USING GENERATIVE ADVERSARIAL NETWORKS TO ASSIST SYNTHETIC POPULATION CREATION FOR SIMULATIONS

Srihan Kotnana

Westfield High School
4700 Stonecroft Blvd
Chantilly, VA, USA
srihank8@gmail.com

David Han

Computer Science and Mathematics
Cornell University
Ithaca, NY, USA
dmh338@cornell.edu

Taylor Anderson
Andreas Züfle

Geography and Geoinformation Science
George Mason University
Fairfax, VA, USA
{tander6,azufle}@gmail.com

Hamdi Kavak

Computational and Data Sciences
George Mason University
Fairfax, VA, USA
hkavak@gmail.com

ABSTRACT

Synthetic populations are heavily used in agent-based simulations and microsimulations to create realistic representations of real-world populations. Many existing techniques rely on duplicating or selecting a sample of disaggregated records captured via surveys to generate the entire synthetic population. The challenge here is the potential bias present in the sample of disaggregated records. This paper posits that such disaggregated records can be improved or replaced by training a generative adversarial network (GAN). We present a case study of a 1.1 million population using iterative proportional fitting (IPF). We illustrate that IPF makes a better fit using GAN-based disaggregated records rather than original census-based disaggregated records. Our results show a promising use of GANs for synthetic population generation.

Keywords: generative adversarial networks, synthetic populations, iterative proportional fitting, agent-based simulation, microsimulation.

1 INTRODUCTION

Computational models in many application domains, including transportation modeling (Moeckel et al. 2003, Huynh et al. 2013, Zhang et al. 2019, Le et al. 2016), computational epidemiology (Laskowski et al. 2012, Xu et al. 2017, Narassima et al. 2020, Bissett et al. 2016, Pesavento et al. 2020), and public policy models (Hafezi et al. 2019, He et al. 2020, Levy et al. 2014), require detailed individual-level representations of the entire population under study. However, in order to maintain the privacy of people, complete demographic data like the U.S. Decennial Census (Bureau 2021) or American Community Survey (ACS) (Bureau 2022) are often available at aggregated counts and percentages. In such cases, researchers

ANNSIM'22, July 18-20, 2022, San Diego, CA, USA;

can generate representative synthetic human populations — a detailed set of artificial persons that fall within some defined geographic area (e.g., country, state, etc.).

There are a vast number of techniques that can be used for the generation of synthetic populations (Jiang et al. 2021, Jiang et al. 2022). These techniques range from Iterative Proportional Fitting (IPF) (Beckman et al. 1996) to more advanced Iterative Proportional Updating (IPU) algorithms (Ye et al. 2009). These approaches typically extrapolate a small disaggregated sample dataset capturing real anonymized individuals' records with attributes such as age, sex, and race. This extrapolation continues until the synthetically generated population is close enough to the aggregated data. However, the challenge here is that the disaggregated sample for certain geographic regions is too small and does not always sufficiently cover minority populations. For instance, the average sample size of the Public Use Microdata Sample (PUMS) dataset from the U.S. Census Bureau (Bureau 2021) is just around 1%.

Therefore, this paper proposes the use of Generative Adversarial Networks (GANs) to replace or enhance the disaggregated sample. GANs are a class of machine learning architectures that integrate deep learning into generative modeling tasks (Goodfellow et al. 2014). The main idea here is that GANs can capture the underlying distribution of the disaggregated sample and generate more representative and potentially diverse samples in existing synthetic population generation algorithms. To our best knowledge, no previous studies have attempted to replace or enhance existing algorithms using GANs. Deep generative modeling (i.e., VAE) was only used as an alternative population synthesis algorithm (Borysov et al. 2019).

The rest of the paper is organized as follows. In Section 2, we summarize the literature regarding synthetic populations and GANs. Our approach is summarized in Section 3, while we present our use case of Fairfax County, Virginia, in Section 4, including datasets and results. Finally, we conclude the paper by highlighting our contributions, some limitations, and future work.

2 LITERATURE REVIEW

2.1 Synthetic Population Generators

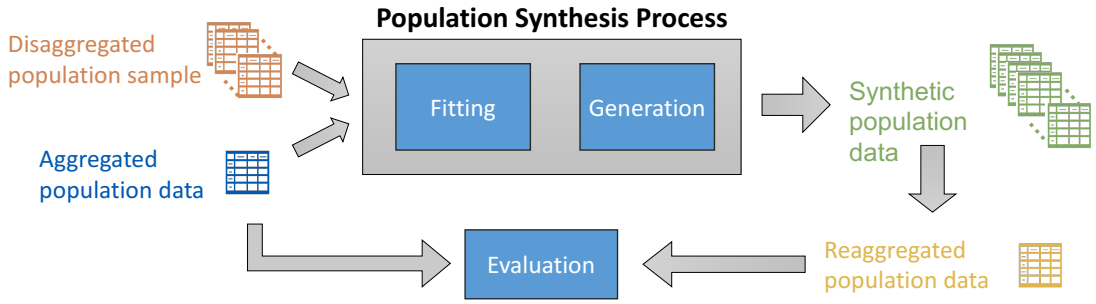

Figure 1: Classical population synthesis approach.

