EVALUATING GROUP-WISE UTILITY FOR SYNTHETIC DATASETS

CLAIRE MORTON, AARON R. WILLIAMS, AND CLAIRE MCKAY BOWEN

390 Jane Stanford Way, Stanford, CA 94305
e-mail address: mortonc@stanford.edu

Abstract. Synthetic data offer immense opportunities for evidence-based policy and program evaluation through balancing individual privacy with the release of disaggregated information. However, current methods to evaluate the utility of synthetic datasets do not provide guidelines for evaluating the utility of these datasets across subgroups to determine whether these groups have similar or different utility. We present methods to evaluate the utility of synthetic datasets by group using discriminant-based metrics. We first conduct a series of simulations to provide general recommendations for model selection for discriminant-based metrics. We then implement these recommendations empirically and explore two methods to evaluate group-wise utility in synthetic datasets. We present two case studies of these methods on microdata from the American Community Survey. Finally, we provide guidelines on implementing these methods to detect when synthetic datasets have acceptable utility for the overall dataset but may have low utility for specific subgroups. Our theoretical and practical findings demonstrate the importance of developing new approaches to critically evaluate synthetic datasets to ensure that they provide high utility to all groups.

1. Introduction

To promote equity in evidence-based policy and program evaluation, former President Biden issued two executive orders during his administration that called for the prioritization of publishing data disaggregated by factors such as race, income, and gender [Biden, 2021, 2023]. However, releasing these disaggregated data could threaten the privacy of millions of people, especially those belonging to minority groups, who are more easily identifiable than those belonging to majority groups [Bowen and Snoke, 2023]. Public policymakers and privacy experts have presented synthetic data as a solution that balances informing policy and preserving privacy [Advisory Committee on Data for Evidence Building (ACDEB), 2022]. Synthetic data consist of pseudo records, often generated based on a model, that can statistically represent confidential data without exactly matching them. This disguises individuals to protect against deidentification [Rubin, 1993, Garfinkel et al., 2015]. Federal organizations like the U.S. Internal Revenue Service [Bowen et al., 2022], Social Security

Key words and phrases: synthetic data, discriminant-based metric, propensity score, classification and classification trees, utility, evaluation, statistical disclosure control.

www.journalprivacyconfidentiality.org
DOI:10.29012/jpc.v0.i0.000

Administration [Abowd et al., 2006], and U.S. Census Bureau [Benedetto et al., 2018, Kinney et al., 2011] have created synthetic data at scale for public policy decision-making.

Synthetic datasets can also be used alongside verification or validation servers [1], which allow researchers to first run analyses on synthetic data [Advisory Committee on Data for Evidence Building (ACDEB), 2022]. In general, we can think of two scenarios in which synthetic data can be used with verification and validation servers. The first is using synthetic data as a tool for exploratory data analysis with a validation server to obtain valid inferences for public release. The second is using synthetic data to debug analyses that may be submitted to validation servers. In either case, it is critical that released synthetic data have a similar structure and range of values to confidential data and adequately preserve utility, defined as the extent to which the synthetic data preserve the statistical information in confidential data.

A failure to preserve utility could dissuade researchers from using such systems because the synthetic data do not report meaningful outputs or lead to inaccurate pre-training of models. Evaluating utility can also inform the choices made in designing synthetic datasets and determine whether synthetic data should be released at all [Drechsler, 2022]. For instance, the synthetic data should preserve marginal distributions and multivariate relationships equally well across subgroups in the data. However, current methods for evaluating utility of synthetic datasets focus on the overall dataset and not disaggregated utility quantification for each subgroup.

In this paper, we consider adapting a set of commonly used utility metrics, discriminant-based metrics, to provide information on the utility for subgroups, such as minority groups, within synthetic and confidential data. This paper proceeds as follows: Section 2 describes various methods for evaluating utility, including discriminant-based metrics. Section 3 outlines our comparison of various models for calculating these metrics. Sections 4 summarizes case studies of two approaches to evaluating discriminant-based metrics by group, and Section 5 offers recommendations, limitations, and concluding thoughts.

2. Evaluating Utility in Synthetic Datasets

Researchers can measure the utility of the synthetic data through three types of metrics: specific, global, and fit-for-purpose [Drechsler, 2022]. Utility is relative, and a dataset may have high utility for one purpose but low utility for another. Thus, a variety of utility metrics are typically used to assess different facets of a dataset's usefulness for a given task [Drechsler, 2022].

[1] Verification servers allow researchers to test whether specific inferences from the synthetic data match the confidential data [Williams et al., 2023]. Validation servers, by contrast, directly provide estimates from the confidential data with certain privacy guarantees once analyses have been prepared with synthetic data [Williams et al., 2023]. For example, to analyze the correlation between disease incidence and social determinants of health using confidential patient data, a verification server might report whether confidence intervals from coefficients of models fit from the confidential data overlap with the associated confidence intervals from the synthetic data. A validation server may instead release noisy versions of these coefficients with the associated errors. Federal statistical agencies including the US Census Bureau [Benedetto et al., 2018, Drechsler and Vilhuber, 2014] have used manual verification and validation servers.

Global Utility Metrics. Global utility metrics attempt to capture the overall utility of a synthetic dataset. Some of these metrics include measures of distance between datasets [Karr et al., 2006], tabulation measures [Nowok and Dibben, 2018, Voas and Williamson, 2001], and a class of metrics known as discriminant-based metrics [Sakshaug and Raghunathan, 2010, Snoke et al., 2018, Woo et al., 2009]. Discriminant-based metrics require fitting a model to predict if an observation comes from the confidential data, then calculating metrics based on the model's propensity scores.

To calculate propensity scores, we first assume for simplicity that the synthetic and confidential data are the same size. In general, however, the synthetic data can be larger or smaller than the confidential data. Let n be the number of observations in the synthetic and confidential data and d be the number of variables being considered for synthesis. Let the synthetic data be $S(n \times d)$ and the confidential data be $C(n \times d)$, where each row of S or C corresponds to one observation. Let $N = 2n$. C is appended to S to create $SC(N \times d)$. Let $y(N \times 1)$ be a vector where

$$\begin{aligned} y_i = & 1 \text{ if } SC(i) \text{ is from the synthetic data} \\ & 0 \text{ if } SC(i) \text{ is from the confidential data} \end{aligned}$$

Define the propensity score $e_i = e(X_i) := \Pr(y_i = 1|X_i)$, where $X_i(1 \times d)$ corresponds to row i of SC. Propensity scores can be estimated for parametric or nonparametric modeling approaches. For parametric models, we assume that the propensity scores can be characterized using a parameter vector θ such that $e_i = \Pr_\theta(y_i = 1|X_i)$. We fit the parametric model to obtain $\hat{\theta}$ and estimate $\hat{e}_i = \Pr_{\hat{\theta}}(y_i = 1|X_i)$.

The main idea of discriminant-based metrics is that a model with a poor fit indicates that it is difficult to discriminate between the synthetic and confidential datasets, so the synthetic data are similar to the confidential data. In contrast, a model with a good fit indicates that it is easy to discriminate between the synthetic and confidential data, so the synthetic data are different from the confidential data. Of course, a model could also poorly identify synthetic observations because it is a poorly developed model. We explore this issue further in Section 3.

Specific Utility Metrics. While general utility metrics are a crucial part of utility evaluation, they are not sufficient to provide a complete picture of the utility of synthetic data. For instance, if we have a certain analysis plan in mind, we should confirm that the synthetic and confidential data will give similar results for this analysis.

Specific utility metrics evaluate how well confidential and synthetic datasets match for specific analyses. These metrics are "specific" because they are tailored to assess the utility of the data for a certain research question or important factor determined by stakeholders.

For example, [Kinney et al., 2011] validated their synthetic Longitudinal Business Database, a census of U.S. businesses with paid employees, by comparing synthetic and confidential means and correlations for several variables including gross employment, gross payroll, establishments per year, and lifetime of establishments. They also compared coefficients from modeling establishment employment using their synthetic and confidential data. [Drechsler and Hu, 2021] created synthetic geocodes for the Integrated Employment Biographies data, an administrative dataset collected by the German Federal Employment Agency. They then compared the distributions of the shares of high-wage earners and foreigners at the zip code level to evaluate which of three potential synthesis methods was most successful at preserving these distributions.

Fit-for-Purpose Metrics. Fit-for-purpose metrics fall between specific and global utility metrics [Drechsler, 2022]. They can involve checking whether synthetic data are logical (e.g., checking that synthetic birth dates are not after synthetic death dates), comparing synthetic and confidential marginal histograms, and calculating global utility metrics based on a subset of variables. Such metrics have been shown to be a critical part of evaluating synthetic data utility throughout the process of generating a synthetic dataset since they provide more general insight than specific utility metrics and a more focused picture of utility than global metrics [Drechsler, 2022].

