Synthetic data disclosure control: Promise and feasibility for SLDS

Jacob M. Schauer
Northwestern University
jms@u.northwestern.edu

Arend M. Kuyper
Northwestern University
a-kuyper@northwestern.edu

Eric C. Hedberg
NORC
hedberg-eric@norc.org

Frank Feinis
Northwestern University
frankfineis2022@u.northwestern.edu

Larry V. Hedges
Northwestern University
l-hedges@northwestern.edu

State longitudinal data systems (SLDS) are legally required to protect the privacy of students and so states have been cautious about sharing data with external researchers. However, other government bodies, such as the Bureau of Labor Statistics, have experimented with releasing synthetic data generated from methods related to multiple imputation. The idea is that the real data is used to generate a series of synthetic datasets on which analyses can be conducted and pooled. Doing so can improve the utility of the released data in that analyses conducted on synthetic data can closely mirror those conducted on the real data. It can also improve privacy, since none of the data is actually real. In this article, we apply these procedures to data from eight states, and assess how feasible these procedures are, how well they preserve the data utility, and how well they protect privacy. We find that while the procedure can be computationally intensive, that the utility of the data is good, and the risk of disclosure is low.

Keywords: data privacy, data disclosure, synthetic data, SLDS

1. INTRODUCTION

US state longitudinal data systems (SLDS) collect data that has the potential to inform improvement in education, but that potential can be limited by concerns over security (Conaway, et. al., 2015). SLDS are legally required to protect the privacy of students under laws like the Family Educational Rights and Privacy Act (FERPA), and so it is not entirely surprising that states have been cautious about sharing data with external researchers. Sometimes this involves denying data requests outright. Other times, states apply various data masking procedures in order to hedge against potential data disclosures. For instance, the few states that offer public-use data apply microsuppression (also called "small cell suppression") rules to delete potentially identifiable records from their data prior to release (e.g., Massachusetts, 2014; North Carolina, 2009). This has even occurred with data shared under secure data use agreements (DUA). However, such procedures can limit the utility of the data that is actually shared. Various researchers have pointed out that microsuppression can lead to substantial bias in important quantities (Seastrom, 2010; Levesque, et. al., 2015) including with education data (Authors, under review). This raises the question of how to expand data access without greatly sacrificing data quality.

A potential solution to this problem is the use of synthetic data. Synthetic data disclosure methods work by releasing multiple datasets that look like the actual data, but that are generated randomly (Rubin, 1993). Part of their appeal is that synthetic datasets can be generated according to the full joint distribution of the variables conditioned on the observed real data, meaning that the joint distribution of variables in the synthetic data will be similar to that of the observed data. Thus, analyses using quality synthetic data should yield very similar results to analyses using the original data. Simulation studies show that such methods can preserve many relationships between variables, though these have necessarily involved rather small samples sizes (typically tens of thousands of individuals) and relatively small numbers of variables (typically dozens) (e.g., Raghunanthan, Reiter, & Rubin, 2003; Reiter, 2005b; Singh, Yu, & Dunteman, 2003). There have, however, been demonstrations of the methods in real, large datasets with many variables (Reiter, 2005a; Dreschsler, et al., 2008; Singh, Yu, & Dunteman, 2003; Prada et al., 2011).

One can also argue that synthetic data methods provide strong protections against disclosure, because none of the actual data gets released. Studies of synthetic data generated from the US Current Population Survey (Reiter, 2005a) and the German establishment survey (Drechsler, et al., 2008) both concluded that disclosure risk was very low. Further, these studies demonstrated that the estimates derived based on a variety of analyses of the synthetic data are quite similar to those run on the real data.

While existing demonstrations and simulations appear promising, less work has been done on the utility and feasibility of these methods for SLDS. This article examines the process of generating synthetic datasets using SLDS data. As part of this examination, we attempted to create synthetic datasets for various cross sections of real state education data. In the following sections, we describe the process we followed and highlight potential barriers for SLDS in generating synthetic data. We then investigate the extent to which important statistics from the real data are preserved in the synthetic data. Finally, we describe potential disclosure risks and limitations of this approach.

2. SYNTHETIC DATA METHODS

The theory of generating and analyzing synthetic datasets emerged from work on multiple imputation (Rubin, 1993). Both synthetic data and multiple imputation methods use a similar machinery, though with one key difference: where multiple imputation fills in missing fields with random draws from some probability distribution, synthetic data methods replace *all* fields with random draws. In this section we provide an overview of the theory for constructing and analyzing synthetic data. We discuss some important considerations for doing so, including how to handle missing data, and the challenges these considerations present to SLDS.

2.1. GENERATING SYNTHETIC DATASETS

Following the notation of Raghunathan et al. (2003), denote a state dataset by $\boldsymbol{P}$, and suppose it contains variables $\boldsymbol{X} = [X_1, \ldots, X_q]$ and $\boldsymbol{Y} = [Y_1, \ldots, Y_p]$, so that we can write $\boldsymbol{P} = [\boldsymbol{X}, \boldsymbol{Y}]$. States may not wish to release $\boldsymbol{P}$ to external researchers because of concerns over student privacy. However, rather than releasing $\boldsymbol{P}$, they could instead release $m > 1$ synthetic datasets $\boldsymbol{P}^{(i)} = [\boldsymbol{X}^{(i)}, \boldsymbol{Y}^{(i)}]$ for $i = 1, \ldots, m$. One reason to release *multiple* synthetic datasets is that this allows for more accurate computation of standard errors. It should be noted that states could also release random samples from each synthetic dataset $\boldsymbol{P}^{(i)}$. In either case, the process of creating and releasing synthetic data involves generating synthetic $\boldsymbol{P}^{(i)}$.

The $\boldsymbol{P}^{(i)}$ can be either fully or partially synthetic. Partially synthetic data may be used when the variables in $\boldsymbol{X}$ do not contain any personally identifiable or sensitive information. If there are no privacy constraints on releasing $\boldsymbol{X}$, then there may be no need to create synthetic $\boldsymbol{X}^{(i)}$, and hence the synthetic data can be written as $\boldsymbol{P}^{(i)} = [\boldsymbol{X}, \boldsymbol{Y}^{(i)}]$. We say these datasets are *partially synthetic* since the true values of $\boldsymbol{X}$ are released in each dataset. Conversely, with fully synthetic data, different synthetic $\boldsymbol{X}^{(i)}$ are released with each dataset $\boldsymbol{P}^{(i)} = [\boldsymbol{X}^{(i)}, \boldsymbol{Y}^{(i)}]$, which may be more appropriate if there are concerns regarding student privacy and the variables contained in $\boldsymbol{X}$. In this article, we will focus on partially synthetic data, where school and district-level indicators are preserved. This can be thought of, loosely, as taking existing schools and populating them randomly with students whose data look similar to the students that actually attend those schools.

The synthetic $\boldsymbol{P}^{(i)} = [\boldsymbol{X}^{(i)}, \boldsymbol{Y}^{(i)}]$ are generated using a similar approach as multiple imputation. For each $i = 1, \ldots, m$, $\boldsymbol{P}^{(i)}$ is drawn from the posterior predictive distribution $f(\boldsymbol{P}^{(i)} \mid \boldsymbol{P})$. However, computing the full multivariate distribution across all variables in $\boldsymbol{P}$ and then drawing $\boldsymbol{P}^{(i)}$ can be difficult, and it is often simpler to do this via a series of conditional probability models (Reiter & Raghunathan, 2007). For a given variable Y_j, this works by modelling the distribution $Y_j^{(i)} \mid \boldsymbol{X}, \boldsymbol{Y}_{(j)}$, where $\boldsymbol{Y}_{(j)}$ is the data in $\boldsymbol{Y}$ excluding Y_j. The posterior predictive distribution $f(Y_j^{(i)} \mid \boldsymbol{X}, \boldsymbol{Y})$ is computed for this model and the values of Y_j in the data are replaced with draws from $f(Y_j^{(i)} \mid \boldsymbol{X}, \boldsymbol{Y})$. This process cycles through each variable in $\boldsymbol{P}$ multiple times, and under certain conditions is equivalent to drawing $\boldsymbol{P}^{(i)}$ from $f(\boldsymbol{P}^{(i)} \mid \boldsymbol{P})$ (Liu et al., 2014).

