 0007 assigned with following weights by sex:<br>F: 0.6, 0.2, 0.2<br>M: 0.2, 0.6, 0.2<br>Remaining observations uniformly distributed using combinations of agencies EEEE, FFFF, GGGG, HHHH and occupations 0001, 0002, 0003, 0004 |
| Education | 12, 14, 16, 18, 20 | Dependent on occupation<br>Occ 0008, ed 18, 20 assigned with weights 0.6, 0.4<br>Occ 0007, ed 16, 18, 20 assigned with weights 0.25, 0.65, 0.10<br>Occs 0005, 0006, ed 14, 16, 18, 20 assigned with weights 0.25, 0.5, 0.2, 0.05<br>Occs 0003, 0004, ed 14, 16, 18 assigned with weights 0.2, 0.75, 0.05<br>Remaining occs, ed 14, 16, 18 assigned with weights 0.6, 0.3, 0.1 |
| y (pay) | | Function of above attributes, generally increases by FY, age, and education, decreases for females, adjusted by race |

## 4 Random Decision Tree Properties and Construction

- Construction method as described in Jagannathan et al.
- Node attributes used: *FY, sex race, age, agency, occupation,* and *education*
- Tree heights evaluated: 3, 4, 5, 6, and 7
- A tree is assigned a classification attribute for which label distributions are computed from frequencies observed in D0 for associated attribute, label combinations in the node path leading to a given leaf
- Level one of the tree consists of a single node for each label of a randomly selected 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
- **An important RDT feature is that only the structure of $D_0$ (attributes and associated labels) is needed for construction - no observations or privacy revealing distributions or covariate relationships are used - $\epsilon$ usage for construction is 0**
- An ensemble of RDTs consists of multiple RDTs, each with a randomly selected sequence of attributes at each level
- Each observation in $D_0$ (or the hybrid SD/$D_0$ data set used to construct an RDT) maps to a single node path and leaf in an RDT constructed from it
- **Using leaf distributions to generate synthetic records can produce combinations of attributes and labels that do not exist in $D_0$ (or SD/$D_0$ hybrid)** - for example, given an RDT that classifies *age* (leaves contain *age* distributions) and where *race* does not appear in a given node path, then *race* and *age* are confounded (along the path), so that classification of an observation using the given node path may produce a synthetic observation with a *race*, *age* combination of A, 35 when at least one $D_0$ observation exists that maps to the given path and has *age*=35, even if none contain *race* of A
- The RDTs used here are constructed recursively using the following algorithm

```
function constructNode(data, classAttribute, treeHeight, nodeLevel, nodeAttribute, attCandidates, obsIndices)
  if(nodeLevel < treeHt)
    // Create a node for each label of a randomly selected attribute that has not been used in this node path
    // Note that observation indices are filtered for the current attribute and each label so that only
    // observations related to the node path under construction remain when the corresponding leaf is
    // composed
    for each currentLabel in nodeAttribute {
      remove nodeAttribute from attCandidates
      currAttribute = randomly selected attribute form attCandidates
      identify observation indices k such that currentAttribute of data(obsIndices) = currentLabel
      return[constructNode(data, classAttribute, treeHeight, nodeLevel+1, currAttribute, attCandidates, k)]
    }
  else
    // Create a leaf node (probability distribution) using observation indices for current node path
    compute table of observation frequencies by classAttribute label for data(obsIndices) subset
    convert table to proportions
    apply Laplace mechanism to proportions
    return(labels with corresponding proportions)
```

## 5 RDT Analysis

- **Leaf distributions may contain mass for attribute and label combinations that do not exist in original data.** Figure 1 Demonstrates this with an ensemble of RDTs, each of height five and constructed from seven attributes, such that two attributes are omitted in every node path. When either *sex* or *race* are omitted, *age* synthesis is accomplished without regard to the actual *sex* and *race* labels that appear in an observation, causing a confounding of these attributes with possible generation of label combinations that do not appear in the source data. This condition is identified in the plot by original data mass (green bars) appearing without corresponding synthesized mass (blue bars).
- Full-height tree (height = number of attributes in $D_0$) leaf distributions contain mass solely for attribute, label combinations that exist in the source data, since all observed attribute, label combinations have an exact corresponding node path in each tree. Figure 2 demonstrates this with an ensemble of

RDTs, each of height seven, constructed from seven attributes. Each original data bar (green) has a corresponding synthesized data bar (blue).

