Australian Government
Department of Infrastructure, Transport, Regional Development and Communications
Bureau of Infrastructure, Transport and Regional Economics

Population synthesis for travel demand modelling In Australian Capital Cities

Author: Poh Ping Lim

Abstract

This paper introduces an efficient and practical population synthesis routine that could be readily used to create microdata in small geographies for Australian capital cities. The aim is to ease the challenging process of preparing the necessary geocoded microdata for microsimulation and incentivise further development of activity-based microsimulation models for travel demand in Australia. In this study, the proposed synthesis routine has been used to generate full size synthetic populations of households and individuals for Greater Sydney, Melbourne and Brisbane. Two heuristic algorithms have been formulated for data treatment before and after the synthesis process to improve the representation of the synthesised populations. The procedure proposed for data treatment before the synthesis routine ensures the consistency of the input data, whereas the procedure proposed for data treatment after the synthesis routine extends under-synthesised estimates to a complete synthetic population. The synthesis process was tested for its efficacy and the synthesised populations were validated extensively. This paper presents an overview of the synthesis routine with examples of validated experimental results for the generated synthetic populations.

Introduction

To date, the Four Step model remains a prevalent framework adopted for assessing and projecting the impact of transport policies in Australia. The model generates aggregate trips and forecasts aggregate travel demand. It is useful for evaluating large scale infrastructure projects and major capacity improvements in the transport system. However, they are less sensitive to management and control of existing transport services and infrastructure. As policy measures in transport shift from "predict and provide" to "manage and control", there is a need to supplement the existing travel demand model system with models that offer better representation and prediction of travel behaviour.

Spatial microsimulation in activity-based travel demand modelling provides an important basis to analyse travel behaviour of individuals in spatial context. A spatial microsimulation framework consists of households and individuals with relevant attributes in fine geographical zones. The model design is well suited for the representation of complex travel behaviour and simulation of spatial interaction. The increasing demand for spatial microsimulation analysis in transport research has been driven firstly by, the need for policy makers to assess distributional effects from transport policy changes across different sub-groups of population. Spatial microsimulation models contain the necessary links between individuals and the geographical information that is useful for evaluating fine grain distributional and spatial effects of transport initiatives in urban and regional planning. Secondly, there is a need for more accurate and detailed projections of travel

demand to facilitate decision making in transport provision. The accuracy of travel demand projections defines the effectiveness of urban planning and infrastructure investment. Currently, the commonly used Four Step model for transport modelling in Australia is lacking in its capability to examine travel behaviour and accurately predict spatial variations in travel demand. The inadequacies of the Four Step model are well documented (Bhat et al. 2003, Mladenovic & Trifunović, 2014, Recker, McNally & Root 1986, McNally 2007). An activity-based microsimulation model for travel demand operates at household and person level. The model is able to capture interdependencies between households and individuals, linkages between activities and trips, and the underlying behaviour that lead to activity participation and creation of trips (McNally & Rindt, 2007). It is a flexible and responsive analytical tool that can be used to simulate the effects of policy decisions under alternate social demographic condition, transport options and land configuration scenarios.

Thus far, Australia is yet to have a fully validated, operational, and open access microsimulation activity-based travel demand model that is readily to be used for transport modelling. An integral and critical part to building a microsimulation travel demand model is to obtain microdata of relevant attributes at fine geographical zones. Information at this level of detail is usually collected in Australian censuses. However, such comprehensive information cannot be made available in the public domain due to privacy reasons. Despite the increasing availability of national censuses, surveys and administrative data, geocoded microdata generally lacks small area demographic or geographic depth for microsimulation analysis. The requirement of spatial microdata remains a major barrier to developing a fully operational microsimulation travel demand model in Australia.

What is population synthesis?

Population synthesis is concerned with estimating unknown information at fine geographical level based on known aggregate information.

The basic idea behind population synthesis in transport research is to construct a complete population with spatial micro units that is statistically representative of the actual population (Beckman, Baggerly & Mckay 1996). Population synthesis is a process of expanding disaggregate sample data to a full size synthetic population based on known distributions in the actual population. A synthetic population basically represents a reconstruction of one possible set of "best estimates" that mirrored the distributions of the actual population; where relevant attributes pertaining to every synthetic person and household in the entire study population are fully enumerated at detailed geographical level (Ryan, Maoh & Kanaroglou 2007).

Population synthesis methods have been developed as viable alternatives to supplement the lack of completeness in spatial microdata for microsimulation analysis. These methods aim to generate synthetic microdata that is statistically sound enough for microsimulation while preserving the confidentiality of the actual population.

Main Methods in Population Synthesis

There are two main methods to population synthesis that are commonly applied in travel demand modelling:

- Synthetic reconstruction (SR)
- Combinatorial optimisation (CO)

Both methods attempt to recreate a complete list of persons and households in a population that are consistent with the known aggregate distributions (Huang & Williamson 2001).

Synthetic reconstruction (SR)

The SR method primarily relies on the Iterative Proportional Fitting (IPF) procedure. The IPF procedure is an iterative scaling method for estimating cell probabilities in a contingency table, subject to marginal constraints (Fienberg 1970a). The application of was first introduced by Deming and Stephan in 1940 IPF using data from the US census of population (Fienberg & Meyer 1981, Lomax & Norman, P. 2016). The idea was to adjust contingency table cells from a sample data such that: (i) its row and column totals align with selected marginal totals obtained from cross tabulations of the actual population data, and (ii) the correlation structure of the sample data is retained after adjustments (Deming & Stephan 1940, Müller 2017). The IPF proof of convergence and its properties have been well established by several researchers after Deming and Stephen, including Bishop (1967, 1969), Brown (1959, 1976), Csiszar (1975), Fienberg (1968), Fienberg & Gilbert (1970), Ireland & Kullback (1968), Haberman (1974, 1984), Mosteller (1968) and Rüschendorf (1995). Parallel to Bishop (1967), Ireland and Kullback (1968) demonstrated that Brown's proof of convergence for IPF could be extended to multi-dimensional contingency tables. In addition, they have proven that IPF produces unique maximum likelihood estimates for the table cell values given the imposed constraints, which represents a maximum entropy or minimum relative entropy (i.e. minimum discrimination information) solution (Birkin & Clarke 1988, Bishop 1967, Ireland & Kullback 1968). Fienberg (1970b) has provided a detailed account of the mathematical development involved in IPF.

Most of the population synthesisers used in travel demand modelling today are based on the IPF procedures proposed by Beckman, Baggerly & McKay in 1996 (Auld & Mohammadian 2010, McBride et al. 2016). They were regarded as pioneers in generating individual records at fine geographical level to reconstruct a synthetic population for transport modelling (Müller & Axhausen 2011). The generated synthetic population was first applied in the TRansportation ANalysis SIMulation System (TRANSIMS) project (Los Alamos National Laboratory 2005). The TRANSIMS microsimulation model is an activity-based travel forecasting microsimulation model that simulates the travel behaviour of each synthetic person over 24 hours based on representative activities derived from survey data (Lee, KS et al. 2014).

The TRANSIM population synthesiser consists of two stages: fitting and generation, which are typically the two principal stages for most population synthesisers that were developed after Beckman (Bowman 2004, Müller & Axhausen 2011, Pritchard & Miller 2012).

At the fitting stage, two main type of data source are required: (i) disaggregate sample data and (ii) aggregate constraints. A disaggregate sample data is a representative sample file that usually consists of unit records drawn randomly from a population census. In TRANSIMS, the Public Use Microdata Sample (PUMS) provides an ideal matrix base for a seed dataset. The 1% and 5% sample from the US census of population offers relevant demographic attributes for households and persons in a collection of small geographical census areas. The sample file inheres reasonably reliable joint probability distributions of multiple attributes (McBride et al., 2016). The selected attributes used to construct a synthetic population in the sample file are referred to as control variables. The joint distributions of these control variables create a multidimensional contingency table referred to as a seed dataset (Müller & Axhausen 2011). Marginal totals from the seed dataset are basically Cartesian products of the control variables with each multi-dimensional cell represents a unique marginal total cross tabulated from two or more control variables. These marginal totals provide the number of households or persons with the same combination of demographic characteristics that define a homogenous group.

Aggregate constraints are a collection of demographic summary tables extracted from the fully enumerated population census or other sources of known aggregate data. These are one dimensional summary tables whereby each table contains univariate distributions in small geographical areas. Aggregate constraints by geographical areas assembled for each selected control variable also referred to as control marginal totals.

Beckman, Baggerly and McKay (1996) used the IPF procedure to complete a multidimensional contingency for each geographical analysis zone. The IPF iterative process started with the seed data. The procedure provides a weighting mechanism whereby cell values in the seed data are repeatedly adjusted to create weights or cell probabilities that produce marginal totals closely matched the control marginal totals. Upon convergence, the weight or cell probabilities for the seed data were estimated for each homogenous group in each geographical zone.

At the generation stage, the number of households for each homogenous group in the seed data is expanded for each geographical area. These numbers are allocated, either by multiplying the total number of households by the group weights estimated in the fitting stage, or by randomly drawing households according to the corresponding estimated weights until the expected number is reached. The estimated weights act as a fractional expansion factor for each household to grow the sample size into a full-size population. Once the allocation process is finalised, a synthetic population with the full properties and attributes of the sample data is constructed.

The repeated probabilistic selection with replacement is the most commonly used method at the generation stage (Müller & Axhausen 2011, Bowman, 2004). Selection with replacement means that once a person or household is selected to be in the synthetic population, the sample unit is placed back in the sample to possibly be sampled again. In the selection without replacement method, once an individual or household is sampled, the sample unit is not placed back in the sample for resampling. This method is usually not suitable for small samples (Choupani and Mamdoohi 2016). There are a variety of selection methods proposed for the SR method, such as conditional Monte Carlo sampling (Pritchard & Miller 2009), deterministic selection (Srinivasan, Ma & Yathindra 2008), and altered selection probability (Auld & Mohammadian2010).

The method used by Beckman, Baggerly and McKay (1996) to create the synthetic population was validated by creating pseudo census tracts from PUMS sample and compared the joint distribution of the household size and the number of vehicles in the households to the actual population. They have shown that the joint distributions created do not differ substantially from the true values of the actual population (Beckman, Baggerly & McKay 1996).

Combinatorial optimisation (CO)

The CO method generally generates a synthetic population by randomly allocating individuals from a disaggregate sample file into each geographical zone. The iterative algorithm is initiated by a random assignment of households from a disaggregate sample, matching the population size of each geographical zone. A goodness of fit statistic that indicates the extent to which control marginal totals are matched is calculated to measure the fit of the random selection set of households to the known distributions of the control variables in the zone. The assigned household is retained if the replacement improves the goodness of fit. Otherwise, the assigned household is replaced with another household. This process repeats until a given termination criterion to find the best fit synthetic population is reached (Cho et al. 2014). Williamson, Birkin and M, Rees (1998) first suggested using the CO approach to build a synthetic population based on the Samples of Anonymised Records (SAR) disaggregated by the smallest geographical units in UK (Voas & Williamson 2000). The approach was further improved by Voas and Williamson

(2000). They developed a 'sequential fitting procedure' whereby the least represented table for a given analysis zone is fitted first to the known aggregate controls, followed by the next least represented table and so on. At each stage, an assigned household that favour the fit of the later table at the expense of the preceding tables cannot be replaced. It was found that it is possible for this new solution to satisfy a level of minimum acceptable fit for every table used to constrain the selection of households from SAR (Huang & Williamson 2001). In Australia, Melhuish, Blake and Day, (2002) used CO to generate the socio-demographic profiles of synthetic households for each census collection district (CDs) in Australia. They benchmarked and validated the sociodemographic profile in the synthesised population with the census Basic Community Profile (BCP) by the Australian Bureau of Statistics (ABS) (Cho et al. 2014, p51).

In recent years, there have been several researchers who used CO to generate synthetic populations. Harland et al. (2012) described using combinatorial optimisation with 'simulated annealing' method for population synthesis. The annealing threshold or terminal criterion set for the iterative process allows an assigned household to be randomly replaced if the replacement leads to an improvement in goodness of fit or if the deterioration in goodness of fits is within a set limit. The annealing thresholds are decreased with every replacement and the procedure stops when the threshold becomes zero. This empirical study demonstrates that the procedure does lead to better fit in generating a synthetic population, but it was computationally an intensive process (Ma & Srinivasan 2015). Abraham, Stefan and Hunt (2012) developed a software to match controls at both household and person level while accounting for constraints at multiple spatial resolutions (Konduri et al. 2016). They used the "stochastic hill climbing" method that is similar to the "simulated annealing" method but without the possibility of backsteps. That is without the option that could potentially deteriorate the goodness of fit in the iterative process (Ma & Srinivasan 2015). The algorithm is reasonably fast with high degree of accuracy (Abraham, Stefan & Hunt 2012). Namazi-Rad, Mokhtarian and Perez (2014) from Smart Infrastructure at the University Wollongong applied a CO algorithm using a quadratic function of population estimators to generate a dynamic synthetic population while considering a two-fold nested structure for individuals and households in the study area. The study used the Confidentialised Unit Record Files (CURFs) and 2006 Australian census tables.

Both SR and CO methods are classified under the static spatial microsimulation approach, which simulate cross-sectional population data at a specific point in time (Lambert et al. 1994, Tanton 2014).

Other Emerging Methods in Population Synthesis

In addition to SR and CO methods, there are other emerging methods in population synthesis for transport modelling, such as model-based generation of synthetic data. For example, Farooq et al. (2013) and Saadi et al. (2015) proposed a simulated-based model to create a synthetic population by using the Markov Chain Monte Carlo (MCMC) methods (Zhuge et al. 2018). MCMC methods simulate a sequence of random draws using partial or/and full conditional probabilities from the actual population to create a synthetic population (Saadi et al. 2015). Another example of a model-based generation method is to synthesise a synthetic population based on Bayesian networks. A Bayesian network is a graphically representation of a joint probability distributions that encoding probabilistic relationships among a set of variables (Sun & Eratha 2015).

Validation

The accuracy or goodness of fit of a synthetic population is assessed differently, depending on the synthesis method and the data involved (Müller, 2017). Broadly, there are two type of validation for evaluating the representation of a synthetic population: internal and external validation

(Edwards & Clarke 2009). Internal validation generally involves comparing the synthesised estimates in a synthetic population to the marginal constraints used in the model. In external validation, synthesised estimates are compared to external data that was not used in the model. Internal validation aims at measuring errors introduced in the synthesis and selection process, whereas the purpose of external validation is to demonstrate synthesised estimates agree with the existing aggregate data from sources that were not used in the synthesis process (Caldwell & Keister 1996).

Most synthetic populations generated from population synthesisers such as PopGen, PopSynWin CEMDAP, FSUTMS and TRANSIMS were validated internally (Choupani & Mamdoohi 2014). Internal validation usually examines how the magnitudes of the percentage differences vary for each synthesised population groups compared to the actual data. There are different statistical tests used to validate a synthetic population using IPF procedure and its variants, such as total absolute error and standardised absolute error, distribution χ^2, the normal and modified Z score. Lovelace et al. (2015) discussed various validation techniques for verifying the integrity of a synthetic population generated using IPF procedure. Voas and Williamson (2001) presents an excellent discussion of different statistical tests to evaluate the fit of synthetic microdata estimates (Rose & Nagle 2016). The challenge of conducting an external validation to a synthetic population is that there are rarely confirmatory data by which to validate against. The limitation in validating a synthetic population is highlighted in several literature sources (Ballas & Clarke 2001, Birkin 2013, Edwards & Tanton 2013, Morrissey & O'Donoghue 2013, Ruther et al. 2013, Williamson et al. 1998).

Population Synthesis Procedure

The synthesis procedure adopted in this study is based on the Iterative Proportional Updates (IPU) procedure. The IPU procedure is a modification of the IPF procedure developed by Ye et al in 2009. The classic IPF procedure can either conform to constraints imposed at household level or person level but not for both simultaneously (Guo & Bhat 2007, Müller & Axhausen 2011). It is an issue that may significantly diminish the representativeness of the synthesised population (McBride et al. 2016). The IPU procedure addresses the issue in the IPF algorithm by fitting person and household constraints simultaneously during the fitting stage (Müller, 2017).

