Australasian Transport Research Forum 2024 Proceedings
27-29 November, Melbourne, Australia
Publication website: https://australasiantransportresearchforum.org.au/

A pool-based approach to population synthesis in transport modeling

Duc Minh La[1], Hai L. Vu[1]
[1]Institute of Transport Studies, Department of Civil Engineering, Monash University
Email for correspondence (presenting author): duc.la@monash.edu

1 Introduction

Population synthesis (PopSyn) is an essential component in transport modeling, particularly in activity-based models (ABM) that require detailed and realistic representations of individuals and households. These models enable transport planners to simulate and forecast mobility patterns, which are crucial for infrastructure development and policy planning. PopSyn typically relies on two types of input data: aggregated marginal data, such as census statistics, and disaggregated seed data from surveys. The challenge lies in reconciling these two data sources, which often exhibit inconsistencies due to sampling biases, privacy constraints, and data sparsity (Ma and Srinivasan 2015).

Traditional approaches to PopSyn, such as Iterative Proportional Fitting (IPF) (e.g., (Beckman et al. 1996; Ye et al. 2009)) and Bayesian Networks (e.g., (Ilahi and Axhausen 2019; Sun and Erath 2015), attempt to align synthesized populations with both data sources but suffer from several limitations. IPF, for example, struggles with the curse of dimensionality and the zero-cell problem, where some combinations of attributes are entirely missing in the sample data. Bayesian Networks, while more flexible, can produce implausible combinations that do not accurately represent real-world populations and cannot match well with the marginal data (Rahman and Fatmi 2023). These shortcomings highlight the need for a more robust methodology that can better integrate diverse data sources and maintain realistic population characteristics.

2 Methodologies

To address these challenges, we propose a novel methodology called Sequential Attribute Adjustment (SAA), which utilizes a pool-based approach to enhance the accuracy and explainability of population synthesis. Unlike traditional methods that directly adjust weights or probabilities, SAA constructs a pool of synthetic agents from seed data and then iteratively adjusts each attribute to match the marginal data while maintaining the original distribution of the seed data. This approach ensures that each attribute adjustment is consistent with previously adjusted attributes, thereby preserving the interdependencies between different characteristics of the synthetic population. Furthermore, the pool helps boost the reliability and explainability of the solution as one can directly check and manipulate the pool (e.g., remove implausible combinations).

The pool-based strategy allows for the integration of various data sources and improves the model's flexibility and robustness. By leveraging a generative model, such as a Bayesian Network, to create the initial pool, SAA can effectively address the zero-cell problem by generating plausible attribute combinations that may not exist in the seed data. This makes it possible to synthesize populations that are more comprehensive and representative of the actual population distribution, even in cases where certain data points are missing or underrepresented.

3 Main Findings

To validate the effectiveness of the Sequential Attribute Adjustment (SAA) method, we conducted a case study focused on household synthesis at the postcode level (POA) for the state of Victoria, Australia. The data used for the analysis consisted of both aggregated and disaggregated sources. The aggregated data was derived from the 2021 Census conducted by the Australian Bureau of Statistics (ABS), while the disaggregated seed data came from the Victorian Integrated Survey of Travel and Activity (VISTA), collected between 2012 and 2018. This case study synthesized five key household attributes across all 691 POAs in Victoria: household size (hhsize), total vehicles (totvehs), household income (hhinc), dwelling type (dwelltype), and dwelling ownership status (owndwell).

The performance of the SAA method was evaluated against two established PopSyn techniques: Iterative Proportional Fitting (IPF) (Beckman et al. 1996) and Bayesian Networks (BN) (Sun and Erath 2015). Given IPF's well-known challenges with integerization, the Truncate, Replicate, Sample (TRS) method (Lovelace and Ballas 2013) was applied to manage this issue. To evaluate the performance of SAA against IPF and BN, two metrics were employed: the Root Mean Squared Error (RMSE) to assess accuracy in matching marginal data, and the Jensen-Shannon Divergence (JSD) to compare the distributional fit of synthesized data to the seed data. SAA applied sequential adjustments to the attributes in the following order: hhsize, totvehs, hhinc, dwelltype, and owndwell.

The results demonstrated that SAA outperformed both IPF and BN across all attributes. For example, in the attribute of total vehicles, SAA achieved an RMSE of 2.31, significantly lower than IPF's 8.81 and BN's 208.46, indicating a more precise alignment with the real-world distribution. SAA also maintained an acceptable JSD score in most cases, indicating the ability to preserve the original seed data distributions. Notably, the SAA method was the only approach that successfully addressed the zero-cell problem, generating realistic data for categories such as households with negative income, which were absent in the seed data.

Overall, the SAA method demonstrated a substantial improvement over traditional techniques in terms of both accuracy and robustness. Its ability to integrate diverse data sources and generate plausible synthetic populations makes it a promising tool for future transport modeling applications.

4 Conclusion and future works

The proposed Sequential Attribute Adjustment (SAA) method offers a significant advancement in population synthesis for transport modeling. By adopting a pool-based approach, SAA effectively addresses the limitations of traditional methods, including the zero-cell problem and the challenge of balancing multiple data sources. The case study in Victoria, Australia, highlighted SAA's superior performance in accurately replicating population attributes and preserving the distributional characteristics of the seed data. These results suggest that SAA is not only a viable alternative to existing PopSyn methods but also a substantial improvement, offering greater accuracy, flexibility, and applicability in complex modeling scenarios.

Future work will focus on extending the SAA framework to incorporate additional attributes and enhancing its scalability for large-scale applications. Integrating domain-specific knowledge into the pool creation process could further refine the method's accuracy and reduce computation time. With these enhancements, SAA can become a standard population synthesis approach in transport and urban planning.

5 Acknowledgements

Duc Minh La's research is sponsored by the SPARC Hub at Monash University and funded by the Australian Research Council (ARC) Industrial Transformation Research Hub Scheme (Project ID: IH180100010). The financial and in-kind support of Monash University is also gratefully acknowledged.

References

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

Ilahi, A. and Axhausen, K. W. (2019), 'Integrating Bayesian network and generalized raking for population synthesis in Greater Jakarta', *Regional Studies, Regional Science* **6**(1), 623–636. Publisher: Routledge.

Lovelace, R. and Ballas, D. (2013), ''Truncate, replicate, sample': A method for creating integer weights for spatial microsimulation', *Computers, Environment and Urban Systems* **41**, 1–11. arXiv: 1303.5228.

Ma, L. and Srinivasan, S. (2015), 'Synthetic population generation with multilevel controls: A fitness-based synthesis approach and validations', *Computer-Aided Civil and Infrastructure Engineering* **30**(2), 135–150.

Rahman, M. N. and Fatmi, M. R. (2023), 'Population Synthesis Accommodating Heterogeneity: A Bayesian Network and Generalized Raking Technique', *Transportation Research Record: Journal of the Transportation Research Board* p. 036119812211442.
URL: *http://journals.sagepub.com/doi/10.1177/03611981221144289*

Sun, L. and Erath, A. (2015), 'A Bayesian network approach for population synthesis', *Transportation Research Part C: Emerging Technologies* **61**, 49–62. Publisher: Elsevier Ltd.
URL: *http://dx.doi.org/10.1016/j.trc.2015.10.010*

Ye, X., Konduri, K. C., Pendyala, R. M., Sana, B. and Waddell, P. (2009), 'Methodology to Match Distributions of Both Household and Person Attributes in Generation of Synthetic Populations', *Transportation Research Board Annual Meeting 2009* **9601**(206), 1–24.