Downscaling Synthetic Populations to Realistic Residential Locations

Joseph V. Tuccillo
Geospatial Science and Human Security Division
Oak Ridge National Laboratory
Oak Ridge, TN, United States
tuccillojv@ornl.gov

***Abstract*—High-fidelity pattern of life (PoL) models require realistic home/origin points for predictive trip modeling. This paper develops and demonstrates a method using open data to match synthetic populations generated from census surveys to plausible residential locations (building footprints) based on dwelling attributes. This approach presents promise over extant methods based on housing density, particularly for small neighborhood areas with a non-uniform building and land-use mix.**

***Index Terms*—census, synthetic population, housing, pattern of life**

I. Introduction

Agent-based pattern of life (PoL) models seek to enhance understanding of spatial accessibility and human behavior. PoL models require realistic residential anchor points for modeling origin-destination trajectories to routine activities including work and school, as well as non-obligate social, civic, and recreational activities. Further, incorporating residential information situates agents in plausible locations relative to different modes of travel (e.g., high-density apartments adjacent to a transit stop in a suburban neighborhood).

Oak Ridge National Laboratory's (ORNL) UrbanPop project [1] serves as a baseline for PoL modeling on curated point of interest (POI) data [2]. UrbanPop leverages the American Community Survey (ACS) and its Public-Use Microdata Sample (PUMS) to produce attribute-rich synthetic populations characterized by demographic, economic, housing, and mobility factors [1]. Likeness, the Python software stack supporting UrbanPop, includes utilities for agent creation, network-based routing, and activity allocation [3]. A key facet of Likeness's development to date has been formulating a residential downscaling capability for agent trip generation. This provides home locations for agents beneath the census block group scale (600 - 3000 people) at which they are natively produced.

Contrasted with existing downscaling methods that rely largely on housing density [4], those in Likeness attempt to place agents in plausible locations based on available housing stock within their assigned block groups. This is achieved by conflating dwelling type labels between synthetic households and building footprint data. The remainder of this paper outlines and demonstrates the current Likeness downscaling routine and concludes with outlook for its future development.

II. Methods

The basic premise of the Likeness downscaling routine is that synthetic households in each block group "select" a building in which to reside from a local supply of residential structures. Synthetic household types include one-unit dwellings like single-family detached housing and mobile homes, multi-family residences, and group quarters (e.g., college dormitories, nursing homes). While residential structures are currently sourced from the Federal Emergency Management Agency's (FEMA) open USA Structures inventory [5], other building footprint data may also be used.

A. Conflating Synthetic Households with Residential Structures

While downscaling is carried out differently for each dwelling type, the general task for a household of any dwelling type is to 1) subset matching residential structures (e.g., synthetic "Single-Family Residential Detached" household $\rightarrow$ USA Structures "Single-Family Dwelling"), then 2) score compatibility by conflating features of households and structures, h for some household attribute (e.g., income) and s for some structure attribute (e.g., building floor area).

The conflation score between household attribute h and a structure attribute S is computed as

$$\Gamma = \frac{1}{exp(|h - s|)} \Big| \, s \in S \tag{1}$$

such that higher values of Γ reflect smaller differences between h and s. Next, convert Γ to a sampling probability as

$$p = \frac{\Gamma}{\sum_1^N \Gamma} \tag{2}$$

B. Downscaling Routine

The Likeness downscaling routine involves 1) matching synthetic households to structures labeled by specific dwelling types (single-family detached residential, mobile homes, institutional/non-institutional group quarters), then 2) matching any multi-family residential and unmatched households from (1) from within a general housing pool to multi-unit structures.

- **One-Unit Housing.** Synthetic households labeled as "Single-Family Residential Detached" and "Mobile Home" are first sampled one-to-one into available USA

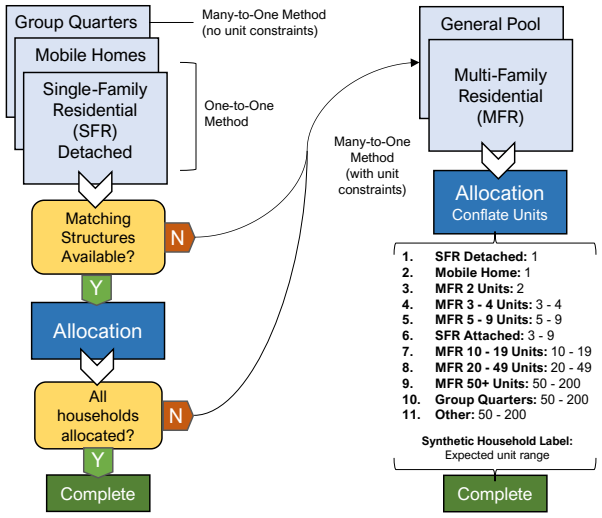

Fig. 1. Workflow for the Likeness downscaling routine.

Structures labeled as "Single-Family Dwelling" and "Manufactured Home" (respectively) weighted by a combination of synthetic household income and building floor area. In Equation 1, let h represent household income and let s represent building floor area, converted to ranks for compatibility with the scoring procedure described in Equation 1.