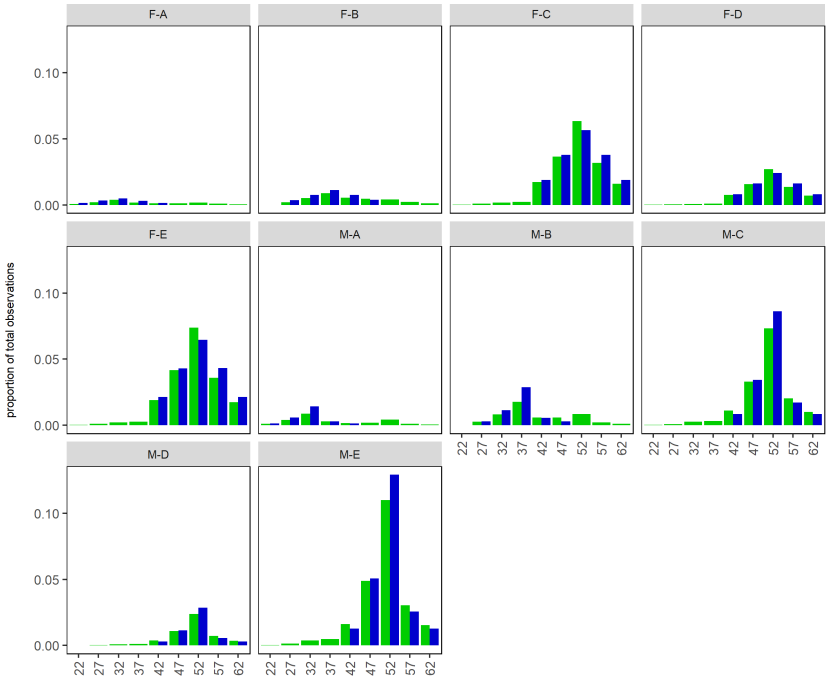


Figure 1: Distributions of classified *age* compared to that of original *age*. RDT height = 5, number of attributes = 7. Original data in green, synthesized data in blue. Laplace mechanism not applied.

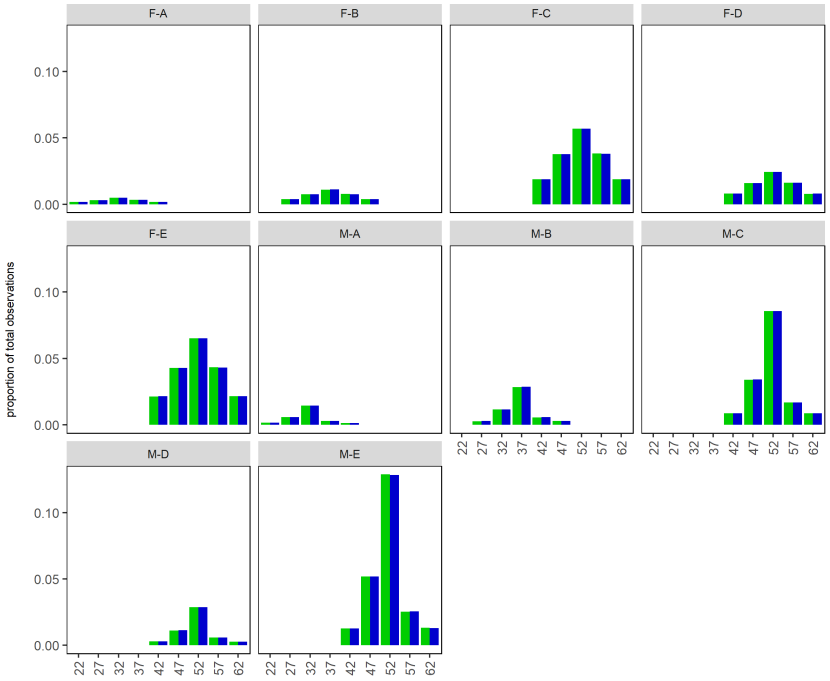


Figure 2: Distributions of classified *age* compared to that of original *age*. RDT height = number of attributes = 7. Original data in green, synthesized data in blue. Laplace mechanism not applied.

## 6 Global Sensitivity and Laplace Mechanism

- The global sensitivity parameter, $\Delta(f)$, is defined as $\max_{(D_1,D_2)\in D_0} \parallel f(D_1) - f(D_2) \parallel_1$, where $D_1$ and $D_2$ are subsets of the original data, $D_0$, differing by a single record (one of $D_1$ or $D_2$ have one record that does not appear the other, while all remaining records are identical). Note that each $f(D)$ has a unique $\Delta(f)$.
- When RDT leaves contain probability mass distributions (as ours do), $\Delta(f)$ can be computed by identifying, for each observation in $D_0$, its classifying distribution, computing the difference in the initial proportion for the classification label and the proportion with quantity of 1.0 subtracted, then reporting the maximum difference of all observations. Figure 3 shows the distribution of leaf proportion differences for all observations in an ensemble of five *age* classification RDTs of height five, composed from seven attributes. The Laplace mechanism is not applied. Reasonable values of $\Delta(f)$ would be taken from the upper region of the x-axis; values of 0.025 and 0.05 are used in this study.
- **The Laplace parameter = $\epsilon/\Delta(f)$, which varies by classification attribute - maintains $\epsilon$-DP**
- **$\Delta(f)$ imposes an upper bound on the Laplace parameter while constructing an RDT for a given attribute (a smaller value could be used, but this would unnecessarily increase the interval of selected random values). Computation of individual $\Delta(f)$ for each attribute during RDT construction (based on observations synthesized to that point) could improve utility while maintaining DP, although the $\epsilon$ in $\epsilon$-DP becomes variable or composite.**
- **Queries of $D_0$ during RDT construction are executed in parallel, retrieving disjoint subsets of observations, one for each node path - total accumulated $\epsilon$ should be computed**
- **Effect of tree height and Laplace parameter on MSE of proportion observation between classified attribute and label combinations in $D_0$ and SD.** Figures 4 and 5 show the relationship of tree height, global sensitivity, and $\epsilon$ (along with attribute synthesis order) to mean squared error of joint distributions (represented by node paths) in $D_0$ and SD. For figure 4, the Laplace mechanism was applied to proportions after computation and for figure 5 the Laplace mechanism was applied to leaf frequencies, then rounded to the nearest non-negative integer, prior to proportion computation. **Important features are:**
  - **MSE is reduced as tree height increases**
  - **With the Laplace mechanism applied to proportions, MSE decreases with increase in $\epsilon$, with Laplace mechanism applied to frequencies prior to proportion computation, $\epsilon$ patterns are ambiguous**
  - **MSE is considerably larger for combinations of attributes appearing later in the synthesis sequence**
- **Compare application of Laplace mechanism to frequencies vs. proportions computed from frequencies**
- **Random generation of proportions uses a Laplace distribution with lower bound = $\mathbf{F}^{-1}$(observed $p$) to avoid negative $p$ values**
- The Laplace mechanism is applied to all leaves in all RDTs of an ensemble (line 17 of the RDT construction algorithm)