General IPU Algorithm

IPU extends from IPF by adjusting the household weights based on the person weights obtained from IPF (Ye, et al 2009). The IPF procedure is well established and documented as referenced in the previous section. The mathematical description of IPF is not included in this paper.

The IPU algorithm begins by creating a frequency matrix D (Table 1). The matrix frequency D shows the household type u and the frequency of different person types T within each household for the sample data P_s. P_s is constructed from P; where P is a matrix of households obtained from the merging of a set of n persons X into a set of m households Y. Each person x is characterised by t_x from q different person types T, where T denotes attributes of the person. Each household y is characterised by u_y from p different household types U, where U denotes attributes of the household. The number of persons in each person type is defined as $n_T = \{n_{tk}\}_{1\leq k\leq q}$ and the number of households in each household type is defined as $n_U = \{n_{ul}\}_{1\leq l\leq p}$. The dimension of D is therefore $|P_s| \times (p+q)$ (R). An element d_{ij} of D represents the contribution of household i to the frequency of person/household type j. The purpose of P_s is to reconstruct t_x by

estimating a weight ω_i associated with each person and each household of the sample that match the total number of each person type in X and households Y.

Table 1 IPU Table

Household ID	Household Type u_1	…	Household Type u_p	Person Type t_1	…	Person Type t_q	Weight
1	d_{11}	…	d_{1q}	d_{1q+1}	…	d_{1q+p}	$\boldsymbol{\omega_1}$
…	…	…	…	…	…	…	…
$\boldsymbol{[P_s]}$	$d_{\lvert P_s\rvert 1}$	…	$d_{\lvert P_s\rvert q}$	$d_{\lvert P_s\rvert q+1}$	…	$d_{\lvert P_s\rvert q+p}$	$\boldsymbol{\omega_{\lvert P_s\rvert}}$
WS	$\boldsymbol{\omega s_1}$	…	$\boldsymbol{\omega s_p}$	$\boldsymbol{\omega s_{p+1}}$	…	$\boldsymbol{\omega s_{p+q}}$	
E	$\boldsymbol{e_1 = \widehat{n_{u_1}}}$	…	$\boldsymbol{e_p = \widehat{n_{u_p}}}$	$\boldsymbol{e_{p+1} = \widehat{n_{t_1}}}$	…	$\boldsymbol{e_{p+q} = \widehat{n_{t_q}}}$	
$\boldsymbol{\delta}$	$\boldsymbol{\delta_1}$	…	$\boldsymbol{\delta_p}$	$\boldsymbol{\delta_{p+1}}$	…	$\boldsymbol{\delta_{p+q}}$	

Source: After Lenormand & Deffuant 2013, p 3

When the match between the weighted sample constraints converged to the pre-specified threshold, the algorithm stops. During the generation stage, the procedure randomly draw household from P_s with probabilities corresponding to the estimated weights (Lenormand & Deffuant 2013). Household selection probabilities are estimated using rounded weights. As the IPU procedure considers joint distributions of household and person level, households which are in the same household type may have different selection probabilities. Households from the sample data are randomly drawn until the number of households in the synthetic population matches the frequencies of households in the rounded joint distribution table for all household types. The drawing process is repeated until a synthetic population with the best possible fit is generated (Ye et al. 2009).

Geometric Interpretation of IPU Algorithm

The underlying logic behind the IPU algorithm can be explained using a two-dimensional graph. Assuming that there two households of the same household type (household 1 and household 2) that are subjected to a set of control variables. In table 2, household 1 has no individual in person type 1 while the household 2 has one individual. Suppose the household type 1 constraint is 4 and the person type 1 constraint is 3, then the weights for satisfying both person and household constrains can be resolved by finding solution a for two simple linear equations.

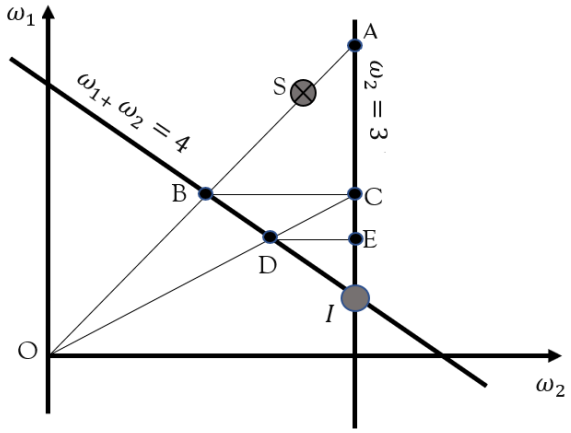
Table 2 Geometric Interpretation of IPU Algorithm

Household Id	Household Type 1	Person Type 1	Weights
1	1	0	ω_1
2	1	1	ω_2
Constraints	4	3	

Source: Ye et al. 2009

As shown in Figure 1, ω_1 on the vertical axis and ω_2 on the horizontal axis denote weights for household 1 and 2 respectively. The iterative process begins at point S, adjusted for household type 1 constraint to point B, then adjusted for person type 1 constraint to point C. These adjustments continue to point D and E until finally both household and person type constraints are met at intersection I.

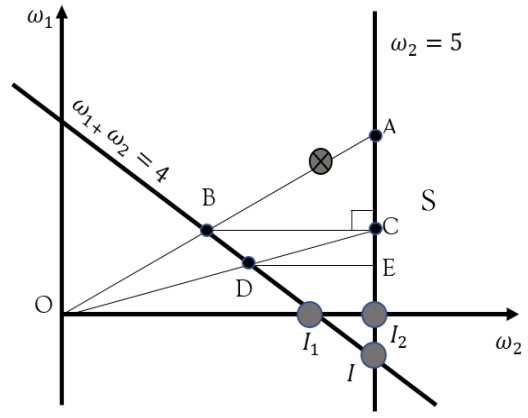
Figure 1 Geometric Interpretation of IPU Algorithm – Feasible Solution Case



Source: Ye et al. 2009

However, it is not always feasible to have a perfect solution where both household and person type constrains are met exactly, especially when there are many constrains imposed. Figure 2 shows a case of infeasible solution. The solution I where both constraints are met is outside the first quadrant. In this instance, the algorithm will attempt to move the coordinates closer to I from one iteration to another by alternating the adjustment between household and person type constraints. The algorithm will eventually move the coordinates back and forth between I_1 and I_2, where the two constraints intersect with the horizontal axis. As the coordinates can never reach I, one can choose to adopt a corner solution of I_1 for matching the household constraint or I_2 for matching the person constraint. Or to adopt a solution between I_1 and I_2, which is a trade-off between matching one constraints to another. The IPU algorithm adopts the corner solution (I_1) corresponding to household constraints (Ye et al. 2009).

Figure 2 Geometric Interpretation of IPU Algorithm – Infeasible Solution Case



Source: Ye et al. 2009

IPU Zero-cell and Zero -Marginal Corrections

One of the most commonly encountered issues of generating a multidimensional contingency table in small geographies for IPU or IPF using sample data is the presence of zero values in some cells. These zero cells exist when there were no respondents who satisfied the parameters of the cell in the contingency table. However, not all zero cells in the table are true zeros. There may be actual

counts of persons or households in those zero cells that represent a particular set of population characteristics when referred to the known population. The occurrence of zero cells is inflated when many control variables and high geographical resolutions are used to build a multidimensional contingency table for the IPF procedure. The execution of an IPF or IPU procedure on any zero cell values will prevent the iterations from ever reaching the constraint totals. As a result, the procedure fails to converge (Guo & Bhat 2007). This is referred to as the zero-cell problem in the literature (Müller & Axhausen 2011). Ye et al. (2009) correct the zero-cell problem in IPU by substituting zero cells in the zone level seed with associated probabilities computed from the region level seed. However, this process poses a risk of over-representing the demographic group, which are infrequent homogenous groups as evidenced by the zero cell from the start. These inconsistencies are corrected by imposing an upper bound threshold for estimated probabilities obtained at regional level when applying to zero cells. This is to ensure that the estimated frequencies in zero cells are not over-estimated. If the sum of cells of the demographic group is greater than unity after substitution, all non-zero cell probabilities are scaled down linearly to reach unity (Ye et. al, 2009, Choupani & Mamdoohi 2016).

The zero-marginal problem occurs when the marginal value of a control variable for a homogenous group is zero from the census data. In the IPU context, the initial IPF procedure will assign a zero to all household/person type cells in the zero-marginal category. However, when the denominator takes a zero value, the weight adjustments fail to proceed. The solution implemented in IPU is to introduce a small value of 0.001 to the zero-marginal categories. This allows the IPU algorithm to proceed with computing the corresponding weights to meet with both household and person constraints (Ye et al. 2009).

Research Scope

In this study, Greater Sydney, Melbourne and Brisbane were selected for the synthesis routine. These three cities have consistently shown the highest estimated traffic volume and congestion cost among all major cities (BITRE 2015).

The geographical coverage for each city is defined by the ABS Greater Capital Cities Statistical Areas (GCCSAs). According to the ABS, GCCSAs are designed to provide a stable and consistent boundary that represent the social economic extent of capital cities in Australia (ABS 2011). The definition of GCCSAs includes people who regularly socialise, shop or work within the city who live in the small towns and rural areas surrounding the city. GCCSAs are built from aggregation of Statistical Area Level 4 (SA4), based on the ABS Australian Statistical Geography structure (ASGS) (See Appendix 1). In fact, SA4s are specifically devised to reflect labour market and are used to facilitate the output of Labour Force Survey data (ABS, 2011). Therefore, these statistical areas capture a large portion of the commuting population within each state and territory. This is relevant in the context of travel demand modelling.

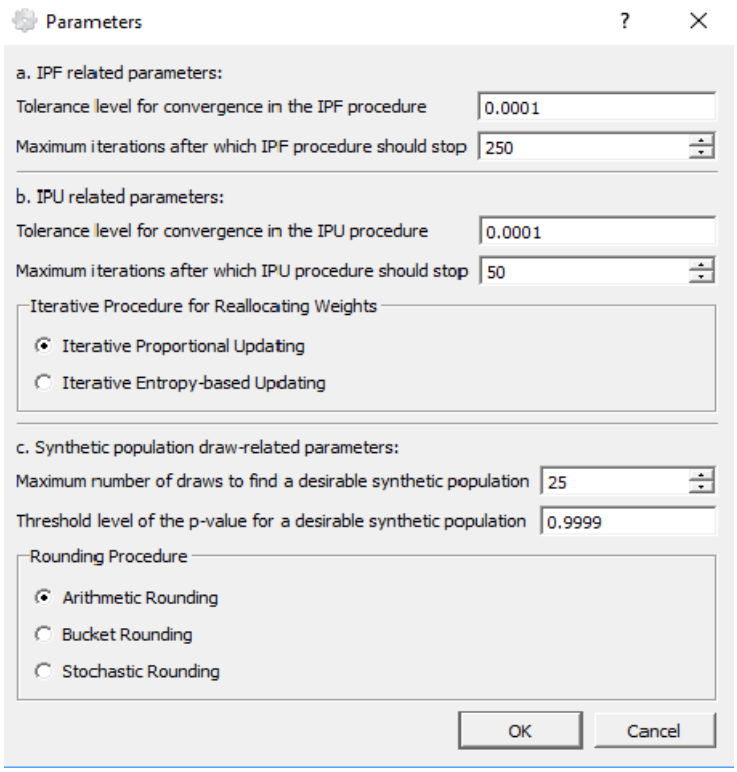
This study synthesised households and individuals in occupied private dwellings at Statistical Area Level 1. Based on ASGS, Statistical Area Level 1 (SA1) is the smallest unit of the released census data. An SA1 generally contains between 200 and 800 people with an average population size of 400 people. SA1s can be directly built up to SA2, SA3 and SA4 for coarser geographical zones. The research scope for this study include 10845 SA1s for Greater Sydney, 9658 SA1s for Greater Melbourne and 5485 SA1s for Greater Brisbane. Table 3 provides a summary of household and person counts from CURF and census data from the ABS table builder for the three major cities. Synthetic households and persons populations were generated at SA1 level based on data from the Although the latest Population Census was conducted in 2016, its microdata was not released during the course of this research study.

Table 3 Summary Statistics of Person and Household Counts in Greater Sydney, Greater Melbourne and Greater Brisbane, 2011

	Greater Sydney		Greater Melbourne		Greater Brisbane	
	Census	1% CURF	Census	1% CURF	Census	1% CURF
Household counts	1,598,439	16,030	1,491,729	14,939	814,364	8,153
Person counts	4,308,248	42,995	3,912,141	38,650	2,155,966	21,292

Overall, there were 840 household marginal constraints and 144 person marginal constraints involved in each synthesis process of the three cities. The input parameters for the synthesis process are shown in Figure 3. The synthesised results presented in the next section were based on the default input parameter setting in PopGen. Under the default parameter setting, the maximum iterations for IPF is 250 and the maximum iterations for IPU is 50. The tolerance level for convergence for both IPF and IPU procedures is set at 0.0001. The number of specified iterations affects the convergence speed and the achievable accuracy of a synthetic population. There is a trade-off between the computational cost and reconstruction accuracy. The question is whether it is possible to modify the number of iterations given the tolerance level to improve the representation of the synthetic population without increasing the computational cost. There has been emerging research exploring a faster convergence to an approximated set of minimised errors by neural network and deep learning (Giryes 2016). This relatively new approach is still evolving, and it is beyond the scope of this research study. As most iterations converged well below the default limit, these settings remained unchanged for all synthesis processes performed in this study. The configuration of input parameters for the rounding procedures were further tested later to examine the effects of these procedures on the performance results of the generated populations.

Figure 3 Input Parameters in PopGen



Data Preparation and Software Implementation

Data sources

Two types of data from the ABS Population and Housing Census are required to prepare the input data for population synthesis:

- 1% microdata from the Confidentialised Unit Record File (*CURF*)
- Aggregate cross tabulation data from the ABS Table Builder, Population and Housing Census

The CURF microdata provides the necessary seeds matrix for household and person samples, while the aggregate census data extracted from the ABS Table Builder provides known household and person constraint or marginal totals. This research study is based on the ABS Census and Housing population data from 2011. While the latest census was conducted in 2016, the microdata (CURF) 2016 was not released until mid-2019.

Data Preparation for Population Synthesis

Five input data files are required to generate a synthetic population using PopGen:

- Geographical correspondence data files
- Sample data files at household level
- Sample data files at person level
- Constraints total or marginal distribution data files at household level
- Constraints total or marginal distribution data files at person level

These data files need to be assembled and pre-treated before feeding into the population synthesiser. As PopGen was primarily developed and designed to be used in the United States, it is necessary to reformat and adapt all input data from the Australian population census data files to generate synthetic populations for Australian cities. There are a few common features in all input data files:

- The first row specifies the variable names
- The second rom specifies the variable types:
- Integers – bigint
- Floating point value – double
- Strings – text
- There is a fix format for the first few columns in each input data file. Field specifications in the first few columns are compulsory. Specifications of subsequent columns are optional.

The following subsections outline the necessary steps involved to produce these data files.

Geographical Correspondence

The geographical correspondence data file provides links between the geographical classifications used in CURF microdata, census marginal totals from the ABS table builder and the existing built-in geographical classification used in PopGen.

The preparation of the geographical correspondence data files begins by extracting for each city based on GCCSA from the ABS ASGS Main Structure and GCCSA. This step defines the study area or geographical coverage for each city at SA1 level. The next step is to link all included SA1s

in each GCCSA to area codes in CURF. Although the geographical areas assigned to CURF microdata and census marginal totals from the table builder are both based on ASGS, they were coded at different geographical hierarchical level. Microdata from CURF 2011 was coded at Statistical Region (SR), whereas the census marginal totals from table builder were coded based on Statistical Area Level (SA). SR is allocated based on SA4. Hence, it is possible to link both data sources and ensure that the geographical coverage in CURF and in the marginal totals from census data are well aligned. Appendix 2 shows the geographical correspondence between CURF and census data for the three Australian cities at SA4. As mentioned, the build in geographical structure in PopGen was designed for PUMA. PUMA is generally based on the US Census geographical hierarchy by block group, census tract, county and state. The compulsory fields in the first few columns of the geographical correspondence file are in the order of county, tract, bg (block group), state, pumano (PUMA number), stateabb and county name. These columns are matched with the ASGS geographical classifications for Australian cities (Table 4).

