Paper No. 18-03141

An integration of population synthesis methods for agent-based microsimulation

Nicholas Fournier
(corresponding author)
University of Massachusetts
Amherst, Massachusetts 01003, USA
Email: nfournie@umass.edu

Eleni Christofa, Ph.D.
University of Massachusetts
Amherst, Massachusetts 01003, USA
Phone: +1 (413) 577-3016, Fax: +1 (413) 545-9569
Email: christofa@ecs.umass.edu

Arun Prakash Akkinepally, Ph.D.
Massachusetts Institute of Technology,
Cambridge, Massachusetts 02139, USA
Email: arunprak@mit.edu

Carlos Lima Azevedo, Ph.D.
Massachusetts Institute of Technology,
Cambridge, Massachusetts 02139, USA
Phone: +1 (617) 452-2482, Fax: +1 (617) 253-1130
Email: cami@mit.edu

An extended abstract submitted to the
Transportation Research Board, 97th Annual Meeting
Washington, D.C.
January 7–11, 2018
Word Count: 1,680

ACKNOWLEDGMENTS

This research was funded in part through the United States Department of Energy Advanced Research Projects Agency-Energy (ARPA-E) under the project Traveler Response Architecture using Novel Signaling for Network Efficiency in Transportation (TRANSNET). The authors would like to thank the Boston Metropolitan Planning Organization and the Boston Metropolitan Planning Organization's Central Transportation Planning Staff for providing access to their GIS files

Revised November 15, 2017

INTRODUCTION

Agent-based microsimulation has become a mainstay in transportation and land-use planning, using an ever growing array of large-scale modeling platforms such as MATSim (1), UrbanSim (2), SimMobility (3), ILUTE (4, 5), and MUSSA (6). These models can inform a variety of decisions, such as policy changes or investment structures. The underlying foundation of these models is an accurate disaggregated dataset of individual agents for which activity-based models can be estimated from. This means that the accuracy of the population is absolutely paramount. Unfortunately, a complete disaggregate data set of persons in a population is not available either due to privacy constraints or costs. To resolve this problem, transport modelers typically generate a synthetic population using available data sources, such as population samples and census totals (7).

Over the past three decades, population synthesis for microsimulation has made many significant advances. Much of this can be attributed to advances in computational power, but many methods have broken off into new branches entirely, such as simulation-based synthesis (8, 9) or land-use models (10, 11). This is due to the varying quantity and type of data available in different locations. As a result, there is no definitive method superior in all regards, but rather an array of problem solving techniques each developed to address a particular challenge.

This research presents a process using a combination of methods for when both aggregated and disaggregated data are available. The data used to generate the population is from the United States (U.S.) Census Bureau, making the framework transferable to anywhere in the United States or where similar census data is available. The process itself is based on Iterative Proportional Fitting (IPF) of households and persons which are reweighted using Iterative Proportional Updating (IPU) in a sparse matrix of household-person groups. The process includes accurate integerization techniques for improved computation performance and a multi-level region seeding technique to better address the zero cell problem. The population in this paper is generated for the Greater Boston Metropolitan Area (GBA) of approximately 4.6-million persons and 1.7-million households.

METHODOLOGY

The synthesis framework builds on several of the methods, techniques, and approaches that have been introduced thus far in the field of population synthesis. The framework can be broken into five steps:

1. *Seeding algorithm*
2. *Iterative Proportional Fitting*
3. *Integerization*
4. *Iterative Proportional Updating*
5. *Monte Carlo Sampling*

1. Seeding Algorithm

Creating the seed data in this population synthesis framework uses the U.S. Census Bureau's Public Use Microdata Sample (PUMS) (12). In order to try and capture spatial differences in the population, a seed is created for each of the smaller census tracts by using data from a larger Public Use Microdata Areas (PUMAs) that the census tracts are located in. However, the PUMS is roughly a five percent sample of the population, which is likely to result in sampling zeros. This is done by first creating a seed using the PUMA which a census tract is within (13, 14). Then if any zero cells exist they are replaced by borrowing proportionally adjusted representatives from the entire PUMS. This method preserves structural zeros in the microdata, but also helps retain the relative proportions of the PUMA without introducing sampling zeros.