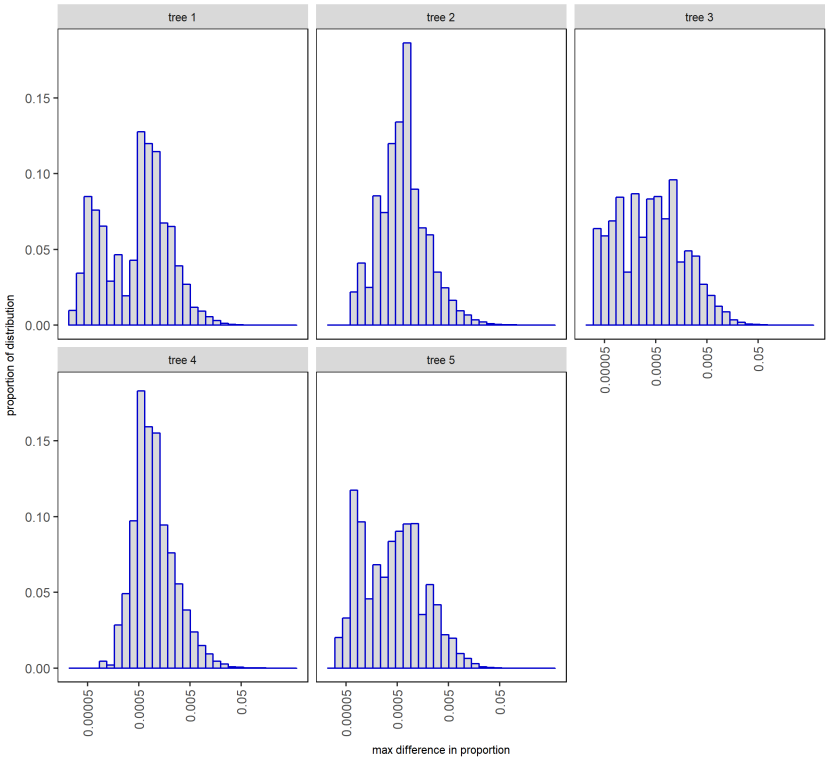


Figure 3: Distribution of $\Delta(f)$ for classification attribute *age*. From an ensemble of five RDTs, each of height five (from seven total attributes). Laplace mechanism not applied.

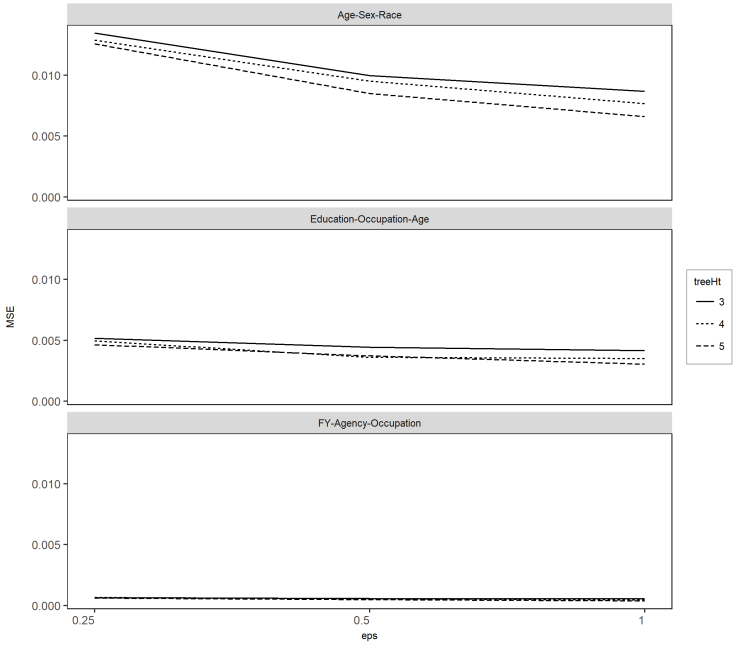


Figure 4: Mean squared error in attribute, label combination proportions between $D_0$ and SD. Laplace mechanism applied to proportions after computation. Synthesis sequence of *FY, agency, occupation, education, age, sex, race*. MSE decreases with tree height and $\epsilon$ and increases with placement in synthesis sequence. Global sensitivity = 0.025. Graphs represent relationship of $\epsilon$ and tree height.