There are many considerations that go into specifying these conditional probability models that can affect the utility and security of synthetic data. An important consideration is the model used for each variable. One could use standard parametric models for $Y_j \mid \boldsymbol{Y}_{(j)}, \boldsymbol{X}$, including normal linear regression, Poisson regression, or logistic regression, depending on the nature of Y_j (Reiter & Raghunathan, 2007). Particularly if linear models are used, one will need to specify in advance any interactions or basis expansion terms (e.g., polynomial terms). Conversely, models for $Y_j \mid \boldsymbol{X}, \boldsymbol{Y}_{(j)}$ can be semiparametric, such as classification and regression tree (CART)-based methods (Little, Liu, & Raghunathan, 2004).

Finally, imputation/synthesis models can incorporate auxiliary variables $\boldsymbol{Z}$, where $\boldsymbol{Z}$ will neither be synthesized nor released, but rather can be used to improve the accuracy of the posterior predictive distributions (see Hsu, 2007). For instance, if a state has internal evaluations of schools that it will not make public, it can still use that information in generating synthetic data.

2.2. HANDLING MISSING DATA

Since synthetic data methods are built around the same theory as multiple imputation, they can naturally handle situations where data are missing from $\boldsymbol{P}$ (Reiter, 2004). Reiter (2005a) argues that multiple imputation can be done prior to synthesis, which would mean that states would be in charge of imputing missing values. In this framework, states would use standard multiple imputation methods to generate n imputed datasets $\boldsymbol{P}_1$, …, $\boldsymbol{P}_\mathrm{n}$ based on $\boldsymbol{P}$. Then, m_n synthetic datasets $\boldsymbol{P}_\mathrm{k}^{(\mathrm{i})}$ ($i = 1$, …, m_n) are generated for each imputed $\boldsymbol{P}_\mathrm{k}$. For example, if a SLDS wanted $m = 50$ synthetic datasets but the original dataset $\boldsymbol{P}$ contains missing data, then they would create $n = 5$ imputed datasets and then generate $m_n = 10$ synthetic datasets from each imputed dataset.

It is worth noting that the choice of imputation model can affect what the imputed data $\boldsymbol{P}_\mathrm{k}$ look like, and hence the validity of inferences based on analyses of the synthetic data $\boldsymbol{P}_\mathrm{k}^{(\mathrm{i})}$. While this is true for any analysis involving missing data, in this context, analysts do not necessarily have direct control over the manner of imputation. Instead, Reiter (2005a) sees this as part of the workflow of the state agencies. Put another way, in the process of imputing missing data, states are inherently making assumptions about how and why it is missing, which can impact later analyses if these assumptions are not met, or if subsequent analytical models are not congenial with the imputation model.

While this article examines the imputation-synthesis framework, this is not the only way to handle missing data. One option would be to generate missing values in synthetic data. This would also be appropriate for scenarios where the missingness is of interest in and of itself, such as in studies of students who opt out of standardized tests. Alternatively, if missing data is merely a nuisance, then generating synthetic data with missingness would give the analyst greater control over the methods used to handle it. For instance, given synthetic data $\boldsymbol{P}^{(1)}$, …, $\boldsymbol{P}^{(\mathrm{m})}$ each with missing values, an analyst could multiply impute complete datasets $\boldsymbol{P}_\mathrm{k}^{(\mathrm{i})}$. They could even use more complex imputation methods that rely on fewer assumptions.

2.3. ANALYZING SYNTHETIC DATA

Suppose an external researcher, which we call the *analyst*, wishes to estimate some quantity Q from the synthetic data. The general approach to doing so is to pool estimates from each synthetic dataset $\boldsymbol{P}^{(i)}$. Let q_i be the estimate of Q based on dataset $\boldsymbol{P}^{(i)}$, and let v_i be the variance of that estimate. Under certain conditions (see Raghunathan, 2003), valid inferences regarding Q can be obtained with the following quantities:

$$\bar{q} = \sum_{i=1}^{m} \frac{q_i}{m}, \qquad B = \sum_{i=1}^{m} \frac{(q_i - \bar{q})^2}{m-1}, \qquad \bar{v} = \sum_{i=1}^{m} \frac{v_i}{m}$$

Here, $\bar{q}$ is just the mean of the q_i, $\bar{v}$ is the mean of the v_i, and B is the variance between the q_i. The pooled estimate $\bar{q}$ is used as an estimate for Q, and its variance can be estimated by:

$$V_q = \left(1 + \frac{1}{m}\right) B - \bar{v}$$

While it is possible for $V_q < 0$, this is less common when m is large. However, Raghunathan et al. (2003) propose a more complicated variance estimator that will always be positive.

It is worth noting at this point that the variance V_q depends on a term $(1 + 1/m)\underline{B}$ that decreases as m increases. Thus, releasing more synthetic datasets can lead to more precise estimates. However, the following sections demonstrate that generating additional datasets is more computationally expensive. Thus, states must determine how large an m yields adequately precise estimates without overburdening their computational resources.

3. EMPIRICAL INVESTIGATIONS: KERNEL DATASETS AND METHODS

While the previous sections described the theory behind synthetic data disclosure, it is worth asking how feasible implementations of this theory would be for SLDS. To that end, we attempted to create synthetic datasets using actual SLDS data from different states. Throughout this process, we investigated resources that might be required by states to construct synthetic data, software that is publicly available, and potential bottlenecks in this process. In this section, we describe the methods we used, and discuss the implications and limitations of those methods.

3.1. KERNEL DATASETS

As part of this investigation, we obtained data from eight states. Some of these datasets were longitudinal, while others were large cross sections. To focus our efforts, we created two cross sections of data for each state: one of fourth graders (for a given year) and one of eighth graders (for the same year). We refer to these as the *kernel datasets*. In total, these kernel datasets contained over 960,000 distinct students. Since the data furnished covered different time frames, the years of the cross sections vary by state. The resulting cross sections range from 2009 to 2012, and contain basic demographic information, including a student's school, race, gender, if they receive free or reduced-priced lunch (FRL), and whether they have limited English proficiency (LEP). They also include state achievement test scores in various subjects, though all datasets (across states and grades) have scores for math and reading. All of the kernel datasets involved some missing fields.

Across the 16 kernel datasets, we encounter a range of possible data collection, analysis, and disclosure issues. The states range from small to large, diverse to relatively homogenous, and vary in the share of urban and rural schools. Moreover, each state collects slightly different demographic information. This involves states tracking different variables; for instance, some states document student homelessness or Native Americans' tribal affiliations. Alternatively, some states have different types of designations for the same variables; for example some states indicate whether a student is proficient in English, while others describe that proficiency from low to high. Overall, we feel that these kernel datasets provide a suitable test of the pipeline we used to generate synthetic data.

4. PIPELINE: FROM REAL TO SYNTHETIC DATA

Our general approach was to impute missing values prior to generating synthetic data as discussed in a previous section. Thus, for each kernel dataset, we used multiple imputation to construct five (5) complete datasets. Then, for each imputed dataset, we generated four (4) synthetic datasets for a total of $m = 20$ datasets per kernel. In theory, it would be possible for states to release all 20 datasets, release random samples from them, or release random samples from a subset of them. However, we chose 20 datasets in order to improve the likelihood that relationships between variables are preserved and estimated well.

To study different possible configurations of this pipeline, we used two different approaches to imputing and synthesizing data: one that imputes and synthesizes data with a series of *parametric* models and one that uses a series *semiparametric* classification and regression tree (CART) models. For the parametric approach, both imputation and synthesis rely on a linear model to predict the values of a given variable. A standard OLS was used for continuous variables, and multinomial logistic regression was used for categorical variables. These models use every variable (including school and district) and some pre-specified interactions at each synthesis step. Draws from the posterior predictive distribution are generated from these models.

Conversely, with the semiparametric approach, we used CARTs to predict a given variable using every other variable (except school and district). Note that one need not specify a functional form or interactions for CARTs. Moreover, draws from the posterior predictive distribution from CARTs are taken via Bayesian bootstrap samples. Unlike the parametric approach where values are generated at random according to a given parametric distribution, the Bayesian bootstrap involves re-sampling actual values in the data, and then choosing one at random to fill in a specific cell (see Kim, 2002; Siddique et al., 2008).

