CREATING SYNTHETIC DATA FOR VIRGINIA LONGITUDINAL DATA SYSTEM

Sean Pili, Kyle Morgan, Ronnie Fesco, and Lata Kodali (Virginia Tech) with Aaron Schroeder (SDAL)
Sponsor: Tod Massa (SCHEV – State Council for Higher Education in Virginia)

Introduction

- Virginia Longitudinal Data System (VLDS) has sensitive information that cannot be released to the public.
- Third parties try accessing the data by guessing which variables they need, which leads to error.
- Machine learning (ML) methods can determine which variables are important, but third parties cannot access the whole data set

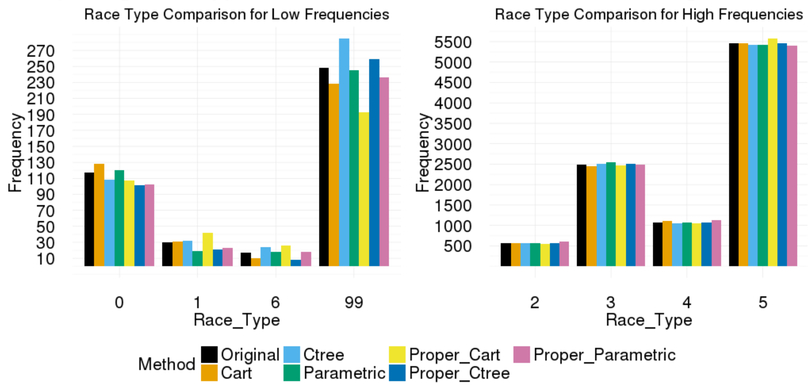
Research Goal: Create a synthetic dataset that is both usable for machine learning purposes and sufficiently privatized.

Methods for Synthetic Generation

The most widely used and modern techniques have now been condensed into an R package called Synthpop. We determine the best method for synthetic generation by comparing the following methods:

Methods:	Cart (Classification and Reg Trees)	Ctree (Conditional Inference Trees)	Parametric	Proper* Cart	Proper* Ctree	Proper* Parametric
Quick Description	Diagram based off decision splits	Reduces selection bias in Cart	Uses reg models	*Proper means joint distribution over all variables in dataset is stable.		

Comparing race type variable across all 6 methods to original:



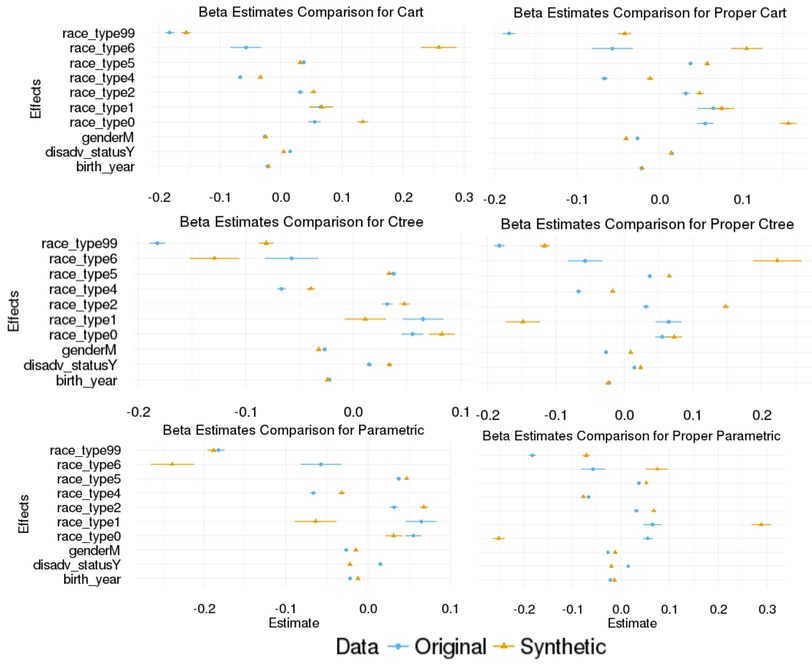
Race Type Comparison Plot: Each method has a trade-off between preserving the original frequencies of the variable race-type.

Usability Analysis with Regression Models

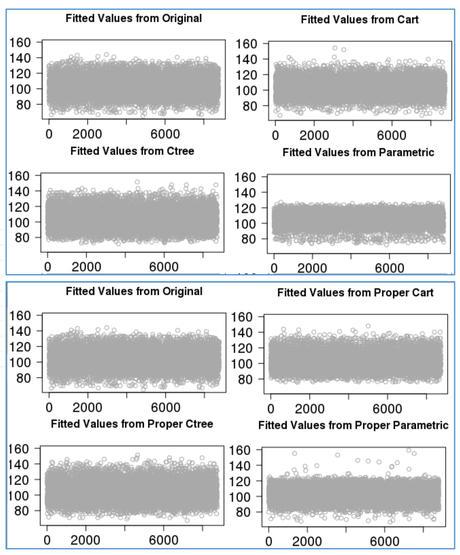
Question: Which relationships are preserved between original and synthetic datasets?

Regression Model Comparisons on Original vs Synthetic:

Response	Predictors	Model
aggregate days present	gender, race type, birth year, disadvantaged status	GLM with log link



GLM comparisons plot: Shows model effects for all chosen methods. Some effects are close between original and synthetic, but some are off.



Fitted values plot: Shows fitted values for all chosen methods. Methods such as Proper Cart and Proper Ctree seem to preserve structure of original fitted values.