Most synthetic population generators work on a simple principle that involves two data inputs and a population synthesis process, as summarized in Figure 1. These two inputs are (1) the aggregated population data, which provides counts for different attributes such as sex, age groups, and race summarized to various spatial resolutions (census block group, census tract, county); and (2) a disaggregated sample (<5%), which is anonymized data of real people and household with their attributes. In the U.S., both input data are available by the U.S. Census Bureau through the Decennial Census (Bureau 2021) and the American Community Survey (Bureau 2022) data sets, respectively. There are two general steps for any population synthesis techniques (Sun et al. 2018). In the **fitting** process, a model or data structure captures the joint distributions of

people's attributes (i.e., learning), which are of interest to the study. In the **generation** process, a synthetic population is selected/duplicated (i.e., sampling) using the disaggregated sample to fit aggregated population counts. Once the generation process is completed, the newly generated synthetic population records are reaggregated to compare agaisnt the real world aggregated population data. Our aim is to make the difference between these two data minimal. The majority of the techniques proposed over the past three decades can be categorized under (1) Synthetic Reconstruction (SR), (2) Combinatorial Optimization (CO), and (3) Statistical Learning (SL).

Synthetic Reconstruction: SR relies on calculating the suitability of each disaggregated record for all the geographic zones according to the attributes of interest (Chapuis and Taillandier 2019). Most SR techniques are based on Iterative Proportional Fitting (IPF) (Deming and Stephan 1940), which is a deterministic procedure for adjusting data in a table in such a way so that the marginal totals, the sums of each row and column, remain the same. In the case of population synthesis, IPF is used to allocate disaggregated records into zones, where a non-integer weight determines how representative an individual is for these zones (Beckman et al. 1996). In a process called integerization, these non-integer values are then converted into integers. In the generation process, each individual is replicated to create the synthetic population for a zone.

Efforts are proposed to improve IPF to account for multiple levels of marginals. The Iterative Proportional Updating (IPU) algorithm (Ye et al. 2009) accounts for both household and individual attributes simultaneously by updating cross-categorization weights until a fit is achieved. Later IPU was updated to account for different geographical resolutions simultaneously while being computationally efficient (Konduri et al. 2016). An alternative approach called Hierarchical IPF (Müller and Axhausen 2011) was developed to consider proportional fitting for households and individuals dependent on an entropy-optimizing process.

Combinatorial Optimization: CO is similar to SR in that it assigns weights to disaggregated records, but these weights are assigned to fit areas independently rather than across all regions (Williamson et al. 1998). Each attribute of individuals is considered separately, and this deterministic process runs iteratively until reaching the desired fit (Harland et al. 2012). While CO-based techniques can take more time to converge, they perform comparatively well against popular SR techniques (Huang and Williamson 2001). In essence, CO aims to select a subset of individuals from disaggregated data that maximizes the fit between the simulated data and the actual data. Metaheuristics (e.g., Hill Climbing) have been used to find a near-optimal subset from disaggregated data (Williamson et al. 1998). In contrast, the effectiveness of the metaheuristics has seen contradicting performance reports (Durán-Heras et al. 2018, Harland et al. 2012), making it necessary to cross-compare more objectively.

Statistical Learning: SL relies on the fact that disaggregated data have underlying distributions which can be inferred/learned/fit using probabilistic techniques. Based on such inferred distributions, these techniques can be used to generate synthetic populations. Markov Chain Monte Carlo (Farooq et al. 2013) is a technique used in population synthesis with noticeable success. Bayesian Networks (Sun and Erath 2015) were used to infer joint distributions between population attributes graphically. Similarly, a Hidden Markov Model-based technique (Saadi et al. 2016) and a Variational Autoencoder-based technique (Borysov et al. 2019) were proposed to infer joint distributions from disaggregated data. All the methods mentioned above assume a flat hierarchy between the attributes of individuals. Some recent studies (Hu et al. 2018, Sun et al. 2018) enabled researchers to infer the hierarchical nature of populations, such as the relationship between households and individuals. While SL techniques can generate new synthetic individuals (not replicated), the learning process is especially challenging when there are many attributes with different data types (Xu 2020).

2.2 Generative Adversarial Networks

Generative adversarial networks (GANs) are a class of machine learning architectures that integrate deep learning into generative modeling tasks (Goodfellow et al. 2014). GANs have the ability to discover latent patterns in data based on their unique generator/discriminator mechanism. In the vanilla GAN setting (Goodfellow et al. 2014), the generator model is trained to evolve to learn compact representations of data distributions to generate new unseen records. On the other hand, the discriminator model is trained to distinguish between real training data records from fake records generated by the generator model. This generator/discriminator mechanism is a zero-sum game with a proximal equilibrium (Farnia and Ozdaglar 2020). Many successful contributions have helped GANs evolve (Zhang et al. 2019, Mao et al. 2017, Liu and Tuzel 2016) and be used in various applications areas, while the main emphasis has mostly been on computer vision (Pan et al. 2019).