A desirable feature of imputation and synthesis models is that they ought to be at least as general as the analytic models used, often referred to as congeniality (see Meng, 1994). Thus, any relationships between variables, including interaction terms in analytic models that one might want preserved in the synthetic data should be included in the imputation/synthesis models. For instance, if we want to preserve the relationship between math achievement for gender and FRL status (i.e., for males who receive FRL), then an interaction between gender and FRL should be in the model used to synthesize math scores. This is done explicitly when specifying models for the parametric approach. However, for the CART models, important interactions are not modelled explicitly, but rather are discerned from the data.

In total across two cross sections for eight states (16 kernel datasets) and two imputation/synthesis models (parametric vs. semiparametric), we wound up with 640 synthetic datasets. Table 1 shows the parametric models used for both imputation and synthesis.

Table 1: Models used for imputation/synthesis

Variable Type	**Variables**	**Parametric Model**	**Semiparametric Model**
Continuous	Test scores for reading, math, etc.	Normal linear regression	Regression tree
Binary	Gender, migrant, gifted, LEP (for some states), FRL (for some states)	Logistic regression	Binary classification tree
Categorical with multiple classes	Race, LEP (for some states), FRL (for some states)	Multinomial logistic regression	Multiclass classification tree

5. Reflections on synthetic data pipeline

Our general pipeline follows directly from the literature on synthetic data disclosure (e.g., Raghunathan et al., 2003; Reiter, 2005b; Reiter & Raghunathan, 2007). We chose to use *m* = 20 synthetic datasets to help ensure quantities of interest are estimated as precisely as possible without overburdening computation. Since the process of imputing and then synthesizing data has been studied and implemented repeatedly in the literature, we consider this to be an empirical test not just of the theory, but also the feasibility for states. It is worth noting that while Reiter (2004, 2005a,b) points out that enhanced privacy can be attained by releasing random samples of synthetic data, we did not take this step in our investigation. While this is certainly something states could do, we opted not to for a few reasons. First, a random sample of synthetic records will contain less information than the full synthetic dataset itself. Since we were interested in the utility of synthetic data, we felt that using the full dataset, rather than a sample, would provide a clearer picture of that. Moreover, precisely how the random sample should be generated is not a settled matter, particularly in education.

An important implication of this pipeline is that using the same general approach for each kernel dataset requires a replicable and streamlined workflow. The data we obtained varied in quality and content, and it is entirely possible to tailor a synthetic data pipeline to each individual dataset. However, we attempted to minimize such adaptations for a few reasons. First, we felt that demonstrating that a procedure is repeatable and does not require substantial adaptation from state to state had some value. In particular, we focused on building a reproducible pipeline using free, open-source software (see the following section), which would limit the cost to states should they use synthetic data methods in the future. Moreover, standardizing this procedure allowed us to identify general bottlenecks, but also investigate how limitations differed across states and grades.

5.1. SOFTWARE

We wanted to use software that may be openly or freely available to states to minimize the additional cost of creating synthetic data. To that end, we focused on two potentially useful libraries. The first is SRCware, a stand-alone version of the IVEware library from SAS, developed by the University of Michigan (https://www.src.isr.umich.edu/software/). This efficiently runs its own imputation-to-synthetic data pipeline based on the parametric models described in the previous sections.

We also used the 'mice' and 'synthpop' packages in the R programming language. At the time of our writing, there was no standalone R package that did both imputation and synthesis. Thus, we imputed datasets using 'mice', and then constructed synthetic datasets using 'synthpop' (Nowok et al., 2016). While both packages are designed to use a variety of models, we used them to implement CART-based methods.

5.2. COMPUTING REQUIREMENTS

An important barrier to using synthetic data methods involves technical capabilities. Between the two data synthesis methods and software tested (SRCware's impute/synthesize functionality and R's MICE imputation and synthpop data synthesis), there is a wide range of computational bottlenecks and concerns. For reference, the State 6 fourth-grade kernel dataset contained just over 75,000 students with 19 variables; SRCware was able to first create imputed datasets, and then 20 synthetic replicates of that dataset in 75 minutes. When running in parallel, the semiparametric methods took 42 minutes to run an equivalent imputation and synthesis in R.

All methods were implemented on a 24-core machine with 2400 MHz CPUs, which is considerably more powerful than a standard desktop computer. We only utilized five (5) of the 24 cores when exploring the methods since most desktop and laptop computers will not have considerably more than five cores.

There are several factors that will affect the time and resources required to create synthetic data. Two obviously important factors are the size of the original dataset and the types of data it contains. Large datasets will require greater computing time and memory. Moreover, if the data contain several categorical variables each with many different levels, this can increase the time required to fit imputation/synthesis models.

A very important determinant of computational demand and runtime is the specification of imputation/synthesis models. Both MICE and synthetic data methods use fully conditional models for each variable in the original dataset, so that a given column Y_{j} is being predicted based on other variables in the data. In our approach, we used all of the other variables in the data to train both the imputation and synthesis models. One may use more parsimonious models in exchange for quicker computation, however this can result in more variation in the synthetic data, and hence less precise or accurate analyses.

The class of imputation/synthesis models is also a key driver of computational complexity. For example, there is the option to fit random forest models for synthesis using the synthpop package in R. By default, the random forest model is an ensemble of 500 CARTs. This means that for each variable being synthesized with a random forest within each imputed dataset, the process would require 500 bootstrapped samples and 500 fitted CARTs, making this method wholly unscalable for a dataset with greater than about 50,000 students, as evidenced in the plots below:

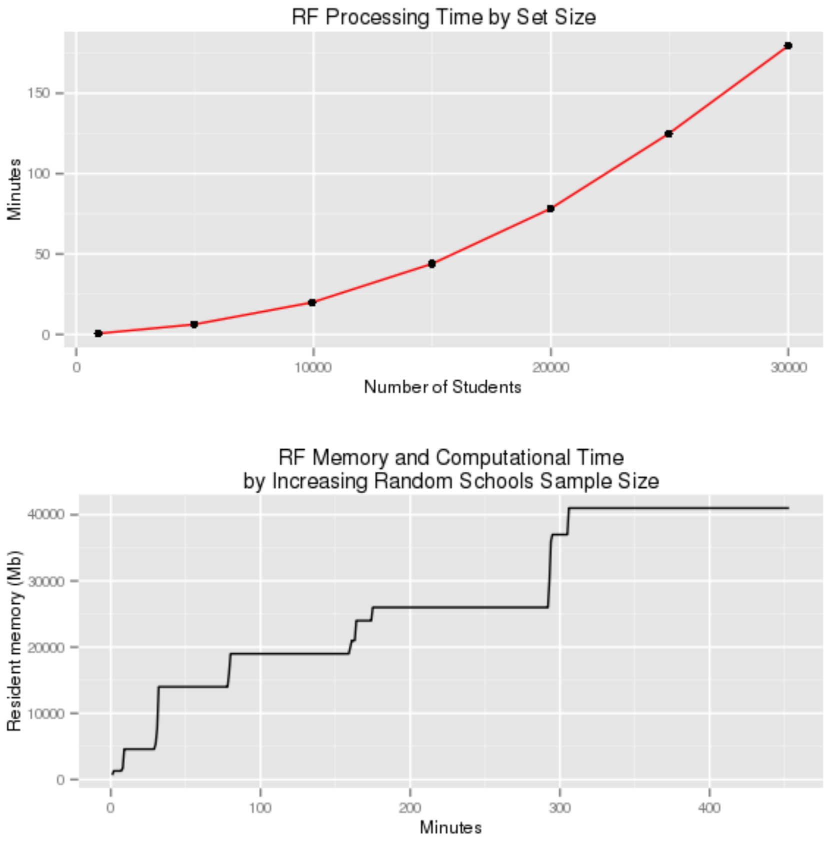


Figure 1: Computing Requirements of Imputation/Synthesis with Random Forests

Using random forests for imputation/synthesis required 40GB of resident memory (our servers have 120GB) to process 30,000 students, and computation time grew in a polynomial manner. This means that each additional record added to the dataset requires an even greater increase in computation time. Using random forests on all 75,000+ students and 20 features (including each student's school identifier) crashed our pipeline. Thus, a simpler semiparametric model based on a single CART (rather than an average of several) was chosen for its speed, low memory demand, and for a tree's tendency to "over-fit" the observed data. CARTs and parametric models, including those run in SRCware, typically consumed well under 10GB of resident memory during the imputation and synthesis stages. Put simply, the key tradeoff is that more complex models, like random forests, may require considerably more computational resources than states may be willing or able to dedicate, however they are likely to result in much better predictions and hence potentially more useful synthetic data.