- **Group Quarters.** For group quarters, both non-institutional (college dormitories, military) and institutional (nursing homes/assisted living facilities, prisons) households are matched to available structures based on building floor area alone (many-to-one, no capacity constraints), assuming units of uniform size. In Equation 1, let $h = 0$ and let s represent building floor area.
- **Multifamily Residential and General Housing Pool.** Finally, households in multifamily dwellings, as well as unassigned one-unit/group quarters households, are allocated (many-to-one, with capacity constraints) to USA Structures of "Multi-Family Dwelling" type. This is achieved by conflating expected unit counts from household dwelling labels with unit counts per structure estimated from facility occupancy models [6], [7], following the row order in Figure 1. In Equation 1 let h represent a number of housing units drawn from the ranges in Figure 1, and let $s \in S$ represent estimated unit counts based on building floor area. Unit capacities per structure are initially set as $U = S$, then decremented with each household assignment. If $\sum U$ is less than the total number of synthetic households, n, available units in multi-family structures are incremented based on a weighted random sample of building floor area until parity is reached with n. This procedure only modifies U and not S. To avoid placing high unit count households in small structures, a subtask is performed during household matching to 1) check for any candidate structures with unit counts less than the lower bound of the target household's unit range, then 2) omit those structures and reassign any remaining available units to larger ones, assuming that the recipient structures have higher housing density.

III. Illustration

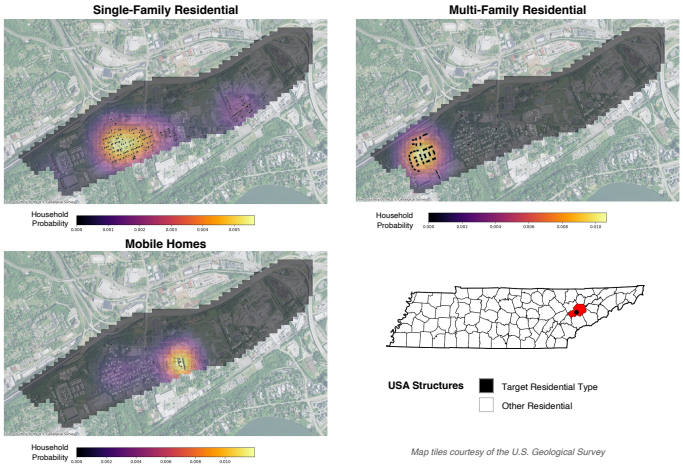

Fig. 2. Comparison of synthetic household occurrence probabilities to USA Structures types for a single block group.

An initial proof of concept for the Likeness downscaling routine was developed for the Knoxville, TN metropolitan statistical area (MSA) based on the ACS 2019 5-Year Estimates. A synthetic population was generated from ACS variables linked to household composition and socioeconomic status, as well as built environment and housing characteristics.

Figure 2 compares downscaled synthetic households to residential USA Structure locations for a single block group. The probability density estimates provide an aggregate representation of household locations in the synthetic population, which are compared to observed single-family, multi-family, and mobile homes. This illustration reveals how household placement closely aligns with the building and land-use mix within the block group.

IV. Conclusion and Outlook

While the Likeness downscaling routine is preliminary, it presents promise for producing more realistic PoL simulations by increasing precision in the synthesis of agent home locations. The home synthesis module included in Likeness is written to be general-purpose, supporting both custom building footprints and modifications of the downscaling routine demonstrated in Sections II and III.

A key limitation of this approach lies in problems of uneven allocation (e.g., sparse or empty structures). This occurs largely because synthetic populations and USA Structures are from disparate sources. Ancillary data sources like the U.S. Department of Transportation's National Address Database[1] may assist in model calibration as their spatial coverage increases.

[1]https://www.transportation.gov/gis/national-address-database

ACKNOWLEDGMENT

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy (DOE). The U.S. government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (https://energy.gov/downloads/doe-public-access-plan).

Supported by the Intelligence Advanced Research Projects Activity (IARPA). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOE, or the U.S. Government.

REFERENCES

[1] J. Tuccillo, Stewart R., Rose A., Trombley N., Moehl J., Nagle N., and Bhaduri B, "UrbanPop: A spatial microsimulation framework for exploring demographic influences on human dynamics," *Applied Geography*, vol. 151, pp. 102844, February 2023.

[2] Thakur, G.S., Bhaduri, B.L., Piburn, J.O., Sims, K.M., Stewart, R.N. and Urban, M.L., "PlanetSense: a real-time streaming and spatio-temporal analytics platform for gathering geo-spatial intelligence from open source data." In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 1–4, November 2015.

[3] J. Tuccillo, and J. Gaboardi, "Likeness: a toolkit for connecting the social fabric of place to human dynamics." In *Proceedings of the 21st Python in Science Conference*, pp. 125–135, Meghann Agarwal, Chris Calloway, Dillon Niederhut, and David Shupe, Eds., July 2022.

[4] A. Burger, Oz, T., Crooks, A., and Kennedy, W.G., "Generation of realistic mega-city populations and social networks for agent-based modeling," In *Proceedings of the 2017 International Conference of The Computational Social Science Society of the Americas*, pp 1–7, October 2017.

[5] H.L. Yang, Lunga, D., and Yuan, J., "Toward country scale building detection with convolutional neural network using aerial images." In *In 2017 IEEE International Geoscience and Remote Sensing Symposium*, pp. 870–873, July 2017.

[6] M. Urban, Stewart, R., Basford, S., Palmer, Z., and Kaufman, J, "Estimating building occupancy: a machine learning system for day, night, and episodic events," *Natural Hazards*, vol. 116, pp. 2417–2436, January 2023.

[7] M. Deru, Field, K., Studer, D., Benne, K., Griffith, B., Torcellini, P., Liu, B., Halverson, M., Winarski, D., Rosenberg, M., Yazdanian, M., Huang, J., and Crawley, D., "US Department of Energy commercial reference building models of the national building stock," National Renewable Energy Laboratory, 2011.