Table 4 Data Structure for Geographical Correspondence File

US	county	tract	bg	state	pumano	stateabb	countyname
	int	int	int	int	int	text	text
Australia	GCCSA	SA4	SA1	State Code	CURF Area Code	State	GCCSA Name

Record linkages and matching control categories

Before preparing for the sample and marginal files at household and person level, the corresponding categories to each selected control variable in these data files must be aligned. Generally, variables in CURF are categorised in more aggregated groups with fewer categories than the actual census data. It is necessary to regroup or collapse some of these groups, either in CURF, census data or in both to ensure that the grouping of categories for the selected control variables are consistent throughout. Appendix 3 and 4 show how the selected control variables were linked by common categories assigned at both household and person level.

The data structure for household and person sample files are shown in Table 5 and Table 6 respectively. In PopGen, the first four columns of these sample files, that is state, pumano, hhid and serialno, are compulsory fields. Additional control variables or attributes can be added in the subsequent optional fields. State and pumano codes in these sample files are assigned in accordance to the geographical correspondence file prepared earlier. The household identifier (hhid) in Table 5 and person identifier (pnum) in Table 6 are allocated by assigning a unique identifier to each housing or person unit. In this study, hhid is generated by creating a sequence increment by one. The serialno is obtained from the truncated ABS household identifier (ABSHID) and person identifier (ABSPID) in CURF accordingly. Both sample files also contain a unique identifier for each household (Hhiduniqueid) and for each person (Personuniqueid) in the last column of the data file. These are unique numbers used to identify households and persons from the original sample file or seed data, to which in this study were assigned by concatenating the serialno with hhid for household sample file or with pnum for person sample file. The assignment of the household and person identifier is required to be handled with care to ensure each person is numbered correctly in the household. This is important for data manipulation before and after the population synthesis process. Finally, there is an extra column in the person sample file to specify the initial weight for each person (pweight), which is equivalent to 100 for each person in CURF.

Table 5 Data Structure for Household Sample File

state	pumano	hhid	serialno	Household variable 1	…	Household variable X	Hhid-uniqueid
bigint	bigint	bigint	bigint	bigint	bigint	bigint	bigint

Table 6 Data Structure for Person Sample File

state	pumano	hhid	serialno	pnum	pweight	Person variable 1	…	Person variable Y	Person-uniqueid
bigint	bigint	bigint	bigint	bigint	bigint	bigint	bigint	bigint	bigint

Balancing census total counts

When preparing for the household- and person-level marginal distributions data files, data are extracted from the ABS Census Table Builder for all selected control variables. Each extracted univariate dataset is a cross tabulated table containing marginal totals of the pertaining control variable by the selected geographical area (i.e. SA1 for this study) of the study region. These datasets are then collated to provide a multi-dimensional marginal data file. At this point, it is essential to ensure that the sum of the marginal totals for all control variables is consistent. Any inconsistencies in marginal totals between the control variables will lead to a breakdown of the IPF and IPU processes in the synthesis algorithm. Census data in the ABS Table Builder are subject to random perturbation to protect the confidentiality of individuals. The randomisation process introduces data inconsistencies and hence inevitably affects the total marginal constraints of the control variables. This is a common practice in census data. However, inconsistent marginal constraints inhibit convergence in a constraint optimisation problem embedded within the synthesis process. The process of overcoming data inconsistencies in census data is known as Census data 'balancing' (Chin and Harding, 2006). This process is one of the most time consuming processes in the creation of small-area weights (Chin and Harding, 2006).

In this study, a sequence of readjustments and redistributions steps have been formulated to achieve consistencies across marginal totals of each selected control variables in each geographical area. In a nutshell, the process involves collating and rearranging all extracted data from the ABS Census Table Builder in such a way that marginal counts for all selected control variables can be realigned across all geographical areas simultaneously. This process can be effectively and efficiently applied to balance the census total counts in seconds using common statistical programs such as SAS, R, SPSS etc.

Below is a summary of the steps taken to realign the imbalance marginal totals across different control variables within each geographical area:

1. First, marginal totals for each selected control variable from the Table Builder were extracted into the selected statistical program separately. This step creates a dataset for each control variable datasets at SA1 level for the study area. Each dataset contains the marginal counts by SA1 for each category of the control variable.

2. Then, for each margin dataset created in step one, a new variable is introduced in each row to identify the associated control variable. For instance, if the margin dataset is for household composition (HHCD) with four categories, then the new variable is labelled HHCD for each SA1. Once the new variable is incorporated for each SA1, the category names of the control variable are renamed to a set of generic variable names; such as Mar1,

Mar2, Mar3 and so on. This step is repeated for all marginal datasets where the same generic variable names are used to rename the categories of each control variables. The purpose of this step is to ensure that when all margin datasets are concatenated, the new identifier created retains the identity of each variable, while the generic variable names allow data comparisons and adjustments simultaneously.

3. At this point, the sum of marginal totals for the same control variable should be the same for each SA1. For example, the sum of marginal total for household composition (HHCD) should be the same as the sum of marginal total for dwelling structure (STRD) within the same SA1. However, as discussed, that is not always true due to the randomisation process by ABS. Table 7 compares the marginal totals of four household control variables added up for all SA1 in Greater Sydney:

Table 7 Comparison of Marginal Totals from ABS Census Table Builders 2011

Marginal totals for household counts				
GCCSA Greater Sydney	HHCD Household Composition	NPRD Number of Persons Usually Resident in Dwelling	STRD Dwelling Structure	VEHD Number of Motor Vehicles
1,601,530	1,598,439	1,593,904	1,597,383	1,590,648

Source: ABS Census Table Builder

While the marginal total for the entire Greater Sydney is 1,601,530, the marginal total for every selected control variable summed up from all SA1 is different. In particular, the difference between the household counts for variable VEHD is almost 11,000 less than the actual household counts for the entire Greater Sydney. The discrepancies inevitably also occur for each control category of corresponding control variable between the marginal totals from summing up all SA1s and from the entire study region. These differences can be substantive.

The third step involves concatenating all margin datasets prepared in step 2 into one single dataset. To adjust the inconsistencies of marginal totals at SA1 level, a marginal total of one control variable is selected to be the benchmark for all other control variables at each SA1 level. Adjustments of marginal totals begin by calculating discrepancies between the benchmark totals and census data totals from the ABS Table Builder in the concatenated table.

4. Then the concatenated table is split up into individual dataset again by control variables. The discrepancies at SA1 calculated in step 3 is redistributed among the variable categories based on the probability distribution for the benchmark control variable. After the redistributions, the marginal total in each SA1 is now equal to the benchmark total. However, the probability redistribution process produces non integer values. These values are rounded up to the nearest integers. The rounding up process slightly distorted the perfectly adjusted marginal totals and required a second round of alignments with the benchmark totals. These misalignments from rounding were eliminated by adjusting the last category of each variable with a non-zero value. The process continues until all marginal totals for each control category is adjusted to equal the benchmark marginal total at SA1 again. Step 4 is repeated for every control variable.

5. The final step concatenates all datasets adjusted in step 4. The balances of marginal totals for each control variables are checked again to make sure that they are consistence at SA1 level. All processed marginal datasets are then merged again by SA1 to form one large dataset. The merged dataset should contain columns with marginal totals of all categories

of the selected control variables and each row represents a geographical area (SA1) of the study region.

The balancing census counts process discussed above are applied to both household and person marginal data files. Once both data files are consistent in marginal totals, each file is merged with the geographical correspondence file created before to fit into the build-in geographical structure of the population synthesiser. In this study, SAS has been used for balancing the total counts of the census data.

The data structure for marginal marginal distribution data files at household level and person level are as shown in Table 8 and 9. The first four columns of these files are compulsory fields, that is state, county, tract and bg. Subsequent fields are marginal totals listed in one column for each category of the selected control variable. When the marginal totals for every control category of all control variables are added in the optional fields, these files provide the marginal distributions or constraints that the synthetic population strives to match at household and person level. Now all input datasets are ready for population synthesis.

Table 8 Data Structure for Marginal Distribution Data File at Household Level

state	county	tract	bg	Household Variable 1 Category 1	…	Household Variable 1 Category N	…	Household Variable X Category 1	…	Household Variable X Category N
bigint	bigint	bigint	bigint	bigint	…	bigint	…	bigint	…	bigint

Table 9 Data Structure for Marginal Distribution Data File at Person Level

state	county	tract	bg	Person Variable 1 Category 1	…	Person Variable 1 Category N	…	Person Variable Y Category 1	…	Person Variable Y Category N
bigint	bigint	bigint	bigint	bigint	…	bigint	…	bigint	…	bigint

The IPU procedure is solely based on a mathematical algorithm. The performance of the IPU algorithm relies on the quality and integrity of input data. Any inconsistencies in the input data would prevent the mathematical algorithm from producing a synthetic population that meets the acceptable validation criteria. Data consistencies must be maintained in population totals for person and household variables, assignments of sample serial numbers, and links in geographical correspondence for all input datasets.

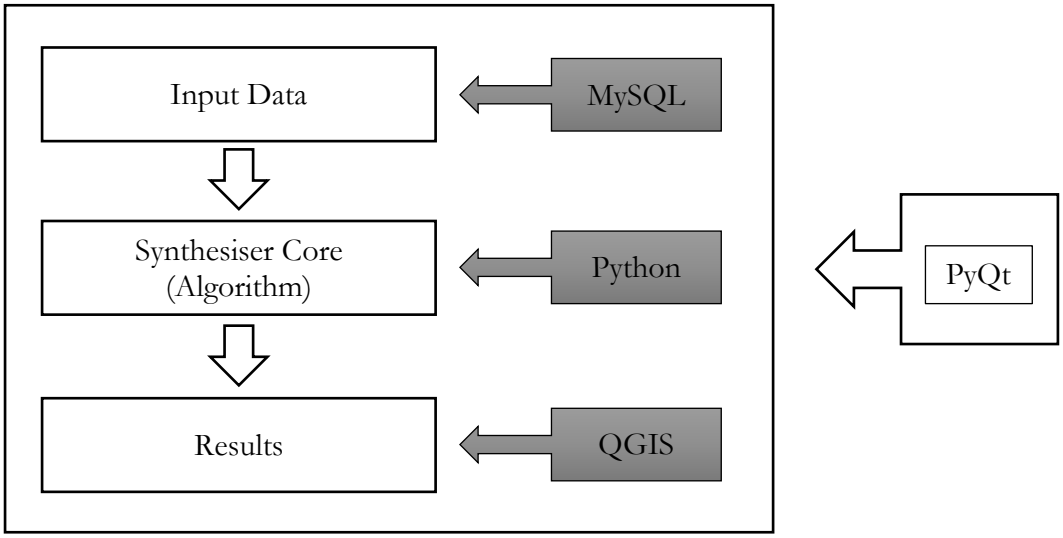
Software Implementations

To date, PopGen is the only standalone software that uses the IPU algorithm. As discussed earlier, the software was developed in 2009 by Karthik Kondari and Bhargava Sana from the School of Engineering at Arizona State University based on the IPU algorithm by Ye et al. (2009).

PopGen was mainly developed and designed to be used in the US. The built-in graphic interface features and spatial data are not applicable to other countries and the mapping of simulated study region outside the US is not possible within the system. As discussed in previous sections, adaptations of input data are necessary as the designated data structure and format were formatted in accordance with PUMA.

PopGen is implemented using a suite of python based open-source software and uses MySQL as the database manager (Figure 4). Essentially, the GUI based PopGen software application integrate multiple open-source software on a platform to deploy multiple interdependent stacks of codes. PopGen uses MYSQL to manage data. PopGen 1.1 generally works with most recent versions of Window operating system. The supplementary software package which support PopGen can be downloaded at https://www.mobilityanalytics.org/popgen.html#Software. The link also provides instruction for installation. However, the application of multiple open-source software on a proprietary platform often creates complex issues, such as transparency, compatibility and reliability of the open-source software. Most open-source software evolves over time, but structured supports and updates are limited. There are many ongoing parallel developments for various open-source software. It is not always clear what functionalities are present or improved upon in each updated version. The assembly of multiple updated versions of open-source software only works when all software versions in PopGen are stable and compatible with the operating system.

Figure 4 PopGen: Open Source Framework



Source: ASU 2009

Below is an overview of PopGen methodology (Figure 5). The IPF procedure is first applied to the priors adjusted household/person type sample data and also household /person marginal distributions. These seed data are then adjusted to account for the zero-cell problem and also the zero-marginal problem. The IPF procedure proceeds to estimate number of person/household type in each cell. Next, the IPU procedure is used to estimate household level sample weights that satisfy constraints obtained from IPF. After the IPU procedure, the IPU computed weights are used as the selection probabilities to construct a synthetic population. Monte Carlo drawing procedures are used in the household selection process.

PopGen is menu driven. In summary, there are six major steps in the setup phase:

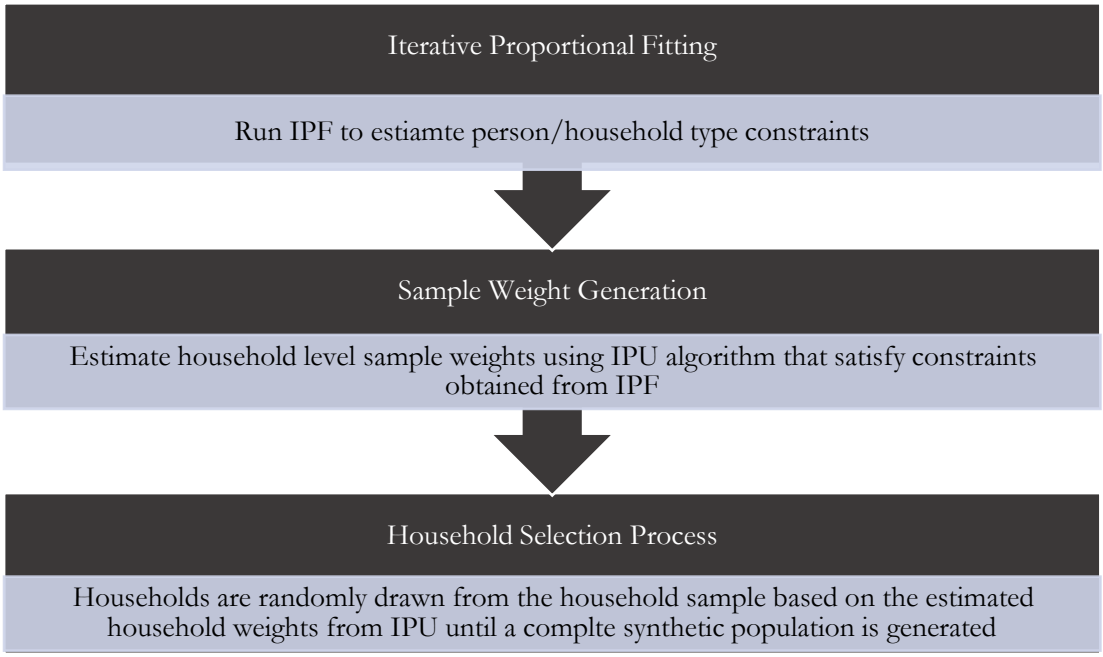
Step 1 Create a new project for a region. This includes assign a project name, select a project file location and enter the project descriptions.

Step 2 Specify a geographical resolution. For the case of Australian cities, the only option is to choose Traffic Analysis Zone (TAZ), which is a geographical resolution that offers the flexibility to specify your own geographical correspondence other than PUMA.

Step 3 Specify sample files for household and person level and their file locations.