There is a growing body of literature in recent years designing GANs to generate tabular data (Xu et al. 2019, Kim et al. 2021, Kunar 2021). These recent developments pave a promising path for using GANs in synthetic population data generation.

3 OUR APPROACH

Our approach aims to integrate GANs into the population synthesis process. We extended the classical approach by adding a GAN trained with the real disaggregated population sample. With this addition, we no longer need to use the real disaggregated population sample; instead, we can use the GAN-generated disaggregated population sample, which should still capture the statistical features of the real disaggregated population sample. Figure 2 summarizes the essence of our approach.

With the addition of a GAN, our approach includes a learning component that makes it similar to the statistical learning (SL) techniques presented in Section 2.1. At the same time, we can still use the classical population synthesis processes of Synthetic Reconstruction (SR) and Combinatorial Optimization (CO). This makes our approach widely applicable by keeping the merits of different techniques.

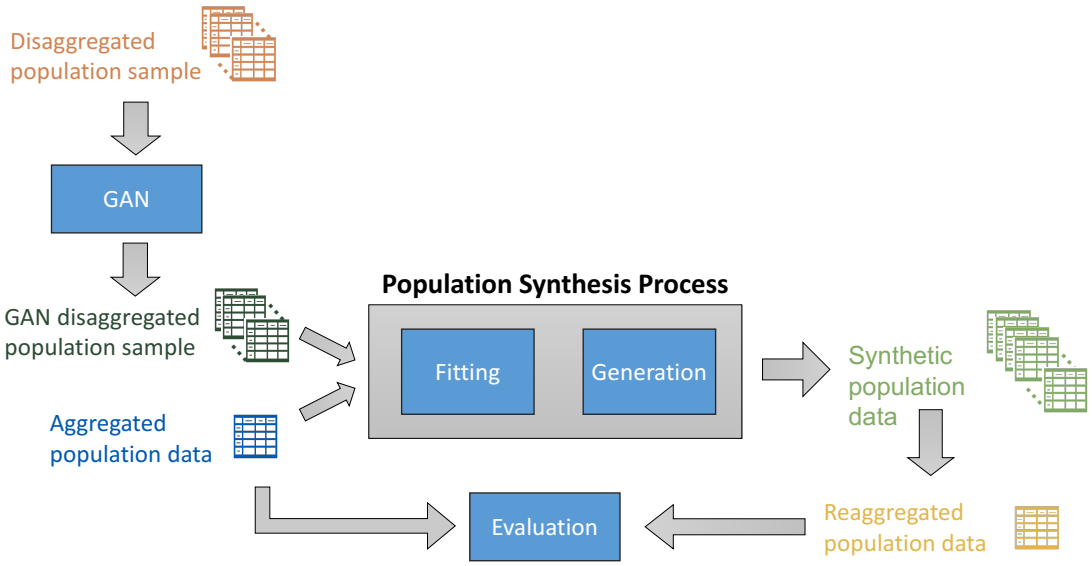

Figure 2: Our proposed population synthesis approach.

3.1 Training GANs

We experimented with many different GANs. First, a TGAN, or tabular GAN, was used with deep neural networks. The TGAN outperforms standard statistical generative models and generates the data using an advanced version of Recurrent Neural Networks (RNN) called Long Short-Term Memory (LSTM) (Xu and Veeramachaneni 2018).

A CTGAN, or conditional tabular GAN, was used to generate the synthetic data. The CTGAN is based on the GAN-based Deep Learning data synthesizer. It offers greater flexibility when modeling distributions compared to statistical or deep neural network models, especially with imbalanced columns (Xu et al. 2019).

Finally, a variation of the CTGAN, the CopulaGAN, was examined. The CopulaGAN aims to improve the accuracy of the CTGAN by using Gaussian Copulas. It takes advantage of Cumulative Distribution Function (CDF) based transformations (Kamthe et al. 2021). Various scores were used to compare the different models as well as metrics from the Synthetic Data Vault (SDV) Project library (Patki et al. 2016). The CTGAN and the CopulaGAN clearly outperformed the tabular GAN while the CTGAN narrowly outperformed the CopulaGAN's average similarity metric from the SDV Evaluation Framework by 2%.

3.2 Population Synthesis Process

For the purpose of this paper, we use IPF for the "Population Synthesis Process". In IPF, a matrix is created that matches the original individual table data with totals closest to the aggregated census tract data possible; after many iterations, the matrix slowly gets closer to the census tract aggregated sums. The inputs needed to run IPF are the disaggregated population sample and the aggregated population data. Generally, with the IPF algorithm, there has to be a unique attribute in which the individual data can be grouped up to get broader level counts. Since the data can be described as spatial microdata, the unique attribute to group the individual data is location, or census tracts (Lovelace and Dumont 2016).