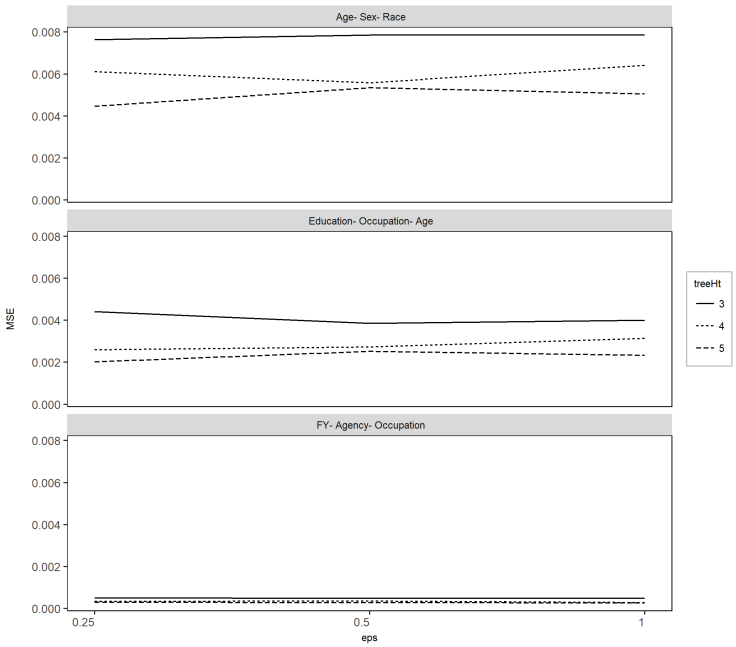


Figure 5: Mean squared error in attribute, label combination proportions between $D_0$ and SD. Laplace mechanism applied to leaf distribution frequencies prior to proportion computation. Synthesis sequence of *FY, agency, occupation, education, age, sex, race*. MSE decreases with tree height and increases with placement in synthesis sequence. Global sensitivity = 1. Graphs represent relationship of $\epsilon$ and tree height.

## 7 Attribute Synthesis

- SD attributes replaced with column generated by RDT ensemble constructed from all prior synthesized attributes and unsynthesized attributes from $D_0$
- **Sequence of attributes synthesized is important since number of synthesized attributes in RDTs affects degree of agreement with $D_0$**
- Synthesizing an attribute for a given observation consists of locating the leaf in each RDT of an ensemble, accumulating probability mass for each possible label, and sampling one label using the accumulated weights
- Leaf retrieval for a given observation, unlike the efficient binary search of a B-tree, requires at each node level a search of all possible labels for the node attribute, which is inefficient and multiplied by the number of trees in an ensemble. A C algorithm was developed, as an alternative to standard R functions, for leaf identification, which reduces synthesis execution time for a single attribute from approximately one hour to under one minute (1,000,000 observations using an ensemble of five RDTs of height 5).

## 8 Comparison of Select D0 and SD Joint Distributions

Figures 6 through 15 compare the joint distribution of $D0$ and SD observations for various triplets of synthesized attributes using sample data as described in section 3. All SD sets generated with an ensemble of five RDTs of height five. Laplace mechanism applied to frequencies prior to proportion computation or proportions after computation as indicated. Attribute synthesis sequence *FY, agency, occupation, education, age, sex, race*. Global sensitivity and $\epsilon$ appear in captions. Note the apparent departure of agreement of $D_0$ and SD distribution as attributes appear later in the synthesis sequence.

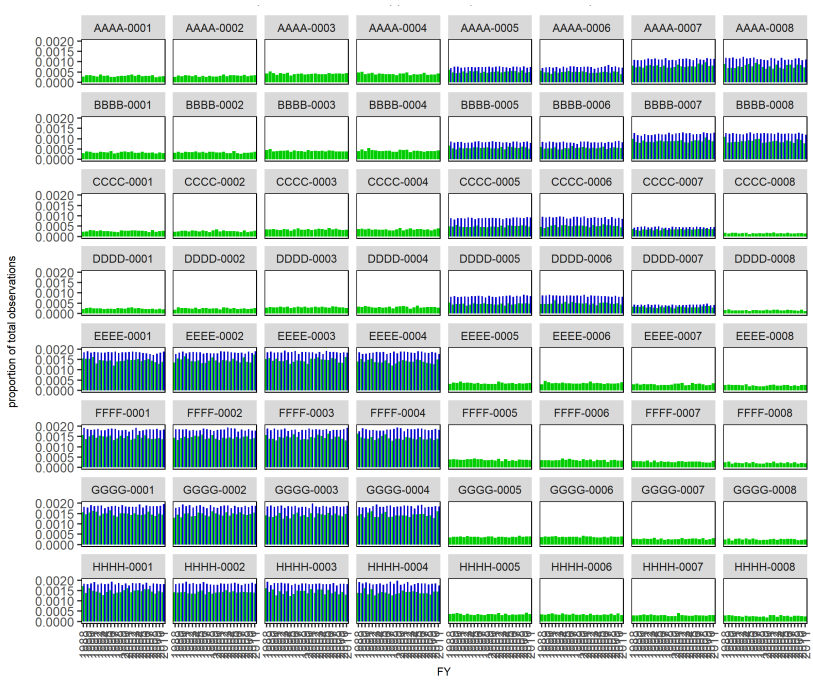


Figure 6: Distribution of *FY* within *agency* and *occupation*. Original data in blue, synthetic in green. Laplace mechanism applied to proportions after computation. Global sensitivity = 0.025, $\epsilon = 1$. Less than full height tree causes synthesis of attribute, label combinations that do not appear in the original data.