Discriminant-Based Metrics. In this paper, we introduce new methods to calculate discriminant-based metrics. We demonstrate these methods using four standard discriminant-based metrics: the Area Under the Receiver Operating Characteristic Curve (AUC), Synthetic Data Generation Propensity Score Matching Empirical CDF Comparison based on Kolmogorov-Smirnov (SPECKS) [Bowen et al., 2021], propensity score mean-squared error (pMSE) [Woo et al., 2009], and pMSE-ratio [Snoke et al., 2018]. We chose SPECKS, pMSE, and pMSE-ratio because of their popularity in existing synthetic data literature [Drechsler, 2022, Raab et al., 2021, Bowen et al., 2021]. We chose AUC because it is a commonly used metric for model selection in machine learning [Yang and Ying, 2022]. We define these metrics below.

- AUC is widely used in machine learning for model selection as a measure of model performance [Bradley, 1997]. The receiving operating characteristic (ROC) curve plots the true positive rate against the false positive rate for varying thresholds of classification using a given classifier. The AUC ranges from 0 to 1, with 1 indicating perfect classification for every threshold, 0.5 indicating classification no better than random, and values less than 0.5 indicating classification worse than random. In the case of discriminant-based metrics, lower AUC values indicate higher synthetic data utility, with a focus on whether or not the model's predictions are correct, since a model with high confidence but incorrect predictions will not have a high AUC.

  To understand the calculation of the AUC, we must define sensitivity and specificity. Let our threshold be t. Let positives be synthetic cases, true positives be positives with $e_i > t$, negatives be confidential cases, and true negatives be negatives with $e_i < t$. Then, let

$$Sensitivity = \frac{\text{Number of true positives}}{\text{Number of positives}}$$

$$Specificity = \frac{\text{Number of true negatives}}{\text{Number of negatives}}$$

  The AUC is calculated through graphing $1 - Specificity$ on the x-axis and $Sensitivity$ on the y-axis, for values of $t \in [0, 1]$. This curve is the ROC curve, and the AUC is the area under the curve [Metz, 1978].

- SPECKS measures the Kolomogorov-Smirnov distance between the empirical CDFs of propensity scores, which has a range of 0 to 1 [Bowen et al., 2021]. Lower SPECKS values indicate higher synthetic data utility.

Given propensity scores e_i for each of the N examples in the combined synthetic and confidential datasets, we can calculate SPECKS through obtaining the empirical CDFs $\hat{F}(e)$ and $\tilde{F}(e)$ for the confidential and synthetic data respectively. Then, we calculate the Kolmogorov-Smirnov distance, d_{KS}, following 2.1 [McKay Bowen et al., 2018]. This distance is the SPECKS.

$$d_{KS} = sup_e|\tilde{F}(e) - \hat{F}(e)| \tag{2.1}$$

- pMSE is calculated following 2.2.

$$\frac{1}{N}\sum_{i=1}^{N}(e_i - \frac{n_1}{N})^2 \tag{2.2}$$

  Similar to the other two metrics, lower values indicate higher utility. The pMSE focuses solely on the model's confidence in its predictions, since the pMSE only considers how extreme propensity scores are. Model correctness is not considered in the pMSE calculation.
- pMSE-ratio is calculated by (1) calculating the pMSE, (2) calculating the null pMSE, and (3) dividing the pMSE by the null pMSE [Snoke et al., 2018]. In this paper, we calculate the null pMSE by first taking a bootstrap sample of size N from the confidential data [Bowen and Snoke, 2019]. We then permute the outcome y_i and recalculate the pMSE. The mean of several bootstrap samples is the null pMSE. [Snoke et al., 2018] calculate the exact form of the null pMSE for logistic regression models; we extend this calculation to dense tree-based models (see the Supplemental Materials). pMSE-ratios close to 1 indicate higher utility, and pMSE-ratios under 10 are generally recognized as acceptable [Raab et al., 2021].

3. Model Selection for Discriminant-Based Metrics

We investigate the extent to which modeling practices influence whether discriminant-based metrics indicate known inadequacies in synthetic datasets for the overall data and within different subgroups. The impact of model type on discriminant-based metrics has largely not been investigated, especially for complex models. Although it is clear that global utility metrics can fail to detect major errors in synthetic data, such as spikes in the data [Drechsler, 2022], the types of errors that various models may be able to detect remain unclear [Hu and Bowen, 2024]. Understanding the impact of model selection for discriminant-based metrics is critical to justify which models are used in the following sections and recommended for future research.

To this end, we compare classification trees, random forest, logistic regression, and LASSO regression on simulated synthetic datasets with known inaccuracies. Based on whether the discriminant-based metrics calculated from these models are different from a control case, we make general recommendations for which models to use. In the past, classification trees have generally been recommended over logistic regression models for this task [Raab et al., 2021]. The model fit also depends on the synthesis method [Drechsler, 2022]. More complex models, such as random forests and regularized logistic regression, have not been compared. This is because these models involve hyperparameter tuning. Traditionally, models for calculating discriminant-based metrics are trained and evaluated on the entire combined synthetic/confidential dataset without hyperparameter tuning or using a train/test split [Raab et al., 2021, Drechsler, 2022]. The potential for overfitting has been recognized [Mendelevitch and Lesh, 2021, Snoke et al., 2018], but, to the best of our knowledge, no studies have used these methods to mitigate it.

Methods. We simulate synthetic and confidential datasets and calculate discriminant-based metrics using several model types. To test the strengths and weaknesses of each model type, our simulated synthetic datasets have specific errors to decrease their utility (Table 1). We introduced these errors to either all the simulated synthetic data points or to only 10% of these data points. The 10% error case mimics a setting in which the synthetic data are a poor match to the confidential data for a minority group or smaller subgroup. We create three types of errors: 1. changing the mean of one variable ("Mean" in Table 1), 2. changing the correlations between one variable and the rest of the variables ("Correlation" in Table 1), and 3. slightly changing the means and increasing the covariances, for all variables ("All" in Table 1). To create a synthetic dataset, S with 10 variables $X_1, ..., X_{10}$, we draw 10,000 samples from multivariate normal distributions with mean vector and covariance matrices shown in Table 1. In the setting where only 10% of the data are altered, we draw 1,000 samples from the multivariate normal distribution of interest, and the other 9,000 samples from the control multivariate normal distribution ("Control/Confidential" in Table 1).

Name	**Mean**	**Covariance**	**Description**
Control/ Confidential	$[0]^{10}$	$cov(x_m, x_{m'}) = 0.7$ for $m' \neq m$ $cov(x_m, x_m) = 1$	Multivariate normal
Mean	$1, [0]^9$	$cov(x_m, x_{m'}) = 0.7$ for $m' \neq m$ $cov(x_m, x_m) = 1$ For $m, m' \in 1, ..., 10$	Mean of X_1 is incorrect
Correlation	$[0]^{10}$	$cov(x_m, x_{m'}) = 0.7$ $cov(x_1, x_{m'}) = -0.7$ $cov(x_{m'}, x_1) = -0.7$ $cov(x_m, x_m) = 1$ $cov(x_1, x_1) = 1$ For $m, m' \in 1, ..., 10, m, m' \neq 1$	Correlation of X_1 and all other variables is reversed
All	Sampled from $[-0.2, 0.2]^{10}$	$cov(x_m, x_{m'}) = 1.4$ for $m' \neq m$ $cov(x_m, x_m) = 4$ For $m, m' \in 1, ..., 10$	All variables in synthetic data have perturbed means and larger variances/covariances.

Table 1: Overview of simulated confidential and synthetic datasets.

We then train models to discriminate between the simulated synthetic and the confidential datasets. Our models were: logistic regression with quadratic terms, LASSO with tuned penalty (penalty $\in 10^{(-(\{0,...,9\}/9)*10)}$) and quadratic terms, classification tree with tuned cost complexity (cost complexity $\in 10^{\{-10,...,-1\}}$), and random forest with 500 trees and a minimum node size of 5. To avoid overfitting, we split both the synthetic and confidential data into $75 - 25\%$ train-test splits. We then tune hyperparameters using 10-fold cross validation on the training data, choosing the best model. This procedure is implemented in the `syntheval` package in R (as used in Seeman et al. [2025]). We then evaluate the discriminant-based metrics on the test set. We fit models and calculate metrics using ten simulations per dataset type.

Results. We determine to what extent our train/test split impacted the model results based on the differences between discriminant-based metrics evaluated on the train and test data (Figures 1, 2, S1, S2). In general, we observe that LASSO and logistic regression have similar train and test metric values across metrics and simulated synthetic train datasets. The random forest model perfectly discriminates under the AUC and SPECKS for all simulated training datasets, including the control. However, the random forest resulted in the expected low control values for the AUC and SPECKS for the test data. The classification tree fit the training datasets better than the test datasets, since the discriminant-based metrics were generally slightly higher for the training data. This difference indicates that the classification tree may overfit to the training data. In this context, we define overfitting to mean that the models fit the training data better than the testing data across the ten simulations.