Perhaps the largest factor driving time/resource demand is the sequential manner in which the parametric and non-parametric methods run, by default. Since we imputed five (5) datasets each of which were used to create four (4) synthetic datasets, we were able to easily run both the imputation and synthesis stages in parallel to realize significant gains in computation time. This worked by first giving each of the five (5) active cores a single copy of the original dataset, and having each core generate a single imputed dataset in parallel. Then, we distributed to the five cores each a single imputed dataset, and used each core to generate four synthetic datasets. For a state like State 6, which had over 75,000 fourth-graders, this process took a grand total of 42 minutes. In contrast, running our pipeline sequentially would impute datasets one at a time, and then synthesize datasets one at a time. For State 6, a single imputation in R took about 30 minutes, and synthesis an addition 3-5 minutes. For five imputed datasets and 20 synthetic datasets, this would take nearly four hours. Truly distributed computing that dispatches imputations and synthesis work across separate servers (e.g. by using the message passing interface, Apache Hadoop, etc.) would further allow states to speed up the imputation/synthesis pipeline even further by enabling parallelization not just across cores but across cores within separate servers.

5.3. DATA QUALITY

Another important finding regarding requirements involves the data we were furnished. The pipeline for generating synthetic data requires that a somewhat "clean" dataset be fed into the process. However, the quality of data and reporting systems tends to vary across states, which can be a very important factor in workflow. Some states have centralized and integrated systems that capture lots of information and store it in relational databases. The data furnished by such systems required very minor effort to develop kernel datasets. Conversely, some states furnished data that required considerable effort to produce something fit for the pipeline. In some cases, the relevant data on a student was distributed across multiple tables. Some of these tables were not designed to be relational, which meant that, for instance, students did not have unique identifying IDs, or their IDs differed across tables. In other words, just getting raw data into a form appropriate for synthetic data methods can be a stumbling block in and of itself; when raw data quality proved to be an issue, often greater time and effort was spent on data preparation than on synthesis.

6. UTILITY OF SYNTHETIC DATA

As an initial evaluation, we were interested in how well the synthetic datasets recaptured important statistics in the real data. This includes marginal means and variances, as well as conditional means and variances, which serve as a proxy for relationships between variables. Since the synthetic data were derived from imputed datasets, and there is no guarantee that the original data (with missing values) and the imputed data have the same moments, we used the imputed data as the benchmark. In other words, we consider the *true mean* (μ_{true}) of a variable as the pooled imputed means. Likewise, the *true variance* (σ_{true}^2) is the average of the variances within imputed datasets.

For continuous variables such as test scores, we calculated two statistics to check utility. First, we calculated a standardized mean difference

$$SMD = \frac{\mu_{\text{syn}} - \mu_{\text{true}}}{\sigma_{\text{true}}}$$

where μ_{syn} is the pooled synthetic mean and μ_{true} and σ_{true} are the true mean and square root of the true variance as described above. A SMD of zero would indicate a perfect recapturing of the true mean. A positive SMD would indicate that the synthetic mean is larger than the true mean while a negative SMD would indicate that the synthetic mean is smaller than the true mean. We also calculated the percent mean difference in the variance as:

$$PD = \frac{\sigma_{\text{syn}}^2 - \sigma_{\text{true}}^2}{\sigma_{\text{true}}^2}$$

where σ_{syn}^2 is the mean of variances within synthetic datasets. A PD of zero would indicate a perfect recapturing of the true variance. A positive PD would indicate the synthetic variance is larger than the true variance while a negative PD would indicate that the synthetic variance is smaller than the true variance.

For categorical variables (race/ethnicity, gender, and FRL, ELL, special education, and gifted status), we calculated the odds ratio by level. For example, for gender, we calculated the odds ratio of male students in the synthetic and imputed data, using

$$OR = \frac{p_{\text{syn}}/(1 - p_{\text{syn}})}{p_{\text{true}}/(1 - p_{\text{true}})}$$

where p_{syn} and p_{true} are the averaged fractions of male students across synthetic and imputed datasets, respectively (as with the means above). An OR of 1 would indicate that the synthetic proportion recaptures the true proportion. An OR greater than 1 would indicate that the synthetic proportion is greater than the true proportion while an OR less than 1 would indicate that the synthetic proportion is less than the true proportion.

We were also interested in how well the synthetic data preserved conditional distributions. Within each categorical variable level, we calculated the standardized mean difference of test scores (e.g., standardized mean difference for males or minority students), and examined the percent difference in the conditional variances.

Finally, since education data is naturally nested, we were interested in how well test scores were preserved within schools. To examine this, we compute the school-level intra-class correlation (ICC), which is the proportion of the variation in test scores attributable to differences between schools (see Raudenbush & Bryk, 2002). We report percent differences between the synthetic and "true" ICCs in the following sections. We also calculated standardized mean differences by school. Here, the true mean is the school mean averaged across imputed datasets, and the true variance is the average within-school-within-dataset variance. In order to obtain sensible results, we excluded schools and districts with five or fewer students when calculating standardized mean differences (but did not exclude them when computing ICCs).

6.1. MARGINAL DISTRIBUTIONS

Table 2 summarizes the similarity between statistics computed on the synthetic data versus the real data using the metrics described above. Differences in the mean, variance, and ICC are computed for each test, state, and grade, and this table reports summaries of those computed differences. For both math and reading achievement, Table 2 shows various quantiles of (1) standardized mean difference in test scores, (2) the percent difference of the total variation in test scores, and (3) the percent difference of the school-level intra-class correlations. Results

Table 2: Marginal distritubtions of synthetic data compared to real data

Utility of Synthetic Test Scores: Marginal Distributions

		Parametric Models (SRCware)					Semiparametric Models (synthpop)				
Variable	**Metric**	**Min**	**25%**	**Median**	**75%**	**Max**	**Min**	**25%**	**Median**	**75%**	**Max**
Math Achievement	Standardized mean difference	-0.10	0.00	0.00	0.00	0.15	-0.01	0.00	0.00	0.00	0.01
	% Difference in Variance	-22%	-3%	0%	1%	12%	-12%	-2%	-1%	0%	0%
	% Difference in School ICC	-45%	-20%	-11%	-4%	0%	-100%	-100%	-100%	-100%	-98%
Reading Achievement	Standardized mean difference	-0.05	-0.02	0.00	0.00	0.01	-0.01	0.00	0.00	0.00	0.01
	% Difference in Variance	-16%	-4%	-1%	1%	5%	-12%	-2%	-1%	-1%	1%
	% Difference in School ICC	-52%	-14%	-4%	0%	10%	-100%	-100%	-100%	-100%	-99%

Utility of Synthetic Categorical Variables: Odds Ratios

		Parametric Models (SRCware)					Semiparametric Models (synthpop)				
Variable	**Category**	**Min**	**25%**	**Median**	**75%**	**Max**	**Min**	**25%**	**Median**	**75%**	**Max**
Gender	Female	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.02
FRL	Yes	0.86	1.00	1.01	1.05	2.30	0.84	0.99	1.00	1.00	1.08
LEP	Yes	0.74	1.02	1.15	1.38	3.04	0.94	0.98	1.00	1.01	1.16
Race	White	0.66	0.92	0.95	0.99	1.00	0.97	1.00	1.00	1.01	1.13
	Hispanic	0.85	0.99	1.00	1.00	1.65	0.86	0.99	1.00	1.00	1.02
	Black	0.75	1.00	1.00	1.01	1.09	0.93	0.99	1.00	1.00	1.04
	Asian	0.64	1.00	1.02	1.50	27.47	0.12	0.99	1.00	1.01	1.10
	Other	1.01	1.03	1.11	1.40	1.72	0.98	1.00	1.01	1.02	1.03

are broken out by synthesis model: parametric versus semiparametric. We note that differences between these in the table are due to the models and not the software used.

What can be seen in the panels on test scores is that both methods tend to preserve the mean test score well, so that the standardized difference between the synthetic and “real” mean is typically smaller than a hundredth of a standard deviation. Both methods also preserve the variance of those distributions as well, so that the synthetic test score variance is within 0-3% of the real test score variance. Thus, the first two moments of the distribution of both test scores tended to be preserved by both synthesis methods.