2. Iterative Proportional Fitting

Iterative Proportional Fitting (IPF), first introduced by Deming and Stephan (15, 16), is used to proportionally fit the cells in an *n*-dimensional contingency table when the marginal totals are known in an iterative process.

3. Integerization

The results of IPF typically contain many possible combinations of person or household type cells. This creates a problem for the following step of Iterative Proportional Updating (IPU), which treats these cells as margins, causing the computational requirements to grow enormously. Fortunately, many of the IPF zero or near-zero cells can be eliminated before IPU begins, reducing the table size. Ye et al. (13) solved this issue by simply rounding the IPF results, but Lovelace et al. (19) showed that this disproportionately eliminates rare person or household types in the population (i.e. all cell frequencies of less than 0.5). The Truncate, Replicate, and Sample (TRS) method mitigates this concern by proportionally sampling the cells using the decimal values as weights. The TRS method achieves the original goal of eliminating excessive near-zero cells, while mitigating the concern for preserving the population proportions as much as possible for IPU.

4. Iterative Proportional Updating

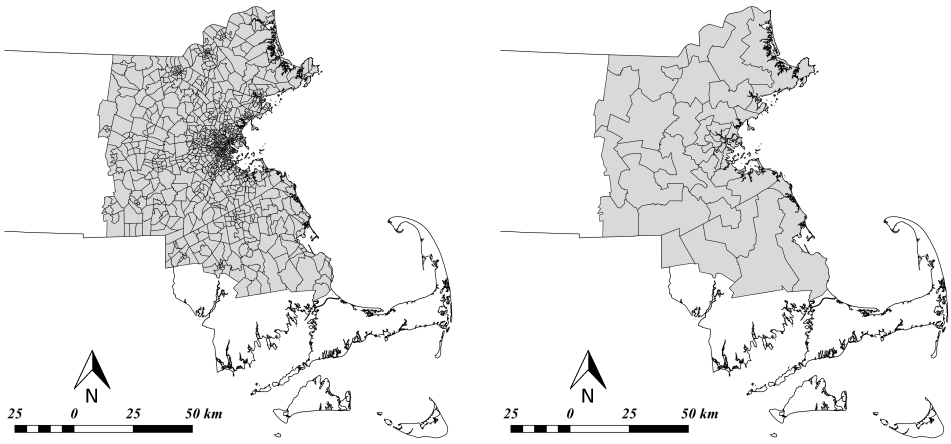
IPU is a very computational intensive step of the synthesis process. In addition to integerizion, the joint table is transformed into a sparse data matrix containing only non-zero values, rather than storing the entire matrix in memory. This was accomplished using the open source 'Matrix' package for the R programming language (20). Furthermore, the IPU process was coded in C++, taking advantage of its inherent speed and memory efficiency benefits.

5. Monte-Carlo Sampling

In order to synthesize a full population, households are drawn randomly from a joint sample and replicated along with their joint person members into a synthetic population. Each census tract has a unique set of household weights determined by the IPU for each tract. The random drawing is performed for each census tract with *n* households drawn from the sample, where *n* is the total number of households in the tract. This random drawing is repeatedly performed, checking each sample to the expected IPF distribution using a root mean square error check, keeping the best fit sample.

Case Study Region: The Greater Boston Area

The study region is the Greater Boston Area as delineated by the Boston Metropolitan Planning Organization (MPO). A synthetic population of 4.6-million individuals and 1.7-million households are generated for the Greater Boston Area (GBA), across 960 census tracts shown in Figure 1(a) and the larger Public Use Microdata Areas (PUMAs), shown in Figure 1(b). Geospatial information systems (GIS) shapefiles provide the spatial geometries for the region, available from MassGIS (21) and the Boston MPO (22).