Evaluating discriminant-based metrics on a holdout test subset of the synthetic and confidential data leads to more accurate assessments of how well the confidential and synthetic datasets can be distinguished from one another using random forests and classification trees. In the absence of a train/test split, the models could produce metric values that are too high. This is a critical insight, especially for classification trees, since the higher metric values achieved using classification trees over logistic regression models have been used as justification to recommend classification trees for fitting discriminant-based metrics [Raab et al., 2021]. If these higher values are due to overfitting, not to actual differences in synthetic and confidential data, then classification trees may produce misleadingly high discriminant-based metric values when fit and evaluated without a train/test split.

We also assess the extent to which various models detected the issues present in simulated synthetic datasets. Models should have lower discriminant-based metric values for the control and higher values for the other synthetic datasets. For SPECKS, pMSE, and AUC, all four models perform similarly on the synthetic dataset where one variable's mean was altered ("Mean") and on the synthetic dataset where bivariate relationships were altered ("Correlation"). However, when all means and covariances were slightly changed ("All"), the logistic regression and LASSO models were noticeably closer to their control values than the random forest or classification tree (Figure 1).

The pMSE-ratio had extremely high values and was occasionally undefined when used with the LASSO model. This is because the LASSO model's regularization can force the optimal discriminating function to be zero for the null pMSE in the denominator. This leads to zeros (or extremely small numbers) in the denominator of the pMSE-ratio, inflating the metric. We omit the LASSO model results and suggest that the pMSE-ratio should be used with different types of regularization or tree-based approaches instead of approaches that drive model coefficients to zero.

Another important note is that the control metric values using classification trees are more variable than other model types. For the synthetic data in which only 10% of the dataset had alterations (Figure 2), the control metric values from the classification tree were occasionally higher than those of low utility synthetic datasets. In the test data, this variability appears primarily for the pMSE and pMSE-ratio. We attribute this to the fact that the pMSE does not actually measure any information related to how accurate discriminant-based metric models are – it solely measures how confident models tend to be in their predictions. Thus, classification trees can have poor accuracy but low pMSE values if they confidently make incorrect predictions. During training, they can create leaves with high confidence that do not correspond to the actual features of the data. When test data examples are funneled through these trees, they are misclassified with high confidence, leading to variability in metric values. This illustrates a key weakness of the pMSE and pMSE-ratio, especially when combined with tree-based modeling approaches. This result is in line with results from previous studies illustrating that the values of the pMSE are highly dependent on the model used to calculate the metrics [Drechsler, 2022].

Discussion/Recommendations. We make the following recommendations for model selection when calculating discriminant-based metrics for data that has minority and majority groups:

- Use hyperparameter tuning and a train/test split when fitting models for discriminant-based metrics. Otherwise, more complex models will overfit, or even memorize, the combined synthetic/confidential data. They will then produce high values for discriminant-based metrics even in cases where the synthetic data have high utility. Simpler models, or models with poorly chosen hyperparameters, may also underfit, resulting in discriminant-based metrics with low values even when the synthetic data are very different from the confidential data.
- LASSO models should not be used with the pMSE-ratio because they can calculate a zero or almost zero denominator, inflating the pMSE-ratio.
- Use tree-based models to calculate discriminant-based metrics if you are undecided on which model to implement. Tree-based models may overfit to training data but are generally more sensitive than parametric models on test data, especially when there are subtle differences between train and test data. For example, if a train-test split is infeasible due to categorical variables with many subgroups that should not be grouped into an "other" category, these models should not be used due to overfitting.
- When assessing global utility with classification tree models, use a suite of metrics, not just the pMSE. Metrics like the AUC, which consider model accuracy rather than just model confidence, should be paired with the pMSE.
- Out of the metrics calculated here, the SPECKS metric most consistently detected low utility when only a small proportion of the synthetic dataset was affected.

In summary, despite somewhat high variability, classification trees balance a comparative lack of over-fitting with the ability to detect a variety of issues in synthetic datasets. Although the random forest is better at this task than the classification tree, we chose to use a classification tree as the model underlying all discriminant-based metric calculations for the remainder of this paper for computational efficiency.

One important advantage of parametric models is that they are highly interpretable, even though they are less successful at detecting issues in synthetic data. Through comparing variable importance scores and model coefficients, we can assess the extent to which a given variable helps a discriminator differentiate synthetic from confidential examples when the discriminator's model is parametric. We can also qualitatively evaluate tree structures for single classification trees; the variables that the tree splits on early are more critical to

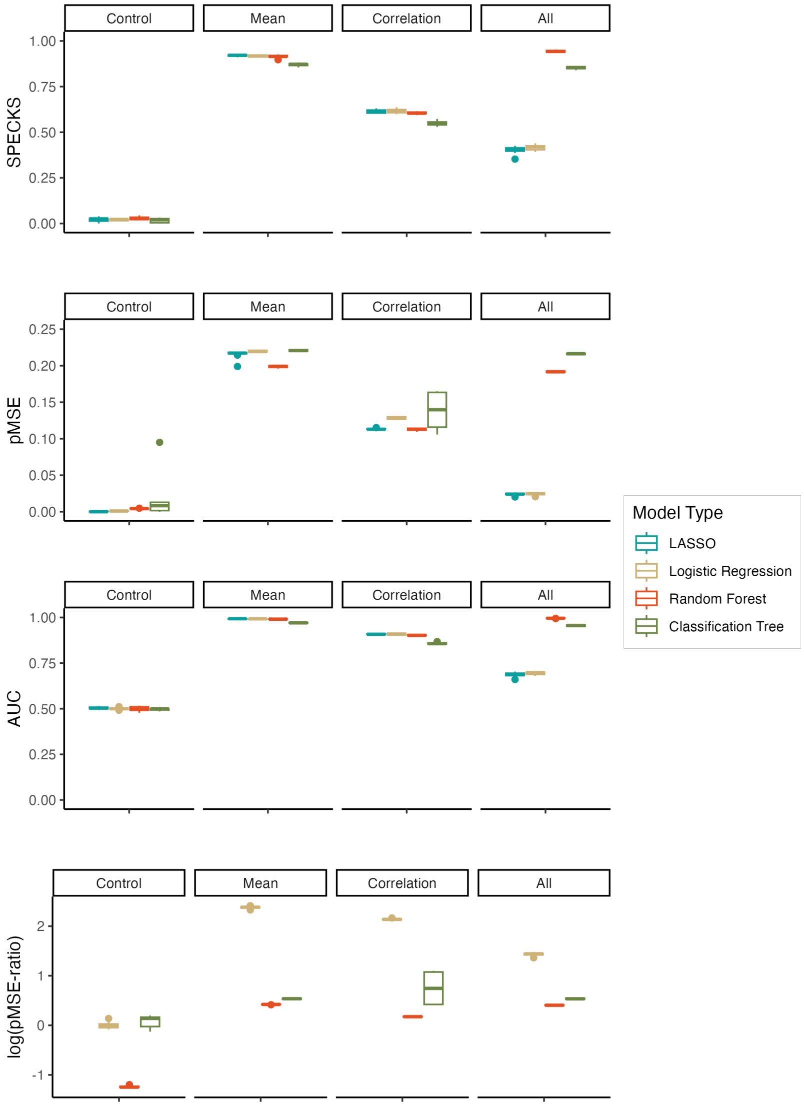

Figure 1: Tree-based models outperform parametric models for syntheses manipulating the entire synthetic dataset. This figure shows four discriminant-based metrics evaluated on test data. Discriminant-based metrics calculated using different model types vary in their ability to detect impairments to the utility of the synthetic data (right three columns). Boxplots show 10 trials from different syntheses.

telling the synthetic and confidential datasets apart. As models get more complex, however, they also get less interpretable, and non-parametric models do not have clear quantitative measures of variable importance.