Not only do the synthetic datasets recapture the mean and variance of test scores, often they retain entire distribution of test scores. Figure 2 shows the distribution of math scores in eighth grade for State 3. The distribution of the synthetic test scores is shown in red, the distribution of the real test scores is shown in blue; their overlap appears purple. In the left panel, we see the results of parametric synthesis models, which produce a reasonable approximation of the true distribution of math achievement, but do not properly cover the upper tail of the achievement distribution. However, in the right panel, the real and synthetic distributions are nearly identical. The difference in these plots may arise from two factors related to the tendency of CARTs to overfit the data. The first is that CART models are more flexible, and hence have lower bias, as discussed in the statistical learning literature (see Hastie et al., 2009). Moreover, because synthetic values generated by CART are from bootstrap samples, the synthetic data contains “real” values that have been swapped between records. Thus, it is not entirely surprising that the distribution of scores is so similar for the semiparametric synthesis approach.

However, both methods tend to understate school-level ICCs. The parametric models tend to produce ICCs that are within 10% of the real ICCs, but the semiparametric models

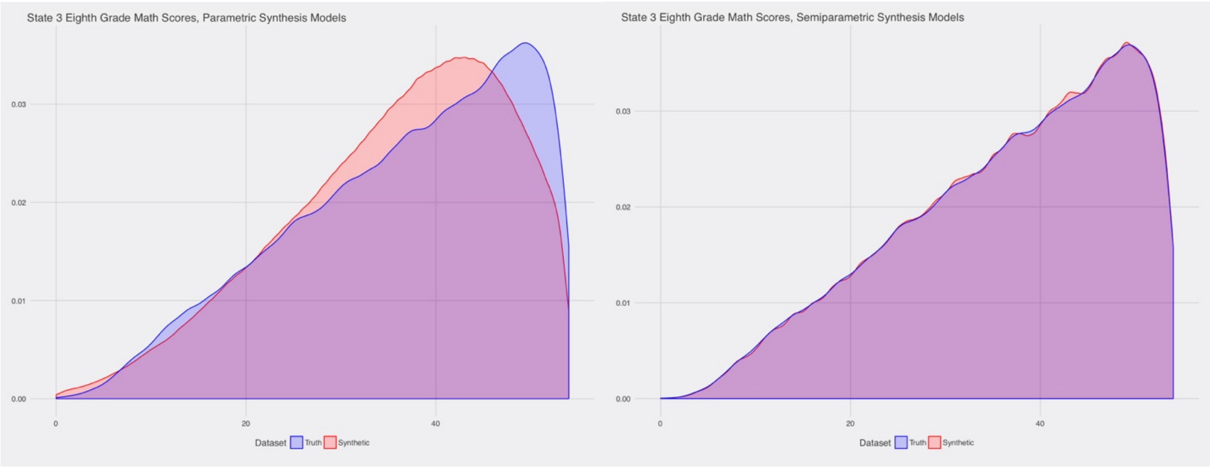
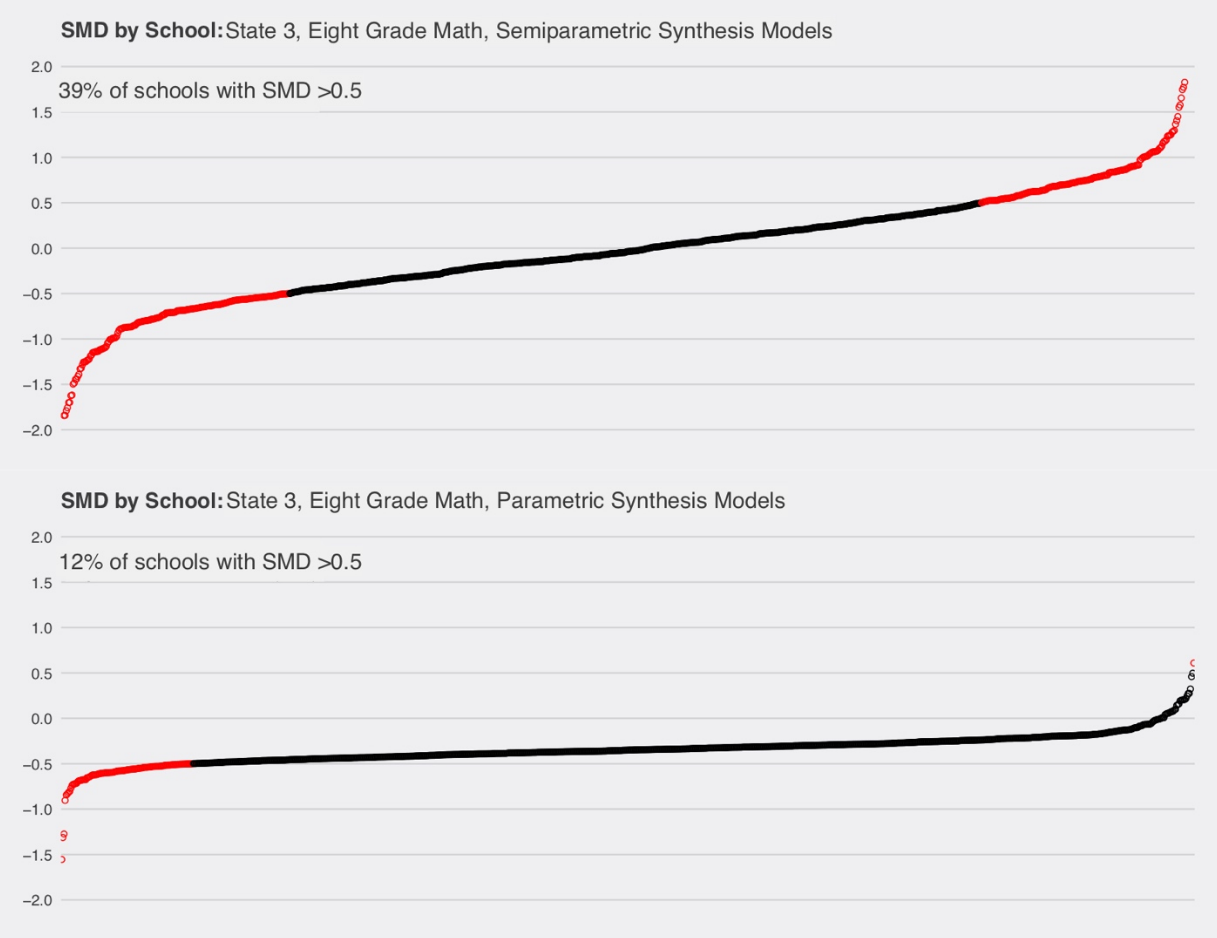


Figure 2: Marginal Distributions of Math Scores. This plot shows the marginal distribution of math test scores for synthetic data (red) and real data (blue) for eighth graders in State 3.

produce ICCs that are considerably smaller. To understand why this is, note that preserving the ICC in synthetic data requires preserving school mean achievement scores in the synthetic data. Since the parametric models explicitly model test scores as a function of school for both imputation and synthesis, this means that school mean test scores are likely to be captured in the synthetic data. Conversely, for CARTs, schools will only be used if they greatly improve predictions (e.g., of test scores). Thus, schools do not always play a substantial role in CART models, and hence school mean test scores in the synthetic data may differ substantially from the true school mean test scores. Further research into using hierarchical linear models as the synthesizer method might help further capture the nested variance structure than relying on a linear or CART model alone.

We can get a clearer sense of this by looking at Figure 3, which shows the standardized mean difference between each school's math achievement in the real versus synthetic data for eighth graders in State 3. Schools are ordered from smallest to largest standardized mean difference. In the top panel of the figure, we see the results from the semiparametric models, which exhibit substantial differences in mean math scores for schools. However, in the bottom panel, which shows the results from the parametric models, the differences are considerably smaller.


SMD by School: State 3, Eight Grade Math, Semiparametric Synthesis Models

39% of schools with SMD >0.5

SMD by School: State 3, Eight Grade Math, Parametric Synthesis Models

12% of schools with SMD >0.5

Figure 3: School Mean Achievement. This plot shows the standardized mean difference in math achievement for schools in State 3.

