Australasian Transport Research Forum 2023 Proceedings
29 November – 1 December, Perth, Australia
Publication website: http://www.atrf.info

Comparing major population synthesis techniques: A case study in Monash, Victoria

Lewen Feng[1], Md. Kamruzzaman[2]

[1]Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia

[2]Monash Institute of Transport Studies, Department of Civil Engineering, Monash University, Clayton, VIC 3800, Australia

Email for correspondence (presenting author): lewen.feng@monash.edu

1. Introduction

Transportation modelling is gradually shifting its focus from zone-based travel flows to individual-based travel trajectories, known as the agent-based modelling technique. This technique helps modellers understand how external factors, such as new policies, impact individual travel behaviour of millions of agents. Significant efforts have been made to develop activity- and agent-based models (Axhausen et al. 2016; Adnan et al. 2016; Mallig et al. 2013; Auld et al. 2016). One crucial input to the activity- and agent-based models is the socio-economics and travel details of agents often collected through travel surveys. However, such data is difficult to obtain from population level due primarily to cost. Population synthesis, a technique to generate individual-level data for the entire population from a smaller sample, has become a key solution to address this problem. Various population synthesis techniques have been developed and applied to transportation models (Ramadan and Sisiopiku 2019). We reviewed the existing literature and summarised three commonly used population synthesis methods.

There are two stages in population synthesis: the fitting stage and the generation stage (Müller and Axhausen 2010). The fitting stage is to iteratively reweight a microsample to match a given set of constraints. The generation stage is to export the entries in the microsample proportional to these weights as the synthetic population. The difference in techniques is how they reweight the microsample. Iterative proportional fitting (IPF) is a procedure that adjusts the distribution of microsample based on marginal totals reported in another dataset (Lomax and Norman 2016). This method has become the primary choice in population synthesis and has undergone significant evolution to address limitations, including zero-cell problems and multi-level constraints (Ramadan and Sisiopiku 2019). The iterative proportional updating (IPU) is one notable improved version of IPF developed by Ye et al. (2009). It is a heuristic approach designed to synthesise the population that can simultaneously satisfy different types of constraints, such as the household-level and person-level totals that do not agree on a single sum. The IPU procedure iteratively adjusts weights among households of a specific type until both household- and person-level constraints are satisfied. Similarly, the combinatorial optimisation (CO) based technique matches household- and person-level constraints by iteratively swapping entries in microsamples to minimise the difference between synthetic populations and given marginal totals. A special case in combinatorial optimisation adopts the simulated annealing (SA) algorithm to determine whether to accept or reject the swap. This integration is further developed into a microsimulation model for population synthesis (Harland 2013).

Various synthesis population techniques have been compared. Huang and Williamson (2001) compared results generated by IPF and CO with the UK 1991 census data. Their evaluation revolved around the goodness-of-fit of each model, leading to the conclusion that CO outperforms IPF. Ryan et al. (2009) applied both IPF and CO techniques to generate synthetic populations using a small population of firms. Attributes of each synthetic population are compared against the real population to evaluate the performance. This analysis also found that the CO method is recommended over IPF. The primary aim of these comparisons is to identify the method that is most effective in generating a synthetic population that mostly mirrors the actual population. Common comparison criteria encompass household characteristics, the socio-economics of individuals, and demographic distributions across subregions. However, it rarely involves travel details, such as methods of travel to work. Given that activity- and agent-based models reply on synthetic populations to provide information about individuals' activities, it is imperative for modellers to consider how different population synthesis techniques affect the representation of individual travel behaviour.

This study aims to address this gap in the literature by testing, validating, and comparing the outputs from the three most commonly applied techniques: iterative proportional fitting (IPF), iterative proportional updating (IPU) and simulated annealing (SA). The study used travel diary data from a representative sample from Monash, Victoria to generate its population-level data, which are compared against the 2016 population census data for validation. The findings of this study will help modellers to make an informed choice about the selection of right population synthesis technique for transportation modelling. Moreover, this research constitutes a fundamental step towards enhancing the robustness of activity- and agent-based models, offering valuable insights for both academic research and real-world applications.