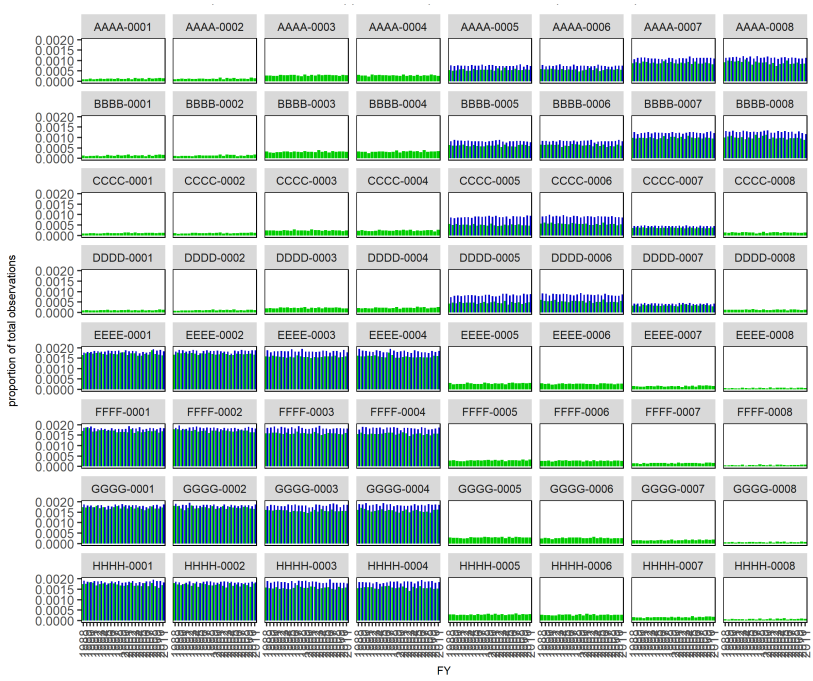


Figure 7: Distribution of *FY* within *agency* and *occupation*. Original data in blue, synthetic in green. Laplace mechanism applied to frequencies prior to proportion computation. Global sensitivity = 1, $\epsilon = 1$. Less than full height tree causes synthesis of attribute, label combinations that do not appear in the original data.

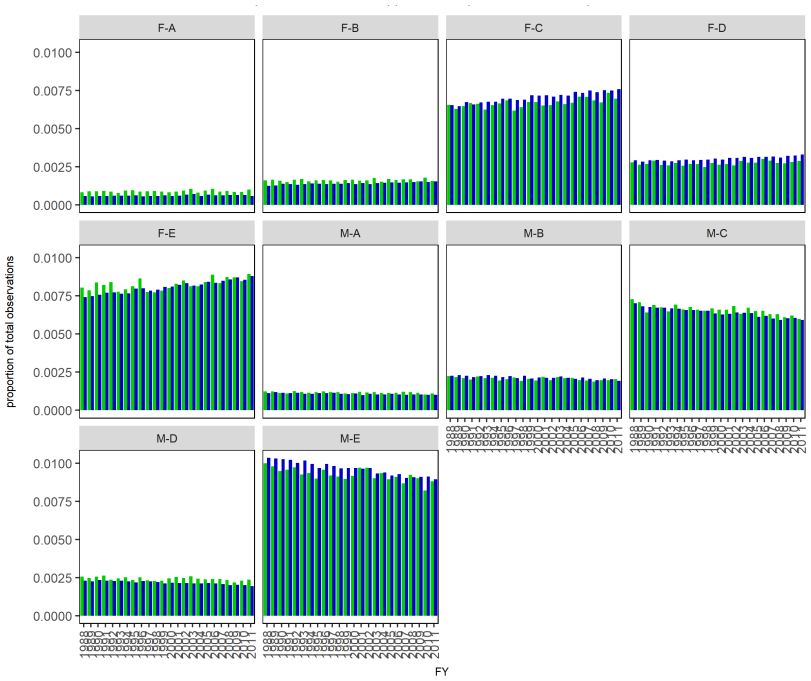


Figure 8: Distribution of *FY* within *sex* and *race*. Original data in blue, synthetic in green. Laplace mechanism applied to proportions after computation. Global sensitivity = 0.025, $\epsilon = 1$. Less than full height tree causes synthesis of attribute, label combinations that do not appear in the original data.

