Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

QUANTITATIVE COMPARISON OF POPULATION SYNTHESIS TECHNIQUES

David Han[1], Samiul Islam[2], Taylor Anderson[3], Andrew T. Crooks[4], and Hamdi Kavak[2]

[1]Dept. of Computer Science, Cornell University, Ithaca, NY, USA
[2]Dept. of Computational and Data Sciences, George Mason University, Fairfax, VA, USA
[3]Dept. of Geography and Geoinformation Sciences, George Mason University, Fairfax, VA, USA
[4]Dept. of Geography, University at Buffalo, Buffalo, NY, USA

ABSTRACT

Synthetic populations serve as the building blocks for predictive models in many domains, including transportation, epidemiology, and public policy. Therefore, using realistic synthetic populations is essential in these domains. Given the wide range of available techniques, determining which methods are most effective can be challenging. In this study, we investigate five synthetic population generation techniques in parallel to synthesize population data for various regions in North America. Our findings indicate that iterative proportional fitting (IPF) and conditional probabilities techniques perform best in different regions, geographic scales, and with increased attributes. Furthermore, IPF has lower implementation complexity, making it an ideal technique for various population synthesis tasks. We documented the evaluation process and shared our source code to enable further research on advancing the field of modeling and simulation.

1 INTRODUCTION

Analysis of human dynamics has long relied on aggregated population data (Heppenstall et al. 2012). Since the 1950s, researchers have criticized the use of aggregated data to draw conclusions about individuals, a problem known as the ecological fallacy (Orcutt 1957; Piantadosi et al. 1988). Limitations in top-down approaches have led to bottom-up modeling methodologies to study populations, including microsimulation, cellular automata, and agent-based models (ABMs). These bottom-up modeling methodologies focus on representing interactions between heterogeneous individuals to model emergent social phenomena like crime (Malleson et al. 2010), traffic and mobility (Kavak et al. 2019), housing market (Geanakoplos et al. 2012), disease spread (Kim et al. 2020; Anderson and Dragićević 2020; Pesavento et al. 2020), and evacuation (D Orazio et al. 2014; Brachman and Dragicevic 2014), among others.

Empirical simulation models aim to realistically represent individual behavior, decision-making, and relationships. Therefore, it is crucial that realistic demographic characteristics and household structures are represented in simulated populations. However, in order to preserve the privacy of individuals, much of the available demographic data such as the U.S. Decennial Census (Bureau 2020b) and the Canadian Census (Canada 2021a) are available only in aggregated counts and percentages or as anonymized individual records of a small sample of the population. In the absence of individual-level data, researchers rely on synthetic population generation methods to construct datasets of artificial individuals that align with specific geographic zones (e.g., state, county, census tract) (Beckman et al. 1996). Ideally, synthetic population data should reflect the statistical characteristics of the actual population and be easily generated across different regions and geographies.

There are many established methods for generating synthetic populations. In that vein, several published studies aim to compare these various approaches and review their advantages and disadvantages. These studies typically focus on reviewing the different methods that synthesize the individuals without hierarchy (Harland et al. 2012) or both the individuals and their respective households (Fabrice Yaméogo et al. 2020).

979-8-3315-8726-0/25/$31.00 ©2025 IEEE

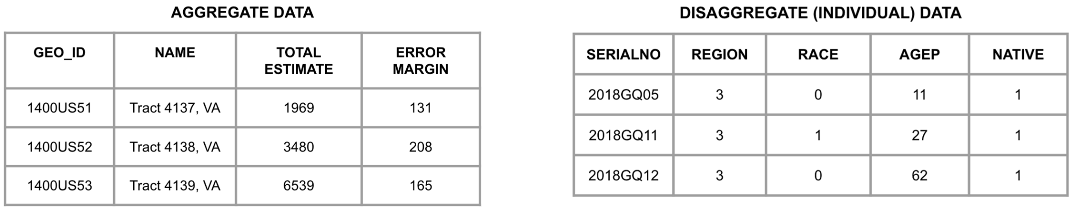
However, many reviews are specific to one geography and do not implement and share the algorithms to quantitatively compare the effectiveness of the different synthetic population methods.

Our study aims to go beyond qualitative comparison by examining the effects of spatial scale, country, and dimensionality. To meet this objective, we employ five methods to generate synthetic populations, including iterative proportional fitting (IPF), hill climbing (HC), simulated annealing (SA), conditional probabilities (CP), and random placement with replacement (RPWR), which serves as the baseline. The selected methods generate synthetic populations for two study areas, leveraging Canadian census data for Metro Vancouver, British Columbia, and United States census data for Fairfax County, Virginia.

2 BACKGROUND

2.1 Definitions

We begin by defining the building blocks of synthetic population generation techniques. Two input data are typically used by a diverse set of techniques. The first is **aggregated population data**, which counts the number of individuals within a defined geographic zone (i.e., census tract) with different attributes such as gender/sex, age, and race. The second is a **disaggregated sample** that provides anonymized data representing real individual people, their households, and their corresponding attributes within a broader geographic zone. Figure 1 shows sample datasets for these two data types. In the U.S., both input data are available through the Census Bureau's Decennial Census (Bureau 2020b), and the Public Use Microdata Sample (PUMS) (Bureau 2020a) data sets. In Canada, both input data are available through Statistics Canada's Census Program (Canada 2021a) and the Public Use Microdata Files (PUMF) (Canada 2021b).



AGGREGATE DATA

GEO_ID	NAME	TOTAL ESTIMATE	ERROR MARGIN
1400US51	Tract 4137, VA	1969	131
1400US52	Tract 4138, VA	3480	208
1400US53	Tract 4139, VA	6539	165

DISAGGREGATE (INDIVIDUAL) DATA

SERIALNO	REGION	RACE	AGEP	NATIVE
2018GQ05	3	0	11	1
2018GQ11	3	1	27	1
2018GQ12	3	0	62	1

Figure 1: An example aggregated and disaggregated data.