Step 4 Specify marginal files for household and person level and their file locations.
Step 5 Establish MYSQL connection.
Step 6 A project summary of all selected options in previous steps is displayed. Changes of these selected setting is possible if there is a need.

Figure 5 Overview of PopGen Methodology



Once a new project is created, the next stage is to import and process all the input files prepared earlier to a MySQL database. This can be carried out under the data menu. The final step of the importing process also involves downloading and extracting the shape files containing the boundaries of the geographies of interest. This is optional. The data menu also offers the options to display the imported data files and modify marginal distributions. The display option allows checking of import data, ensuring that the read in data are formatted correctly and designated to the right file locations.

The final stage before running the synthesis process is to set the corresponding variables and specify the run parameters under the synthesizer menu option. The setting of corresponding variables is for establishing correspondence between the control categories of households and persons. The option for linking the group-quarters is not required for Australian data. This is where the number of household and person control variables for running the population synthesis is specified. Correspondences must set for at least one control variable in household and person. Below is a list of control variables used in running the synthesis routines for this study. Overall, 840 household type constraints and 144 person type constraints have been imposed on the synthesis process in this research study (Table 10).

It is also under the synthesizer menu option, the parameters/setting allows for four type of adjustments:

- IPF related parameter
  For specification of tolerance level for convergence and maximum number of iterations in the IPF procedures.
- IPU related parameter

For specification of tolerance level for convergence and maximum number of iterations in the IPU procedures.

- Synthetic population draw-related parameters
  For specification of the maximum number of draw and the threshold level of the p-value to achieve a desirable synthetic population
- Rounding procedure
  The rounding procedure is used to convert estimated weights with decimal values to integers.

Table 10 Selected Household and Person Control Variables for Population Synthesis

Control Variables at Household Level	
HHCD	**Household Composition**
HHCD1	One family household
HHCD2	Two or more family household
HHCD3	Non- family household
HHCD4	Other groups
STRD	**Dwelling Structure**
STRD1	Separate house
STRD2	Semi-detached, row
STRD3	Flat, unit or apartment
STRD4	Other dwelling
STRD5	Other groups
NPRD	**Number of people (Derived)**
NPRD1	1 person
NPRD2	2 persons
NPRD3	3 persons
NPRD4	4 persons
NPRD5	5 persons
NPRD6	6 persons or more
NPRD7	Not applicable
VEHD	**Number of Motor Vehicles**
VEHD1	None
VEHD2	1 motor vehicle
VEHD3	2 motor vehicles
VEHD4	3 motor vehicles
VEHD5	4 or more motor vehicles
VEHD6	Other groups
Total Number of Household type constraints = 840	

Control Variables at Person Level	
Agep	**Age**
Agep1	0-4 years
Agep2	5-9 years
Agep3	10-14 years
Agep4	15-19 years
Agep5	20-24 years
Agep6	25–29 years
Agep7	30–34 years
Agep8	35–39 years
Agep9	40–44 years
Agep10	45–49 years
Agep11	50–54 years
Agep12	55–59 years
Agep13	60–64 years
Agep14	65–69 years
Agep15	70–74 years
Agep16	75–79 years
Agep17	80–84 years
Agep18	85 years and over
Sexp	**Sex**
Sexp1	Male
Sexp2	Female
Lfsp	**Labour Force Status**
Lfsp1	Employed
Lfsp2	Unemployed
Lfsp3	Not in the labour force
Lfsp4	Other groups
Total Number of Person type constraints = 144	

Note that PopGen allows a project to test up to five different scenarios under the scenario option on the menu bar in the PopGen interface. The scenario option offers the flexibility to test a number of different settings, including the number of household- and person-level control variables, geographies and changes in selection criterion for the IPF and IPU algorithms. After all the necessary configuration and specifications of the new project is completed, the synthesis process can be initiated by proceeding to run the synthesiser under the synthesizer menu option. This step will prompt the selection of geographies included for running the new project. At least one or more geographies can be selected for starting the process of population synthesis. After all selected geographies are highlighted and transferred into the synthesis process, it is time to execute the synthesis algorithm. Permission will be requested for pre-processing the data for any changes in

the control variables and their categories. The progress of the population synthesis can be monitored from the output screen.

Upon completion of the synthesis process, synthesised results can be viewed for the entire study region or at the individual geographical level. There are four options to examine the synthesis results using the results menu:

- Average Absolute Relative Difference (AARD) shows the distribution of the AARD values of geographies selected for population synthesis
- P-value shows the distribution of p-values for all geographies selected for population synthesis.
- Distributions of household attributes compares numbers of the selected control variables at the household level generated from the synthesis process and those from the marginal data files (the actual marginal distributions from census data).
- Distribution of person attributes compares numbers of the selected control variables at person level generated from the synthesis process and those from the marginal data files (the actual marginal distributions from census data).

The built-in analysis of the synthesised results is fairly limited, however in depth analysis of the results is possible using other software. The synthesised households, persons and performance statistics can be easily exported to another software application in CSV or tab-delaminated format.

Synthesised Estimates

This paper mainly features the synthesised estimates and performance results for Greater Sydney as an example to illustrate the validation process, testing of IPU algorithm and data treatment after the synthesis process. The validation process gauges if the synthetic data reflected the source data and the relationships between the control variables were retained (Knight et al. 2017). Further statistical tests were carried out to evaluate the goodness of fit of the generated synthetic populations. To fundamentally assess whether the IPU algorithm produced reasonable spatial allocations of the control variables, the synthesised results were mapped and compared to the same variables from the actual census data. Detailed descriptions and discussions of the synthesis process and performance results for all three cities can be found in Lim, 2019.

Synthesised Households, Greater Sydney

Overall, the synthetic population generated produced almost a virtual match to the actual population at household level for Greater Sydney, with zero percent difference in percentage. There is a discrepancy of -4.4 percent between the actual number of persons and the synthesised number of persons for Greater Sydney (Table 11). The issue of under representation of synthetic persons is investigated further in the next section and the results are rectified by a heuristic algorithm proposed for this study in the later section.

Table 11 Comparison of Overall Synthesised and Actual Population of Greater Sydney, 2011

	Actual	Synthesised	Difference (%)
Households	1,598,439	1,598,433	-0.00
Persons	4,308,248	4,118,543	-4.40

Figure 6 provides overall household aggregate distributions by the selected control variables for the actual and synthesised data. The aggregate comparisons by household attributes show that the

synthesised household distributions of these variables are generally closely matched with the actual distributions.

Figure 6 Household Estimates at SA1 by Control Variable, Greater Sydney 2011

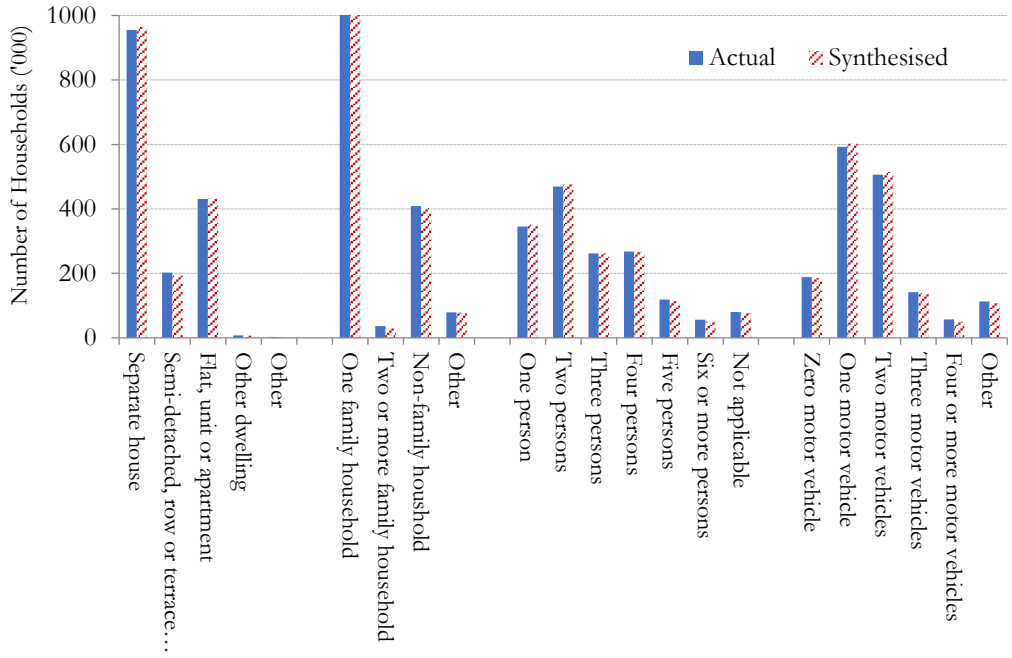


Table 12 presents the distributions by the household attributes in percentage between the synthesised and actual data. The percentage difference for each category of the control variables were calculated. As shown, the discrepancies in percentage points for each sub population group by the control variables were all within ±1%.

Table 12 Distribution of Actual and Synthesised Population at SA1 by Household Attributes, Greater Sydney 2011

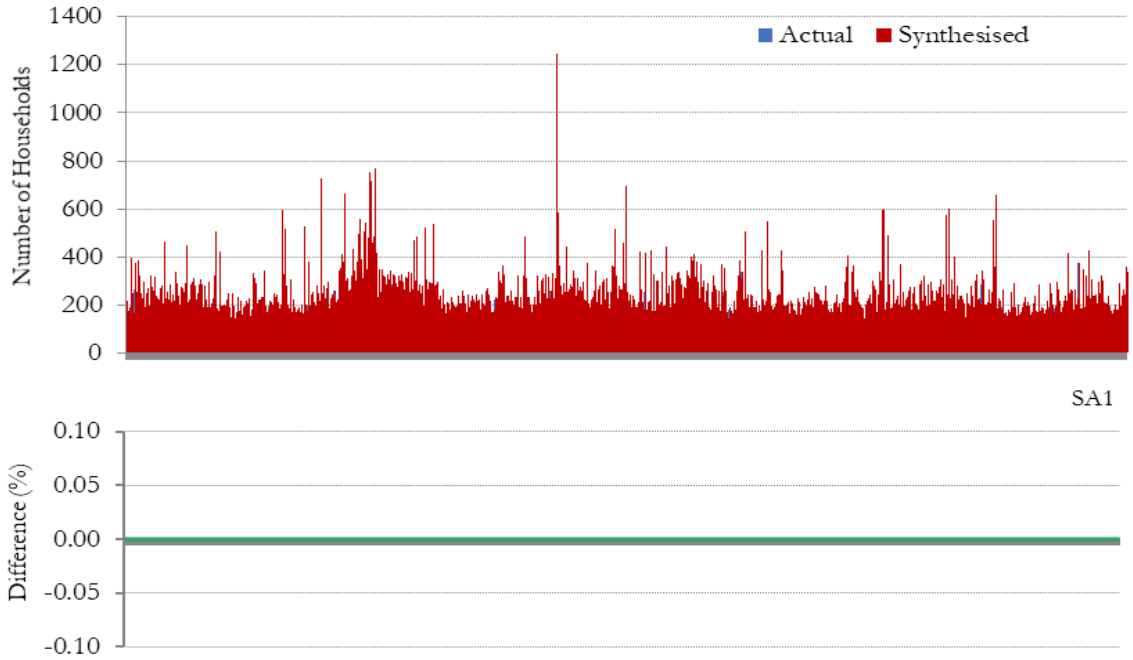
	Actual %	**Synthesised** %	**Difference** % point
Dwelling Structure			
Separate house	59.71	60.36	-0.65
Semi-detached, row or terrace house, town house, etc.	12.68	12.09	0.59
Flat, unit or apartment	26.96	26.99	-0.03
Other dwelling	0.50	0.45	0.06
Other	0.14	0.11	0.04
Household Composition			
One family household	67.19	68.07	-0.88
Two or more family household	2.30	1.83	0.47
Non-family household	25.56	25.18	0.38
Other	4.95	4.92	0.03
Number of Persons Usually Resident in Dwelling			
1 person	21.57	21.96	-0.39
2 persons	29.35	29.82	-0.47
3 persons	16.40	16.39	0.01

Table 12 Distribution of Actual and Synthesised Population at SA1 by Household Attributes, Greater Sydney 2011 (continued)

	Actual %	Synthesised %	Difference % point
4 persons	16.73	16.71	0.02
5 persons	7.42	7.11	0.30
6 persons	3.51	3.16	0.35
7 persons	5.01	4.84	0.17
Number of Motor Vehicles			
None	11.60	11.81	-0.20
1 motor vehicle	37.69	37.06	0.64
2 motor vehicles	32.15	31.64	0.51
3 motor vehicles	8.69	8.85	-0.17
4 or more motor vehicles	3.12	3.57	-0.45
Other	6.74	7.07	-0.33

Figure 7 shows the number of synthesised households compared to the number of actual households at SA1 level. The upper chart shows the aggregate totals between actual and synthesised number of households by SA1, and the lower chart shows the percentage difference between the actual and synthesised households for each SA1. The number of synthesised households were perfectly matched with the actual households at every synthesised SA1 for Greater Sydney. There is no percentage difference between the two distributions. This is expected at household level. The IPU algorithm prioritised matching the distributions of all selected control variables at household level.

Figure 7 Distributions of Actual and Synthesised Households at SA1, Greater Sydney 2011



Synthesised Persons, Greater Sydney

Figure 8 provides overall distributions of persons by the selected control variables between the actual and synthesised data.

The overall underestimated number of persons of the IPU algorithm is reflected in each of the person control variable (Table 13). Table 13 presents the frequency distributions between the synthesised and actual data. It can be seen that the differences in percentage points for the distributions of person control variables are mostly within one percent point from the actual distributions. Comparatively, percent differences between the actual and synthesised persons by person attributes are smaller relative to the number of synthesised households by household attributes. In fact, the spread of the distributions for synthesised persons by gender and age matched the actual distributions very well.