(a) Census tracts (b) Public Use Microdata Areas (PUMAs)

FIGURE 1 Boston Metropolitan Area

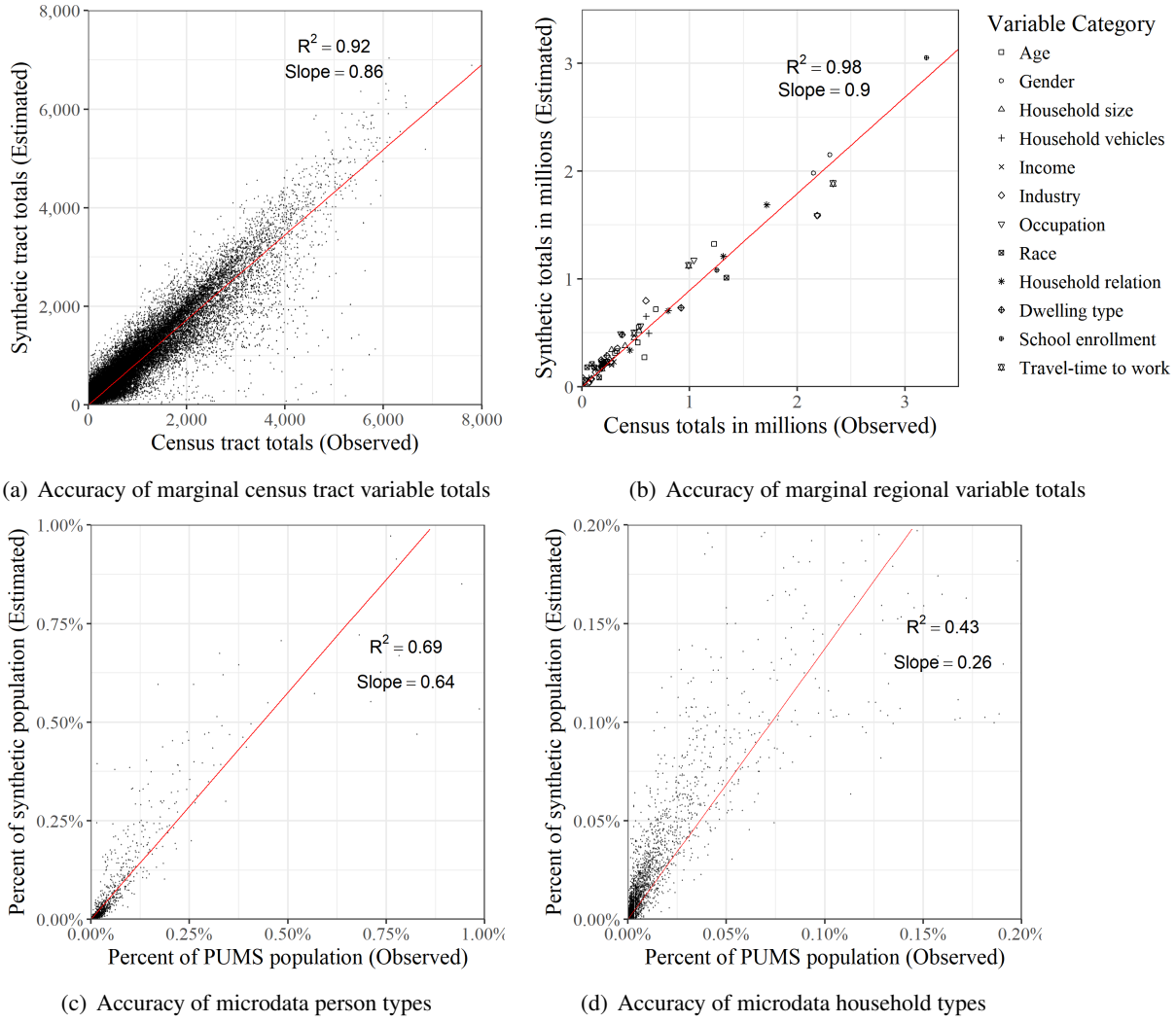
This framework utilized five household control variables and eight person control variables, shown in Table 1. The data for these variables are obtainable in comma separated values (CSV) file type format from the U.S. Census Bureau (12, 17, 18). A synthesis year of 2010 was used as the data year for all tables in order to be consistent with the most recent decennial census.

