Novel framework to generate a synthetic population with diversities in transport modelling

DM. La[a,*], HL. Vu[a]

[a] Civil Engineering Dept., Monash University, Melbourne, Australia
Duc.La@Monash.Edu, Hai.Vu@Monash.Edu
[*] Corresponding author

Extended abstract submitted for presentation at the Conference in Emerging Technologies in Transportation Systems (TRC-30)
September 02-03, 2024, Crete, Greece

April 28, 2024

Keywords: Transport Modelling, Population Synthesis, Activity-based Modelling

1 Introduction

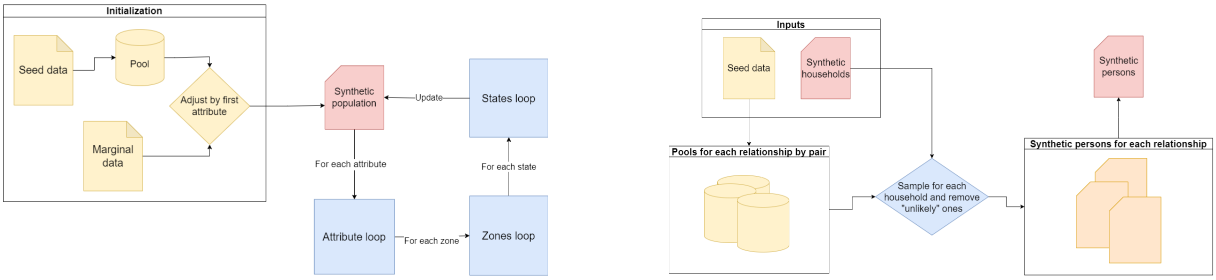
Population synthesis, or PopSyn in short, is a module to realise the full details of all individuals (or agents) to input into Activity-based modelling (ABM) models. It aims to create a realistic population using the limited available data. The realisticity of a synthetic population depends on its diversity compared to the real population. Therefore, maintaining diversities in the synthesized population is the key to a robust and reliable PopSyn. Typically, there are two types of available data for PopSyn: aggregated and disaggregated. Aggregated data (i.e., marginal data) is the aggregated numbers of the population's characteristics in a zone or region which can often be found in census data. Disaggregated data (i.e., seed data) is the detailed data of some individuals with their daily activities and travels found in travel surveys. Generally, the seed data only accounts for around 1% up to 5% of the actual population (Sun & Erath, 2015).

Three main types of diversity exist in the population of a city or large region. The first type of diversity is the unique mobility patterns or spatial traits among different areas or zones in a network. For example, people in regional areas will be more car-reliant than those in urban areas where public transport is more accessible. The next type of diversity is the household-person connection where a household consists of individuals living in the same dwelling. For instance, a household with cars should have people with driving licenses. The last type of diversity is the relationships within a household such as parent and their children or husband and wife etc. This type of diversity is not well-researched in transport modelling (Ho & Mulley, 2015). Generally, the spatial diversity is reflected in the marginal data while the other two are in the seed data. Balancing all 3 types of diversity to create the synthesized population is challenging and still an open question in the current literature.

Existing approaches have attempted to tackle the above diversity problems. A telling example is iterative proportional updating (IPU) from Ye *et al.* (2009). It is a popular method that can heuristically synthesize households and persons for each area while attempting to match the marginal data. However, most literature has only been focused on spatial and household-person connection diversity. Despite being crucial, intra-household relationships receive little attention in existing research on PopSyn (Ho & Mulley, 2015). Currently, most of the existing work in PopSyn considers the relationships within households as uncontrolled variables (i.e. ignored during the synthesis) or generated via assumptions. Examples of assumptions are spouses will be male and female with a minimum age gap (Sun *et al.*, 2018) or fixed household structures sets (Rahman & Fatmi, 2023). However, these assumptions cannot consider special cases such as huge age gaps or same-sex couples etc. Thus, these approaches can be unreliable.

To comprehensively solve the diversities in population synthesis, we propose a framework namely the integrated population synthesis framework (IPSF) consisting of 2 new methodologies: sequential attributes adjustment (SAA) and chained sample pools (CSP). SAA focuses on spatial diversity, while CSP solves both the household-person and household intra-relationship diversities. The proposed novel IPSF then combines them both in a modular framework.