We also note that when a minority of the synthetic data have low utility, the AUC and pMSE have very similar values to the control case where the synthetic data have ideal utility (Figure 2). This suggests that a new set of methods is needed to describe whether

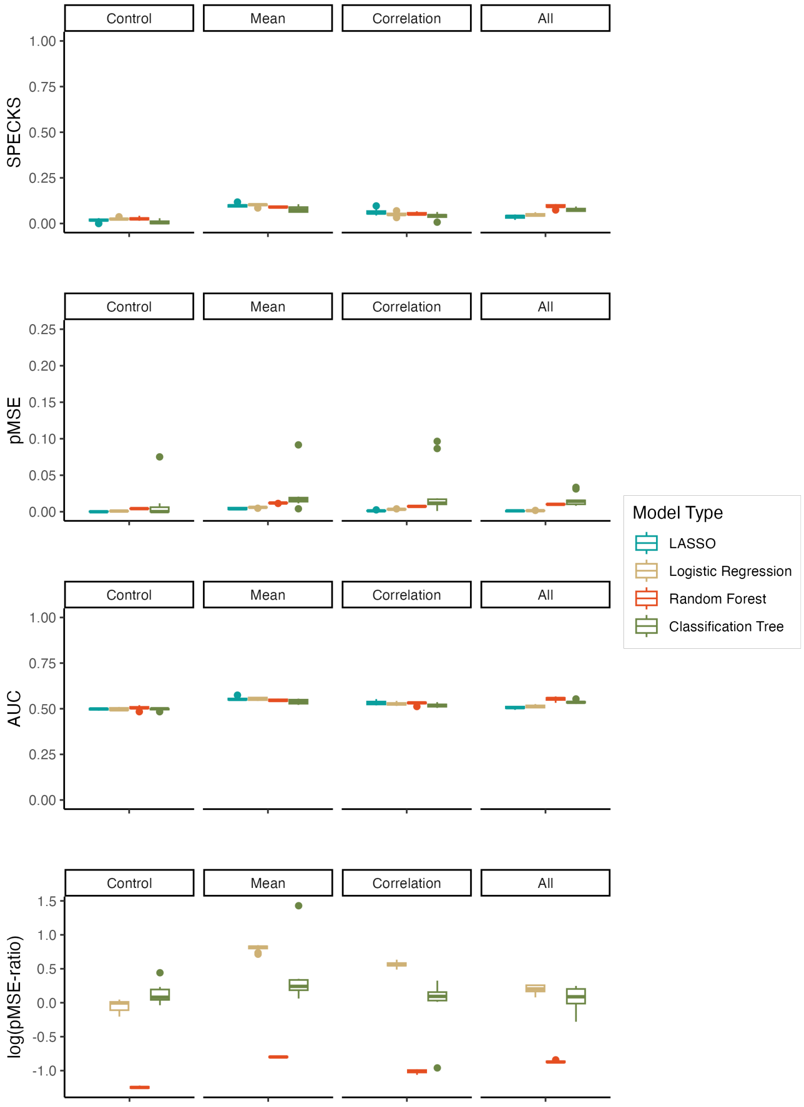

Figure 2: Discriminant-based metrics have generally similar values across model types for syntheses only impacting a minority of the synthetic data. This figure shows four different discriminant-based metrics evaluated on test data for synthesis manipulating 10% of the data. Discriminant-based metrics calculated using different model types vary in their ability to detect that the synthetic data's utility has been impaired (right three columns), but generally have similar values to each other and to the control case, in which there is no manipulation. Boxplots show 10 trials of different syntheses.

minority groups in the synthetic data have different and lower utilities to majority groups. In the following sections, we present case studies to achieve this goal through creating and evaluating synthetic microdata from the American Community Survey, an annual survey conducted by the U.S. Census. We then introduce and demonstrate two methods to assess utility of these synthetic datasets by group.

4. Single-Model and Dual-Model Approaches

We call these two methods the single- and dual-model approaches; the former method involves fitting one model to calculate discriminant-based metrics while the latter method involves fitting two models.

Data. To demonstrate these approaches, we extract the 2022 American Community Survey microdata for heads of household in the state of Michigan along with the sex, age, race, education level, health coverage status, employment status, poverty status, and veteran status for respondents ($n = 46,641$). We restrict the data to adults (over the age of 18), Black/African American and white people only, and complete cases, resulting in 37,890 respondents.

Our synthesis procedure creates ten different synthetic datasets with different levels of utility by majority/minority subgroup. To create these synthetic datasets, we alter the confidential data that inform our syntheses (Table 2). To create the altered confidential datasets, we permute a certain percentage of each column for either the majority or minority group. Depending on the proportion permuted, this interferes with the multivariate relationships within the majority or minority group to varying extents. Each altered confidential dataset was focused on a specific group indicator that split the data into a majority and minority group. The two group indicators that we used were race (Black and African American/white) and poverty status (under/over the federal poverty line). We categorize our syntheses as "good", "fair", and "poor". We did not permute any data for the "good" syntheses. For the "fair" syntheses, we permute 25% of the data. For the "poor" syntheses, we permute 75% of the data.

Using these ten altered confidential datasets, we complete our synthesis procedure using sequential synthesis with the regression and classification trees at a cost complexity $1e-4$, where the synthesis order was determined by the variables with the highest correlation to the grouping variable of interest, with categorical variables preceding continuous variables [Bowen et al., 2020]. We used the `tidysynthesis` package in R to synthesize the data [Williams, 2022].

In reality, we would never alter the confidential data used to inform a synthesis. However, to illustrate our methods, we use these syntheses to represent cases where a synthesized dataset does not have adequate utility for a majority or minority group in the data. In the rest of the paper, we refer to the ten synthesized datasets as "Race-good," "Black-fair," "Black-poor," "White-fair," "White-poor," "Poverty-good," "Under-fair," "Under-poor," "Over-fair," and "Over-poor." Here, for example "Black-fair" corresponds to the dataset synthesized from confidential data where 25% of entries corresponding to Black people in each column were permuted, while "Over-poor" corresponds to the dataset synthesized from confidential data where 75% of entries corresponding to people over the poverty line in each column were permuted (Table 2). The altered confidential datasets are only used to create the ten synthetic datasets. These synthetic datasets are then compared to the original unaltered confidential data to assess utility by group.

Dataset Name	**Description**
Black-poor	Permuted entries in each column for 75% of rows corresponding to Black people in confidential data
Black-fair	Permuted entries in each column for 25% of rows corresponding to Black people in confidential data
Race-good	No permutations
White-fair	Permuted entries in each column for 25% of rows corresponding to white people
White-poor	Permuted entries in each column for 75% of rows corresponding to white people
Under-poor	Permuted entries in each column for 75% of rows corresponding to people living under the poverty line
Under-fair	Permuted entries in each column for 25% of rows corresponding to people living under the poverty line
Poverty-good	No permutations
Over-fair	Permuted entries in each column for 25% of rows corresponding to people living over the poverty line
Over-poor	Permuted entries in each column for 75% of rows corresponding to people living over the poverty line

Table 2: Summary of how the confidential data were changed to create the synthetic data used in the case studies.

Single-Model Approach.

Statistical Approach. To evaluate group-wise utility using discriminant-based metrics, we fit a single model to the combined synthetic and confidential data and calculate propensity scores; just as in normal calculation of the metrics. We then separate the propensity scores into two groups and use the scores to calculate one discriminant-based metric per group. Our model was the classification tree from Section 3, with a tuned cp parameter (using 10-fold cross-validation) fit on training data and evaluated on test data (75/25 train/test split). The algorithm is shown below:

Algorithm 1: The single-model approach

```
Data: The confidential dataset $C = \{c_1, ..., c_n\}$, the synthetic dataset
      $S = \{s_1, ..., s_n\}$
Result: Metric values by group
(1) Concatenate $S$ and $C$ to produce $SC$;
(2) Add indicator column $y$ which is 1 for synthetic examples and 0 for confidential
    examples;
(3) Split into train and test;
(4) Fit classification trees to train data using 10-fold cross-validation;
(5) Choose best hyperparameters;
(6) Fit classification tree with the best hyperparameters to the full training dataset;
(7) Obtain propensity scores on test dataset;
(8) Split the test dataset by group;
(9) Calculate the metrics separately on the two subsets of the test dataset;
```

We hypothesized that the grouped metrics would have similar values if the synthetic data had similar utility across the groups. If one group had lower utility, its group metric should be higher than that of the other group. We evaluated the metrics using this method on ten different syntheses of the American Community Survey microdata following the synthesis procedure described above.

This method differs from previous studies, which have described methods to evaluate the utility of a synthesis based on subsets of variables – simply fit a model using just this subset and calculate the discriminant-based metrics based on the propensity scores [Raab et al., 2021]. However, when we tried this approach with an indicator for group membership, we found the method to be less effective for categorical variables with few categories.

Results. For our synthetic data with varying qualities of synthesis for people above/below the federal poverty line, we found that the AUC and SPECKS generally detected the "poor" syntheses, in which the synthetic data were informed by a confidential dataset with 75% of the data from one of the groups permuted (Figures 3 and 4). The pMSE and pMSE-ratio generally were incapable of distinguishing the differences in utility. Also, the two groups were less distinguishable for all metrics under the "fair" syntheses, indicating that the single-model approach is not sensitive to cases where the synthetic data have slightly lower utility for one of the groups.

Dual-Model Approach.