TABLE 1 Control variables

Household							
Size	Vehicles	Annual Income	Dwelling type	Race			
1	0	<$15k	1 unit	Hispanic or latino			
2	1	$15k-$25k	2-4 units	Black			
3	2	$25k-$35k	5-19 units	Native American			
≥4	3	$35k-$50k	≥20 units	Asian			
	≥4	$50k-$75k		Pacific islander			
		$75k-$100k		White			
		$100k-$150k		Multi			
		>$150		Other			
Person							
Sex	Age	Work hours	School enrollment	Relationship	Travel time (mins)	Industry	Occupation
Male	0-9	0	Yes	Head	0	None	None
Female	10-14	1-34	No	Spouse	1- 14	Nat. res. extraction	Mgmnt./business/science/arts
	15-19	>34		Child	15-34	Trans. and Utilities	Sales/office/admin.
	20-24			Relative	45-59	Construction	Nat. res./const. maint.
	25-44			Nonrelative	>59	Manufacturing	Production/trans.
	45-54					Wholesale trade	Service
	55-64					Retail trade	
	>65					Information	
						Finance/real estate	
						Prof./science/mgmnt.	
						Educ./social work	
						Arts/accommodation	

FINDINGS

The synthetic population is validated at two levels, at the marginal and microdata level using the Normalized Root Mean Square Error (NRMSE). NRMSE is a common measure used in population synthesis and is essentially a simple Root Mean Square Error (RMSE) that is normalized by the mean expected value. NRMSE results that are below 1.0 are considered satisfactory and 0 is perfect. The purpose of normalizing RMSE is that the error can grow very large in situations where there are many values being compared. Overall, the synthetic population achieved excellent results at the marginal level, but error was high at the microdata level, shown in Table 2 and Figure 2.



(a) Accuracy of marginal census tract variable totals

(b) Accuracy of marginal regional variable totals

(c) Accuracy of microdata person types

(d) Accuracy of microdata household types

FIGURE 2 Comparison of synthetic and observed data

TABLE 2 Validation results

Comparison	R^2	*Slope*	*NRMSE*
Marginal regional totals	0.98	0.90	0.27
Marginal tract totals	0.92	0.86	0.47
Microdata Household types	0.43	0.26	6.61
Microdata Person types	0.69	0.64	6.79

CONCLUSIONS

The proposed population synthesis framework utilizes both established methods, such as Iterative Proportional Fitting and Monte-Carlo sampling, with the newer innovative methods of Truncate, Replicate, Sample and Iterative Proportional Updating. The proposed process begins by generating IPF seeds using a proportionally adjusted seed algorithm, reducing the risk of losing structural zeros or the risk of introducing sampling zeros. The IPF results are then preprocessed for IPU by using TRS to integerize the fitted results. This dramatically reduces memory requirements for IPU, rather than conventional approaches of limiting the number of variables synthesized or crudely rounding IPF results. The combined approach allows for population synthesis with many variables, while minimizing loss of accuracy, with computation for the entire process of under two hours (1.9-hours).

Although there is room for improvement, the overall population synthesis framework was able to achieve accurate results, maintain relatively low computational time and resources, and incorporate population variables at or above typical numbers. The framework is also easily transferable to anywhere within the United States or other locations where similar data are available. The next steps are to integrate origin-destination distribution directly into syntheses and to improve the joint re-weighting step of IPU with a computationally more efficient method.

