Evaluation of synthetic population data created using generative adversarial networks

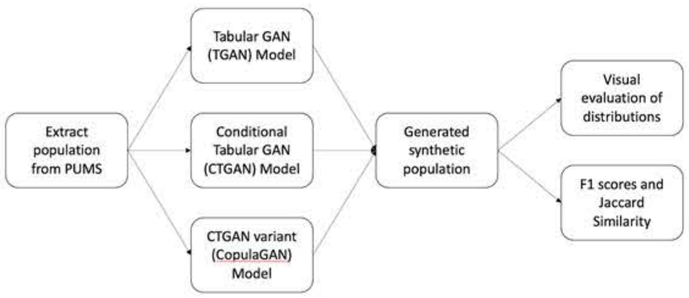
DAVID HAN, SRIHAN KOTNANA, Taylor Anderson (GGS), Andreas Züfle (GGS), Hamdi Kavak (CDS)
Geography and Geoinformation Sciences (GGS) and Computational and Data Sciences (CDS)

The 2021 Aspiring Scientists' Summer Internship Program

Introduction and Purpose

The generation of realistic synthetic populations is an important function for many agent-based models to provide accurate predictions. The problem with synthetic population data lies within the high dimensional data and irregular distributions. However, deep generative models have been proposed to tackle this issue because of their ability to model arbitrary distributions with greater flexibility. The objective of this study is to present a comparison and evaluation of synthetically generated populations from different generative adversarial network (GAN) models. We utilize the Public Use Microdata Sample (PUMS) of the Fairfax County, Virginia population to evaluate the performance of a tabular GAN, conditional tabular GAN (CTGAN), and CopulaGAN, a CTGAN variant.

Methods



Results

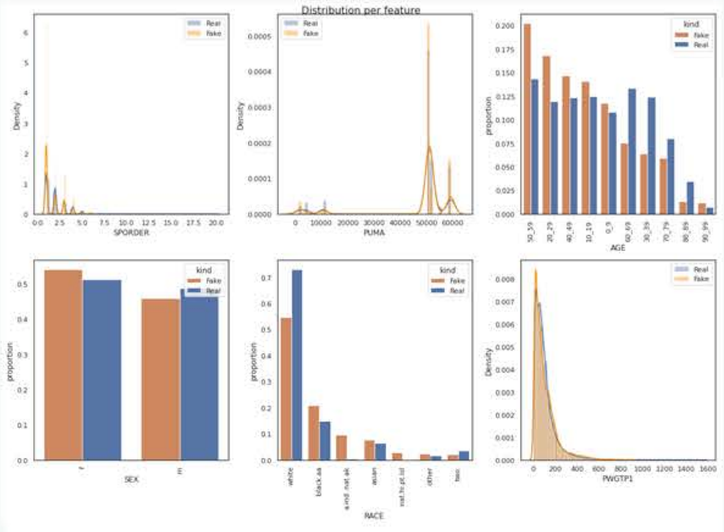


Figure 1

Jaccard Similarity	AGE	SEX	RACE	AVERAGE
CTGAN	0.03	0.37	0.62	0.34
CopulaGAN	0.06	0.30	0.64	0.33

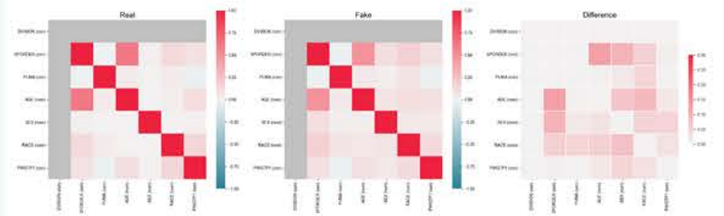


Figure 2

Discussion and Conclusions

Metrics from the TableEvaluator and SDV python libraries were used to measure correlations and probabilistic distributions of population attributes. We found that both the CTGAN and the CopulaGAN outperformed the tabular GAN, while the CTGAN narrowly outperformed the CopulaGAN's average similarity score by 2%. To compare the close models, we used various F1-scores including logistic regression, random forest classifiers, decision trees, and a multi-layer perceptron. The CTGAN also had an average Jaccard similarity of 0.34, a metric used to compute the closeness between the real and synthetic F1 scores for each category. Further analyses show that there is only a slight difference in correlation between synthetic and realistic attributes. Our research can be applied to other regions in the United States and can be used to accurately model populations when only a small sample of the population is available.

Major Citations

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling Tabular data using Conditional GAN. *Advances in Neural Information Processing Systems*, *32*, 7335-7345.
Xu, L., & Veeramachaneni, K. (2018). Synthesizing tabular data using generative adversarial networks.

Acknowledgements

This work is supported by National Science Foundation Grant #2109647 titled "Data-Driven Modeling to Improve Understanding of Human Behavior, Mobility, and Disease Spread" and the George Mason University Aspiring Scientists Summer Internship Program (ASSIP).