Figure 8 Person Estimates at SA1 by Control Variable, Greater Sydney 2011

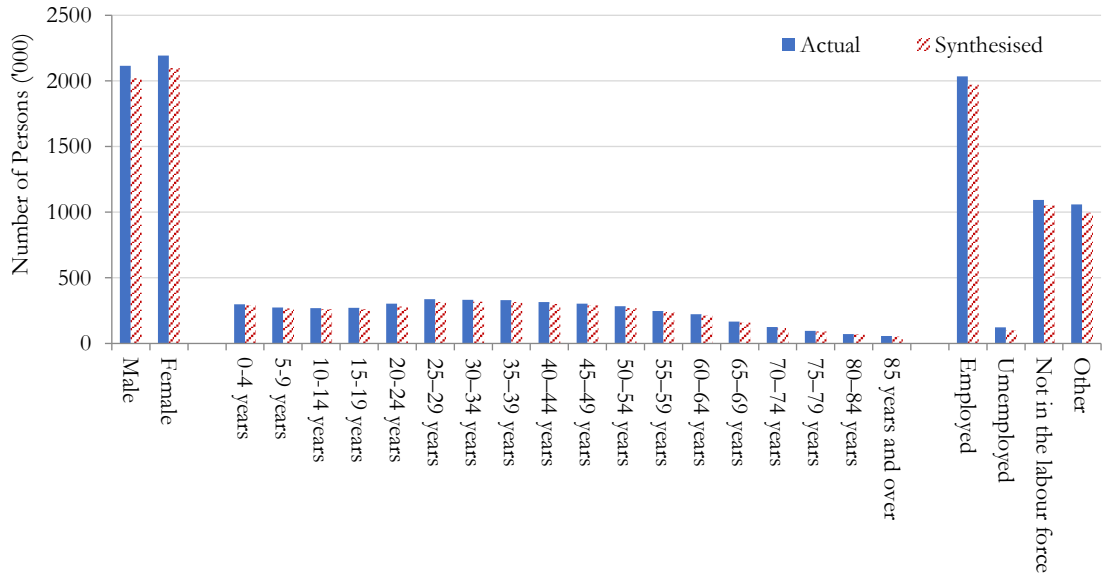


Table 13 Distribution of Actual and Synthesised Population at SA1 by Person Attributes, Greater Sydney 2011

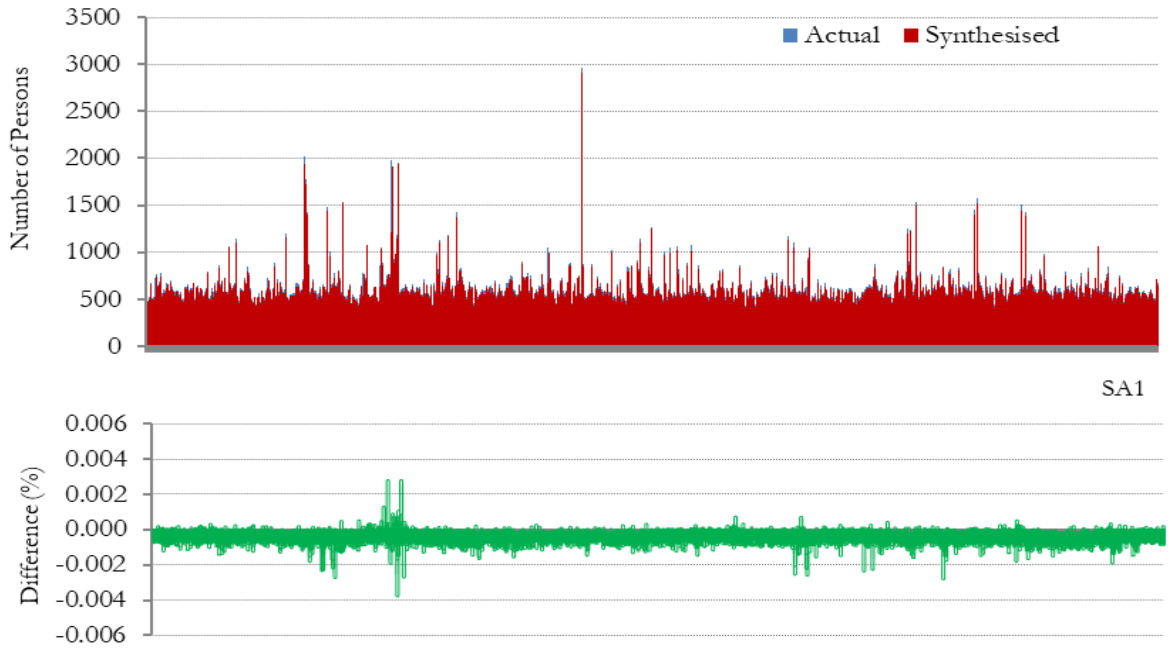
	Actual %	Synthesised %	Difference % point
Gender			
Male	49.09	49.18	-0.06
Female	50.91	50.82	0.06
Age			
0-4 years	6.93	7.09	0.16
5-9 years	6.38	6.47	0.08
0-14 years	6.22	6.34	0.12
15-19 years	6.31	6.36	0.05
20-24 years	7.05	6.82	-0.22
25–29 years	7.84	7.67	-0.17
30–34 years	7.72	7.72	-0.01
35–39 years	7.66	7.66	0.00

Table 13 Distribution of Actual and Synthesised Population at SA1 by Person Attributes, Greater Sydney 2011 (continued)

	Actual %	Synthesised %	Difference % point
40–44 years	7.31	7.32	0.01
45–49 years	7.05	7.07	0.02
50–54 years	6.60	6.56	-0.04
55–59 years	5.76	5.8	0.04
60–64 years	5.19	5.2	0.01
65–69 years	3.84	3.85	0.00
70–74 years	2.91	2.89	-0.02
75–79 years	2.21	2.21	0.00
80–84 years	1.69	1.69	0.00
85 years and over	1.31	1.29	-0.02
Labour Force Status			
Employed	47.21	47.85	0.65
Unemployed	2.85	2.45	-0.39
Not in the labour force	25.39	25.51	0.13
Not stated	24.56	24.18	-0.38

There were small variations by percentage between the actual and synthesised persons at SA1 level. In Figure 9, the top chart shows the aggregate number of synthesised persons versus actual persons. The variations are not prominent and hardly observable as the percentage difference in each SA1 is between ±0.005 percent. In fact, 99.7 percent of the synthesised persons are accurate to 0.001 percent difference when compare to the actual number of persons by SA1. Overall, the number of synthesised persons were reasonably matched with the actual persons at every synthesised SA1 for Greater Sydney.

Figure 9 Distributions of Actual and Synthesised Persons at SA1, Greater Sydney 2011



Performance Measures

Performance Statistics

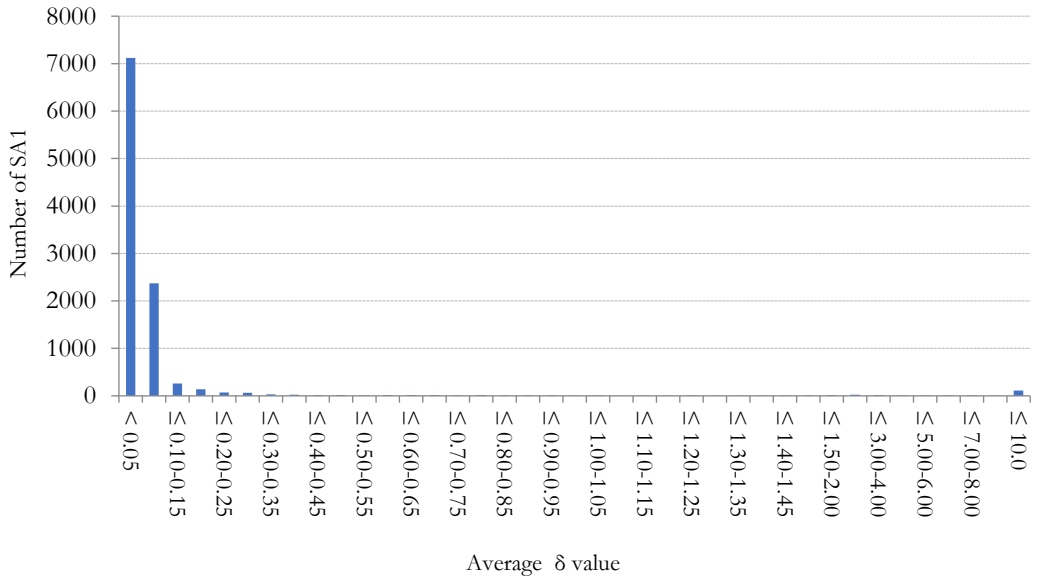
Three performance measures were used to gauge the goodness of fit of the generated synthetic populations:

- Average Absolute Relative Difference
- χ^2 statistics and P-value
- Validation of Spatial Distribution by control variables

Distribution of Average δ Value

The Average Absolute Relative Difference (AARD) measures the average deviation of the weighted sums with respect to the household/person type constraints. The average value across all constraints is denoted by δ Value and serves as an overall goodness of fit measure for each complete iteration in the IPU algorithm. The δ Value is useful in monitoring the convergence of the IPU algorithm. However, it is not an appropriate measure of fit for the synthetic population as the differences in magnitude between the synthesised actual distributions are concealed in the way δ Value is derived. Figure 10 displays a positively skewed distribution of the δ values with long tail, whereby more than 90% of SA1s for Greater Sydney were concentrated on the lower ranges.

Figure 10 Distribution of δ Values at SA1 Level, Greater Sydney 2011



Note that less than one percent of SA1s have been observed to have δ values which fall into the higher end of the spectrum in the three charts above. These outliers arise possibly from the inherent variability of the sample data. It is possible that an outlier was a result of legitimate random sampling from the population. Sample size plays a role in the probability of outlying values (Osborne and Overbay 2004). Within a normally distributed population, a given data point is more likely to be drawn from a highly concentrated area of the distribution, rather than from one of the tails (Evan 1999; Sachs 1982). A large sample tends to resemble more of the population from which it was drawn, and thus the likelihood of the occurrence of outlying values becomes greater. In other words, there is about a one percent chance of getting an outlying data point from a normally distributed population. That means, on average there is about one percent of the sample data points that are three standard deviations away from the mean (Osborne and Overbay 2004).

Distribution of Chi-Square Statistics

An alternate measure of fit for the synthesised data is the Chi-square (χ^2) statistic. χ^2 is commonly used to statistically compare two distribution of interest. The Chi Square distribution is the distribution of the sum of squared standard normal deviates. A standard normal deviate is a random sample from the standard normal distribution. The degrees of freedom of the distribution is equal to the number of standard normal deviates being summed. The χ^2-distribution with k degrees of freedom is the distribution of a sum of squared k independent standard normal random variables. The chi-square curve approaches normal distribution as the degree of freedom increases. A chi-square goodness of fit test determines if a sample matches a population. In the context of this study, the χ^2 statistics serve as an appropriate measure of fit to evaluate the statistical differences between the joint distributions obtained from the IPU algorithm and the actual joint distributions.

Figure 11 Distribution of χ^2 Values at SA1 Level, Greater Sydney 2011

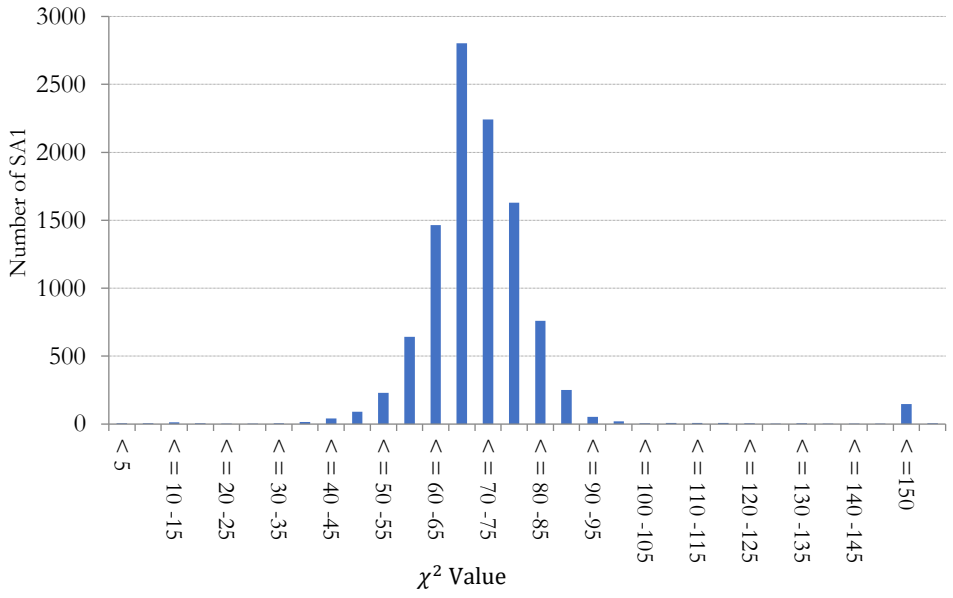


Figure 12 Distribution of p-values at SA1 Level, Greater Sydney 2011

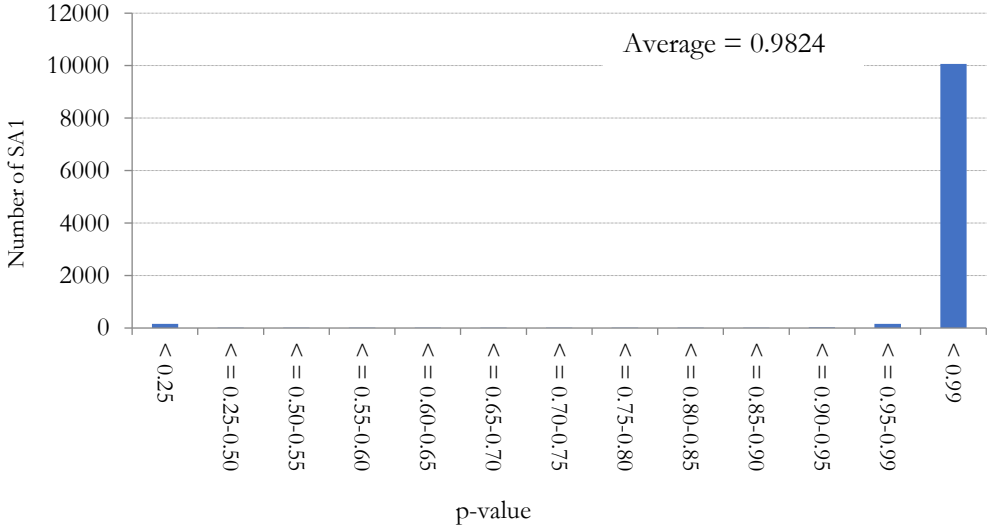


Figure 11 shows the distribution of χ^2 values with 143 degree of freedom and Figure 12 shows the distribution of p-value across all SA1 in Greater Sydney in 2011. The average p-value across 10,487 SA1 in Greater Sydney is 0.9824, indicating a high level of confidence that the synthesised joint distribution matches the actual joint distribution. In fact, 96.4 percent of all SA1 in Greater Sydney has a corresponding p-value of more than 0.99, of which nearly 60 percent of those p-value are very close to unity.

Validation of Spatial Distribution by Control Variable

To fundamentally assess whether the IPU algorithm produced reasonable spatial allocations of the control variables, the synthesised results were mapped and compared to the same variables from the actual census data.

There is a basic feature in PopGen for plotting thematic maps households or persons that is linked to an open source mapping software; Quantum Geographic Information System (QGIS). However, this feature is not available for any input data outside the United States. Hence, the thematic mapping for any non-US cities need to be conducted using other independent mapping tools. MapInfo is used for the analysis presented in this section. The spatial data for Greater Sydney, Melbourne and Brisbane were obtained from the ABS website. These shape files for mapping are readily available from the link below:

https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.001July%202011?OpenDocument

These maps are in the Mapinfo and ESRI (Environmental Systems Research Institute) format. Depending on the software, both formats are commonly used for mapping.

The number of synthesised and actual households and persons by selected subcategories of the control variables at SA1 were aggregated to SA3 for visual comparisons. To obtain percentage difference in number of households and persons generated compared to the actual population, the percentage distribution by the control variables across selected subcategories were calculated for each SA3 and then subtracted from the percentage distribution of the equivalent actual data. As the accumulative differential effects of these distributions showed very minimal differences at finer level of geography (i.e. SA2 and SA1), mapping results at spatial resolution smaller than SA3s could not effectively provide clear visual distinctions between synthesised and actual distributions. This validation process helps to determine whether the population synthesis process have accurately recreated the spatial variation and retained the spatial heterogeneity of the actual data.

The following maps compare the spatial distributions of aggregated synthesised and actual estimates for each selected subcategory of the control variables by SA3. In each figure, two smaller maps on the side provide a visual comparison of the aggregate distribution by SA3 for either the number of households or persons. The bigger map on the right displays the positive and negative accumulative differential effects in percentage for the corresponding subcategory of a control variable. If the percentage differences are positive, indicating over synthesised percentages of households or persons, the range of these percentage differences are shown in modular colour of blue. Whereas if the percentage differences are negative, indicating under synthesised percentages of households or persons, the range of these percentage differences are shown in the modular colour of red.

The synthesised and actual results were mapped for household composition (one family household and multiple family household) and for labour force status (employed and unemployed persons) for Greater Sydney. Figure 13 shows the spatial allocation of IPU estimates for one family

household in Greater Sydney in 2011. The distribution of the synthesised results was nearly identical to the actual distribution, asserting that the spatial variation of the actual distribution was recreated in the synthesised distribution and the spatial heterogeneity is retained throughout the synthesis process. The accumulative percentage differences were within the range of ±0.04 percent. Twelve SA3s were marginally under-synthesised, of which eleven of them are within -0.01to 0.02 percent. The only SA3 with a slightly higher percentage difference of -0.05 percent for the number of synthesised one family household is in Sydney inner city, follow by a percentage difference of -0.02% in both Gosford and Cronulla-Miranda-Caringbah. The rest of SA3 in Greater Sydney were minimally over synthesised between 0.01 to 0.02 percent with two SA3 (Marylands – Guildford and Fairfield) at a differential percentage of 0.03 percent.