As the value names of the disaggregated population sample and the aggregated population data are already matched up, the dimensions of the aggregated data has to match up with the disaggregated data as a prerequisite to IPF. This is accomplished by flattening the individual dataset by changing disaggregated data columns into a matrix, with each new column being an aggregated category name. Once the prerequisites have been met, the IPF runs and successfully generates synthetic individual data.

3.3 Evaluation

As the last step in our approach, we evaluate the performance of the newly generated synthetic population. In this case, we reaggregate our new data at an appropriate level (e.g., census block group, census tract, etc.) and compare that data with the real aggregated population data. Several error measures, including mean absolute error, mean squared error, can be used to make this comparison. Since our synthetic population generation technique has novelty, we also compare our results with the classical approach that uses real disaggregated sample and a combined dataset with real and GAN-generated disaggregated samples. By making these comparisons, we can evaluate the extent to which GAN can replace or improve the synthetic population generation process.

4 USE-CASE: SYNTHETIC POPULATION FOR FAIRFAX COUNTY, VIRGINIA

4.1 Datasets and Pre-processing

A collection of the PUMS disaggregated dataset from the US Census Bureau was used with the CTGAN. The data included 13,053 individuals, around 1.1% of Fairfax County's actual population, with attributes like age, sex, employment, housing situation, income, citizenship, marital status, schooling, disability, work type, insurance, and race. Continuous data variables like age and income were binned into discrete categorical bins to match the census aggregate statistics.

Also from the US Census Bureau, the American Community Survey (ACS) was utilized to get aggregated population data for all 255 census tracts in Fairfax County (Bureau 2022). This census tract data had the general population counts for each census tract on each specific attribute in the disaggregated data. For example, a single row/individual in the disaggregated data could have an age value between 30 and 40 or an age value above 80; the census data, on the other hand, would have how many individuals in the entire census tract have age values between 30 and 40, age values above 80, etc. This ACS data counts sum up to the total Fairfax County population of around 1.1 million.

Some of the individual attributes had to be combined into new attributes since the aggregated level data before did not include these attributes. Finally, the values of the disaggregated data, which were originally numerical values, were relabeled to match the categorical labels from the aggregated data.

The census tract data from the ACS, as we called aggregated population data, functions as constraints for the IPF algorithm. The specific datasets used were as follows.

- S0101 - age and sex
- DP03 - selected economic characteristics
- S2301 - employment status
- DP02 - selected social characteristics in the US
- S1901 - income in the past 12 months (in 2018 inflation-adjusted dollars)
- S1201 - marital status
- S1401 - school enrollment
- S1810 - disability characteristics
- S2406 - industry by class of worker for the civilian employed population 16 years and over
- S2701 - selected characteristics of health insurance coverage in the US
- DP05 - ACS demographic and housing estimates

The data was then processed and relabeled to match the disaggregated data characteristics from the PUMS. In addition, columns with low populations within the census tracts are combined. It is critical that the disaggregated and aggregated data at the census tract level match in order for the IPF algorithm to function. The census aggregated data has 255 rows (one for each census tract) and 56 columns (the number of attributes) like age between 30 and 40 or male, for the 12 attributes, like age or sex.

4.2 Synthetic Population Generation

We leveraged the IPF algorithm, using the mipfp R package (Barthélemy and Suesse 2018), to generate our synthetic population of Fairfax County. In terms of the disaggregated sample needed for IPF, we used three different settings: (1) the PUMS data, (2) a new disaggregated sample based on a GAN trained on the PUMS data, and (3) a merger of the two. The CTGAN, from the sdv library (Patki et al. 2016), successfully

generated a synthetic disaggregated population of the same size and characteristics of the PUMS data. The model ran for 300 epochs and for around an hour in Google's Colaboratory (Google 2022). The output format is 13,053 rows, each signifying a single individual, and 12 unique columns, each signifying a unique attribute of an individual, like marital status. For the aggregated population data, we used the ACS data as described in Section 4.1.

The IPF code is run three times while keeping the aggregated population data the same. The disaggregated sample is changed from the real 13,053 population from the PUMS to the synthetic 13,053 population with the CTGAN, to a combination of real and synthetic data with 26,106 records. The IPF code successfully generates data for all 1.1 million Fairfax County individuals, organized by each census tract. This synthetic data closely matches up with the aggregated census level statistics.

4.3 Results

The three full synthetic datasets generated from the IPF are reaggregated into census tract-level data in order to compare with the real census data from the ACS. These datasets are of the same format: 255 rows (one for each census tract) and 56 columns for the distinct attributes. The summary of the results is provided in Table 1. Here, the general mean absolute error is calculated using (1).

$$\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left|y_{(i,j)}-\hat{y}_{(i,j)}\right|}{m*n} \tag{1}$$

Here, $y_{(i,j)}$ represents the ith population attribute in census tract j, $\hat{y}_{(i,j)}$ represents the ith population attribute in synthetic population tract j, n represents number of census tracts, and m represents number of attributes. In other words, the absolute difference between real and synthetic attribute values is calculated across census tracts, and all attributes are then averaged. The standard deviation calculation used these absolute errors across all census tracts and attributes. The results suggest that our proposed approach outperformed the other two with low general mean absolute error and low standard deviation of the mean absolute error.