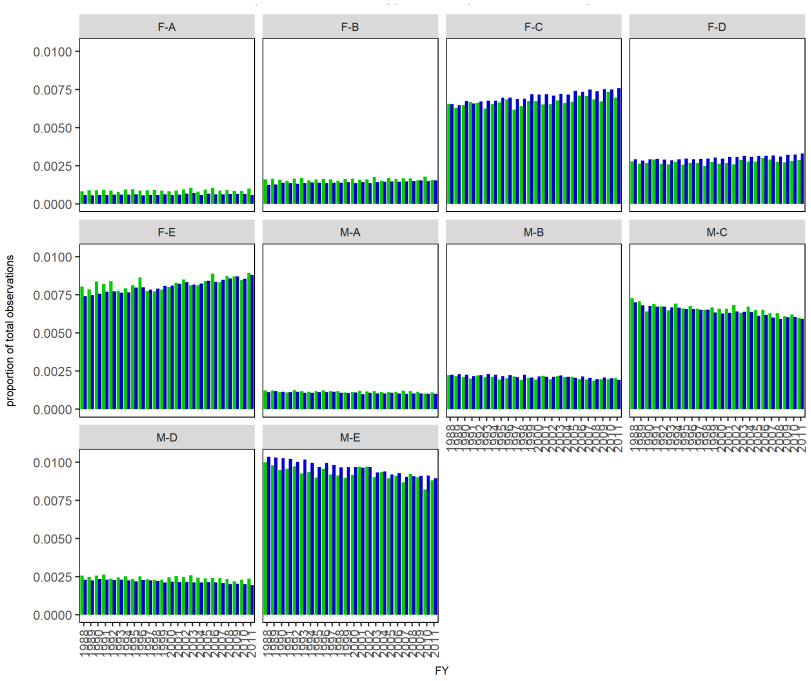


Figure 9: Distribution of *FY* within *sex* and *race*. Original data in blue, synthetic in green. Laplace mechanism applied to frequencies prior to proportion computation. Global sensitivity = 0.025, $\epsilon = 1$. Less than full height tree causes synthesis of attribute, label combinations that do not appear in the original data.

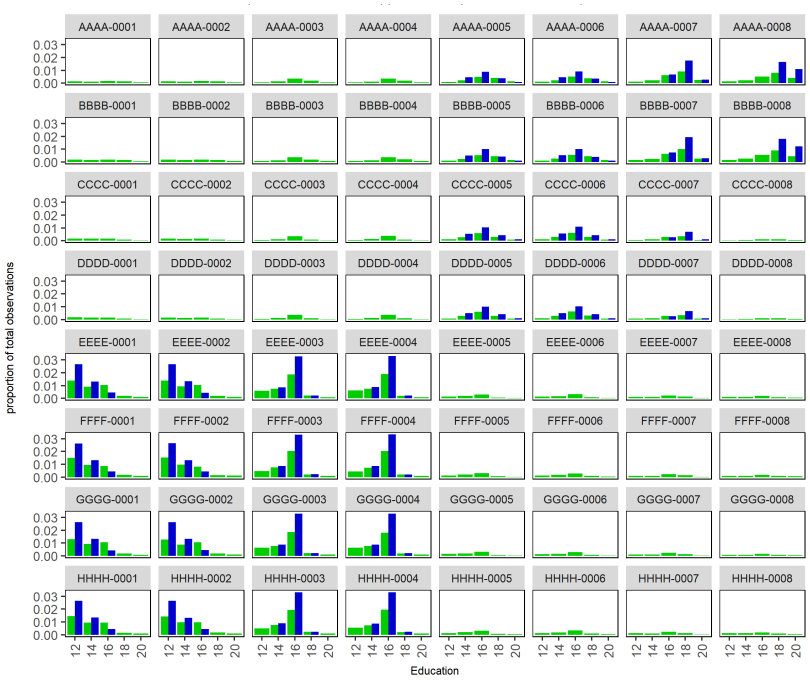


Figure 10: Distribution of *education* within *agency* and *occupation*. Original data in blue, synthetic in green. Laplace mechanism applied to proportions after computation. Global sensitivity = 0.025, $\epsilon = 1$. Less than full height tree causes synthesis of attribute, label combinations that do not appear in the original data.

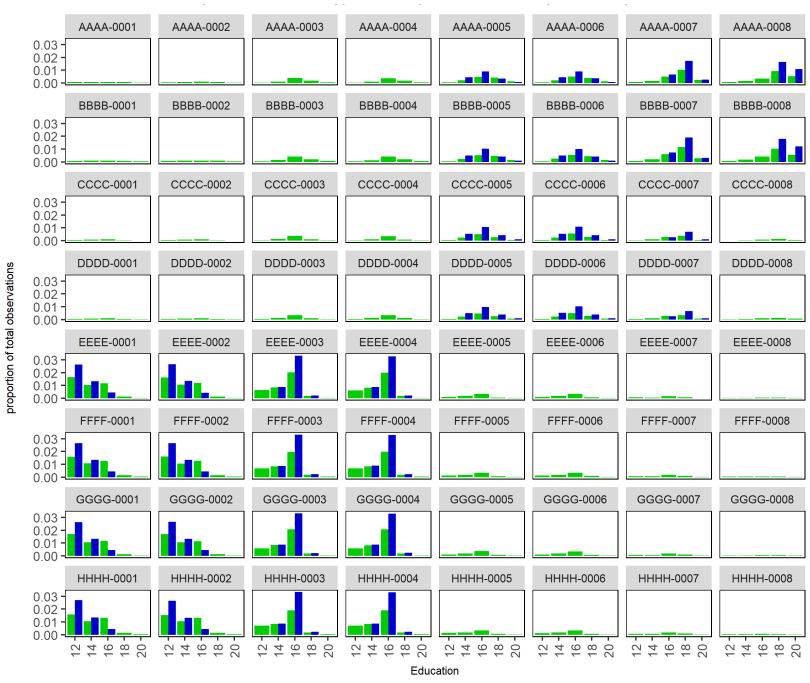


Figure 11: Distribution of *education* within *agency* and *occupation*. Original data in blue, synthetic in green. Laplace mechanism applied to frequencies prior to proportion computation. Global sensitivity = 0.025, $\epsilon = 1$. Less than full height tree causes synthesis of attribute, label combinations that do not appear in the original data.