There are two high-level steps for any population synthesis techniques: fitting and generation (Sun et al. 2018). In the **fitting** step, a model or data structure captures the joint distributions of individual attributes that are of interest to the study, and this distribution is learned to create a synthetic population. In the **generation** step, a synthetic population is selected or duplicated using the disaggregated sample to fit aggregated population data statistics (akin to sampling).

2.2 A Brief Summary of Synthetic Population Generation Techniques

The majority of the synthetic population generation techniques developed over the last three decades can be categorized into three groups based on their underlying generation mechanism: Synthetic Reconstruction (SR), Combinatorial Optimization (CO), and Statistical Learning (SL).

SR methods assess how well each disaggregated record matches various geographic zones based on selected attributes (Chapuis and Taillandier 2019). In essence, SR aims to assign individuals to regions in a way that best aligns with their attributes. Most SR techniques today are based on the 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 sums of each row and column, also known as *marginal totals*, remain the same. Several efforts are proposed to improve IPF to account for the high dimensionality of marginal values resulting from many attributes that individuals hold. For instance, the Iterative Proportional Updating

(IPU) algorithm (Ye et al. 2009) accounts for household and individual characteristics simultaneously by updating cross-categorization weights until a fit is achieved. Later, IPU was updated to account for different geographical resolutions simultaneously in a more computationally efficient manner (Konduri et al. 2016). An alternative approach called Hierarchical IPF (Müller and Axhausen 2011) was developed to consider proportional fitting for households and individuals using an entropy-optimizing process.

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 an individual is considered separately, and this deterministic process runs iteratively until the desired fit is reached (Harland et al. 2012). While CO-based techniques can take longer to converge, they perform comparatively well against popular SR techniques (Huang and Williamson 2001). Since it is intractable to find the optimal combination, several metaheuristics have been developed to find near-optimal solutions. Metaheuristics (e.g., Hill Climbing) have been used to find a near-optimal subset from disaggregated data (Williamson et al. 1998). The effectiveness of the metaheuristics has seen contradicting performance reports (Durán-Heras et al. 2018; Harland et al. 2012), necessitating a more objective cross-comparison.

SL relies on the fact that disaggregated data have underlying distributions which can be inferred using probabilistic techniques. A basic implementation would be *conditional probabilities*, constructing a chain of dependent probabilities based on known data breakdowns (Harland et al. 2012). Based on inferred distributions or probabilities, SL techniques can be used to generate synthetic individuals. Markov Chain Monte Carlo (Farooq et al. 2013) is a technique used in population synthesis with notable success, while Bayesian Networks (Sun and Erath 2015) have also been employed to infer joint distributions between population attributes graphically. Similarly, a Hidden Markov Model-based technique was proposed to infer joint distributions from disaggregated data (Saadi et al. 2016). Recently, generative adversarial models have been used to capture the joint distributions to enrich the disaggregated sample (Kotnana et al. 2022). 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) have enabled researchers to infer the hierarchical nature of populations, such as the relationship between households and individuals. Compared to SR and CO, SL techniques can technically generate new synthetic individuals that are not part of the disaggregated data.

2.3 Selected Techniques

2.3.1 Iterative Proportional Fitting

IPF is a procedure for adjusting data such that marginal totals, or the sums of each row and column, remain fixed. In population synthesis, IPF assigns individuals to geographic zones using non-integer weights that indicate the degree of representativeness of each individual for a given area (Choupani and Mamdoohi 2016). These weights are then converted into integers, as population or household weights must reflect actual counts. The Truncate, Replicate, Sample (TRS) method is a common approach, which handles differences between replication-based and conventional weights (Lovelace and Ballas 2013). Integerisation is a critical step, as it often reduces the statistical fit of the synthetic population. Using these integer weights, individuals are replicated to construct a full synthetic population that conforms to zone-level constraints. This process is implemented using tools such as the `ipfp` package in R and conceptually shown in Figure 2.

Creating a synthetic population has challenges due to inconsistencies and limitations in available data (Durán-Heras et al. 2018). Aggregated data are often used to construct the marginal totals required by IPF. These constraints—such as counts by age, race, or language—are typically drawn from official statistics at the zone level. The initial cell values in the IPF table are usually seeded from a disaggregated sample covering a broader region, which may represent only a fraction of the actual individuals. As a result, IPF plays a key role in generating a representative population for smaller geographic units using limited disaggregated sample (Lovelace and Dumont 2016).