The bottom panel of Table 2 shows that the demographic composition of the synthetic data is quite similar that of the real data. For each demographic category, the table shows the odds ratio of proportions computed on the synthetic data versus the real data. The odds ratios for gender are very close to 1.0, which indicates that there are about the same proportion of females in the real and synthetic datasets. For the other demographic categories most datasets had odds ratios near 1.0 as well, including FRL and LEP status, as well as racial subgroups.

However, not all odds ratios in that panel are small. Indeed, when datasets contain only a few students that have a given demographic profile, models used to predict the proportion of such students become less accurate. As a result, there may be more or fewer such students in the synthetic data than in the real data. This can be seen in the bottom panel of Table 2, which has two large maximum odds ratios, one for LEP status, and one for the proportion of Asian students. Both of these odds ratios arise from synthetic data generated for State 5, which is smaller and more homogenous; there are very few Asian students in State 5, and even fewer students designated as LEP.

6.2. CONDITIONAL DISTRIBUTIONS

An important test of synthetic data is not just whether it preserves marginal distributions, but also whether it preserves relationships between variables. Here, we check relationships by examining conditional distributions of test scores within demographic subgroups. As in Table 2, we examine the standardized mean difference in test scores, as well as the percent difference in variance. The results for racial subgroups are shown in Table 3, which summarizes differences between the real and synthetic means and variances within each racial subgroup across all 16 kernel datasets. For instance, the median standardized mean difference for white students' math scores is 0.00.

What can be seen in Table 3 is that for most states, the conditional means and variances of achievement scores are preserved quite well in the synthetic data. Most differences in mean test scores between synthetic and real data are less than a few hundredths of a standard deviation for nearly all racial subgroups. However, that is not true for all datasets and racial subgroups. In particular, there are particularly large differences for Asian test scores (SMD = -1.47 for math, and -0.94 for reading). These arise in State 5, which had only a few Asian students, and hence the models used to predict their test scores were considerably less accura

Table 3: Utility of Synthetic Data: Conditional Distributions of Test Scores by Race

Math Achievement Scores

		Parametric Models (SRCware)					Semiparametric Models (synthpop)				
Subgroup	**Metric**	**Min**	**25%**	**Median**	**75%**	**Max**	**Min**	**25%**	**Median**	**75%**	**Max**
White	Standardized Mean Difference	-0.16	0.00	0.00	0.03	0.17	-0.02	-0.01	-0.01	0.00	0.01
	% Difference in Variance	-23%	-8%	-2%	5%	47%	-18%	0%	0%	1%	2%
Hispanic	Standardized Mean Difference	-0.04	-0.01	0.00	0.03	0.08	-0.13	-0.01	0.01	0.02	0.06
	% Difference in Variance	-29%	-4%	-1%	5%	36%	-10%	-1%	1%	4%	34%
Black	Standardized Mean Difference	-0.04	0.00	0.01	0.09	0.15	-0.01	0.02	0.06	0.08	0.14
	% Difference in Variance	-65%	-11%	2%	11%	29%	-10%	-2%	2%	7%	27%
Asian	Standardized Mean Difference	-0.94	-0.11	-0.01	0.04	0.36	-0.32	-0.16	-0.10	-0.04	0.46
	% Difference in Variance	-30%	-15%	-6%	4%	240%	-72%	-15%	-2%	6%	29%
Other	Standardized Mean Difference	-0.27	-0.07	-0.03	0.00	0.06	0.00	0.00	0.02	0.03	0.10
	% Difference in Variance	-19%	-5%	-2%	9%	59%	-17%	-2%	1%	3%	15%

Reading Achievement Scores

Subgroup	**Metric**	**Min**	**25%**	**Median**	**75%**	**Max**	**Min**	**25%**	**Median**	**75%**	**Max**
White	Standardized Mean Difference	-0.05	-0.01	0.00	0.01	0.06	-0.03	-0.01	0.00	0.00	0.01
	% Difference in Variance	-19%	-5%	-2%	9%	59%	-17%	-2%	1%	3%	15%
Hispanic	Standardized Mean Difference	-0.06	-0.01	0.00	0.01	0.05	-0.11	-0.01	0.01	0.02	0.07
	% Difference in Variance	-28%	-5%	-1%	5%	36%	-10%	-1%	1%	3%	41%
Black	Standardized Mean Difference	-0.06	-0.01	0.01	0.02	0.07	0.00	0.02	0.03	0.05	0.07
	% Difference in Variance	-62%	-10%	-2%	6%	35%	-8%	-3%	2%	5%	22%
Asian	Standardized Mean Difference	-1.47	-0.09	-0.01	0.05	0.68	-0.20	-0.07	-0.03	0.06	0.20
	% Difference in Variance	-33%	-19%	-6%	8%	185%	-25%	-2%	0%	8%	13%
Other	Standardized Mean Difference	-0.11	-0.06	-0.05	-0.01	0.04	-0.04	-0.02	-0.01	0.02	0.07
	% Difference in Variance	-13%	-9%	-3%	-1%	45%	-24%	-4%	0%	1%	19%

Table 4: Utility of Synthetic Data: Conditional Distributions of Test Scores by FRL and Gender

Conditional Distribution of Achievement Scores by FRL Status

			Parametric Models (SRCware)					**Semiparametric Models (synthpop)**				
Outcome	**FRL**	**Metric**	**Min**	**25%**	**50%**	**75%**	**Max**	**Min**	**25%**	**50%**	**75%**	**Max**
Math Achievement	No	Standardized Mean Difference	-0.16	-0.01	0.00	0.01	0.08	-0.03	-0.02	-0.02	-0.01	0.03
		% Difference in Variance	-18%	-5%	-1%	9%	91%	-22%	-1%	0%	2%	4%
	Yes	Standardized Mean Difference	-0.13	-0.01	0.00	0.02	0.60	-0.04	0.00	0.02	0.04	0.21
		% Difference in Variance	-49%	-6%	-2%	10%	61%	-19%	-1%	1%	5%	44%
Reading Achievement	No	Standardized Mean Difference	-0.04	-0.02	-0.01	0.00	0.02	-0.04	-0.02	-0.02	-0.01	0.03
		% Difference in Variance	-18%	-6%	-1%	3%	83%	-22%	-1%	0%	1%	3%
	Yes	Standardized Mean Difference	-0.08	-0.01	0.00	0.02	0.47	-0.03	-0.01	0.02	0.04	0.25
		% Difference in Variance	-47%	-3%	1%	11%	52%	-8%	0%	2%	6%	33%

Conditional Distribution of Achievement Scores by Gender

Outcome	**Gender**	**Metric**	**Min**	**25%**	**50%**	**75%**	**Max**	**Min**	**25%**	**50%**	**75%**	**Max**
Math Achievement	Female	Standardized Mean Difference	-0.14	-0.01	0.00	0.00	0.15	-0.01	0.00	0.01	0.02	0.05
		% Difference in Variance	-21%	-4%	2%	6%	16%	-12%	-2%	-1%	2%	4%
	Male	Standardized Mean Difference	-0.03	-0.01	0.00	0.00	0.01	-0.01	0.01	0.01	0.01	0.04
		% Difference in Variance	-23%	-4%	-3%	1%	10%	-11%	-3%	-1%	-1%	0%
Reading Achievement	Female	Standardized Mean Difference	-0.07	-0.03	0.00	0.00	0.00	-0.05	-0.02	-0.01	0.00	0.01
		% Difference in Variance	-18%	-7%	0%	4%	12%	-13%	-1%	0%	1%	4%
	Male	Standardized Mean Difference	-0.07	-0.01	0.00	0.01	0.15	-0.03	-0.02	-0.01	-0.01	0.00
		% Difference in Variance	-14%	-5%	-2%	0%	8%	-11%	-3%	-1%	-1%	5%

Similar results can be seen in Table 4, which shows standardized mean differences and percent differences in variance for test scores within subgroups defined by FRL status and by gender. Taken together, Tables 3 and 4 show that the mean and variance of the distribution of test scores within various demographic subgroups are largely preserved in the synthetic data. In general, the differences are somewhat smaller for the data generated with semiparametric models than with parametric models.

It is worth noting that not only are the moments (i.e., mean and variance) of various conditional distributions preserved in the synthetic data, often the entire conditional distribution of test scores is preserved. As an example, consider Figure 4, which takes math scores of eighth graders in State 3 and overlays the distribution of scores for each race in the real data (shaded in blue) and in the synthetic data (shaded in red). These distributions are nearly identical. Moreover, similar plots were produced for every kernel dataset and variable and nearly all of them show that synthesis often preserves many conditional distributions.