Figure 13 Synthesised and Actual Distribution of One Family Households in Greater Sydney 2011

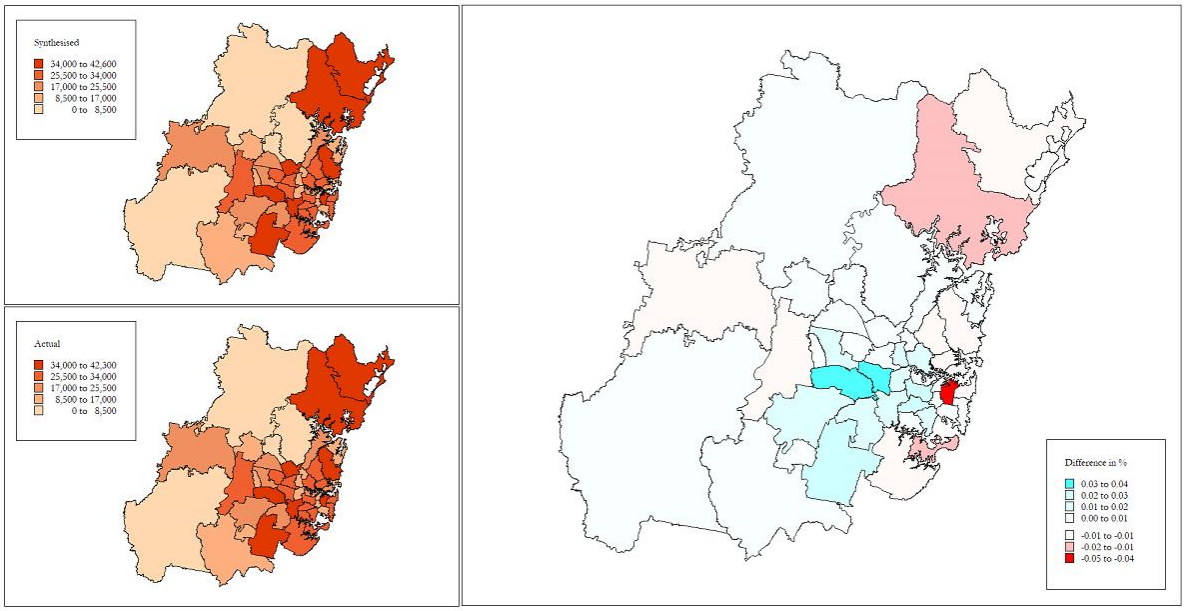


Figure 14 Synthesised and Actual Distribution of Multiple Family Households in Greater Sydney 2011

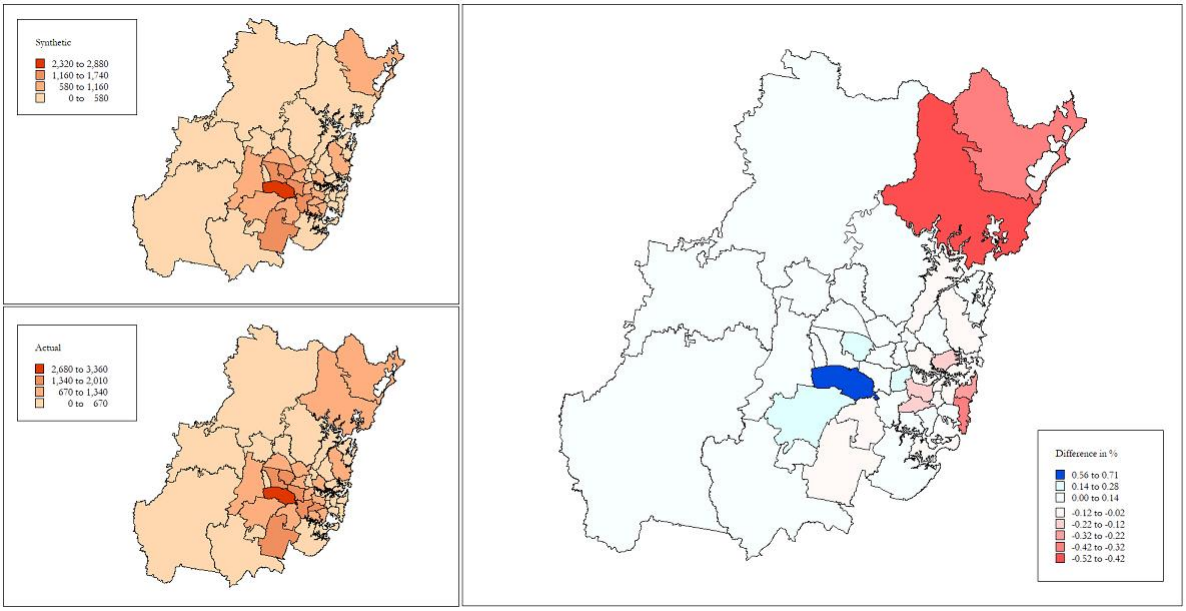


Figure 14 shows nearly similar distributions between synthesised and actual number of multiple family households. Gosford and Wyong appeared to have marginally under synthesised by a percentage of 0.52 percent and 0.8 percent respectively. Fairfield was slightly over-synthesised by 0.71 percent.

Figure 15 Synthesised and Actual Distribution of Employed Persons in Greater Sydney 2011

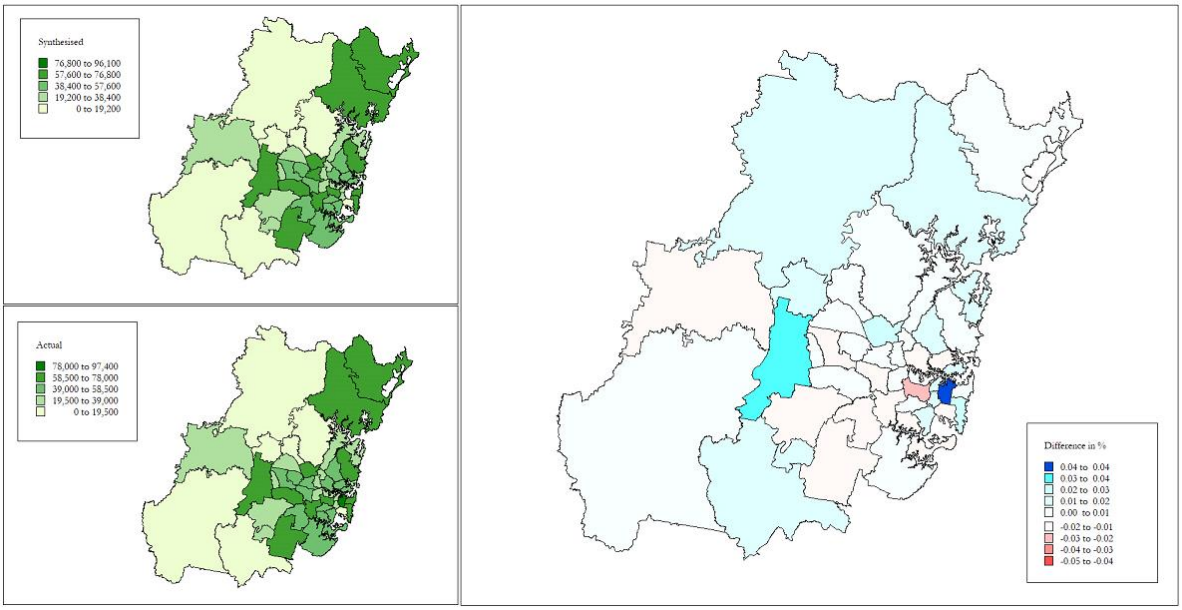
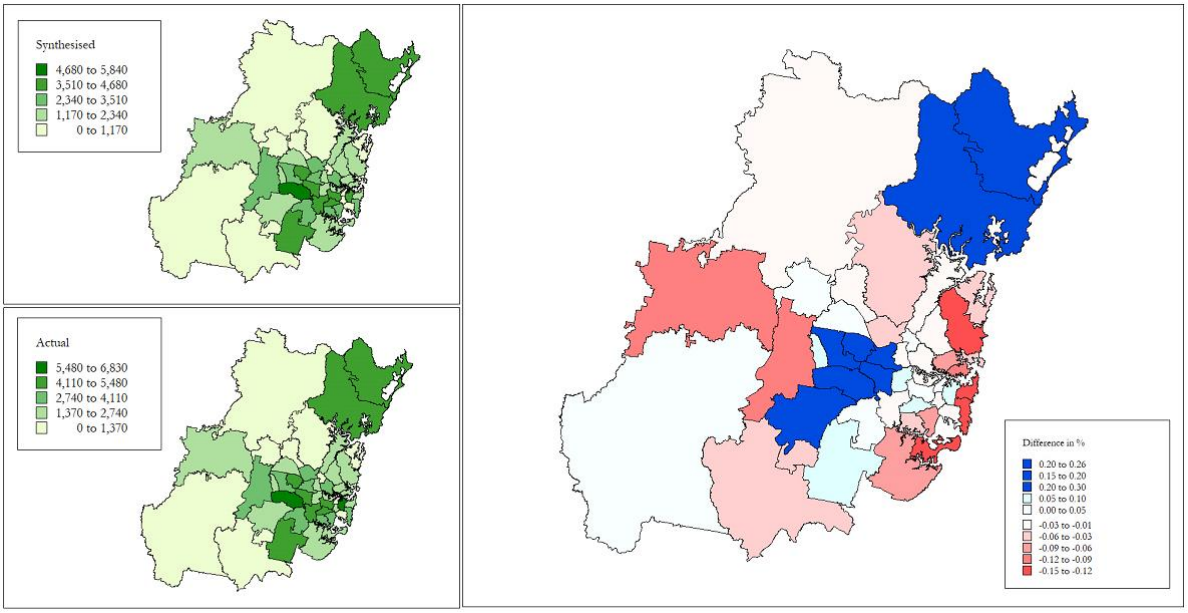


Figure 16 Synthesised and Actual Distribution of Unemployed Persons in Greater Sydney 2011



The spatial distribution for synthesised employed and unemployed persons is also consistent with the actual distribution with marginal differences in percentages (Figure 15 and Figure 16). The highest percentage difference is in Sydney inner city where the employed persons were over synthesised by 0.04 percent and Strathfield-Burwood-Ashfield were under synthesised by 0.03 percent compare to the actual number of employed persons (Figure 15).

Figure 16 displays higher variations in the percentage differences between synthesised and actual unemployed persons. Wyong, Gosford, Blacktown, Mount Druitt, Paramatta, Maryland-Guildford, Fairfield and Bringelly-Green Valley were all over synthesised between 0.20 to 0.26 percent, while Warringah, North and South Eastern Suburbs and Cronulla-Miranda-Caringbah were under synthesised between 0.12 and 0.15 percent. However, these differential in percentages is relatively minimal.

Testing of IPU Algorithm

In this section, four different scenarios were set up to gauge the efficiency and efficacy of the IPU algorithm. These experiments were conducted in all three cities to test the computational time and performance results.

- Scenario 1 Pre-synthesis zero marginal totals adjustment
- Scenario 2 Change in control categories
- Scenario 3 Change in geographical resolution
- Scenario 4 Change in rounding procedure

Scenario 1 Pre-synthesis zero marginal totals adjustment

This scenario tests if eliminating zero marginal totals in the input data for marginal totals prior to the synthesis process would improve the synthesised results and performance. Table 14 shows the SA1 counts before and after the zero marginal corrections for the three cities. In this experimental run, the input data for household and person marginal totals were adjusted to eliminate zero marginal totals before feeding into the synthesis process. Zero marginal totals include:

- marginal totals with zero households and zero persons in the corresponding geographical zone;
- marginal totals with zero households and non-zero persons in the corresponding geographical zone and;
- marginal totals with non-zero households and zero persons in the corresponding geographical zone.

From the ABS Table Builder 2011 census, Greater Sydney consists of 10,845 SA1s, of which 305 SA1s had zero households and zero persons; 51 SA1s had zero households and non-zero persons and two SA1s had households with zero persons. Hence, after eliminating all combination of zero marginal totals by SA1, the total synthesised SA1 for Greater Sydney was 10,487.

Table 14 SA1 Counts for Greater Sydney, Greater Melbourne and Greater Brisbane, 2011

Number of SA1	Greater Sydney	Greater Melbourne	Greater Brisbane
Before Zero Marginal Correction	10,845	9,658	5,485
After Zero Marginal Correction	10,487	9,420	5,333

Scenario 2 Change in control categories

This scenario tests if reduction in control categories affects the performance results and the extent of the computational time. In this experiment, three instead of four household control variables were used. They are dwelling structure, household composition and number of motor vehicles per household. This combination of household control variables reduced the household constraints

from 840 to 120. The number of constraints for person control variables were also reduced from 144 to only 18, where by only gender and age were included in the experimental run.

Scenario 3 Change in geographical resolution

This scenario tests the performance and computational time of changing geographical resolution from SA1 to SA2. For Greater Sydney, the geographical zones were combined from 10,485 SA1s to 270 SA2s, Greater Melbourne from 9,658 SA1s to 278 SA2s and Brisbane from 5,485 to 246 SA2s.

Scenario 4 Change in rounding procedure

This scenario tests if changing the rounding procedure in PopGen affects the performance results and computational time. The synthesis results presented in above for Greater Sydney was based on the arithmetic rounding procedure. Under Scenario 4, two more type of rounding procedures were tested. They are bucket and stochastic rounding procedures. Differences between these rounding procedures were explained in the previous section.

Below are examples of test results for Greater Sydney (Table 15). P-values produced under these scenarios are examined to gauge the performance of the synthesis results at SA1 level (Figure 17 on page 32).

Table 15 Comparison of Experimental Results by Control Variables, Greater Sydney 2011

	Actual	Synthesised[1]	Zero Margin Corrections	Control Categories	Geographic Resolution	Rounding Procedures Scenario 4	
		Base Case	Scenario 1	Scenario 2	Scenario 3	Bucket	Stochastic
Number of Households	1,598,439	1,598,433	1,598,433	1,598,433	1,598,433	1598433	1,598,433
% Difference between synthesised and actual household total		-0.0004	-0.0004	-0.0004	-0.0004	-0.0004	-0.0004
Number of Persons	4,308,248	4,118,543	4,118,660	4,271,266	4,223,172	4,157,035	4,156,469
% Difference between synthesised and actual person total		-4.4033	-4.4006	-0.8584	-1.9747	-3.5098	-3.5230
Distribution by Control Variable	**Actual Distribution**	**% Difference between Synthetic and Actual Data**					
	%	Base Case	Scenario 1	Scenario 2	Scenario 3	Bucket	Stochastic
Dwelling Structure							
Separate house	59.71	-0.65	-0.65	0.21	0.07	0.08	0.08
Semi-detached, row or terrace house, town house, etc.	12.68	0.59	0.59	-0.18	-0.02	-0.03	-0.03
Flat, unit or apartment	26.96	-0.03	-0.03	0.05	-0.02	-0.05	-0.05
Other dwelling	0.5	0.06	0.06	-0.04	-0.02	0.00	0.00
Other	0.14	0.04	0.04	-0.03	-0.01	0.01	0.01
Household Composition							
One family household	67.19	-0.88	-0.88	0.53	0.00	-0.30	-0.30
Two or more family household	2.3	0.47	0.47	-0.20	-0.03	0.01	0.01
Non-family household	25.56	0.38	0.38	-0.03	-0.04	0.08	0.08
Other	4.95	0.03	0.03	-0.29	0.08	0.21	0.21
1 person	21.57	-0.39	-0.39		0.02	0.01	0.01
2 persons	29.35	-0.47	-0.47		0.02	0.04	0.04
3 persons	16.4	0.01	0.01		0.01	0.00	0.00
4 persons	16.73	0.02	0.02		0.00	0.04	0.04
5 persons	7.42	0.3	0.3		-0.02	-0.03	-0.03
6 persons	3.51	0.35	0.35		-0.03	-0.05	-0.05
7 persons	5.01	0.17	0.17		0.00	0.00	0.00

[1] Based on synthesis results from the previous section.