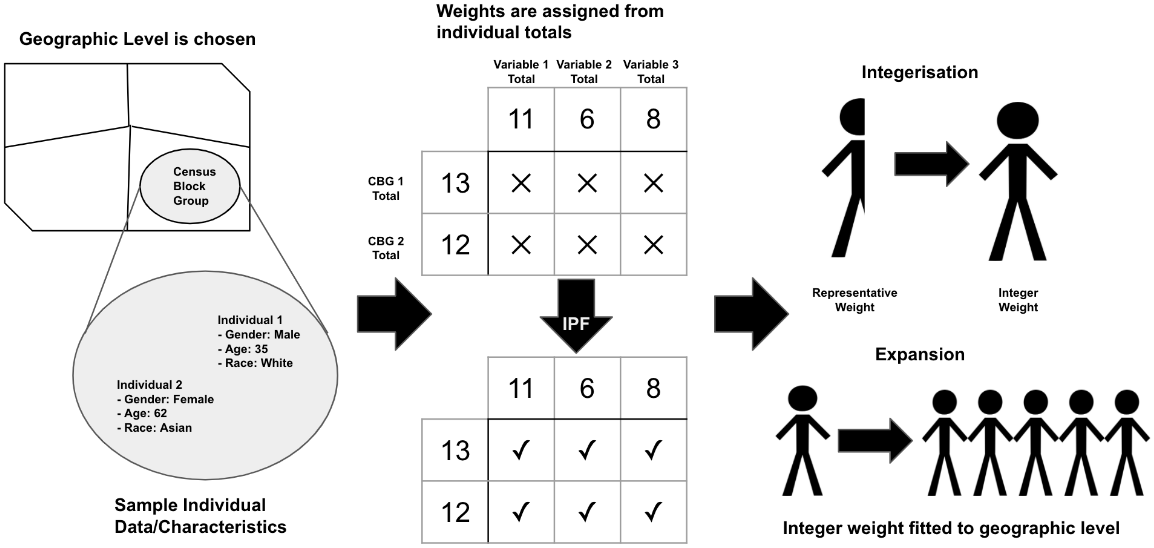


Figure 2: A conceptual depiction of the IPF process for population synthesis.

2.3.2 Hill Climbing and Simulated Annealing

Hill Climbing (HC) and Simulated Annealing (SA) are metaheuristics algorithms, to find near minimum-cost solutions (Kim and Lee 2015). In such a method, we begin with an initial sample population and attempt to allocate a subset of the sample that matches a realistic population. An integer weight, 1 or 0, corresponds to each individual's inclusion or exclusion in the model. In this process, we randomly assign a subset of the sample population for each region where the subset's people count equals the region's population. Then, we calculate the goodness of fit for the given population based on aggregate counts. After selecting a random individual to replace, the goodness of fit is calculated again with the new individual included. If the goodness of fit improves from the previous population, the individuals are replaced. This cycle is repeated until an optional fit or a specific number of attempts is reached. As shown in Figure 3, this nuanced version of HC is also called Stochastic Hill Climbing since randomness is involved.

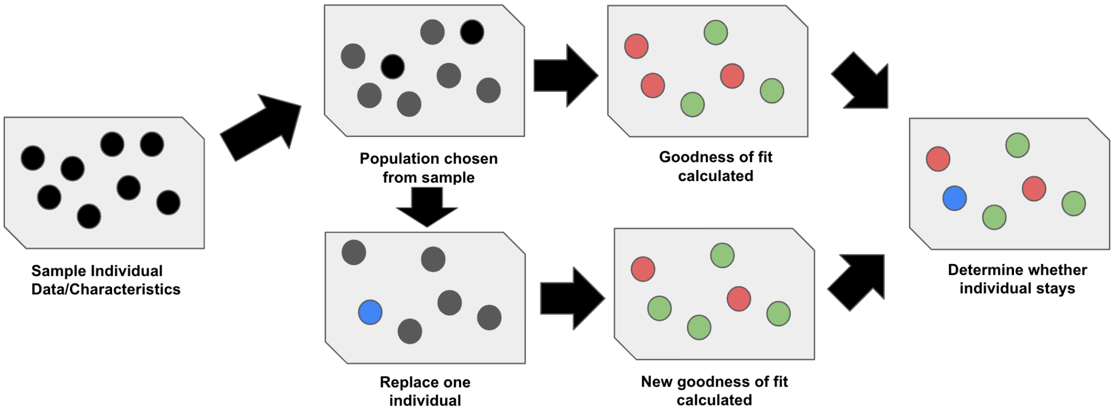

Sample Individual Data/Characteristics
Population chosen from sample
Goodness of fit calculated
Replace one individual
New goodness of fit calculated
Determine whether individual stays

Figure 3: One round of the HC algorithm process.

The simplicity of HC leads to its advantage of being a fast algorithm; however, it can become trapped in a locally optimal solution (Mauša et al. 2013). Using the First Found neighborhood search, the algorithm finds any solution superior to the current one and incorporates it into a new solution in the next iteration. Unlike HC, SA reduces the risk of getting trapped in local minima by occasionally accepting inferior solutions. This acceptance is governed by an annealing factor (Harland et al. 2012), which sets the probability threshold for replacing the current solution with one that has a worse goodness of fit. The annealing factor gradually decreases over time. As a result, the model becomes less likely to accept worse solutions as it approaches an optimal solution.

2.3.3 Conditional Probabilities

Another method used for population synthesis is the Conditional Probabilities (CP) technique (Harland et al. 2012). An element of this technique is that the population is synthetically derived and is built up constraint by constraint based on the aggregated population data provided at each geographic region. Due to the sampling aspect, we can categorize this method under statistical learning. Unlike IPF, HC, and SA, this technique does not require a disaggregated sample. Additionally, the assignment of characteristics is stochastic rather than deterministic, allowing for randomness. Since CP utilizes constraints from the official statistics, modifications to constraints make the process more complex. Especially removing an existing attribute or adding a new one requires a major overhaul in the synthesis source code. Furthermore, the order of constraints matters and should be introduced from most influential to least influential. This means introducing the constraints beginning from the lowest entropy distributions where the population is relatively evenly distributed between categories (e.g., gender). Figure 4 summarizes the concept of CP.

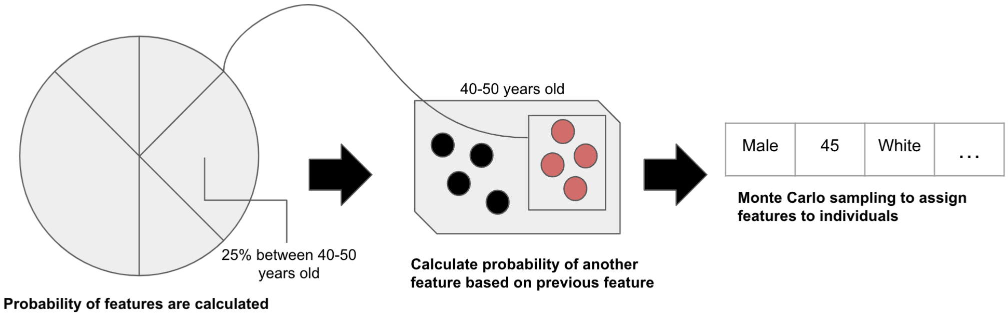

40-50 years old
Male
45
White
...
Monte Carlo sampling to assign features to individuals
25% between 40-50 years old
Calculate probability of another feature based on previous feature
Probability of features are calculated

Figure 4: An illustrative process for the CP algorithm.