Statistical Approach. Because of the observed low sensitivity of the single model approach, we present an alternative method using two separate models. Rather than fitting a single model to all of the combined synthetic and confidential data and then splitting by group prior to calculating the metrics, we start by dividing the synthetic and confidential data by group. We then join the synthetic and confidential datasets corresponding to the majority group and the synthetic and confidential datasets corresponding to the minority group. We fit two separate models to distinguish between synthetic and confidential examples – one model for the majority and one for the minority. We tune the models' hyperparameters separately. We evaluate the discriminant-based metrics with each model, yielding one set of

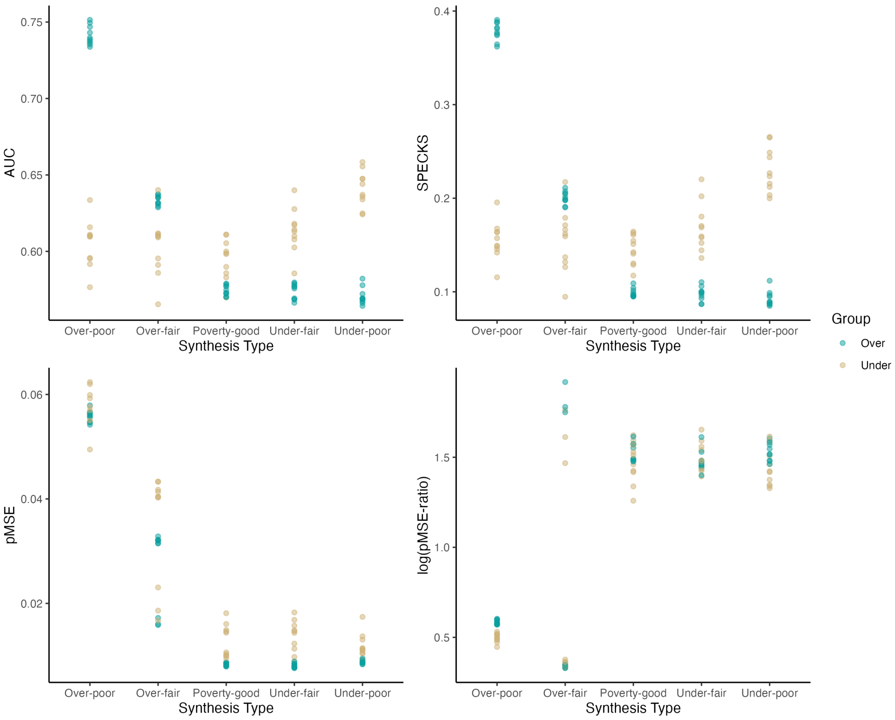

Figure 3: The single-model approach detects when majority groups have lower utility than minority groups but is not sensitive to cases where the difference in utility is slight. The figure shows metric values for synthetic data of varying qualities for people living over and under the poverty line, evaluated using the single-model approach.

metrics per group. Similar to the single-model approach described above, the dual-model approach allows us to compare the utility from group to group. However, the dual-model approach also accounts for the fact that different groups may have different underlying structures that can be used to predict whether an example comes from the confidential data. We applied this method to the synthetic data described in the previous section using the same model specifications. The algorithm is shown below:

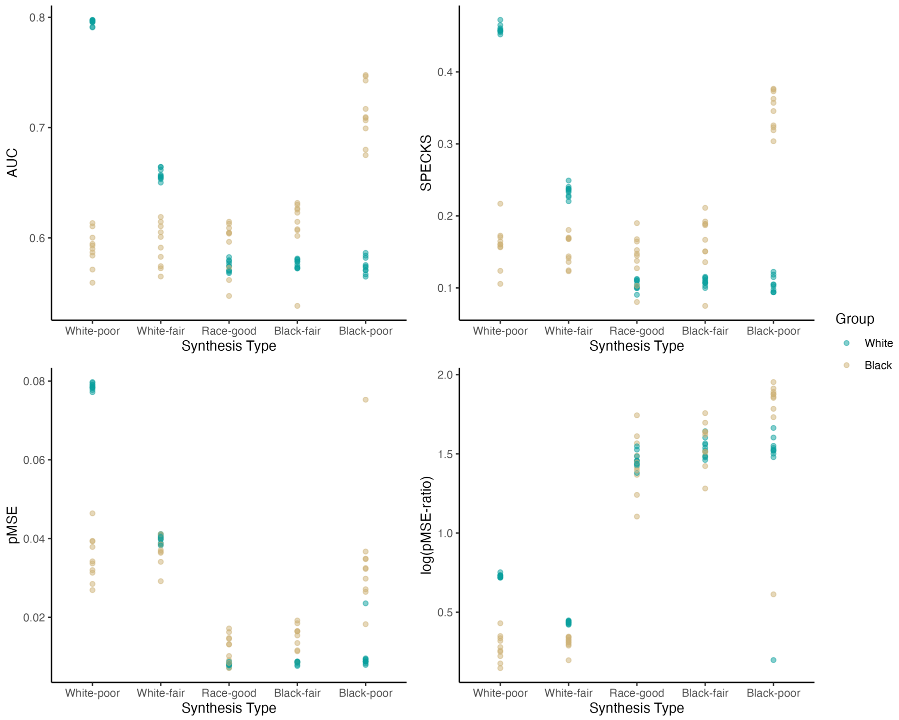

Figure 4: The single-model approach is sensitive to synthetic datasets where one racial group has different utility to another. The figure shows metric values for synthetic data of varying qualities for Black and white people (blue and red, respectively), evaluated using the single-model approach.

```
Algorithm 2: The dual-model approach
  Data: The confidential dataset $C = \{c_1, ..., c_n\}$, the synthetic dataset
        $S = \{s_1, ..., s_n\}$
  Result: Metric values by group
  (1) Concatenate $S$ and $C$ to produce $SC$;
  (2) Add indicator column $y$ which is 1 for synthetic examples and 0 for confidential
      examples;
  (3) Split the combined dataset by group;
  (4) for Each of the two datasets do
          Split into train and test;
          Fit classification trees to train data using 10-fold cross-validation;
          Choose best hyperparameters;
          Fit classification tree with the best hyperparameters to the full training
        dataset;
          Obtain propensity scores on test dataset;
          Calculate the metrics on the test dataset;
      end
```

Results. The grouped AUC and SPECKS in the dual-model approach are much further separated from each other than in the single-model approach, indicating that the dual-model approach has higher sensitivity. This is especially apparent for the “fair” datasets in which 25% of the entries for one group were permuted in the confidential data. The single-model approach shows slight changes from model to model, but the dual-model approach shows clear splits between the permuted and non-permuted group (Figures 5 and 6). However, for the dataset where 25% of the data were permuted for people living above the poverty line, the dual-model approach resulted in metrics with similar values across classes.

Compared to the single-model approach, the dual-model approach pMSE more clearly captures differences in the classes. The pMSE-ratio still fails to capture these differences in utility, even for the “Poor” syntheses – the pMSE-ratio instead consistently finds that the majority group has worse utility than the minority group. The denominator of the pMSE-ratio in the dual-model approach is calculated through bootstrapping by group, then fitting and calculating the pMSE for two separate models, one on each bootstrapped group dataset. We have shown that the null pMSE-ratio is smaller for larger group sizes for dense tree-based models (see Proofs), a result that has also been proven for logistic regression models Snoke et al. [2018]. This may be influencing the pMSE-ratio in the single- and dual-model approaches.

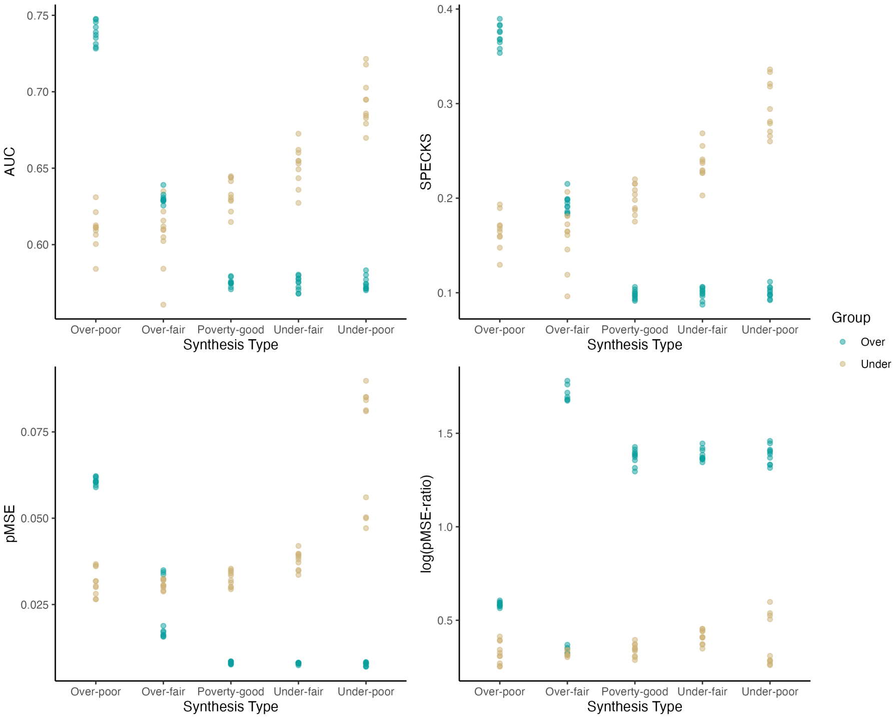

Figure 5: The dual-model approach is sensitive to changes in utility along the poverty line. The figure shows metric values for synthetic data of varying qualities for people living over and under the poverty line, evaluated using the dual-model approach.