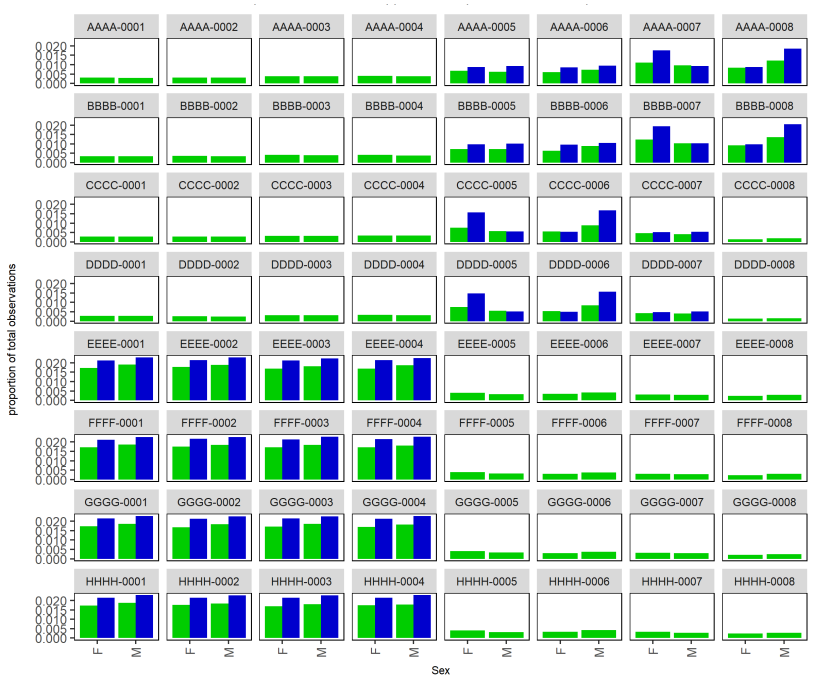


Figure 12: Distribution of *sex* within *agency* and *occupation*. Original data in blue, synthetic in green. Laplace mechanism applied to proportions after computation. Global sensitivity = 0.025, $\epsilon = 1$. Less than full height tree causes synthesis of attribute, label combinations that do not appear in the original data.

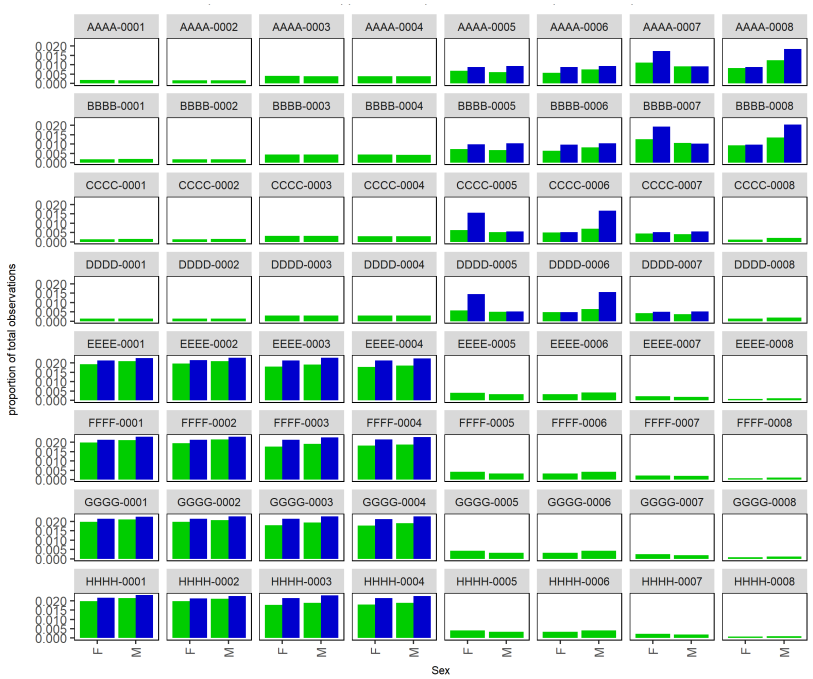


Figure 13: Distribution of *sex* within *agency* and *occupation*. Original data in blue, synthetic in green. Laplace mechanism applied to frequencies prior to proportion computation. Global sensitivity = 0.025, $\epsilon = 1$. Less than full height tree causes synthesis of attribute, label combinations that do not appear in the original data.