Table 15 Comparison of Experimental Results by Control Variables, Greater Sydney 2011 (continued)

	Actual	Synthesised[2]	Zero Margin Corrections	Control Categories	Geographic Resolution	Rounding Procedures Scenario 4	
		Base Case	Scenario 1	Scenario 2	Scenario 3	Bucket	Stochastic
Number of Motor Vehicles							
None	11.60	-0.20	-0.20	0.17	0.19	0.17	0.17
1 motor vehicle	37.69	0.64	0.64	-0.37	-0.61	-0.56	-0.56
2 motor vehicles	32.15	0.51	0.51	-0.44	-0.48	-0.47	-0.47
3 motor vehicles	8.69	-0.17	-0.17	0.10	0.17	0.18	0.18
4 or more motor vehicles	3.12	-0.45	-0.45	0.32	0.43	0.36	0.36
Other	6.74	-0.33	-0.33	0.22	0.31	0.33	0.33
Gender							
Male	49.09	-0.06	-0.05	-0.01	0.00	-0.01	0.00
Female	50.91	0.06	0.05	0.01	0.00	0.01	0.00
Age							
0-4 years	6.93	0.16	0.16	0.02	0.05	0.19	0.17
5-9 years	6.38	0.08	0.08	0.01	-0.01	0.14	0.13
0-14 years	6.22	0.12	0.13	-0.01	-0.02	0.13	0.17
15-19 years	6.31	0.05	0.05	0.03	0.00	0.01	0.01
20-24 years	7.05	-0.22	-0.24	-0.02	-0.01	0.00	-0.02
25–29 years	7.84	-0.17	-0.14	0.01	0.02	0.02	0.04
30–34 years	7.72	-0.01	0.00	0.01	-0.01	0.02	-0.02
35–39 years	7.66	0.00	0.02	0.02	-0.01	-0.01	0.00
40–44 years	7.31	0.01	0.00	0.00	-0.04	-0.04	-0.01
45–49 years	7.05	0.02	0.01	0.04	-0.03	-0.05	-0.06
50–54 years	6.6	-0.04	-0.02	0.03	-0.02	-0.08	-0.08
55–59 years	5.76	0.04	0.03	0.03	0.01	-0.02	-0.02
60–64 years	5.19	0.01	0.01	0.02	0.02	-0.06	-0.06
65–69 years	3.84	0.00	0.02	-0.01	0.03	-0.04	-0.04
70–74 years	2.91	-0.02	-0.02	-0.04	0.03	-0.05	-0.05
75–79 years	2.21	0.00	-0.02	-0.04	0.01	-0.05	-0.05
80–84 years	1.69	0.00	-0.02	-0.03	0.01	-0.04	-0.04
85 years and over	1.31	-0.02	-0.03	-0.03	0.00	-0.05	-0.04
Labour Force Status							
Employed	47.21	0.65	0.71		-0.03	0.43	0.48
Unemployed	2.85	-0.39	-0.41		-0.03	-0.35	-0.35
Not in the labour force	25.39	0.13	0.07		0.02	-0.03	-0.06
Not stated	24.56	-0.38	-0.38		0.03	-0.06	-0.09

The synthesised results have been relatively consistent under the four scenarios for all cities. Although the test results for Greater Melbourne and Brisbane are not included in this paper, the following discussion summarises the test results for the three cities to demonstrate the validity and consistency of the IPU algorithm.

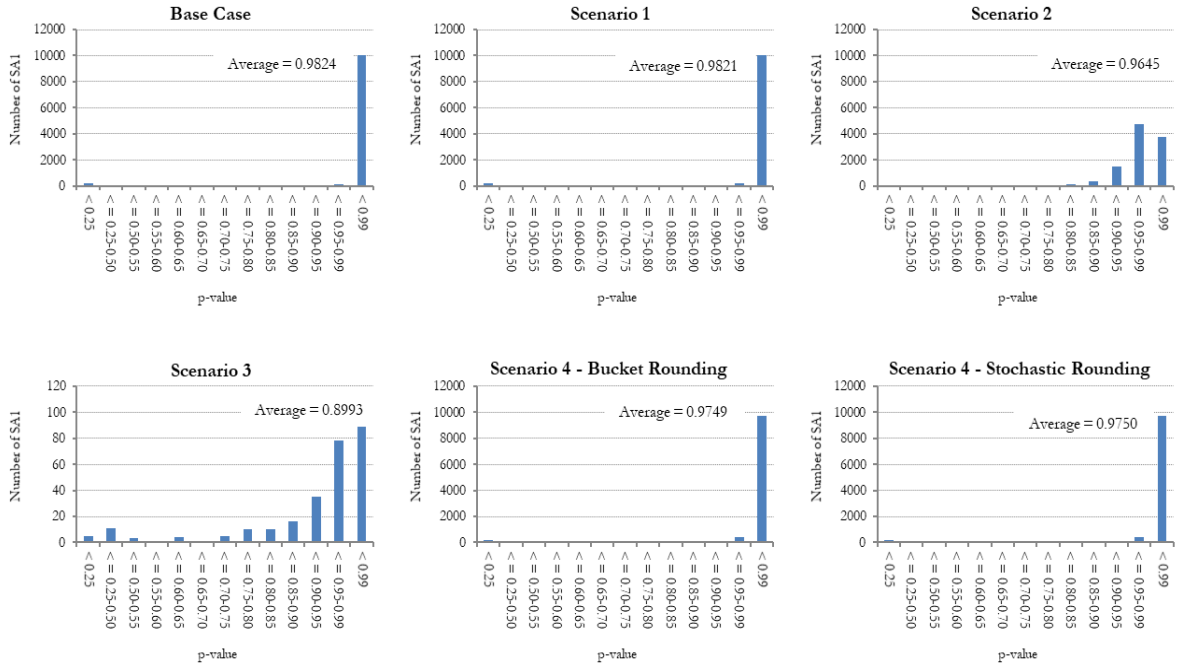
The reduction in the number of control categories and geographical resolution for Scenario 2 and 3 have drastically reduced the computational time for all three cities. This is especially true for Greater Sydney where the number of synthesised geographical zones was reduced from 10,845 SA1 to 270 SA2s for Scenario 3.

Generally, Scenario 2 and 3 generated higher number of synthesised persons compare to other scenarios. However, Scenario 3 consistently produced the lowest average and more scattered p-values for the three cities compare to the other scenario. One possible reason for the lower p-values at SA2 level could be due to the marginal totals used in the synthesis process. These marginal

[2] Based on synthesis results from the previous section.

totals were amalgamated from marginal totals at SA1 level. Theoretically these derived marginal totals should be the same as marginal totals that were extracted directly at SA2 level from the ABS table builder, since they are essentially the same source. However, it is possible that the marginal totals disseminated by ABS at SA2 contains less perturbations and hence less adjustments are required in census balancing. Note that the p-values produced under Scenario 3 are still well within the acceptable range.

Figure 17 Comparisons of P-Values Distribution by SA1, Greater Sydney 2011



The performance results for the base case scenario, Scenario 1 and 4 varied slightly across the three cities. For Greater Sydney, Scenario 4 with bucket rounding procedure has generated the highest number of synthetic persons and better matched distributions by control variables when compare to the arithmetic rounding procedure used in the base case scenario. However, the average p-value at SA1 level is slightly lower comparatively. As for Greater Melbourne, Scenario 4 with the stochastic rounding procedure has generated the highest number of synthetic persons and also exhibits an excellent match for distributions by the control variables. The average p-value for this scenario is 0.985, which is marginally lower compared to 0.991 for the base case scenario. Lastly, Scenario 4 with bucket rounding procedure generated the highest number of synthetic persons for Greater Brisbane. Consistent with Greater Sydney and Greater Melbourne, the synthesised distributions by control variables under Scenario 4 are the best among all scenarios and the distribution at SA level is marginally less fitted compared to the base case scenario and Scenario 1.

Generally, the rounding procedure affects the synthetic estimates in different ways. The estimated cell probabilities or decimal values were converted to integer frequencies using the arithmetic procedure in the base case scenario. This rounding procedure consistently generated better represented distributions by geographical zone. Both Bucket and Stochastic rounding procedures under Scenario 4 consistently produced better matched distributions by control variable but marginally less represented distributions by geographical zones when compare to the base case scenario. For Scenario 1, the pre-adjusted zero marginal totals have not shown any beneficial effects on the performance at variable and SA1 level for all three cities.

Overall, the synthesised results under each scenario are within an acceptable range. The choice of synthesis criteria for creating a synthetic population is very much dependent on its research purpose. For a synthetic population built for travel demand modelling, it is always important to retain as many demographic and geographical details as possible to provide a robust foundation for data linkage to other travel data. This is a critical bridging step for the accuracy of further travel demand simulations.

One of the more prominent issues observed in the synthesised results produced through the IPU procedure is the discrepancies between the synthesised and actual person counts. While all person control variables retained reasonable distributions of the actual population and the distribution of synthesised person at SA1 is within an acceptable range, the number of synthesised persons has been consistently underestimated for all three cities. The number of persons in Greater Sydney, Melbourne and Brisbane were under-synthesised by -4.40, -0.47 and -3.86 percent respectively. These results are consistent with past research. In the case study conducted by Ye et al. (2009), the research group who developed the IPU algorithm, the number of synthetic persons generated for the city of Maricopa County region of Arizona using the IPU algorithm differed by 4.6 percent to the actual number of persons in the population. Another example of a synthesis result using PopGen was by Jain, Ronald and Winter (2016). They obtained a difference of -5.82 percent for the number of synthesised persons relative to the actual persons in Melbourne. It is important to point out that the basis of comparison of synthesised results from this research study and that from the other research studies are not completely parallel. The number of household type constraints used in Ye et al. (2009) and Jain, Ronald and Winter (2016) is notably lower than this study. In Ye et al.'s case study, the synthesis results were based on three household and three person control variables with a total number of household-type and person type constraints of 280 and 140 respectively. Jain, Ronald and Winter also used three household and three person control variables with 180 household type constraints and 144 person type constraints. The synthesised results in this study is based on four household and three person control variables, with 840 household type constraints and 144 person type constraints.

In the IPU algorithm, the multi fitting problem in IPF is overcome by reallocating weights among households, taking into account the person-level distribution as much as possible, without compromising the fit to household-level distribution. Therefore, the total number of household/person constraints, which is the number of cells in the joint distributions, underpins the degree to which the person-level control variables will be matched by IPU algorithm. As the level of aggregation increases from stratification by household type only to household-person type, the frequency in each cell decreases. The IPU algorithm will function in the presence of sparse cells, however there is a trade off in the fit with respect to person-level distributions (Ye et al. 2009).

Although the percent differences for the under-synthesised persons were relatively substantial, a further suggestion or solution for addressing this issue was not found in Ye's research nor subsequent research based on PopGen. This issue is investigated further in next section.

Data Treatment Post Population Synthesis

This section proposed a heuristic algorithm to compensate for the number of under synthesised persons generated from the IPU algorithm. This process rectifies the under synthesised results presented in the previous section. Validation measures are used to evaluate the adjusted synthesised results.

Proposed Person Allocations Algorithm Post Population Synthesis

The number of persons usually resident in dwelling (Nprd) in the sample data is an important household variable in determining the number of persons synthesised in the IPU algorithm. One of the possible reasons for the IPU algorithm to constantly fall short in generating the number of synthetic persons could be due to how this control variable is categorised in the sample or seed data and marginal data. Under Nprd, any household with six or more persons is assigned into one category. Hence, the number of persons is often under synthesised for large families or households. The proposed algorithm compensates under synthesised SA1s based on conditional allocations of additional persons. At household level, the algorithm will only affect the distribution of one variable that is Nprd. As the number of persons randomly allocated into SA1, the number of persons in the household changes, unless they are allocated into the category with six or more persons. The selection process in the proposed algorithm is conditioned to prioritise allocations of person as discussed in the following steps to minimise the adjustment effects on the distribution for Nprd. The algorithm will affect the distributions of all person control variables. All adjusted results will be validated again after implementing the algorithm.

Below are the general steps involved in allocating new persons to compensate for the under synthesised population:

Stage 1 Setting up

Four steps of merging are involved in this step:

1 A dataset is created by merging the total number of synthesised and actual persons by SA1. This dataset should contain identifiers for SA1 and the merged data is used to calculate the difference between the number of actual and synthesised persons for each SA1. The number of observations for this data is equal to the number of SA1s synthesised.

2 The dataset from stage 1 is then match-merged with the associated under synthesised persons generated by the IPU algorithm. The under synthesised dataset is a large file containing information of the control variables on every synthesised person. Each observation in the match-merging data set retained all information read in at SA1 level during the merging, which includes the difference between the actual and synthesised number of persons for each SA1 level. At this stage, the total observation number for this match-merging data set should equal the total number of persons synthesised.

3 The match-merging data set from stage 2 is then merged with the seed data or sample data for household to obtain and link information on the number of persons resided in the dwelling to each person.

4 Finally, the match-merging data set is merged with the synthesised household dataset to obtain and link information on the synthesised number of persons resided in the dwelling to each person. The variable values obtained from step 3 must be the same with those obtained from step 4 since the synthesised households were drawn from the household seed data in the synthesis process.

After four merges, the dataset should contain geographical identifier at SA1, household identifier, person identifier, selected person characteristics (gender/Sexp, age/Agep and labour force status/Lfsp) and two household characteristics (Number of persons usually resident in dwelling/Nprd from the seed data and from the synthesised household dataset) for all synthesised

persons. The observation number should remain the same after merging. The large dataset is now ready for stage 2.

Stage 2 Allocating

1 At the allocation stage, a subset of data from stage 1 is created for all synthesised persons who resided in SA1s that contained less than the actual number of persons. These observations are identifiable using the calculated differences between actual and synthesised persons for each SA1 attached during merging. These differences are the number of under synthesised persons, which also used to set the maximum number of synthesised persons allowable to be allocated in each SA1.

2 The next step is to simulate a random number for each observation or synthesised person in the dataset created in previous step. These generated random numbers can be either uniformly or normally distributed. They are used to randomise the position of synthesised persons within each household by SA1. This dataset must be sorted by geographical zone and then household identifier. The sorting process conditions the randomisation to only within each household and also ensures that the number of additional persons added to each SA1 later does not exceed the discrepancy between the actual and synthesised estimates in each SA1.

3 Upon the completion of the randomisation process, a new person is duplicated from the last observation within each household for an under synthesised SA1. This process limits the allocation process to only one person per household. At this point, new household and person identifiers are created for identifying the first and last person within the household and to accommodate the additional new person after the randomisation process. All newly added persons are then output into a new dataset to form a new pool of observations for allocating additional persons to under synthesised SA1s. Observations that are identified as "Not applicable" in the subcategory of Nprd are deleted. The selection process will exclude this category to minimise possible disturbance on the overall distribution of Nprd. It is also a category with the least flexibility to accommodate newly generated persons.

4 Before the allocation process begins, two conditions are imposed. The new pool of observations is sorted by SA1, then in descending order by a household control variable (Nprd) and nested by a person variable (Agep). This condition imposed on Nprd prioritises the selection to the category of six or more people. In this Nprd category, any additional person generated does not change the status of Nprd in the household population. The person control variable, Agep is ordered from largest to smallest according to the gap between the synthesised and actual estimates. This condition prioritises allocation of persons into the age category with the largest discrepancies between synthesised and actual estimates.

5 During the allocation process, a counter is set up so that the selection of persons for each SA1 stops when the number added reaches the maximum number of synthesised persons. The maximum number of persons for each SA1 is the difference in number between actual and synthesised persons obtained in stage1.

6 At the end of the selection process, the number random of persons added to the initial dataset from the IPU process is adjusted accordingly to align with the actual number of persons.