2.3.4 Simple Random Sampling (Baseline)

Simple Random Sampling (Gallagher et al. 2018) is a baseline technique we adopted to compare the superiority of other techniques. Simply put, the disaggregated sample of an area is assumed representative, and sampling with the replacement of this data should yield statistically similar properties of the area, as illustrated in Figure 5. The advantage of this approach is the simple implementation and interpretation.

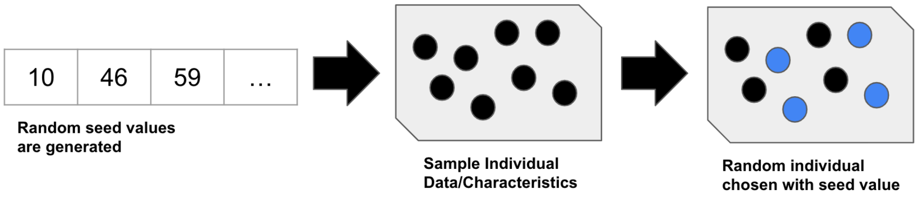
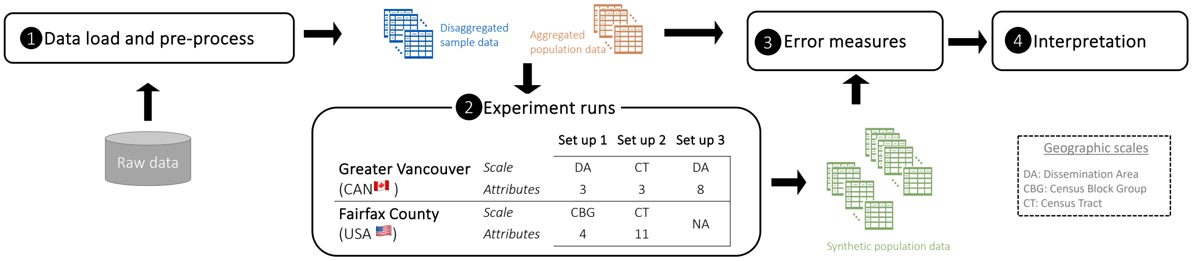

10
46
59
...
Random seed values are generated
Sample Individual Data/Characteristics
Random individual chosen with seed value

Figure 5: Process for RPWR.

3 METHODOLOGY

Our methodology is composed of four main processes, as illustrated in Figure 6. We *first* obtain aggregated population data and disaggregated sample data from publicly accessible official data sources. When these data sets have inconsistent counts at different granularities, we address them by normalizing the counts with known ratios. *Second*, for each use case area, we set up a series of experiments at different geographic scales. We run all five selected techniques at each scale. This step generates all synthetic populations. *Third*, we evaluate the accuracy of the synthetic populations by aggregating and then comparing them with official numbers using two metrics described below. *Fourth*, we interpret the results and report which technique works better under which condition.



		Set up 1	Set up 2	Set up 3
Greater Vancouver (CAN 🇨🇦)	*Scale*	DA	CT	DA
	Attributes	3	3	8
Fairfax County (USA 🇺🇸)	*Scale*	CBG	CT	NA
	Attributes	4	11	

Figure 6: Our four-step process used in this study.

3.1 Error Metrics

Let the set of geographic areas (e.g., CTs) be denoted by $A = \{a_1, a_2, \ldots, a_n\}$. For a given attribute, y_i is the official (true) value in area a_i, and $\hat{y}_i$ is the estimated (synthetic) value in area a_i, $n = |A|$ is the number of geographic areas. The Percent Total Absolute Error ($\%TAE$) is calculated using 1:

$$\%\text{TAE} = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{\sum_{i=1}^{n} y_i} \cdot 100, \tag{1}$$

where $\sum |\hat{y}_i - y_i|$ is the total absolute error and $\sum y_i$ is the total official value, across all areas. This metric is especially useful to uncover general biases.

The Coefficient of Determination (R^2), is calculated using 2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \tag{2}$$

where $\bar{y} = \frac{1}{n} \sum\limits_{i=1}^{n} y_i$ is the mean of the official values across all areas. This second metric is a measure to indicate the preservation of spatial patterns where the value ranges from $-\infty$ to 1, with higher values indicating better model fit.

3.2 Study Areas

As for the study area, we focused on two locations in North America.

- **Fairfax County, VA, USA**. Fairfax County is the most populous county in Virginia with ≈1.16 million inhabitants according to the US Census (Bureau 2020a) numbers as of this writing. We obtained data from the 2018 5-year American Community Survey (ACS), which provides detailed population attributes such as age, sex, and race for communities. The geographic units used in this study are Census Block Groups (CBG) and Census Tracts (CT). The PUMS datasets provide