2. Case study of Monash

2.1. Data

The data used in this case study can be broadly divided into two categories: microsamples of population and control totals. The microsample of population is obtained from the Victorian Integrated Survey of Travel and Activity (VISTA) 2012-2018, and control totals are extracted from the 2016 Australia Census. The VISTA data is an ongoing survey collecting personal information and corresponding travel data across greater Melbourne, Geelong and other key regional centres in Victoria. From 2012 to 2018, the survey has sampled over 27,000 households and 69,000 individuals. The Australian Census is a comprehensive database that captures statistics of demographics, families and dwellings of the entire population. The Census data is aggregated to different geographical levels (e.g. Meshblock, Statistical Area Level 1-4, Greater Capital Cities). This study used data at the Statistical Area Level 3 (SA3), chosen for its adequate sample size and computational feasibility. SA3 is analogous to a regional town in Australia with a population range of 30,000 to 130,000. The case study location, Monash, is an SA3 region located in southeast Melbourne, with its center approximately 15km from Melbourne's central business district.

The microsample of Monash is extracted from the household and person tables of VISTA data, comprising a total of 1,043 household and 2,816 individual records. Control totals for Monash consists of attributes at both the household and person levels. The initial step in population synthesis involves identifying data categories that match between the microsample and control totals. Due to differences in the purposes of VISTA data and the Australia Census, not all data categories are compatible. We identified matching categories and selected some that ensure a converged population synthesis process. Table 1 shows that 5 data categories present in both

VISTA and Australia Census, including dwelling type (describing the building type of dwellings), dwelling ownership (indicating whether individuals fully own their dwellings or are in other conditions), household size (describing the number of people in the household), and the remaining categories are self-explanatory. Similarly, the person data category include age, sex, income description, and employment type (indicating whether the individual is employed or in another employment status). Further details about the attributes of each data category are presented in Table 2.

Table 1: Matching data categories between VISTA and Census data.

VISTA Data Category		2016 Australia Census Data Category	Applied in modelling
Household	Dwelling type	Dwelling structure	Yes
	Dwelling ownership	Tenure type	No
	Household size	Count of persons in family	Yes
	Household income	Household weekly income	No
	Total vehicles	Motor vehicles	Yes
Person	Age	Age	No
	Sex	Sex	Yes
	Employment type	Status in employment	No
	Person income	Person income	No

2.2. Method

Programs to execute all three population synthesis techniques are included in the simPop R package. Each technique is provided with the identical microsample and control total data. Minor changes were made to satisfy code requirements. For the IPF and IPU procedure, simPop uses the *ipu* function to adjust sampling weights to given control totals as shown in Table 1. The difference is that IPF only receives household-level totals, whereas IPU receives household- and person-level totals. The input to the SA procedure requires control margins instead of totals. This control margin describes the frequency of households or people in multivariate joint distribution among attributes (dwelling type × total vehicles). For example, the distribution can be the number of households characterised by living in a separate house and have zero motor vehicle. This data is extracted from Census data using the TableBuilder tool. Once input data is complete, the SA procedure is executed using the *calibPop* function.

3. Results and discussion

Results of the IPF and IPU procedure are fractional weights for each household in the microsample. If such weights are directly applied to generate synthetic households, the household number will be fractional, which is impractical in reality. To avoid this, we rounded each weight to the nearest integer before replicating households. After the household replication process, we obtained the synthetic population by adding the number of people associated with each synthetic household. The result of the SA procedure is a complete synthetic population, and no further process is required.

Table 2 presents the comparison of the synthetic populations generated by three techniques. For an intuitive comparison, all synthetic populations are further grouped into the same categories as control totals. For household- and person-level control totals, the absolute error percentage is calculated to measure the performance of each result. The absolute error is defined as follows:

$$\text{Absolute Error \%} = \frac{\sum_{i=1}^{A}|X_i - x_i|}{\sum_{i=1}^{A} X_i} \times 100 \tag{1}$$