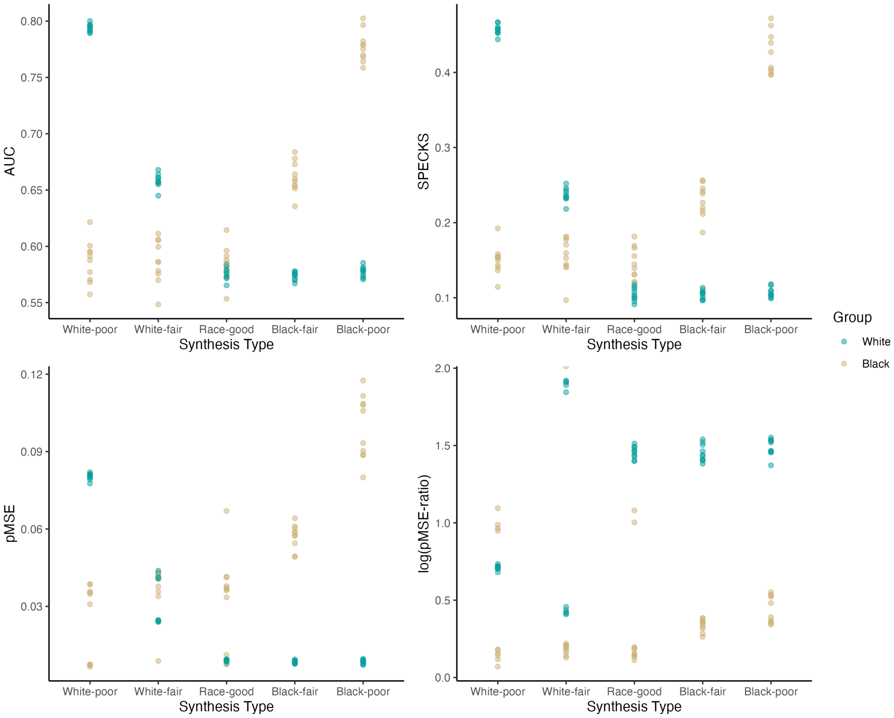

Figure 6: The dual-model approach is sensitive to changes in utility along racial lines. The figure shows metric values for synthetic data of varying qualities for Black and white people, evaluated using the dual-model approach.

Discussion/Recommendations. As expected, the single-model approach shows that the synthetic data generated from the original confidential data have the highest utility and the most equitable utility across the groups. However, this approach is generally not very sensitive, especially for the dataset in which we permute either the individuals in poverty or not in poverty. Although the dual-model approach is more sensitive than the single-model approach, it does not always behave as expected. For example, on the dataset where we permute 25% of the data for people living above the poverty line, the dual-model approach results in similar metric values for those on either side of the poverty line (Figure 5). The approach also shows that the synthesis from the original confidential dataset has a higher utility (lower metric values) for the majority group. This may be due to the fact that the minority group has a lower variance for certain predictors, particularly total family income. For a person to live below the poverty line, total family income must be constrained between zero and low values. The majority group does not have these constraints and thus have higher variance for total family income. When a model predicts whether or not an example comes from confidential data, this higher variance translates to a higher irreducible error and lower values for discriminant-based metrics.

In summary, the existence of a strong correlation between the grouping variable of interest and another variable in the dataset, where the other variable has different variances by group, leads to different levels of complexity for fitting models to calculate discriminant-based metrics. This hypothesis appears to be supported by a comparison with the dual-model approach results for the datasets where either Black or white respondents' data were permuted. In this case, there are no variables with extremely strong correlations to the group indicator. Thus, the irreducible errors remain similar across the groups. The discriminant-based metrics indicate that the synthesis from unpermuted confidential data is then the most equitable, since it led to the best balance of discriminant-based metric values.

All of these insights are available only from the AUC and SPECKS for the single-model approach, and for the AUC, SPECKS, and pMSE for the dual-model approach.

We provide several recommendations for the use of the dual-model approach below:

- Use the single-model approach as an efficient check to see if the utilities across groups are drastically different.
- Use the dual-model approach over the single-model approach if computational resources are available, but be aware that the results may be influenced by whether or not there are other variables present with strong correlations to the grouping variable of interest.
- Do not use the single-model approach to calculate the pMSE or pMSE-ratio, nor the dual-model approach to calculate the pMSE-ratio, since these metrics are unable to detect differences in synthesis quality by group.

5. Conclusion

In this paper, we provide a set of methods and recommendations to evaluate group-wise utility in synthetic data using discriminant-based metrics. This is critical because without a reliable approach to determine whether synthetic data have equal utility across different groups, we risk releasing synthetic data with lower utility for some subgroups of the population, negatively impacting research and public policymaking. Our focus on discriminant-based metrics enables general measures of utility to capture whether synthetic data have equal utility across subgroups.

We also provide general guidelines for the use of discriminant-based metrics, regardless of whether users seek to evaluate utility by group. We emphasize the importance of using a train-test split when calculating discriminant-based metrics. To the best of our knowledge, no existing studies have used train-test splits or cross-validation to tune hyperparameters or calculate these metrics. This impairs our ability to use more complex models to discriminate between synthetic and confidential data and risks overfitting the model, leading to falsely high discriminant-based metric values. We explore a variety of models for calculating discriminant-based metrics and find that tree-based models generally balance a lack of overfitting with the specificity to detect both broad and specific issues in synthetic datasets.

In two use case studies from the American Community Survey microdata, we apply our recommendations for model selection. We use both single-model and dual-model approaches to detect when synthetic data have low, medium, or high utility for either the majority or minority group. Based on the ease of implementation of the single-model approach, we recommend that future work with discriminant-based metrics calculate the AUC, SPECKS, and pMSE by group as well as for the entire dataset. This can be easily added into existing workflows since it does not require any additional models to be fitted and simply involves grouping calculations from existing propensity scores. Evaluating these metrics by group is an efficient way to check whether one of the groups has dramatically lower utility than another, though it does not have high sensitivity. The dual-model approach, paired with the AUC, SPECKS, or pMSE, can detect small issues in synthetic datasets better than the single-model approach. However, it is more complex and requires more computing power and time. We recommend that future research investigating group-wise utility in detail employ the dual-model approach.

Our work suggests several future directions. First, the tools and approaches we suggest should be applied to existing synthetic datasets to evaluate whether they have equitable utility across groups. Additionally, in order to fit better models for calculating discriminant-based metrics, future studies should continue to test alternative model types and, in particular, determine whether there are cases not tested here that impair the fit of tree-based models.

In the dual-model approach, we are uncertain whether the gap in utility for the synthesis from unpermuted confidential data was due to an actual disparity in synthesis quality or to the difference in variance across the groups of a highly correlated variable. Future work could examine this question further to determine whether there is a correction that can be made to these metrics when they are calculated using the dual-model approach.

Acknowledgment

This research was funded by the National Science Foundation National Center for Science and Engineering Statistics [ADC-2023-606-01].

The authors also thank Art Owen and members of the Owen lab at Stanford University for helpful discussions about the work.

References

J. Abowd, M. Stinson, and G. Benedetto. Final report to the social security administration on the sipp/ssa/irs public use file project. 2006.

Advisory Committee on Data for Evidence Building (ACDEB). Advisory committee on data for evidence building: Year 2 report. Technical report, 2022.

G. Benedetto, J. C. Stanley, E. Totty, et al. The creation and use of the sipp synthetic beta v7. 0. *US Census Bureau*, 2018.

J. Biden. Executive order on advancing racial equity and support for underserved communities through the federal government. 2021.

J. Biden. Executive order on advancing racial equity and support for underserved communities through the federal government. 2023.

C. Bowen and J. Snoke. Do no harm guide: Applying equity awareness in data privacy methods. 2023.

C. M. Bowen and J. Snoke. Comparative study of differentially private synthetic data algorithms from the nist pscr differential privacy synthetic data challenge. *arXiv preprint arXiv:1911.12704*, 2019.

C. M. Bowen, V. Bryant, L. Burman, S. Khitatrakun, R. McClelland, P. Stallworth, K. Ueyama, and A. R. Williams. A synthetic supplemental public use file of low-income information return data: methodology, utility, and privacy implications. In *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2020, Tarragona, Spain, September 23–25, 2020, Proceedings*, pages 257–270. Springer, 2020.

C. M. Bowen, F. Liu, and B. Su. Differentially private data release via statistical election to partition sequentially: Statistical election to partition sequentially. *Metron*, 79(1):1–31, 2021.

C. M. Bowen, V. Bryant, L. Burman, J. Czajka, S. Khitatrakun, G. MacDonald, R. McClelland, L. Mucciolo, M. Pickens, K. Ueyama, et al. Synthetic individual income tax data: Methodology, utility, and privacy implications. In *International Conference on Privacy in Statistical Databases*, pages 191–204. Springer, 2022.