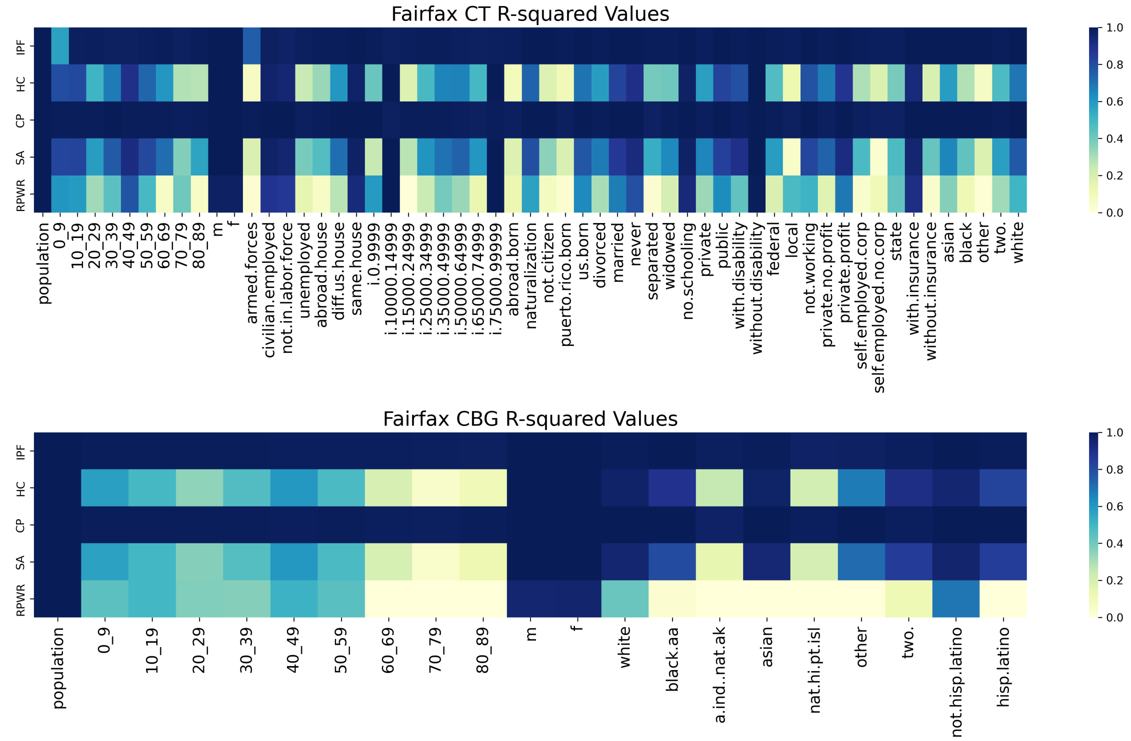


Figure 7: R-squared values of Fairfax at the CT (upper panel) and CBG (lower panel) levels.

anonymized information on individuals or households, and we used a PUMS dataset consisting of 84,755 individuals located in the study area.

- **Metro Vancouver, BC, Canada**. Metro Vancouver is a census metropolitan area in the southwest corner of mainland British Columbia (BC). It is composed of thirteen of the most populous municipalities in BC. According to the 2016 Census, the total population of Metro Vancouver was 2.46 million and with a population density of 854 individuals per square kilometer. The census data for Metro Vancouver for 2016 was obtained from Statistics Canada who provides population, age, gender, and race data at various geographic units. The geographic units used in this study are Dissemination Areas (DA) and Census Tracts (CT).

4 RESULTS

As discussed earlier, all constraint variables in the synthetic population should portray the attributes of the population of interest. The R-squared measure allows comparisons between the effectiveness of the different algorithms across geographic regions.

Figure 7 highlights the R-squared values in Fairfax CTs and CBGs. Darker shades of blue represent little to no deviation between the synthetically produced and the realistic constraint attributes, while lighter colors implicate a lower performance. We can qualitatively infer from these visuals that *IPF* and *CP* tend to produce a well-represented population for all constraint attributes. The metaheuristic algorithms *HC* and *SA* performed poorly, especially for secondary attributes. The lowest performing algorithm was the baseline (*PRWR*), which only constructed a few suitable constraint attributes. Compared to the Fairfax CT-level results, the Fairfax CBG R-squared values were higher for all algorithms, noticeably in *HC* and *SA*. While *PRWR* performed slightly better in Fairfax CT, the overall performance was still the worst.

In parallel with Fairfax regions, Figure 8 illustrates the R-squared values in the Metro Vancouver DA and CT. Like the Fairfax case, when limited to the same attribute set, *IPF* and *CP* performed well at the DA and CT levels. Qualitatively, *HC* and *SA* algorithms perform slightly better than their Fairfax region counterpart while not performing as well as *IPF* or *CP*. To test the reliability of the algorithms, we added more constraints, including knowledge of official languages, the official language of work, marital status, dwelling, and mobility, which were only available at the DA level in the Metro Vancouver dataset. The performance does not seem to be affected by these additional attributes. The difference between the algorithms' performance on the DAs and CTs is more noticeable. The baseline (*PRWR*) remains the worst-performing algorithm.