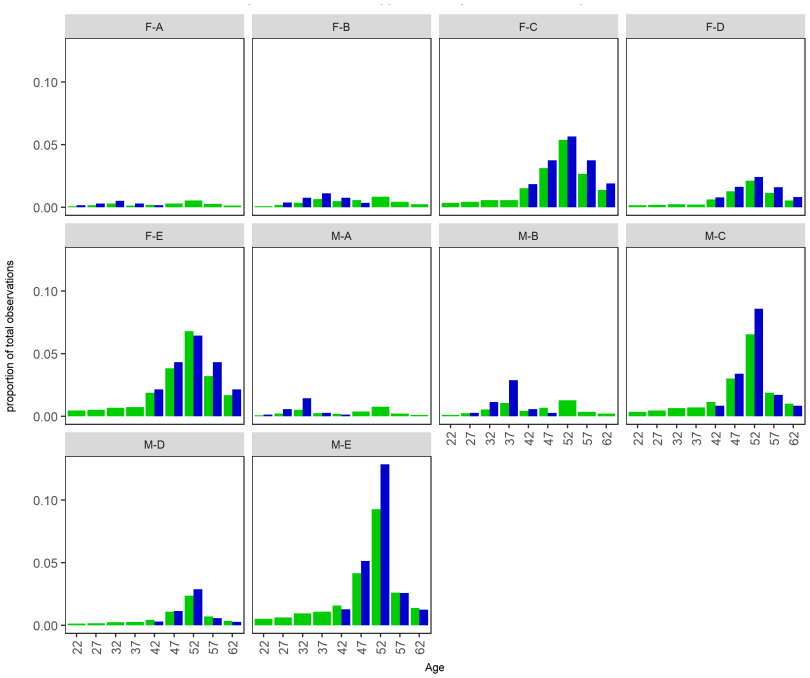


Figure 14: Distribution of *age* within *sex* and *race*. Original data in blue, synthetic in green. Laplace mechanism applied to proportions after computation. Global sensitivity = 0.025, $\epsilon = 1$. Less than full height tree causes synthesis of attribute, label combinations that do not appear in the original data.

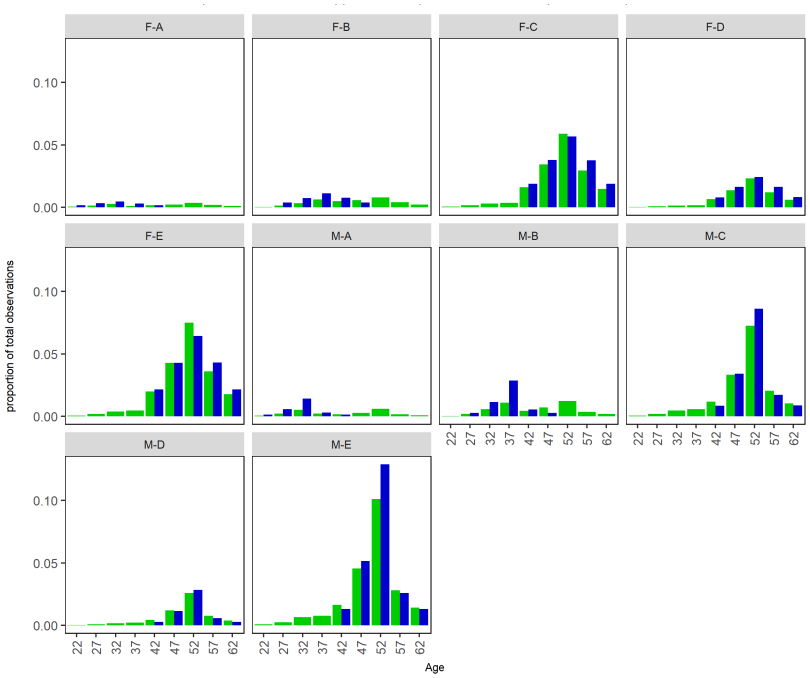


Figure 15: Distribution of *age* within *sex* and *race*. Original data in blue, synthetic in green. Laplace mechanism applied to frequencies prior to proportion computation. Global sensitivity = 0.025, $\epsilon = 1$. Less than full height tree causes synthesis of attribute, label combinations that do not appear in the original data.

## 9 Comments and Questions

- An RDT smooths distribution of observation mass by collapsing all labels for attributes that are absent in a node path into a single label (assuming a less than full height tree). Averaging leaf distributions from an ensemble of trees further smooths distribution. **All labels of all attributes that appear in node paths are available to generate synthetic records, enabling synthesis of combinations that do not appear in the source data. Is this desirable?** Careful selection of tree height controls smoothing.
- Smoothing masks covariate relationships. **What can we develop to measure loss of important attribute relationships?** It might be possible to establish a functional relationship between RDT properties, $\epsilon$, and information loss.
- The Laplace mechanism further perturbs distribution of mass. **Are RDT smoothing and Laplace adjustment both necessary to achieve desired results?** Note that the Laplace mechanism has numerical measure, whereas loss of fidelity due to smoothing does not (perhaps MSE of joint category proportion differences, or some measure of distance between $D_0$ and SD).
- An over-fit method that would limit synthesized records to attribute, label combinations appearing in the source data is to generate SQL style aggregation tables with Laplace adjusted proportions. Additional hypothesized attribute, label combinations could be added to synthesize unobserved observations in a controlled manner. The efficiency of SQL is expected to be much greater than RDT construction with subsequent sampling.
- Current results are limited to discrete attributes. Methods of recording continuous pdf parameters in leaves, along with synthesis sampling methods, must be developed.