A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

J. Drechsler. Challenges in measuring utility for fully synthetic data. In *International Conference on Privacy in Statistical Databases*, pages 220–233. Springer, 2022.

J. Drechsler and J. Hu. Synthesizing geocodes to facilitate access to detailed geographical information in large-scale administrative data. *Journal of Survey Statistics and Methodology*, 9(3):523–548, 2021.

J. Drechsler and L. Vilhuber. Synthetic longitudinal business databases for international comparisons. In *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2014, Ibiza, Spain, September 17-19, 2014. Proceedings*, pages 243–252. Springer, 2014.

S. Garfinkel et al. *De-identification of Personal Information:*. US Department of Commerce, National Institute of Standards and Technology, 2015.

J. Hu and C. M. Bowen. Advancing microdata privacy protection: A review of synthetic data methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 16(1):e1636, 2024.

A. F. Karr, C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil. A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, 60(3):224–232, 2006.

S. K. Kinney, J. P. Reiter, A. P. Reznek, J. Miranda, R. S. Jarmin, and J. M. Abowd. Towards unrestricted public use business microdata: The synthetic longitudinal business database. *International statistical review*, 79(3):362–384, 2011.

C. McKay Bowen, F. Liu, and B. Su. Differentially private data release via statistical election to partition sequentially. *arXiv e-prints*, pages arXiv–1803, 2018.

O. Mendelevitch and M. D. Lesh. Fidelity and privacy of synthetic medical data. *arXiv preprint arXiv:2101.08658*, 2021.

C. E. Metz. Basic principles of roc analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier, 1978.

B. Nowok and C. Dibben. Putting synthetic people in place: creating synthetic data for spatial analysis at the individual level. *QCumber-EnvHealth project: WP3 Accessible health data*, 1(1):1–12, 2018.

G. M. Raab, B. Nowok, and C. Dibben. Assessing, visualizing and improving the utility of synthetic data. *arXiv preprint arXiv:2109.12717*, 2021.

D. B. Rubin. Statistical disclosure limitation. *Journal of official Statistics*, 9(2):461–468, 1993.

J. W. Sakshaug and T. E. Raghunathan. Synthetic data for small area estimation. In *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2010, Corfu, Greece, September 22-24, 2010. Proceedings*, pages 162–173. Springer, 2010.

J. Seeman, A. R. Williams, and C. Bowen. Synthetic data for the nebraska statewide workforce & educational reporting system. Technical report, Urban Institute, 2025.

J. Snoke, G. M. Raab, B. Nowok, C. Dibben, and A. Slavkovic. General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181(3):663–688, 2018.

D. Voas and P. Williamson. Evaluating goodness-of-fit measures for synthetic microdata. *Geographical and Environmental Modelling*, 5(2):177–200, 2001.

A. Williams. The tidysynthesis r package. 2022.

A. Williams, J. Snoke, C. Bowen, and A. Barrientos. Disclosing economists' privacy perspectives: A survey of american economic association members on differential privacy and data fitness for use standards, 2023.

M.-J. Woo, J. P. Reiter, A. Oganian, and A. F. Karr. Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*, 1(1), 2009.

T. Yang and Y. Ying. Auc maximization in the era of big data and ai: A survey. *ACM computing surveys*, 55(8):1–37, 2022.

Proofs

We derive the expected value of the null pMSE for tree-based models. We consider tree-based models where each leaf is its own data point–in other words, the most dense trees possible. To calculate the null pMSE, following SOURCE, we use our confidential data with n rows. Let x_j be the j^{th} point in the confidential dataset, $j \in \{1, ..., n\}$. We assume the synthetic data also have size n, though this assumption can be relaxed. We bootstrap a dataset of size $2n$ from the confidential data. Let z_i be the i^{th} point in the bootstrapped dataset, $i \in \{1, ..., 2n\}$, and label these data randomly as confidential (0) or synthetic (1), such that $y_i = \mathbb{1}(z_i \text{ synthetic})$. Then, fit the dense tree-based model and calculate the null pMSE:

$$\frac{1}{2n}\sum_{i=1}^{2n}(\hat{p_i} - \frac{n}{2n})^2 \tag{5.1}$$

Where $\hat{p_i}$, the estimated propensity score, can be calculated as $\hat{p_i} = \frac{\sum_{k=1}^{2n}\mathbb{1}(z_k=z_i)\cdot y_k}{\sum_{k=1}^{2n}\mathbb{1}(z_k=z_i)}$. The expected value of the null pMSE is then

$$\begin{aligned}\mathbb{E}[pMSE] &= \mathbb{E}\left[(\frac{1}{2n}\sum_{i=1}^{2n}(\hat{p_i} - \frac{n}{2n})^2)\right] \\ &= \frac{1}{2n}\sum_{i=1}^{2n}\mathbb{E}\left[(\hat{p_i} - \frac{1}{2})^2)\right] \\ &= \frac{1}{2n}\sum_{i=1}^{2n}[\mathbb{E}(\hat{p_i}^2) - \mathbb{E}(\hat{p_i}) + \frac{1}{4}]\end{aligned}$$

where

$$\begin{aligned}
\mathbb{E}[\hat{p_i}] &= \sum_{j=1}^{n} P(z_i = x_j)\mathbb{E}[\hat{p_i}|z_i = x_j] \\
&= \sum_{j=1}^{n} \frac{1}{n}\mathbb{E}[\hat{p_i}|z_i = x_j] \\
&= \mathbb{E}\left[\frac{\sum_{k=1}^{2n} \mathbb{1}(z_k = z_i)\cdot y_k}{\sum_{k=1}^{2n} \mathbb{1}(z_k = z_i)}|z_i = x_1\right] \text{ since all } j \text{ terms are identical} \\
&= \sum_{t=0}^{2n} \mathbb{P}\left[\left(\sum_{i=1}^{2n}(z_i = x_1)\right) = t\right] \mathbb{E}\left[\frac{\sum_{k=1}^{2n} \mathbb{1}(z_k = z_i)\cdot y_k}{\sum_{k=1}^{2n} \mathbb{1}(z_k = z_i)}| \left(\sum_{k=1}^{2n} \mathbb{1}(z_i = z_k) = t\right)\right] \\
&= \sum_{t=1}^{2n} \binom{2n}{t}\left(\frac{1}{n}\right)^t \left(\frac{n-1}{n}\right)^{2n-t}\left(\frac{1}{t}\right)\left(\frac{t}{2}\right) \\
&= \frac{1}{2}\sum_{t=1}^{2n} \binom{2n}{t}\left(\frac{1}{n}\right)^t \left(\frac{n-1}{n}\right)^{2n-t} \\
&= \frac{1}{2}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[\hat{p_i}^2] &= \sum_{j=1}^{n} \mathbb{P}(z_i = x_j)\mathbb{E}\left[\hat{p_i}^2|z_i = x_j\right] \\
&= \sum_{j=1}^{n} \frac{1}{n}\mathbb{E}\left[\hat{p_i}^2|z_i = x_j\right] \\
&= \mathbb{E}\left[\left(\frac{\sum_{k=1}^{2n} \mathbb{1}(z_k = z_i)\cdot y_k}{\sum_{k=1}^{2n} \mathbb{1}(z_k = z_i)}\right)^2 |z_i = x_1\right] \text{ since all } j \text{ terms are identical} \\
&= \sum_{t=1}^{2n} \mathbb{P}\left[\left(\sum_{i=1}^{2n}(z_i = x_1)\right) = t\right] \mathbb{E}\left[\left(\frac{\sum_{k=1}^{2n} \mathbb{1}(z_k = z_i)\cdot y_k}{\sum_{k=1}^{2n} \mathbb{1}(z_k = z_i)}\right)^2 | \left(\sum_{k=1}^{2n} \mathbb{1}(z_i = z_k) = t\right)\right] \\
&= \sum_{t=1}^{2n} \binom{2n}{t}\left(\frac{1}{n}\right)^t \left(\frac{n-1}{n}\right)^{2n-t}\left(\frac{1}{t^2}\right)\left(\frac{t+t^2}{4}\right) \\
&\text{since } \left(\frac{\sum_{k=1}^{2n} \mathbb{1}(z_k = z_i)\cdot y_k}{\sum_{k=1}^{2n} \mathbb{1}(z_k = z_i)}\right)^2 | \left(\sum_{k=1}^{2n} \mathbb{1}(z_i = z_k) = t\right) \sim Binom(t, \frac{1}{2}) \\
&= \frac{1}{4}\sum_{t=1}^{2n} \binom{2n}{t}\left(\frac{1}{n}\right)^t \left(\frac{n-1}{n}\right)^{2n-t}\left(\frac{t+1}{t}\right)
\end{aligned}$$

Thus

$$\begin{aligned}\mathbb{E}[pMSE] &= \frac{1}{2n}\sum_{i=1}^{2n}\left[\mathbb{E}\left[\hat{p_i}^2\right] - \mathbb{E}\left[\hat{p_i}\right] + \frac{1}{4}\right] \\ &= -\frac{1}{4} + \frac{1}{4}\sum_{t=1}^{2n}\binom{2n}{t}\left(\frac{1}{n}\right)^t\left(\frac{n-1}{n}\right)^{2n-t}\left(\frac{t+1}{t}\right)\end{aligned}$$

which is decreasing in n. Thus, $\mathbb{E}[pMSE]$ is decreasing in n, as required.