2 Methodologies



(a) *Sequential attributes adjustment (SAA)*

(b) *Chained sample pools (CSP)*

Figure 1 – *Methodologies*

In a population, there are attributes for each agent such as age, sex, or income, and each attribute will have a set of states like male or female for sex. The marginal data will likely have the total number of each state in an area while the seed data contains records (i.e. agents) having those states. Classical methods such as Iterative Proportional Fitting (IPF) from (Beckman *et al.*, 1996) attempt to find the appropriate weights for each agent in the seed data to match with the marginal data. As a result, if a state does not exist in the seed data, it cannot be synthesized, i.e., zero-cell problems in PopSyn literature. This incentivizes the development of statistical-learning-based (SL-based) methods such as Bayesian Network (BN) from Sun & Erath (2015). However, SL-based methods do not perform well in marginal data matching (i.e., spatial diversity).

We propose the framework IPSF, consisting of the SAA and CSP algorithms to tackle all past methods' drawbacks. In particular, we use SAA to synthesize the households which will be input into CSP to get the persons. The proposed methods are based on pools which are lists of detailed samples. To create them, we can use the seed data directly or derive from a generative model whose parameters are learned from the available seed data. For example, later in our case study, we used BN the same way as Sun & Erath (2015) to generate the pools. The pools are the source of samples used to generate the population for the SAA and CSP proposed algorithms. This pool-based approach is explainable and helps maintain the seed data distributions, and tackle the zero-cell issues while still being able to match the marginal data (via our proposed SAA method).

Specifically, SAA (Figure 1.a) adjusts each state per attribute for each zone sequentially to match the marginal data. It starts with adjusting the first attribute to create the first synthetic population (i.e., the initialization step). Being the first adjustment, this attribute will perfectly match the marginal data. For the following attribute adjustments, the synthetic population will be updated. At each adjustment, the population will be checked against the marginal data to detect which states need to be replaced while maintaining the previously adjusted attributes. The replacements are sampled repeatedly from the pool, hence, preserving the seed data distributions.

Subsequently, CSP (Figure 1.b) creates a person population based on synthetic households. It is based on having a Main person in a household, so it can create chained pools by pairing with the Main person. It starts with the households and the Main person pair, then the Main person with each possible household relationship such as Main and Spouse or Main and Child. Consequently, we can sample the persons in each household considering the relationships by sampling each pool, starting with the household and Main. By sampling from these multiple pools, "unlikely" results (i.e., results that exist in one pool but not in another) can also be eliminated to improve the process. Additionally, this multi-pools approach makes CSP suitable for parallel computing and, hence, more efficient with improved run times.

3 Results

3.1 The case study

To validate the proposed methodologies, we performed a full-scale study for Victoria, Australia. We aim to synthesize the full population of households and persons across 691 postcodes (POAs) using the Australian 2021 Census data as marginal data and the Victorian Integrated Survey of Travel and Activity (VISTA) data as seed data. We studied 5 attributes of households: household size (hhsize), household income (hhinc), total vehicles (totvehs), type of dwelling (dwelltype), and dwelling ownership status (owndwell); 5 attributes of persons: age, sex, personal income (persinc), having a diver license, and relationship. Specifically for relationships, we have 5 types: Main, Spouse, Child, Grand-Child, and Others. The Main person is the oldest person by age (e.g. for a household with a Grand-Child, the Main would be a grandparent) and every household will have at least 1 Main person.

We compared IPSF against IPU (Ye *et al.*, 2009), a popular method currently used in industry and practice. We also tested the performance of CSP and its ability to combine with another method by replacing SAA in IPSF with BN (Sun & Erath, 2015). We named it simple chained Bayesian networks (SCBN). To compare, we used the common root mean squared error (SRME) for marginal-data-related comparisons (i.e., spatial diversity) and the Jensen-Shannon divergence (JSD) for distribution-related comparisons (i.e., other two diversities).