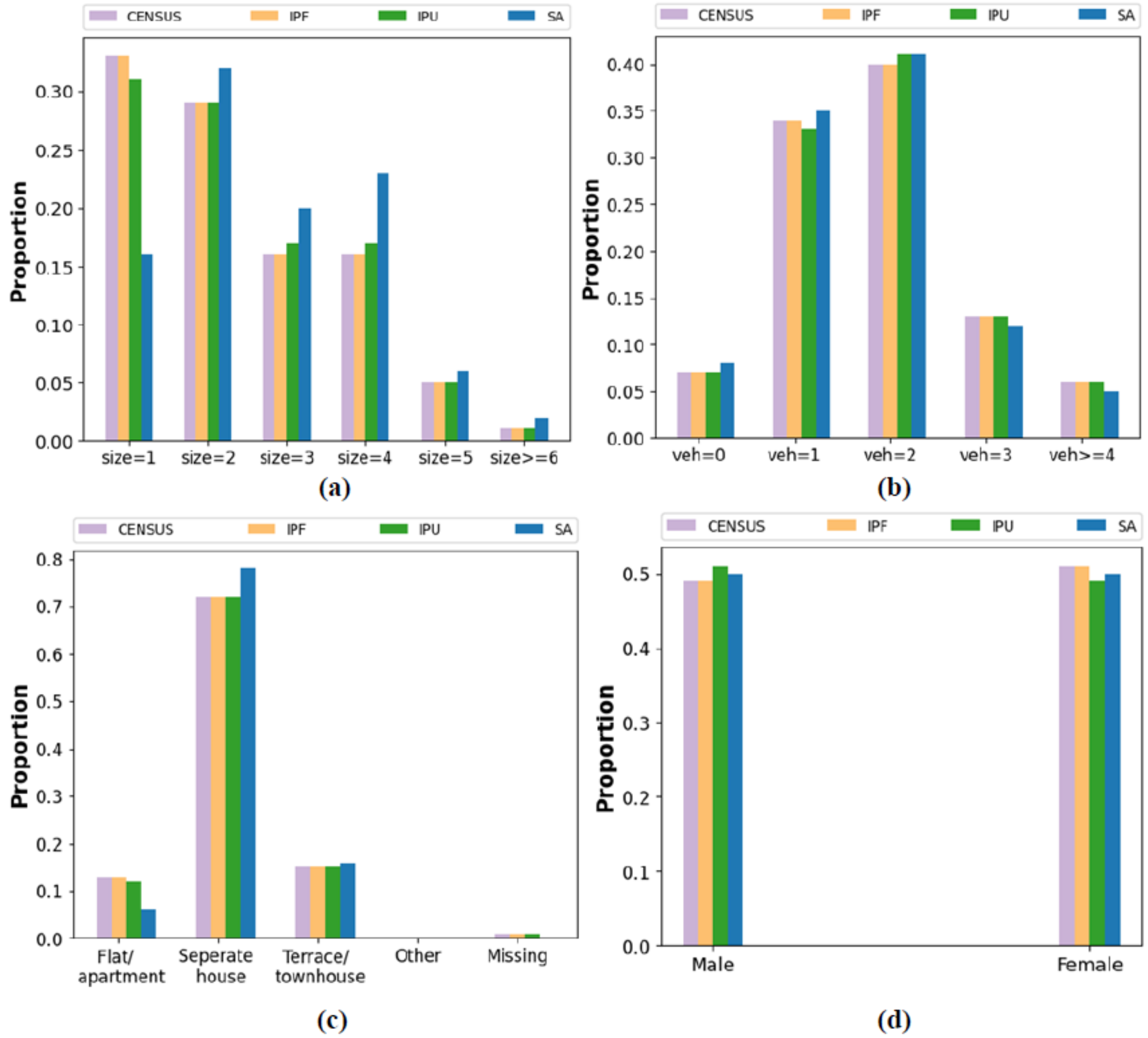
where *A* denotes the attribute type of interest, *X* denotes the known value of each attribute i, and *x* denotes the estimated value for attribute *i*. The IPF method has the lowest absolute error percentage in household-level attributes. This is predictable, as IPF only needs to control the household-level totals, and converges more easily than the other two methods. Since IPF does not use person-level control totals, the person-level absolute error percentage exceeds 10%. Both IPU and SA methods constrain household- and person-level totals, and both fit person-level totals better than household-level totals. The IPU method performs worst on household-level attributes, but at the same time performs best on person-level attributes.