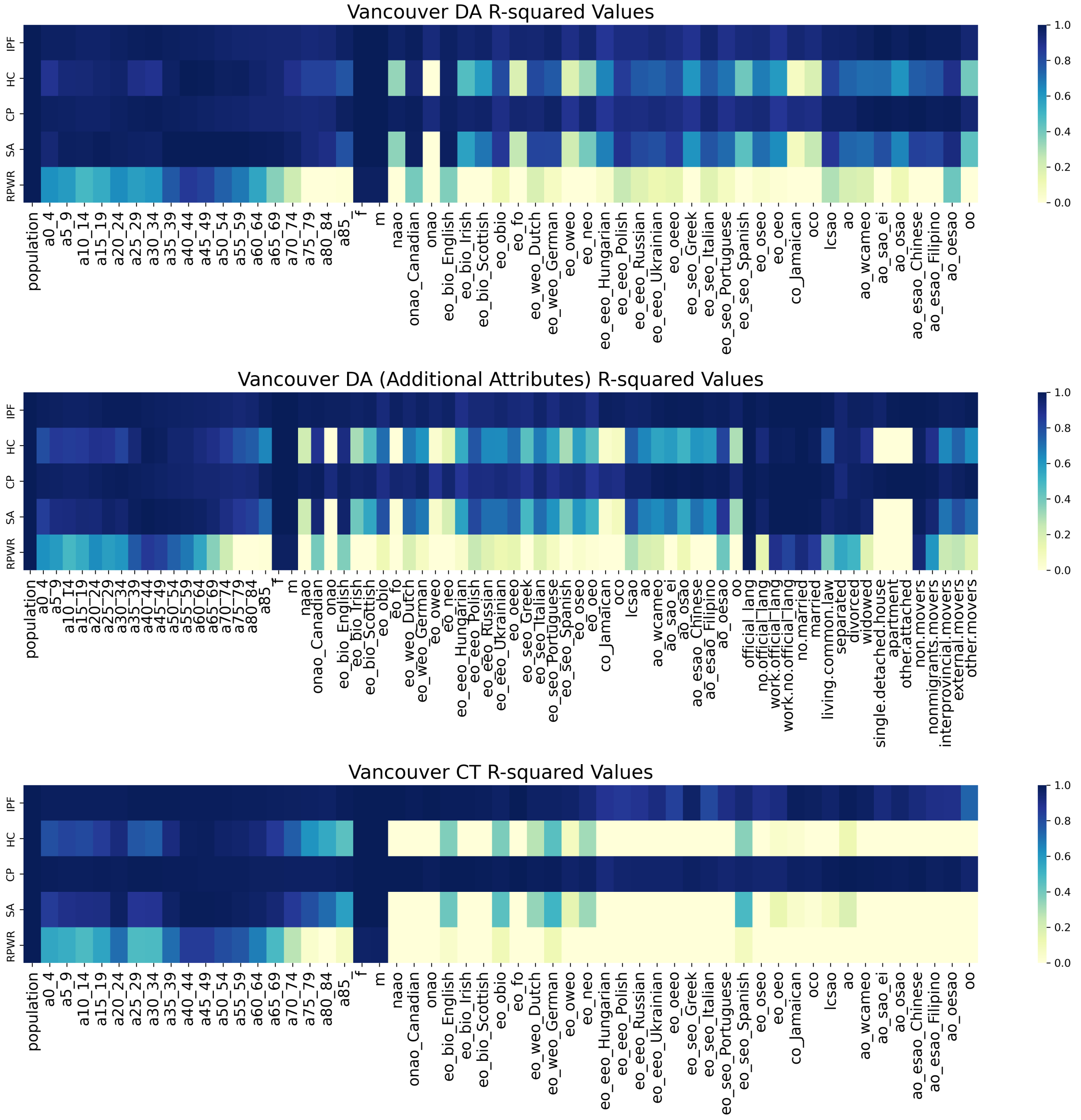


Figure 8: R-squared values of Metro Vancouver at the DA (upper panel), DA with additional attributes (middle panel), and CT (lower panel) levels.

Table 1 summarizes how each method performed per geographic level by comparing the average and standard deviations of R-squared values. These results numerically confirm the qualitative evaluation done per area and technique. IPF and CP performed nearly flawlessly across different geographic levels with low standard deviation. A noteworthy pattern is that HC and SA captured smaller but still positive R-squared values, while the Vancouver CT level had a negative R-squared value, as did RPWR, indicating a lack of fit. It is challenging to pinpoint the reason for this misfit by just considering the plots.

Table 1: Average R^2 values by geographic level and method (standard deviations in *italics*).

Geographic Level	**IPF**	**HC**	**CP**	**SA**	**RPWR**
Fairfax CBG	0.99 (*0.01*)	0.62 (*0.33*)	0.99 (*0.01*)	0.61 (*0.33*)	0.26 (*0.44*)
Fairfax CT	0.98 (*0.06*)	0.59 (*0.30*)	0.99 (*0.01*)	0.66 (*0.27*)	0.44 (*0.33*)
Vancouver CT	0.96 (*0.05*)	-17.75 (*65.79*)	0.98 (*0.02*)	-16.55 (*61.66*)	-22.06 (*80.52*)
Vancouver DA	0.95 (*0.03*)	0.73 (*0.28*)	0.95 (*0.04*)	0.77 (*0.28*)	0.23 (*0.38*)
Vancouver DA w/ Addl. Attrs	0.97 (*0.02*)	0.56 (*0.79*)	0.96 (*0.04*)	0.62 (*0.72*)	0.10 (*1.18*)

Next, we evaluate % TAE values across different variables to highlight the cause of the lack of fit. Table 2 and Table 3 showcases these results for Fairfax and Greater Vancouver, respectively. These results clearly show that IPF and CP consistently performed well, with the exception that IPF performed slightly lower for Greater Vancouver's DA and CT level attributes, Ethnicity, and Dwelling. When it comes to other attributes in Fairfax, SA, and HC have challenges capturing Age at CBG and CT levels, as well as Income, Citizenship, Profitability, and Race at the CT level. SA and HC have performed worse for certain variables, including ethnicity, for the Greater Vancouver area at all geographic levels. More specifically, the low R2 score was due to the variable "Ethnicity" at the CT level.

Table 2: % Total absolute error (% TAE) comparison by attribute for Fairfax County.

Fairfax County					
Attribute	**IPF**	**HC**	**CP**	**SA**	**RPWR**
CBG Level (% TAE)					
Population	0.00	0.00	0.00	0.00	0.00
Age	0.53	8.85	0.26	8.68	14.95
Gender	0.17	0.00	0.20	0.00	1.67
Race	1.85	3.11	0.12	3.56	32.78
Hispanic Descent	1.20	3.41	0.06	3.41	17.88
Mean	**0.75**	**3.07**	**0.13**	**3.13**	**13.46**
CT Level (% TAE)					
Population	0.00	0.00	0.00	0.00	0.00
Age	0.74	8.67	0.38	7.34	10.81
Gender	0.08	0.26	0.12	0.16	1.19
Employment	2.72	5.16	0.09	5.09	1.85
Location	0.16	1.34	0.04	0.70	3.68
Income	2.57	51.12	0.18	48.12	55.37
Citizenship	0.20	14.89	0.17	13.28	11.52
Marriage	0.57	6.70	0.13	5.25	9.44
School	0.54	0.94	0.14	0.83	3.03
Disability	0.06	0.36	0.02	0.06	1.11
Profitability	1.56	14.41	0.18	13.64	21.15
Insurance	0.20	7.43	0.09	5.92	9.26
Race	0.30	11.21	0.18	9.70	12.66
Mean	**0.74**	**9.58**	**0.13**	**8.37**	**10.66**

Table 3: % Total absolute error (% TAE) comparison by attribute for Greater Vancouver.