Table 1: Summary statistics about three synthetic population datasets.

	General Mean Absolute Error	Standard Deviation of Mean Absolute Error
Synthetic population using real PUMS disaggregated data only	18.75	42.30
Synthetic population using CTGAN-based disaggregated data only	13.51	14.97
Synthetic population using both real PUMS and CTGAN-based disaggregated data	18.44	23.35

Ultimately, the error between the real aggregated data and the reaggregated synthetic data was low, as most of the error for each attribute is below 50, while the average census tract population is 4,484. The data is averaged for each distinct attribute across 255 census tracts to get a comprehensive graph comparing the three IPF- generated data as shown in Figure 3. From all these comparisons, we can see that when using a CTGAN-only disaggregated sample helped IPF perform better. The PUMS data-only version and the combined version almost consistently performed worse.

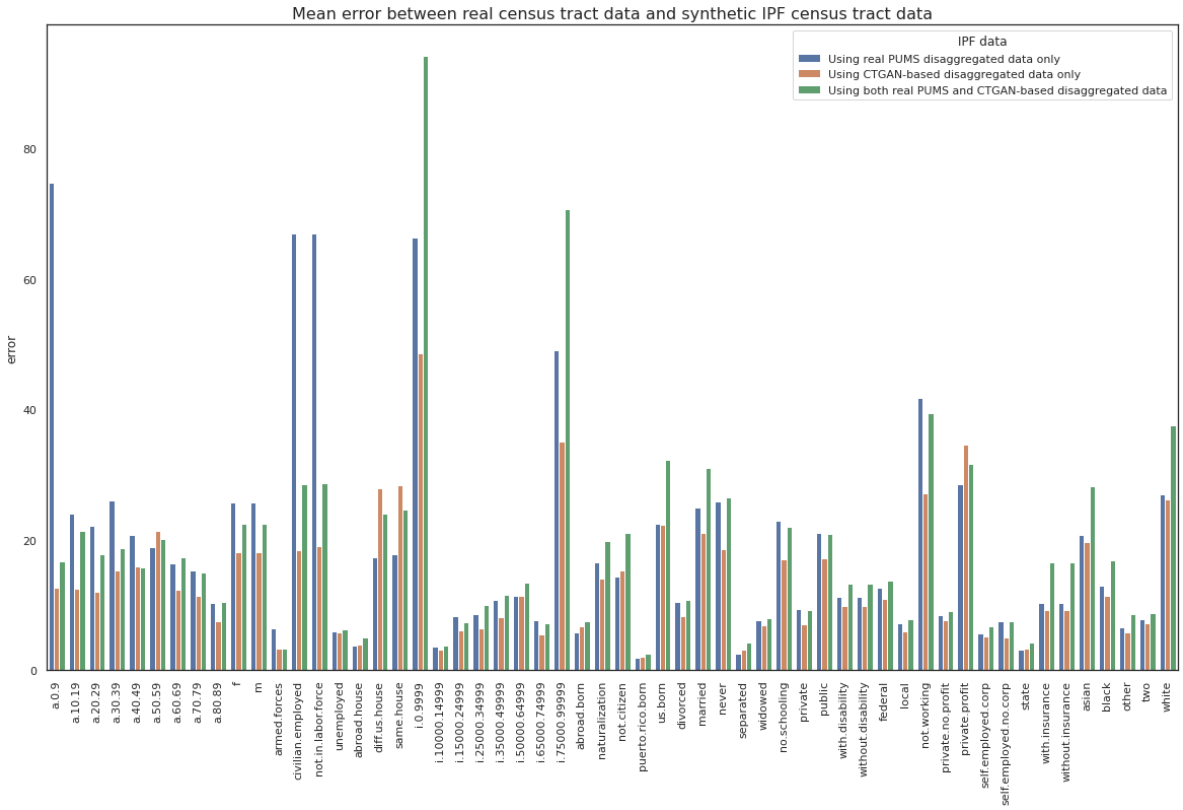

Figure 3: Attribute-specific mean absolute error between real and synthetic data across all 255 census tracts.

In addition, the real and synthetic data can be compared for a specific attribute like age or income, showing which attributes are modeled well by which type of IPF data. As shown in Figure 4, the IPF data based on just the real PUMS data generates the age attribute poorly as there lie many datapoints significantly off the line of best fit, where the real census tract and synthetic census tract populations differ significantly. On the other hand, the IPF data based on the synthetic CTGAN data and IPF data based on the real PUMS and synthetic CTGAN data capture the age characteristic well with no points off of the line of best fit. From both census track and attribute-based results, we can posit that this CTGAN-based technique has a promising contribution to the modeling and simulation community.