Table 2: Comparison between Census data and three synthetic populations.

Attribute type	**Attribute**	**2016 Census data (Control totals)**	**IPF Results**		**IPU Results**		**SA Results**	
			Value	Abs error %	Value	Abs error %	Value	Abs error %
Household	Total household	67500	67489	0.016	77306	14.53	60963	9.68
	Hhsize=1	22111	22095		23958		10003	
	Hhsize=2	19267	19259		22590		19694	
	Hhsize=3	10958	10981		12880		12446	
	Hhsize=4	11056	11038		13036		14008	
	Hhsize=5	3330	3338		3928		3537	
	Hhsize>=6	778	778		914		1275	
	Flat/apartment	8643	8647		9645		3484	
	Separate house	48308	48308		55580		47303	
	Terrace/townhouse	10187	10190		11666		9985	
	Other	20	2		13		45	
	Dwelling type missing	342	342		402		146	
	Vehicle=0	4989	4998		5488		4676	
	Vehicle=1	22733	22667		25204		21071	
	Vehicle=2	27050	27099		31684		25099	
	Vehicle=3	8623	8624		10146		7041	
	Vehicle>=4	4105	4101		4784		3076	
Person	Total resident	178768	159284	10.90	185306	3.66	168579	5.70
	Male	87615	77794		94027		83707	
	Female	91153	81490		91279		84872	

Figure 1 depicts the marginal distributions of household- and person-level attributes between Census data and synthetic populations. All three methods produce reasonable marginal distributions across most attributes. However, there are notable discrepancies in the marginal distribution of household size when comparing SA population with the Census data. Specifically, there is an approximately 15% difference in the distribution of households with a size equal to 1 and a 10% difference in the distribution of households with a size of 4.

Figure 1: Household- and person-based marginal distribution comparison: (a) household size, (b) number of motor vehicles, (c) dwelling type, and (d) sex.



4. Conclusion and future work

This study generated the Monash population using three major population synthesis techniques: IPF, IPU and SA. Results show that the IPF method is best for matching household-level control totals and the IPU method is best for matching control totals at the person level. All three methods maintain an acceptable difference in the marginal distribution of attributes, except for the age category. We have yet to conclude which method is the best because we still need to compare the results of travel details with the ground truth. The next research phase is to obtain travel information from each synthetic population and compare it to Census data. Our research outcomes provide modellers with valuable insight into the most suitable population synthesis method for transportation modelling.

5. References

Adnan, M., Pereira, F.C., Azevedo, C.M.L., Basak, K., Lovric, M., Raveau, S., Zhu, Y., Ferreira, J., Zegras, C. and Ben-Akiva, M., 2016, January. Simmobility: A multi-scale integrated agent-based simulation platform. In *95th Annual Meeting of the Transportation*

Research Board Forthcoming in Transportation Research Record (Vol. 2). Washington, DC: The National Academies of Sciences, Engineering, and Medicine.

Auld, J., Hope, M., Ley, H., Sokolov, V., Xu, B. and Zhang, K., 2016. POLARIS: Agent-based modeling framework development and implementation for integrated travel demand and network and operations simulations. *Transportation Research Part C: Emerging Technologies, 64*, pp.101-116.

Harland, K., 2013. Microsimulation Model user guide (flexible modelling framework).

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

Lomax, N., & Norman, P. (2016). Estimating population attribute values in a table:"get me started in" iterative proportional fitting. *The Professional Geographer, 68(3)*, 451-461.

Mallig, N., Kagerbauer, M. and Vortisch, P., 2013. mobitopp–a modular agent-based travel demand modelling framework. *Procedia Computer Science, 19*, pp.854-859.

Müller, K. and Axhausen, K.W., 2010. Population synthesis for microsimulation: State of the art. *Arbeitsberichte Verkehrs-und Raumplanung, 638*.

Ramadan, O.E. and Sisiopiku, V.P., 2019. A critical review on population synthesis for activity- and agent-based transportation models. *Transportation Systems Analysis and Assessment*.

Ryan, J., Maoh, H. and Kanaroglou, P., 2009. Population synthesis: Comparing the major techniques using a small, complete population of firms. *Geographical Analysis, 41(2)*, pp.181-203.

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