REFERENCES

[1] Balmer, M., M. Rieser, K. Meister, D. Charypar, N. Lefebvre, and K. Nagel, MATSim-T: Architecture and simulation times. In *Multi-agent systems for traffic and transportation engineering*, IGI Global, 2009, pp. 57–78.

[2] Waddell, P., UrbanSim: Modeling urban development for land use, transportation, and environmental planning. *Journal of the American planning association*, Vol. 68, No. 3, 2002, pp. 297–314.

[3] Adnan, M., F. C. Pereira, C. M. L. Azevedo, K. Basak, M. Lovric, S. Raveau, Y. Zhu, J. Ferreira, C. Zegras, and M. Ben-Akiva, SimMobility: A Multi-scale Integrated Agent-based Simulation Platform. In *95th Annual Meeting of the Transportation Research Board Forthcoming in Transportation Research Record*, 2016.

[4] Salvini, P. and E. J. Miller, ILUTE: An operational prototype of a comprehensive microsimulation model of urban systems. *Networks and Spatial Economics*, Vol. 5, No. 2, 2005, pp. 217–234.

[5] Wagner, P. and M. Wegener, Urban land use, transport and environment models: Experiences with an integrated microscopic approach. *disP-The Planning Review*, Vol. 43, No. 170, 2007, pp. 45–56.

[6] Martinez, F. and P. Donoso, The MUSSA II land use auction equilibrium model. In *Residential Location Choice*, Springer, 2010, pp. 99–113.

[7] Müller, K. and K. W. Axhausen, Population synthesis for microsimulation : State of the art. *90th Annual Meeting of the Transportation Research Board*, , No. August, 2011, p. 21.

[8] Farooq, B., M. Bierlaire, R. Hurtubia, G. Flotterod, and G. Flötteröd, Simulation based population synthesis. *Transportation Research Part B: Methodological*, Vol. 58, 2013, pp. 243–263.

[9] Casati, D., K. Müller, P. J. Fourie, A. Erath, and K. W. Axhausen, Synthetic Population Generation by Combining a Hierarchical, Simulation-Based Approach with Reweighting by Generalized Raking. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2493, 2015, pp. 107–116.

[10] Hensher, D. A. and T. Ton, TRESIS: A transportation, land use and environmental strategy impact simulator for urban areas. *Transportation*, Vol. 29, No. 4, 2002, pp. 439–457.

[11] McBride, E. C., A. W. Davis, J. H. Lee, and K. G. Goulias, Incorporating Land Use into Methods of Synthetic Population Generation and of Transfer of Behavioral Data. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2668, 2017, pp. 11–20.

[12] U.S. Census Bureau, *2011-2015 5-year American Community Survey Public Use Microdata Sample*, 2011-2015.

[13] Ye, X., K. Konduri, R. M. Pendyala, B. Sana, and P. Waddell, 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*, 2009.

[14] Guo, J. and C. Bhat, Population synthesis for microsimulating travel behavior. *Transportation Research Record: Journal of the Transportation Research Board*, , No. 2014, 2007, pp. 92–101.

[15] Deming, W. E. and F. F. Stephan, On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *The Annals of Mathematical Statistics*, Vol. 11, No. 4, 1940, pp. 427–444.

[16] Stephan, F. F., An Iterative Method of Adjusting Sample Frequency Tables When Expected Marginal Totals are Known. *The Annals of Mathematical Statistics*, Vol. 13, No. 2, 1942, pp. 166–178.

[17] U.S. Census Bureau, *5-year American Community Survey Tables*, 2010.

[18] U.S. Census Bureau, *2010 Decennial Census Tables*, 2010.

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

[20] Bates, D. and M. Maechler, Matrix: sparse and dense matrix classes and methods. *R package version 0.999375-43, URL http://cran. r-project. org/package= Matrix*, 2010.

[21] Massachusetts Office of Geographic Information (MassGIS), *Geographic Census/Demographic Data*, 2010.

[22] Boston Region Metropolitan Planning Organization Central Transportation Planning Staff, *Traffic Analysis Zones*, 2012.