5 SUMMARY AND FUTURE WORK

Synthetic populations are used in creating realistic population representations for agent-based and microsimulation modeling approaches. In this work, we made a promising contribution to an existing and popular synthetic population generation technique called IPF. We trained a deep generative model called CTGAN using the US Census' PUMS disaggregated sample. Once the patterns were learned, the CTGAN was able to generate a novel disaggregated sample to be used in IPF. Overall, the IPF algorithm is leveraged in three settings: (1) data based on the real PUMS, (2) data based on the synthetic CTGAN, and (3) a combination of the two.

We found that the disaggregated sample generated using the CTGAN-based technique outperforms the other two consistently by generating more realistic synthetic populations. Since disaggregated samples are used in many different synthetic population generation algorithms (other than IPF) as discussed in Section 2, we can make larger contributions to the modeling and simulation community. Our future work includes testing

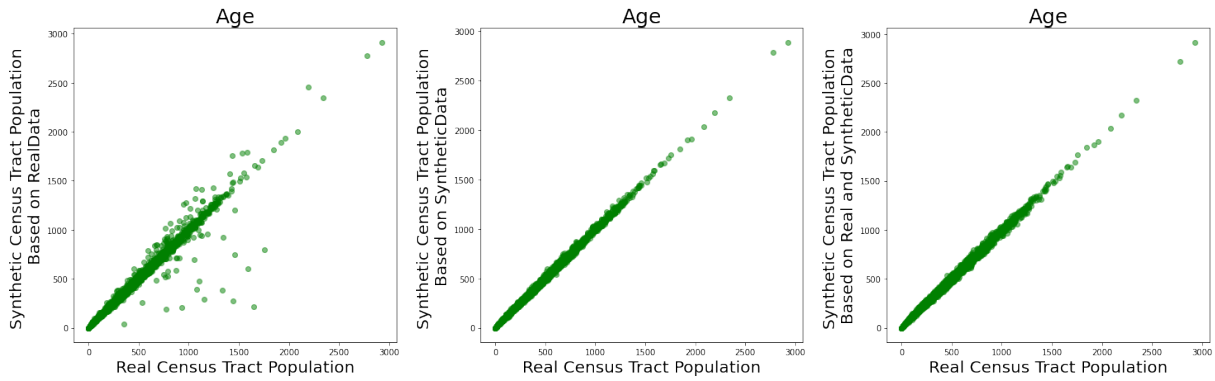

Figure 4: Comparison of real and synthetic census data on the age attribute using different disaggregated samples.

different synthetic population generation algorithms and different geographic areas to evaluate the extent of the results found in this paper.

NOTES

This research has been supported by National Science Foundation grants (DEB-2109647 and DEB-2030685) and is conducted as part of the 2021 Aspiring Scientists Summer Internship Program at George Mason University. All code and data used in this paper are available at https://github.com/srihan-kotnana/spatial-microsimulation.

REFERENCES

Barthélemy, J., and T. Suesse. 2018. "mipfp: An R Package for Multidimensional Array Fitting and Simulating Multivariate Bernoulli Distributions". *Journal of Statistical Software* vol. 86, pp. 1–20.

Beckman, R. J., K. A. Baggerly, and M. D. McKay. 1996. "Creating Synthetic Baseline Populations". *Transportation Research Part A: Policy and Practice* vol. 30 (6 PART A), pp. 415–429.

Bissett, K., J. Cadena, M. Khan, C. J. Kuhlman, B. Lewis, and P. A. Telionis. 2016. "An Integrated Agent-based Approach for Modeling Disease Spread in Large Populations to Support Health Informatics". In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 629–632. IEEE.

Borysov, S. S., J. Rich, and F. C. Pereira. 2019. "How to Generate Micro-agents? A Deep Generative Modeling Approach to Population Synthesis". *Transportation Research Part C: Emerging Technologies* vol. 106, pp. 73–97.

Bureau, U. S. C. 2021 (accessed September 14, 2021). *Decennial Census of Population and Housing*. https://www.census.gov/programs-surveys/decennial-census.html.

Bureau, U. S. C. 2022 (accessed March 18, 2022). *American Community Survey (ACS)*. https://www.census.gov/programs-surveys/acs.

Chapuis, K., and P. Taillandier. 2019. "A Brief Review of Synthetic Population Generation Practices in Agent-based Social Simulation". In *SSC2019, Social Simulation Conference*.

Deming, W. E., and F. F. Stephan. 1940. "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known". *The Annals of Mathematical Statistics* vol. 11 (4), pp. 427–444.

Durán-Heras, A., I. García-Gutiérrez, and G. Castilla-Alcalá. 2018. "Comparison of Iterative Proportional Fitting and Simulated Annealing as synthetic population generation techniques: Importance of the rounding method". *Computers, Environment and Urban Systems* vol. 68, pp. 78–88.