7. DISCLOSURE RISKS

Accurately assessing disclosure risk is particularly difficult for synthetic data because none of the records are "real." If an intruder were to obtain any of the synthetic records, they would only have randomly generated quantities, rather than real data. However, that does not mean that synthetic data pose no disclosure risk. Even though the data are generated randomly, there is a chance (particularly with CART-based methods) that records are perfectly (or almost perfectly) recreated. While an intruder may have no way of knowing whether a given record was identical to a real one, it may be of interest to states whether they might be releasing such records. Thus, in this section we address the extent to which this might have occurred in the synthetic data we generated.

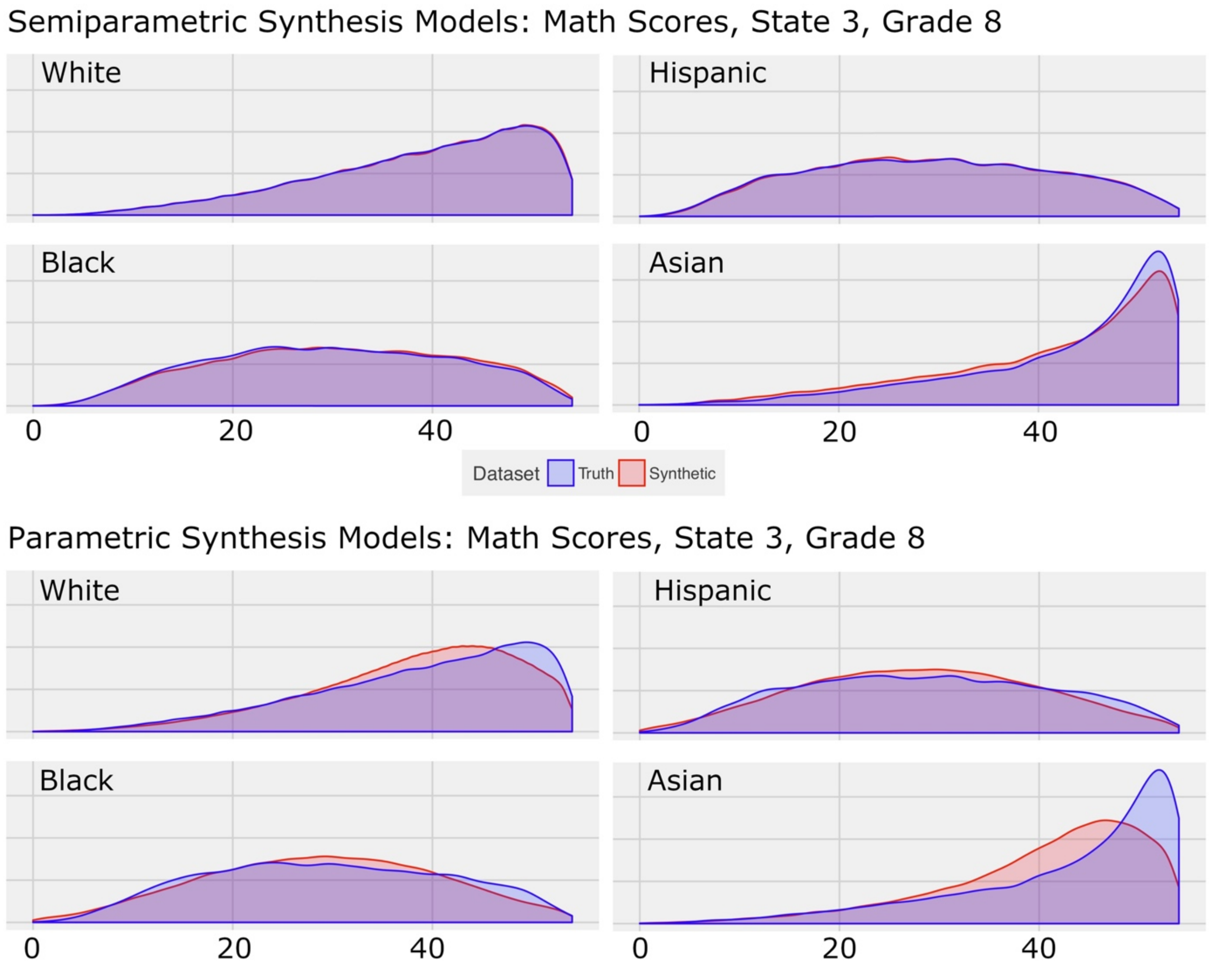


Figure 4: Conditional Distributions of Math Achievement. This plot shows the distribution of math achievement for synthetic data (red) and real data (blue) within racial subgroups of eighth graders in State 3.

Typically, federal agencies describe disclosure risk in terms of some sort of intruder attack for the data at hand (see Reiter, 2019). These attack scenarios often assume an intruder is attempting to identify a target individual in the dataset. The intruder may know some covariates about that individual (e.g., their race and gender) that are typically a subset of the variables in the dataset, and which are not considered particularly sensitive. They then attempt to match those covariates to records in the data. There may be many records that match, in which case, the intruder could have a difficult time discerning which record refers to their intended target.

However, if there are only a few matches, then the intruder will have a greater probability of figuring out which refers to their target (see Duncan & Lambert, 1986).

This logic can be seen as a basis for the type of microsuppression (i.e., small-cell suppression) that states already engage in. Under microsuppression, states divide students into risk strata based on certain demographic information, and strata with few students are deleted. While the exact procedures followed differ according to the data to be released, there appear to be some rules of thumb (see Johnson, 2007; Levesque, et. al., 2015). Often, strata are defined on school, grade, race, and gender (Massachusetts, 2014) and "few" typically means fewer than five or 10 (e.g., Oregon, 2016; Wisconsin, 2018). Given guidance in various data governance manuals (e.g., Montana, 2018; Nebraska, 2013), it would seem that states, at least legally, have accepted the risk to records *not* suppressed as reasonable.

Thus, in order for synthetic data to pose the same or greater risk, the following would need to be true about a synthetic record. First, it would need to be identical to a real record. Second, it would need to be in a small risk stratum. This is because if it were not, then states might live with releasing such data to researchers, and some have accepted the risk of releasing such data for public use (Massachusetts, 2014).

To get a sense of this, we took the real data and identified records in risk strata smaller than five, which we call "at-risk" records. Across the 16 kernel datasets, this totaled over 247,000 at-risk records. Then, we attempted to match these records in the synthetic data. Of the 16 sets of synthetic datasets (a total of 320 datasets), there were only 17 at-risk records that were matched. Most of these were from the datasets generated by CART-based methods: one in State 3 (fourth grade) and 15 in State 4 (5 in fourth grade, 10 in eighth grade). One additional match occurred among the datasets generated by parametric methods in State 4 (fourth grade). In sum, these matches were quite rare, though not nonexistent.

7.1. STRENGTHENING PRIVACY

This section highlights that it is possible for such methods to release records that states would otherwise suppress prior to releasing data. However, there are several steps that states can take in order to further reduce disclosure risks. The first, and most obvious, would be to conduct a similar risk analysis as the previous section and simply delete the offending records from the data.

A related approach would be to release subsamples of each synthetic dataset, rather than releasing each synthetic dataset in its entirety. Methods for doing so have been discussed by Reiter (2005b). The utility of such data may not match the promising results presented here, primarily because releasing less data means that quantities will be estimated with greater uncertainty. Reiter et al. (2009) and Reiter (2019) discuss systems of differential privacy wherein subsamples of synthetic data are released, researchers use those data to train models, and then submit their code to data administrators who can verify analytic results on the real data.

8. REFLECTIONS & CONCLUSIONS

In this article, we set out to generate and evaluate synthetic datasets with an eye on how feasible and replicable this process might be for SLDS. What we found is that with only modest tuning of model specifications, we obtained synthetic data that preserved many marginal and conditional distributions fairly well. Moreover, it tended to afford at least as much disclosure protection as microsuppression and could potentially offer much stronger protections depending the steps take to post-process the synthetic. Thus, from a utility and privacy standpoint, synthetic data methods offer a useful alternative the practice of microsuppression.