3.2 Selected results

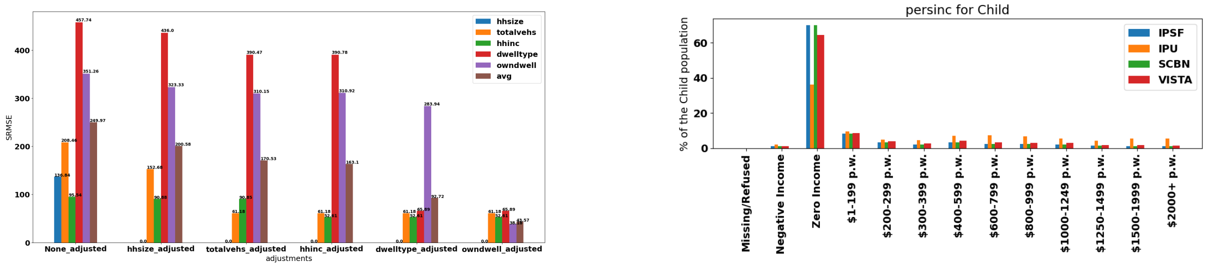
Table 1 – *Overall results*

Ground-truth	**Description**	**IPSF**	**IPU**	**SCBN**
Census	RMSE for households overall	2.39	9.23	231.6
Census	RMSE for persons overall	637.14	490.16	644.79
VISTA	JSD for households' attributes	0.12	0.12	0.04
VISTA	JSD for persons' attributes	0.03	0.08	0.01
VISTA	JSD for relationships	0.04	0.09	0.02

The overall results can be seen from Table 1. It can be seen that IPSF outperforms both IPU and SCBN for household RMSE by around 74% and 99%, respectively. For person RMSE, IPSF performs worse than IPU but by only around 30%. It should be noted that IPSF does not observe person marginal data at all, yet has been able to achieve remarkable results. Regarding distribution-related comparisons, we can see that IPSF and SCBN (both resulting from CSP) exceed the performance of IPU. Thus, the overall results have shown that IPSF has improved performance compared to that of the existing state-of-the-art methods.

We also plot results in Figure 2.a to show how SAA reduces the RMSE for each attribute after each adjustment. The order in the figure's legend is the order of adjustment in our case study. It can be seen that the RMSE was reduced significantly after each adjustment and the previous adjustments are not affected by the new adjustment. To view CSP performance in synthesizing household relationships, Figure 2.b shows the personal income distributions for *Child* against the

VISTA data (i.e., seed data). It shows that IPSF and SCBN learn the distribution in the seed data well while IPU tends to underestimate the lower incomes and overestimate the higher ones.



(a) *Comparing RMSE for each attribute after each adjustment*

(b) *Comparing personal income distributions for Child*

Figure 2 – *Selected results*

4 Discussion

In this work, we proposed a comprehensive framework (IPSF) for population synthesis with two novel methods (SAA and CSP) to improve the realism and diversities of the synthesized population used in the ABM transport models. Our synthesized population can effectively balance the 3 diversities that exist in practice where the SAA algorithm focuses on spatial diversity while the CSP implements households-persons connection and household relationships.

Our results showed that IPSF has the best overall performance compared to that of IPU and SCBN. The implementation of SCBN also proved the flexibility of the framework in combining IPSF with other methods. Furthermore, CSP is the first method to exclusively include the relationships within a household during the modelling process. Our analysis showed that CSP provides a reliable and robust method to create realistic persons for each household relationship type which filled an important yet often ignored gap in transport modelling.

The impact of relationships within a household on activity generation, especially joint activities, can be an exciting future research direction. Furthermore, as part of CSP, rare or undesirable combinations can be removed to continuously improve the sampling process which can be designed smartly, potentially with additional available data.

References

Beckman, Richard J, Baggerly, Keith A, & McKay, Michael D. 1996. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, **30**(6), 415–429.

Ho, Chinh, & Mulley, Corinne. 2015. Intra-household interactions in transport research: a review. *Transport Reviews*, **35**(1), 33–55.

Rahman, Md. Nobinur, & Fatmi, Mahmudur Rahman. 2023. Population Synthesis Accommodating Heterogeneity: A Bayesian Network and Generalized Raking Technique. *Transportation Research Record: Journal of the Transportation Research Board*, Jan., 036119812211442.

Sun, Lijun, & Erath, Alexander. 2015. A Bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies*, **61**, 49–62. Publisher: Elsevier Ltd.

Sun, Lijun, Erath, Alexander, & Cai, Ming. 2018. A hierarchical mixture modeling framework for population synthesis. *Transportation Research Part B: Methodological*, **114**, 199–212. Publisher: Elsevier Ltd.

Ye, Xin, Konduri, Karthik Charan, Pendyala, Ram M, Sana, Bhargava, & Waddell, Paul. 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.