Farnia, F., and A. Ozdaglar. 2020. "Do GANs Always Have Nash Equilibria?". In *International Conference on Machine Learning*, pp. 3029–3039. PMLR.

Farooq, B., M. Bierlaire, R. Hurtubia, and G. Flötteröd. 2013. "Simulation Based Population Synthesis". *Transportation Research Part B: Methodological* vol. 58, pp. 243–263.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. "Generative Adversarial Nets". In *Advances in Neural Information Processing Systems*, pp. 2672–2680.

Google 2022 (accessed March 14, 2022). *Welcome To Colaboratory*. https://colab.research.google.com/.

Hafezi, M. H., N. S. Daisy, L. Liu, and H. Millward. 2019. "Modelling Transport-related Pollution Emissions for the Synthetic Baseline population of a large Canadian university". *International Journal of Urban Sciences* vol. 23 (4), pp. 519–533.

Harland, K., A. Heppenstall, D. Smith, and M. Birkin. 2012. "Creating Realistic Synthetic Populations at Varying Spatial Scales: A Comparative Critique of Population Synthesis Techniques". *JASSS* vol. 15 (1), pp. 1–24.

He, B. Y., J. Zhou, Z. Ma, J. Y. Chow, and K. Ozbay. 2020. "Evaluation of City-scale Built Environment Policies in New York City with an Emerging-mobility-accessible Synthetic Population". *Transportation Research Part A: Policy and Practice* vol. 141, pp. 444–467.

Hu, J., J. P. Reiter, Q. Wang et al. 2018. "Dirichlet Process Mixture Models for Modeling and Generating Synthetic Versions of Nested Categorical Data". *Bayesian Analysis* vol. 13 (1), pp. 183–200.

Huang, Z., and P. Williamson. 2001. "A Comparison of Synthetic Reconstruction and Combinatorial Optimisation Approaches to the Creation of Small Area Microdata". (October).

Huynh, N., M.-R. Namazi-Rad, P. Perez, M. Berryman, Q. Chen, and J. Barthelemy. 2013. "Generating a Synthetic Population in Support of Agent-based Modeling of Transportation in Sydney". In *20th International Congress on Modelling and Simulation (MODSIM 2013)*, pp. 1357–1363, The Modelling and Simulation Society of Australia and New Zealand.

Jiang, N., A. Crooks, H. Kavak, and W. G. Kennedy. 2021. "Generation of Reusable Synthetic Population and Social Networks for Agent-Based Modeling". In *Proceedings of the 2021 Annual Modeling and Simulation Conference*.

Jiang, N., A. T. Crooks, H. Kavak, A. Burger, and W. G. Kennedy. 2022. "A Method to Create a Synthetic Population with Social Networks for Geographically-explicit Agent-based Models". *Computational Urban Science* vol. 2 (1), pp. 1–18.

Kamthe, S., S. Assefa, and M. Deisenroth. 2021. "Copula Flows for Synthetic Data Generation". *arXiv preprint arXiv:2101.00598*.

Kim, J., J. Jeon, J. Lee, J. Hyeong, and N. Park. 2021. "OCT-GAN: Neural ODE-based Conditional Tabular GANs". In *Proceedings of the Web Conference 2021*, pp. 1506–1515.

Konduri, K. C., D. You, V. M. Garikapati, and R. M. Pendyala. 2016. "Enhanced Synthetic Population Generator that Accommodates Control Variables at Multiple Geographic Resolutions". *Transportation Research Record* vol. 2563 (1), pp. 40–50.

Kunar, A. 2021. "CTAB-GAN: Effective Tabular Data Synthesizing". *Delft University of Technology, Master Thesis*.

Laskowski, M., B. C. P. Demianyk, J. Benavides, M. R. Friesen, R. D. McLeod, S. N. Mukhi, and M. Crowley. 2012. "Extracting Data from Disparate Sources for Agent-Based Disease Spread Models". *Epidemiology Research International* vol. 2012 (Ili), pp. 1–18.

Le, D.-T., G. Cernicchiaro, C. Zegras, and J. Ferreira Jr. 2016. "Constructing a Synthetic Population of Establishments for the Simmobility Microsimulation Platform". *Transportation Research Procedia* vol. 19, pp. 81–93.

Levy, J. I., M. P. Fabian, and J. L. Peters. 2014. "Community-wide Health Risk Assessment Using Geographically Resolved Demographic Data: A Synthetic Population Approach". *PloS ONE* vol. 9 (1), pp. e87144.

Liu, M.-Y., and O. Tuzel. 2016. "Coupled Generative Adversarial Networks". In *Advances in Neural Information Processing Systems*, pp. 469–477.

Lovelace, R., and M. Dumont. 2016. *Spatial Microsimulation with R*. CRC Press.

Mao, X., Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley. 2017. "Least Squares Generative Adversarial Networks". In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2794–2802.