Greater Vancouver

Attribute	**IPF**	**HC**	**CP**	**SA**	**RPWR**
DA Level (% TAE)					
Population	0.09	0.00	0.00	0.00	0.00
Age	0.64	1.28	0.32	0.83	5.19
Gender	0.14	0.00	0.06	0.00	5.12
Ethnicity	1.71	19.44	0.29	18.71	51.16
Mean	**0.65**	**5.18**	**0.17**	**4.88**	**15.37**
DA w/ Addl' Attributes (% TAE)					
Population	0.00	0.00	0.00	0.00	0.00
Age	0.63	2.15	0.32	1.54	5.19
Gender	0.33	0.01	0.06	0.00	5.12
Ethnicity	2.44	25.36	0.29	24.30	51.16
Knowledge of Off. Languages	0.04	0.11	0.05	0.22	5.41
Official Language at Work	0.15	0.25	0.03	0.02	4.06
Marital Status	0.84	1.15	0.10	0.78	10.39
Dwelling	3.16	97.94	0.09	96.71	136.87
Mobility	0.80	3.15	0.04	2.82	6.63
Mean	**0.82**	**14.90**	**0.11**	**13.82**	**25.65**
CT Level (% TAE)					
Population	0.05	0.00	0.00	0.00	0.00
Age	1.26	2.91	0.30	1.98	5.29
Gender	0.05	0.02	0.09	0.00	5.06
Ethnicity	5.24	123.38	0.32	120.08	139.64
Mean	**1.65**	**31.58**	**0.18**	**30.52**	**37.50**

5 CONCLUSION

In this study, we generated synthetic populations with our own implementation of five popular techniques as introduced in section 2.3, in hopes of understanding their effectiveness through quantitative comparisons. As demonstrated in our methodology, we evaluate these approaches in Fairfax County, VA, USA, and Metro Vancouver, BC, Canada, concerning their geographic measures. The divisions in geographic regions reveal how these techniques perform with changes in area size. We note that our population generation techniques do not capture family structures such as households.

Through the quantitative comparison of these techniques, we provide a stronger basis for choosing certain methods over others for generating synthetic populations with regard to a geographic domain. Our data and results can be reproduced through a public code repository shared under the Notes section. We hope to make a greater contribution to the modeling and simulation field by providing a baseline for quantitatively comparing techniques with the prospect of finding more suitable ways to model individual behavior in synthetic populations. Additionally, our paper and code provide an entry point for new researchers interested in exploring synthetic population generation approaches.

NOTES

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 by the Aspiring Scientists Summer Internship Program (ASSIP) at George Mason University. All code and data are available to interested readers in our GitHub repository located at https://github.com/kavak-lab/synthetic-pop-comparison.

REFERENCES

Anderson, T., and S. Dragićević. 2020. "NEAT approach for testing and validation of geospatial network agent-based model processes: case study of influenza spread". *International Journal of Geographical Information Science* 34(9):1792–1821.

Beckman, R. J., K. A. Baggerly, and M. D. McKay. 1996. "Creating synthetic baseline populations". *Transportation Research Part A: Policy and Practice* 30(6 PART A):415–429.

Brachman, M. L., and S. Dragicevic. 2014. "A spatially explicit network science model for emergency evacuations in an urban context". *Computers, Environment and Urban Systems* 44:15–26.

Bureau, U. S. C. 2020a. *American Community Survey (ACS)*. https://www.census.gov/programs-surveys/acs, accessed 17 June.

Bureau, U. S. C. 2020b. *Decennial Census of Population and Housing*. https://www.census.gov/programs-surveys/decennial-census.html, accessed 17 June.

Canada, S. 2021a. *Census Program*. https://www12.statcan.gc.ca/census-recensement/index-eng.cfm, accessed 17 June.

Canada, S. 2021b. *Public Use Microdata Files*. https://www.statcan.gc.ca/eng/help/microdata, accessed 17 June.

Chapuis, K., and P. Taillandier. 2019. "A brief review of synthetic population generation practices in agent-based social simulation". In *Advances in Social Simulation - Proceedings of the 15th Social Simulation Conference*, edited by P. Ahrweiler and M. Neumann, 189–200. Mainz, Germany: Springer.

Choupani, A.-A., and A. R. Mamdoohi. 2016. "Population synthesis using iterative proportional fitting (IPF): A review and future research". *Transportation Research Procedia* 17(0):223–233.

D Orazio, M., L. Spalazzi, E. Quagliarini, and G. Bernardini. 2014. "Agent-based model for earthquake pedestrians evacuation in urban outdoor scenarios: Behavioural patterns definition and evacuation paths choice". *Safety Science* 62:450–465.

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* 11(4):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* 68:78–88.

Fabrice Yaméogo, B., P. Gastineau, P. Hankach, and P.-O. Vandanjon. 2020. "Comparing methods for generating a two-Layered synthetic population". *Transportation Research Record* 2675(1):136–147.

Farooq, B., M. Bierlaire, R. Hurtubia, and G. Flötteröd. 2013. "Simulation based population synthesis". *Transportation Research Part B: Methodological* 58(C):243–263.