Supplementary Figures/Tables

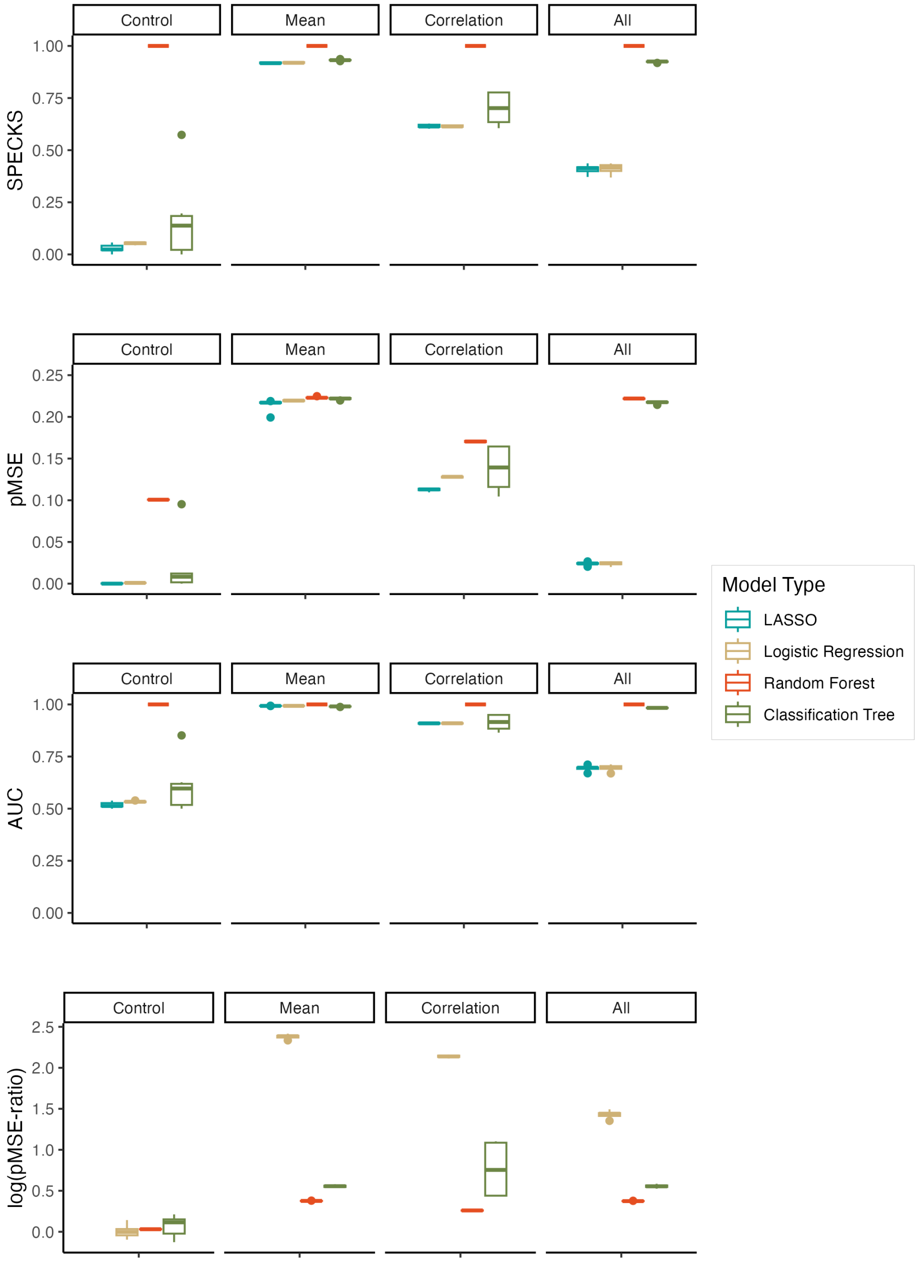

Figure S1: This figure shows four discriminant-based metrics evaluated on train data for synthesis manipulating 100% of the data. When evaluated on the training data, discriminant-based metrics calculated using different model types vary in their overfitting, but generally LASSO, logistic regression, and classification tree models do not overfit to the control, in which there is no manipulation. Boxplots show 10 trials of different syntheses.

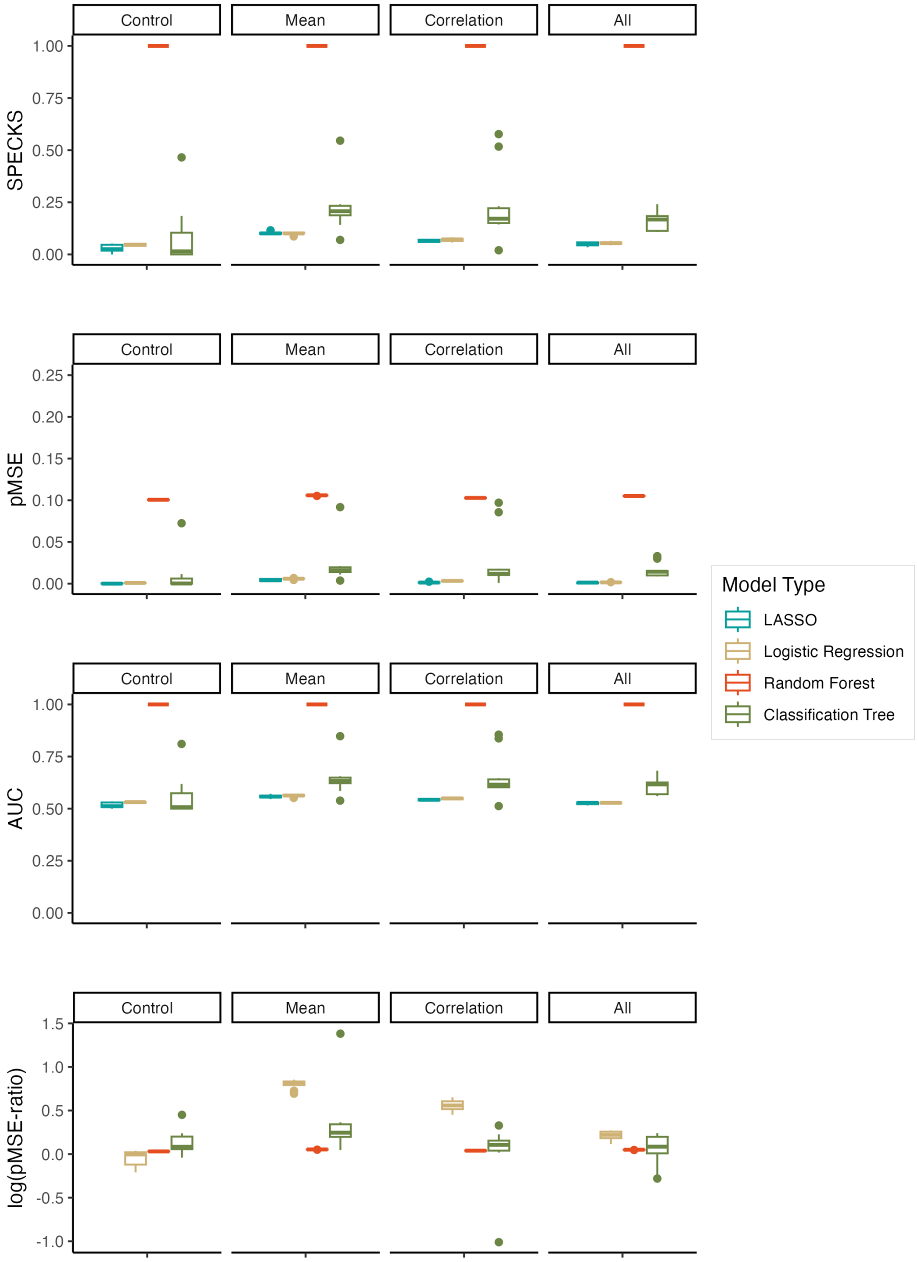

Figure S2: This figure shows four discriminant-based metrics evaluated on train data for synthesis manipulating 10% of the data. When evaluated on the training data, discriminant-based metrics calculated using different model types vary in their overfitting, but generally LASSO, logistic regression, and classification tree models do not overfit to the control, in which there is no manipulation. Boxplots show 10 trials of different syntheses.