Moeckel, R., K. Spiekermann, and M. Wegener. 2003. "Creating a Synthetic Population". In *Proceedings of the 8th International Conference on Computers in Urban Planning and Urban Management (CUPUM)*, pp. 1–18.

Müller, K., and K. W. Axhausen. 2011. "Hierarchical IPF: Generating a Synthetic Population for Switzerland". *Arbeitsberichte Verkehrs-und Raumplanung* vol. 718.

Narassima, M., G. R. Jammy, R. Pant, L. Choudhury, R. Aadharsh, V. Yeldandi, S. Anbuudayasankar, and P. Rangasami. 2020. "An Agent Based Model Methodology for Assessing Spread and Health Systems Burden for COVID-19 Using a Synthetic Population from India". *medRxiv*.

Pan, Z., W. Yu, X. Yi, A. Khan, F. Yuan, and Y. Zheng. 2019. "Recent Progress on Generative Adversarial Networks (GANs): A survey". *IEEE Access* vol. 7, pp. 36322–36333.

Patki, N., R. Wedge, and K. Veeramachaneni. 2016, Oct. "The Synthetic Data Vault". In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 399–410.

Pesavento, J., A. Chen, R. Yu, J.-S. Kim, H. Kavak, T. Anderson, and A. Züfle. 2020. "Data-driven Mobility Models for COVID-19 Simulation". In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Advances in Resilient and Intelligent Cities*, pp. 29–38.

Saadi, I., A. Mustafa, J. Teller, B. Farooq, and M. Cools. 2016. "Hidden Markov Model-based Population Synthesis". *Transportation Research Part B: Methodological* vol. 90, pp. 1–21.

Sun, L., and A. Erath. 2015. "A Bayesian Network Approach for Population Synthesis". *Transportation Research Part C: Emerging Technologies* vol. 61, pp. 49–62.

Sun, L., A. Erath, and M. Cai. 2018. "A Hierarchical Mixture Modeling Framework for Population Synthesis". *Transportation Research Part B: Methodological* vol. 114, pp. 199–212.

Williamson, P., M. Birkin, and P. H. Rees. 1998. "The Estimation of Population Microdata by Using Data from Small Area Statistics and Samples of Anonymised Records". *Environment and Planning A: Economy and Space* vol. 30 (5), pp. 785–816. PMID: 12293871.

Xu, L. 2020. "Synthesizing Tabular Data using Conditional GAN". Master's thesis, Massachusetts Institute of Technology.

Xu, L., M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni. 2019. "Modeling Tabular Data Using Conditional GAN". *Advances in Neural Information Processing Systems* vol. 32, pp. 7335–7345.

Xu, L., and K. Veeramachaneni. 2018. "Synthesizing Tabular Data using Generative Adversarial Networks". *arXiv preprint arXiv:1811.11264*.

Xu, Z., K. Glass, C. L. Lau, N. Geard, P. Graves, and A. Clements. 2017. "A Synthetic Population for Modelling the Dynamics of Infectious Disease Transmission in American Samoa". *Scientific reports* vol. 7 (1), pp. 1–9.

Ye, X., K. Konduri, R. M. Pendyala, B. Sana, and P. Waddell. 2009. "A Methodology to Match Distributions of Both Household and Person attributes in the Generation of Synthetic Populations". In *88th Annual Meeting of the Transportation Research Board, Washington, DC*.

Zhang, D., J. Cao, S. Feygin, D. Tang, Z.-J. M. Shen, and A. Pozdnoukhov. 2019. "Connected Population Synthesis for Transportation Simulation". *Transportation Research Part C: Emerging Technologies* vol. 103, pp. 1–16.

Zhang, H., I. Goodfellow, D. Metaxas, and A. Odena. 2019. "Self-attention Generative Adversarial Networks". In *International Conference on Machine Learning*, pp. 7354–7363. PMLR.

AUTHOR BIOGRAPHIES

SRIHAN KOTNANA is a junior at Westfield High School in Virginia, USA interested in Computer Science. Specifically, his current interests are data science and algorithms/data structures. His email address is srihank8@gmail.com.

DAVID HAN is an undergraduate student working towards a bachelor's degree in computer science and mathematics at Cornell University. His current interests lie in software engineering and data science. His email address is dmh338@cornell.edu.

TAYLOR ANDERSON is an Assistant Professor in the Department of Geography and Geoinformation Science. Her research leverages GIScience and geosimulation for better understanding, predicting, and responding to diseases in human and ecological systems. Her email address is tander6@gmu.edu.

ANDREAS ZÜFLE is an Associate Professor in the Department of Geography and Geoinformation Science. His research interest includes data mining, spatial computing, spatial database management, and geosimulation. His email address is azufle@gmu.edu.

HAMDI KAVAK is an Assistant Professor in the Computational and Data Sciences Department and co-director of the Center for Social Complexity at George Mason University. His research interests lie at the intersection of data science and modeling and simulation. His email and website addresses are hkavak@gmu.edu and http://www.hamdikavak.com.