Gallagher, S., L. F. Richardson, S. L. Ventura, and W. F. Eddy. 2018. "SPEW: Synthetic populations and ecosystems of the world". *Journal of Computational and Graphical Statistics* 27(4):773–784.

Geanakoplos, J., R. Axtell, J. D. Farmer, P. Howitt, B. Conlee, J. Goldstein, *et al*. 2012. "Getting at systemic risk via an agent-based model of the housing market". *American Economic Review* 102(3):53–58.

Harland, K., A. Heppenstall, D. Smith, and M. H. Birkin. 2012. "Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques". *Journal of Artificial Societies and Social Simulation* 15(1):1–15.

Heppenstall, A., A. Crooks, M. Batty, and L. See. (Eds.) 2012. *Agent-Based Models of Geographical Systems*. New York, NY: Springer.

Hu, J., J. P. Reiter, and Q. Wang. 2018. "Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data". *Bayesian Analysis* 13(1):183–200.

Huang, Z., and P. Williamson. 2001. "A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small area microdata". Technical report, Department of Geography, University of Liverpool, Liverpool, UK.

Kavak, H., J.-S. Kim, A. Crooks, D. Pfoser, C. Wenk, and A. Züfle. 2019, August. "Location-based social simulation". In *Proceedings of the 16th International Symposium on Spatial and Temporal Databases*, 218–221. Vienna, Austria.

Kim, J., and S. Lee. 2015. "A Reproducibility analysis of synthetic population generation". *Transportation Research Procedia* 6(0):50–63.

Kim, J.-S., H. Kavak, C. O. Rouly, H. Jin, A. Crooks, D. Pfoser, *et al*. 2020. "Location-based social simulation for prescriptive analytics of disease spread". *SIGSPATIAL Special* 12(1):53–61.

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* 2563(1):40–50.

Kotnana, S., D. Han, T. Anderson, A. Züfle, and H. Kavak. 2022, July. "Using generative adversarial networks to assist synthetic population creation for simulations". In *2022 Annual Modeling and Simulation Conference (ANNSIM)*, 1–12. San Diego State University, San Diego, CA, USA: IEEE.

Lovelace, R., and D. Ballas. 2013. "'Truncate, replicate, sample': A method for creating integer weights for spatial microsimulation". *Computers, Environment and Urban Systems* 41:1–11.

Lovelace, R., and M. Dumont. 2016. *Spatial microsimulation with R*. Boca Raton, FL, USA: CRC Press.

Malleson, N., A. Heppenstall, and L. See. 2010. "Crime reduction through simulation: An agent-based model of burglary". *Computers, Environment and Urban Systems* 34(3):236–250.

Mauša, G., T. G. Grbac, B. D. Bašić, and M.-O. Pav˘ević. 2013, July. “Hill climbing and simulated annealing in large scale next release problem”. In *Eurocon 2013*, 452–459. Zagreb, Croatia: IEEE.

Müller, K., and K. W. Axhausen. 2011. “Hierarchical IPF: Generating a synthetic population for Switzerland”. Arbeitsberichte Verkehrs- und Raumplanung 718, Eidgenössische Technische Hochschule Zürich, Institut für Verkehrsplanung und Transportsysteme (IVT), Zurich, Switzerland.

Orcutt, G. H. 1957. “A New Type of Socio-Economic System”. *The Review of Economics and Statistics* 39(2):116–123.

Pesavento, J., A. Chen, R. Yu, J.-S. Kim, H. Kavak, T. Anderson *et al.* 2020, November. “Data-driven mobility models for COVID-19 simulation”. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Advances in Resilient and Intelligent Cities*, 29–38. Seattle, WA, USA.

Piantadosi, S., D. P. Byar, and S. B. Green. 1988. “The ecological fallacy”. *American Journal of Epidemiology* 127(5):893–904.

Saadi, I., A. Mustafa, J. Teller, B. Farooq, and M. Cools. 2016. “Hidden Markov model-based population synthesis”. *Transportation Research Part B: Methodological* 90(3):1–21.

Sun, L., and A. Erath. 2015. “A Bayesian network approach for population synthesis”. *Transportation Research Part C: Emerging Technologies* 61:49–62.

Sun, L., A. Erath, and M. Cai. 2018. “A hierarchical mixture modeling framework for population synthesis”. *Transportation Research Part B: Methodological* 114:199–212.

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* 30(5):785–816.

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*. Paper 09-2096, 24 pages.

AUTHOR BIOGRAPHIES

DAVID HAN is currently a Masters of Engineering student studying Computer Science at Cornell University. He specializes in software engineering with a focus on distributed systems and machine learning. His email address is dmh338@cornell.edu and his website is https://david-han.dev/.

SAMIUL ISLAM is a Computational Sciences and Informatics PhD candidate in the Department of Computational and Data Sciences at George Mason University. His research focuses on county-level analysis of mental health and socioeconomic factors across the United States, aiming to identify improved determinants and extend existing studies using machine learning techniques. His email address is sislam22@gmu.edu.

TAYLOR ANDERSON is an Associate Professor in the Department of Geography and Geoinformation Science at George Mason University. Her research focuses on modeling the spread of diseases in human and ecological systems. Her e-mail address is tander6@gmu.edu and her website is https://science.gmu.edu/directory/taylor-anderson.

ANDREW CROOKS is a Professor in the Department of Geography at the University of Buffalo. His research interests include geographical information science and agent-based modeling. His email address is atcrooks@buffalo.edu and his website is https://www.gisagents.org/.

HAMDI KAVAK is currently Associate Professor in the Department of Computational and Data Sciences and Co-Director of Center for Social Complexity at George Mason University. His research combines data science with modeling and simulation to investigate challenges in urban and social systems. His email address is hkavak@gmu.edu and his website is https://hamdikavak.com/.