Usability Analysis with Machine Learning

Question: Which clusters are preserved between original and synthetic datasets?

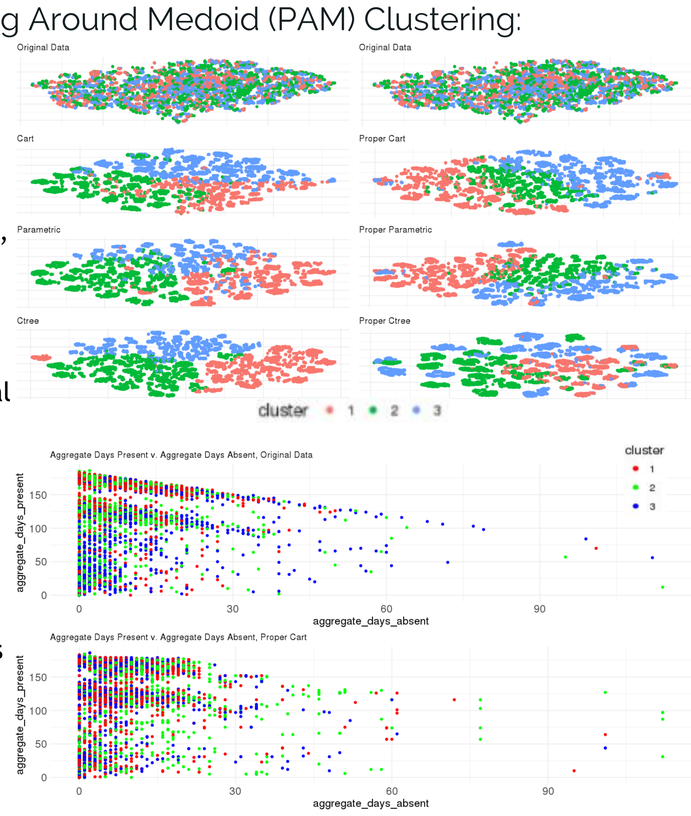
ML Comparison using Partitioning Around Medoid (PAM) Clustering:

PAM Clustering with Gower Distances finds natural splits in a dataset using character and numeric variables as split criteria.

Hypothesis: if the statistical properties of a dataset are preserved in a synthetic dataset, both datasets will exhibit similar clusters.

Fit with 3 clusters. **Plot1 (Top Right)** reduces the clusters to 2 dimensions. The Proper Ctree clusters best mirror the original clusters because their points are the most dispersed.

Plot2 (Bottom Right) plots the relationship between the days students did not attend school (points colored by cluster). Proper Cart seems to capture this relationship best as it assigns all but 1 student to two clusters when they are absent more than roughly 60 days. Additionally, all of the non-cart plots slope upward, which is not desirable.



Privacy Options

Goal: Determine how well the synthetic data protects sensitive information.

Method: Use the SDC function in R to detect how many times a row in the synthetic data exactly matches a row in the original data, and determine whether this happens infrequently enough to remove exact-copy rows when they happen. Repeat 100 times to ensure accuracy.

Plot 3 (Below): How many rows, out of 10,000, showed up unchanged in the synthetic data for each run of syn(). In this instance, the Proper Ctree method is used as an example.

Plot 4 (Right): How many times, in 100 runs, each unique row showed up unchanged in the synthetic data. In this instance, the Proper Ctree method is used as an example.

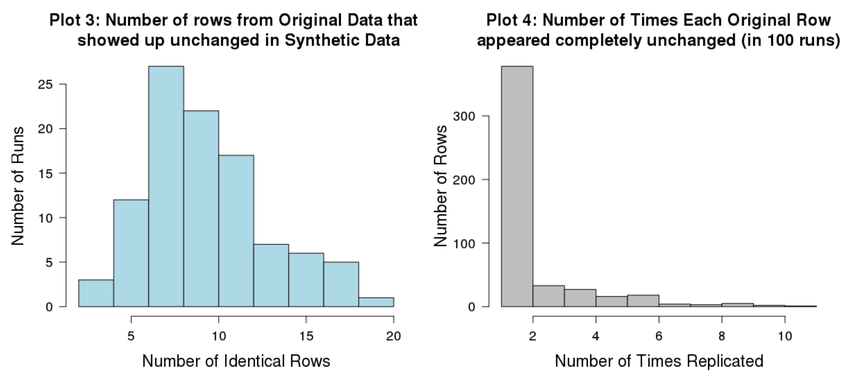


Table 1 (below): A comparison of how well each method protects private information

Methods:	Cart	Ctree	Parametric	Proper Cart	Proper Ctree	Proper Parametric
Average # of Synthetic Rows that exactly match original rows	48.02	8.7	1.8	46.99	9.78	1.42
Maximum # of times (out of 100) any one row exactly matched an original row	32	14	4	12	11	5

The first row shows the average number of rows that each method exactly copies from the original dataset (the idea is to minimize this number). The second row represents the maximum number of times that any individual row is exactly copied in 100 iterations, which indicates whether the method is specifically more likely to give away the information of individuals with certain characteristics.

Future Work

Summary: Different methods will preserve different features of the original data. Based on preliminary analysis, proper ctree is a method that works well.
For Machine Learning: Extend the number of clusters and re-examine their demographics.
For Data Privacy: Evaluate the effectiveness of top and bottom coding for protecting entries with unique attributes, and evaluate whether smoothing is more effective when dealing with numerical variables. Lastly, ensure anonymity for release to public.