That said, the workflow established in this article was not always smooth, and certainly has its share of bottlenecks. First, it requires reasonably high-quality data to begin with. States that already have centralized data systems capable of generating "clean" datasets that require minimal pre-processing will be well-positioned to leverage such a pipeline.

States will also need some capacity to implement the methods we used. We relied on open source software, which means it costs nothing to use them. However, that does not mean they are free; in order to use them, SLDS would need personnel capable of coding the imputation and synthesis stages. This is important, since higher-quality synthetic data are possible through more accurate models, and the software used in this demonstration is amenable to custom modeling functions, not just linear models, random forests, and CARTs, that might be able to rectify observed differences in conditional distributions. However, using more complex models requires greater computational resources. With most kernel datasets, even our simple models required a server capable of distributing computations in order to generate synthetic datasets in a reasonable amount of time.

Finally, a key part of this pipeline involves checking utility and disclosure risks on the synthetic data. We consider this to be a key step in the pipeline, and an important evaluation that can ensure the synthetic data are useful without posing inordinate risks to student privacy. We have described several approaches states might take in order to check both utility and privacy, and consider those to be a useful starting point or baseline analysis.

There are additional limitations worth noting, both of synthetic data methods and how we implemented them. Perhaps one of the biggest limitations of synthetic data methods is that while they preserve relationships between variables within the observed data fairly well, they tend to shrink relationships between exogenous variables and variables in the data. For instance, suppose a state generates synthetic data in a manner similar to our pipeline (with similar datasets that include race and test scores), and a researcher then joins information from the U.S. Census Bureau (e.g., mean earnings, mean family size, etc.) to the synthetic datasets. Relationships between variables used to generate the synthetic data (e.g., race and test scores) should be preserved in the synthetic data, but relationships between variables in the Census data (e.g., mean earnings) and variables in the synthetic data (e.g., race or test scores) will be attenuated toward zero.

There are several limitations to this study. First, we focused on a replicable workflow pipeline and models simple enough to reduce computational demands. But, it is entirely possible that data quality could be affected by using more customized methods, including modifying the pipeline and employing more flexible imputation/synthesis models. Such steps could, in theory, improve the utility of synthetic data, but could also (depending on the models used) increase the time and resources needed to generate it, as well as potentially increase disclosure risks.

In addition, the type of datasets we generated may not reflect the many needs of SLDS and educational researchers. Data types may be more complex, and many analyses in education are longitudinal and follow students over multiple years. The approaches described here would need to be extended in various ways in order to ensure that temporal relationships between variables are maintained. Moreover, the types of models used, for instance, to model student achievement growth are more expansive and hence present additional challenges both in terms of model and computational complexity.

9. ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (award DGE-1437953) and the Institute for Education Sciences (awards R305B140042 and R305D140045). In addition, we would like to thank the state data systems that worked to furnish data for this project, and for the National Opinion Research Center (NORC) for providing the secure computational infrastructure for completing this work.

REFERENCES

CONAWAY, C., KEESLER, V., & SCHWARTZ, N. (2015). What research do state education agencies really need? The promise and limitations of state longitudinal data systems. *Educational Evaluation and Policy Analysis*, *37*(1S), 16S–28S.

DRECHSLER, J., BENDER, S., RÄSSLER, S. (2008). Comparing fully and partially synthetic datasets for statistical disclosure control in the german IAB establishment panel. *Trans. Data Privacy*, 1(3), 105–130.

DUNCAN, G. T. & LAMBERT, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association*, *81*, 10–28.

HASTIE, T., TIBSHIRANI, R., & FRIEDMAN, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction.* 2nd ed. New York: Springer.

HSU, C., TAYLOR, J. M., MURRAY, S. & COMMENGES, D. (2007). Multiple imputation for interval censored data with auxiliary variables. *Statistics in Medicine*, *26*, 769–781.

JOHNSON, C. (2007). Safeguarding against and responding to the breach of personally identifiable information. Memorandum for the Heads of Executive Agencies. U.S. Office of Management and Budget.

KIM, J. K. (2002). A note on approximate Bayesian bootstrap imputation. *Biometrika*, *89*(2), 470–477.

LEVESQUE, K., FITZGERALD, R., & PFEIFFER, J. (2015). A guide to using state longitudinal data for applied research. National Center for Education Evaluation and Regional Assistance Report # NCEE 2015–4013. U.S. Institute for Education Sciences.

LITTLE, R. J. A., LIU, F., & RAGHUNATHAN, T. E. (2005). Statistical disclosure techniques based on multiple imputation. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, edited by W. A. Shewhart, S. S. Wilks, A. Gelman, and X. Meng.

LIU, J., GELMAN, A., HILL, J., SU, Y.-S., & KROPKO, J. (2014). On the stationary distribution of iterative imputations. *Biometrika*, *101*, 155–173.

Massachusetts Department of Elementary and Secondary Education. (2014). Researcher's guide to Massachusetts state education data. Retrieved from http://sites.bu.edu/miccr/files/2015/12/Researcher-Guide-to-Massachusetts-State-Education-Data.pdf.

MENG, X-L. (1994). Multiple imputation inferences with uncongenial sources of input. *Statistical Science*, *9*(4), 538–558.

Montana Office of Public Instruction. (2018). Montana's consolidated state plan under the Every Student Succeeds Act. Retrieved from http://opi.mt.gov/Portals/182/Page Files/ESSA/Goodbye NCLB,Hello ESSA/Accessible ESSA Submission Jan 2018 Updated Date.pdf.

Nebraska Department of Education. (2013). Data access and use policy and procedures including research and evaluations. Retrieved from https://2x9dwr1yq1he1dw6623gg411-wpengine.netdna-ssl.com/wp-content/uploads/2017/07/Nebraska_Data_Access_and_Use_Policy_and_Procedures.pdf.

NOWOK, B., RAAB, G., & DIBBEN, C. (2016). synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software, 74*(11), 1–26.

North Carolina Department of Education. (2009). Data management group policy: Reporting on data in small cells or extremes. Retrieved from http://www.ncpublicschools.org/docs/data/management/policies/security/dmg-2009-004-se.pdf.

Oregon Department of Education. (2016). A summary to the Legislature of the annual report to the Legislature on English language learners 2014-2015 Oregon Department of Education. Retrieved from https://www.oregon.gov/ode/reports-and-data/LegReports/ Documents/ell-report-summary-1415-final.pdf.

PRADA, S. I., GONZALEZ, C., BORTON, J., ET AL. (2011). *Avoiding disclosure of individually identifiable health information: A literature review*. University Library of Munich, Germany.

RAGHUNATHAN T. E., REITER J. P., & RUBIN, D. B. (2003). Multiple imputation for statistical disclosure limitation. Journal of Official Statistics, 19(1), 1–16.

RAUDENBUSH, S. W., & BRYK, A. S. (2002). *Hierarchical Linear Models. Applications and Data Analysis Methods* (*2nd ed.*). Thousand Oaks, CA: Sage Publications.

REITER, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, *30*, 235–242.

REITER, J. P. (2005a). Estimating Risks of Identification Disclosure in Microdata. *Journal of the American Statistical Association, 100*(472), 1103–1112.

REITER, J. P. (2005b), Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. Journal of the Royal Statistical Society – Series A, 168, 185–205.

REITER, J. P. (2019). Differential privacy and federal data releases. *Annual Review of Statistics and Its Application*, 6(1), 85–101.

REITER, J. P., OGANIAN, A., & KARR, A. F. (2009). Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics & Data Analysis*, 53(4), 1475–1482.

REITER, J. P. & RAGHUNATHAN, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102, 1462–1471.

RUBIN, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, *9*, 462–468.

SIDDIQUE, J., & BELIN, T. R. (2008). Using an approximate bayesian bootstrap to multiply impute nonignorable missing data. *Computational Statistics & Data Analysis*, *53*(2), 405–415.

SINGH, A. C., YU, F., & DUNTEMAN, G. H. (2004). MASSC: A new data mask for limiting statistical information loss and disclosure. In *Work Session on Statistical Data Confidentiality 2003*, Linden, H., Riecan, J., & Belsby, L. (Eds.). Eurostat, Luxemburg. Monographs in Official Statistics, 373–394.

Wisconsin Department of Public Instruction. (2018). Student Privacy. Retrieved from https://dpi.wi.gov/assessment/student-privacy.