Performance of Post-treated Synthesised Results

The analysis of the synthesised results selected to treat the under synthesised persons is based on the performance results from testing the IPU algorithm. As Scenario 2 contains reduced number of control variables and Scenario 3 contains coarser geographical zones, the synthetic population generated under these scenarios were not considered for post-synthesised data treatment.

Overall, six generated populations have been chosen for testing the regularities of the proposed algorithms. The average p-value is the highest for all cities under the base case scenario and hence these three set of synthesised populations are selected for post data treatment. All base case scenarios were based on arithmetic rounding. The distributions by control variables for these three case studies are consistently aligned with the actual distributions, however the synthetic populations generated under Scenario 4 seem to perform relatively better. Particularly for Greater Melbourne, the synthetic population generated under Scenario 4 with Stochastic Rounding procedure was selected for further data treatment. The dataset has the highest number of synthesised persons, average p-value of 0.99 by SA1 and excellent matches in distributions by the control variables. For Greater Sydney and Brisbane, post-synthesised data treatments were carried out based the synthetic estimates under Scenario 4 with Bucking rounding procedure. Both synthetic populations have marginally higher average p-value compare to Scenario 4 with Stochastic rounding procedure.

These case studies test whether the proposed algorithm is able to improve the overall synthesised results whilst to retain or improve the synthesised distributions by variables and geographical zones.

Table 16 presents the pre- and post-treated synthesised person results for the three cities. The synthesised number of persons was under by 4.4 percent compare to the actual number of persons for Greater Sydney under the base case scenario. After the data treatment, the gap is reduced to only 0.0065 percent. Note that base case scenario applied arithmetic rounding procedure in the synthesis process. The synthesised results based on bucket rounding procedure further reduced the data gap to 0.0007 percent post data treatment, which is a near complete set of synthetic persons for Greater Sydney. As for Greater Melbourne, the difference between synthesised and actual person is reduced from -4.7084 to 0.0004 percent post data treatment for the synthetic population generated under the base case scenario and almost zero percent under Scenario 4 with stochastic rounding. The gap between synthesised and actual person for Greater Brisbane also reduced substantially after data treatment. The discrepancy of synthetic population for base case scenario has reduced from -3.8 percent to -0.0044 percent. This is also consistent under Scenario 4 with Bucket rounding procedure where the discrepancy reduced from -3.1581 percent to merely -0.0015 percent. At city level, the number of synthesised persons relative to actual persons have improved to near complete synthetic populations for all three cities post data treatment. The following two sections investigate how these post-treated results preformed in terms of distributions by the control variables and at SA1 level.

Table 16 Comparison of Pre- and Post-Treated Synthesised Person Results

	Greater Sydney		**Greater Melbourne**		**Greater Brisbane**	
	Persons	Difference %	Persons	Difference %	Persons	Difference %
Actual	4,308,248		3,912,141		2,155,966	
Arithmetic Rounding						
Pre-treated results	4,118,543	-4.4033	3,727,942	-4.7084	2,072,706	-3.8618
Post-treated results	4,308,526	0.0065	3,912,156	0.0004	2,155,872	-0.0044

Table 16 Comparison of Pre- and Post-Treated Synthesised Person Results (continued)

	Greater Sydney		Greater Melbourne		Greater Brisbane	
	Persons	Difference %	Persons	Difference %	Persons	Difference %
Stochastic/Bucket Rounding	Bucket		Stochastic		Bucket	
Pre-treated results	4,157,035	-3.5098	3,761,908	-3.8402	2,087,878	-3.1581
Post-treated results	4,308,280	0.0007	3,912,140	0.0000	2,155,933	-0.0015

At variable level, the actual data is compared to pre-treated synthesised and post-treated synthesised estimates. Figure 18 shows the comparison of treated and untreated synthesised number of persons in usual resident for Greater Sydney. This is the only household variable that is affected by the proposed algorithm. Distributions for all other control variables at household level remained unchanged. The top chart shows the aggregate total by Nprd and the bottom chart shows the percent difference between the synthesised and actual aggregate total. The post-treated base case scenario (post-treated arithmetic) generally displays slightly higher differences of percentages in aggregate totals than the pre-treated arithmetic estimates, except for the category with six persons or more. However, the post-treated bucket estimates generally showed greater improvements in reducing the percent difference for each category. Table 17 shows that the distribution by Nprd were almost identical between pre- and post-treated arithmetic estimates but better matched for post-treated bucket data.

As the proposed algorithm is mostly processed at person level, all distributions of person control variables were affected. Figure 19 to 21 show that percent differences of aggregate totals for all person control variable have improved for post-treated data compare to pre-treated arithmetic estimates. Generally, the post-treated bucket estimates have out-performed the post-treated arithmetic estimates in reducing the percent differences between the synthesised and actual data.

Figure 18 Comparison of Household Estimates by Number of Persons Usually Resident in Private Dwellings between Benchmark, Adjusted and Unadjusted Synthesised Results, Greater Sydney 2011

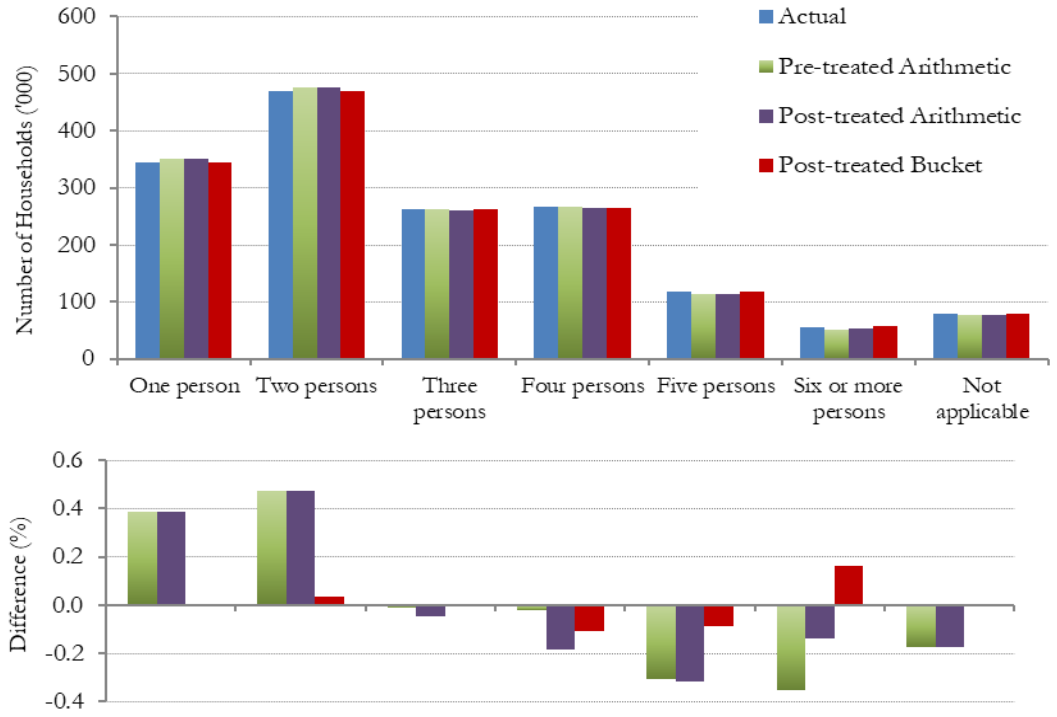


Table 17 Distributions of Actual and Adjusted Synthesised Household Estimates by Number of Persons Usually Resident in Private Dwellings, Greater Sydney 2011

	Actual Distribution (%)	% Difference between Synthetic and Actual Data		
		Pre-treatment	Post-treatment	
Number of Persons Usually Resident in Dwelling		Arithmetic	Arithmetic	Bucket
1 person	21.57	0.39	0.39	0.00
2 persons	29.35	0.47	0.47	0.04
3 persons	16.40	-0.01	-0.05	-0.01
4 persons	16.73	-0.02	-0.19	-0.11
5 persons	7.42	-0.30	-0.32	-0.09
6 persons	3.51	-0.35	-0.14	0.17
7 persons	5.01	-0.17	-0.17	0.00

Figure 19 Comparison of Person Estimates by Gender between Actual Data, Adjusted and Unadjusted Synthesised Results, Greater Sydney 2011

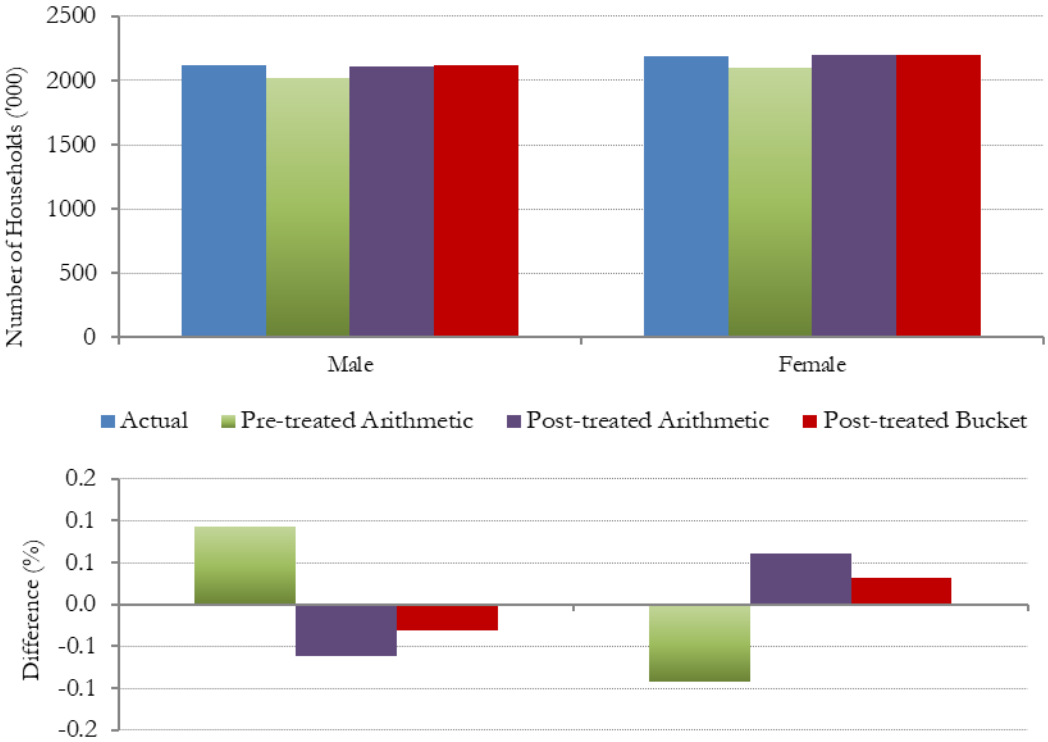


Figure 20 Comparison of Person Estimates by Age between Actual Data Adjusted and Unadjusted Synthesised Results, Greater Sydney 2011

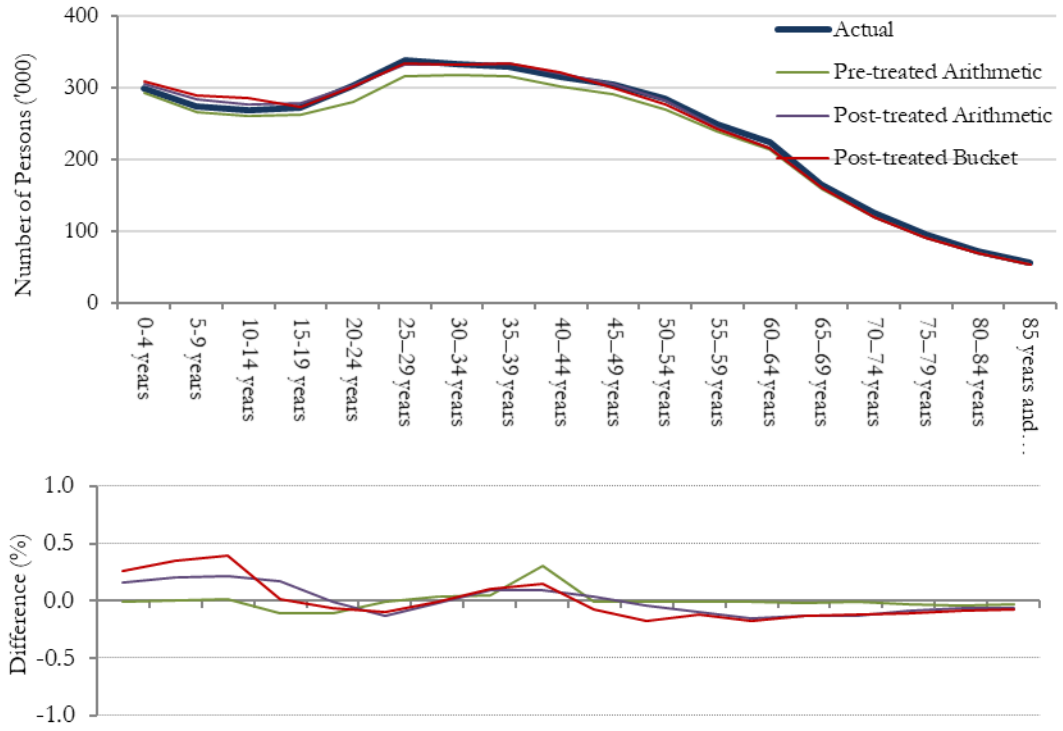


Figure 21 Comparison of Person Estimates by Labour Force Status between Actual Data Adjusted and Unadjusted Synthesised Results, Greater Sydney 2011

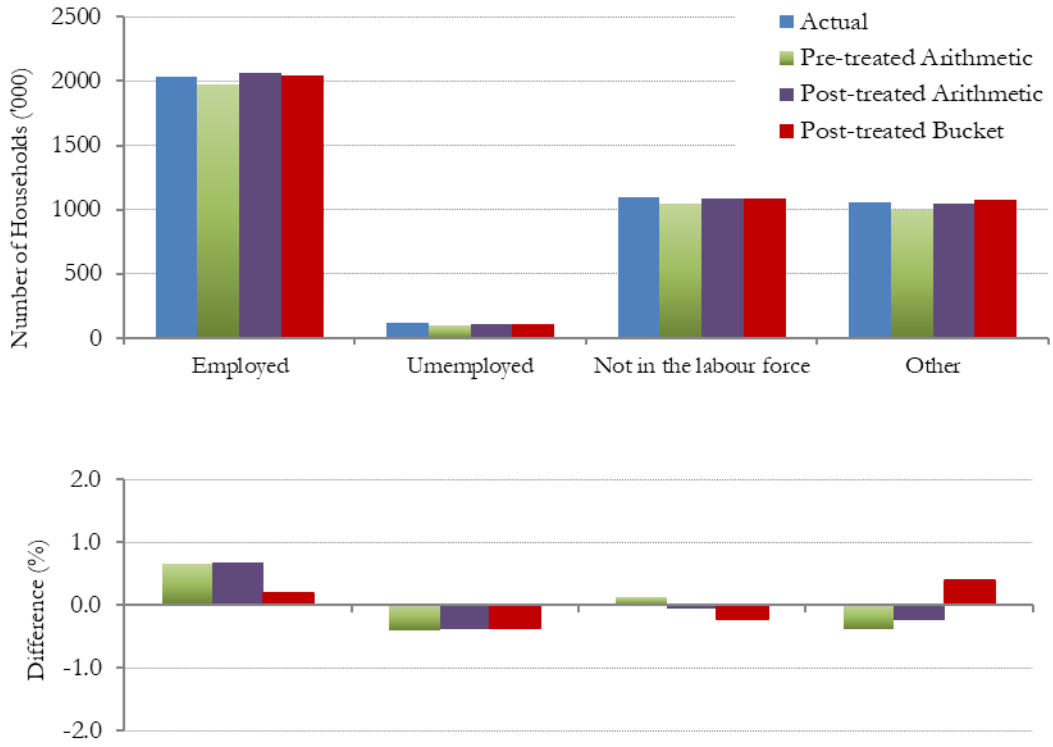


Table 18 shows that distributions by gender and age were close between pre- and post-treated arithmetic estimates but marginally better matched by labour force status for post-treated

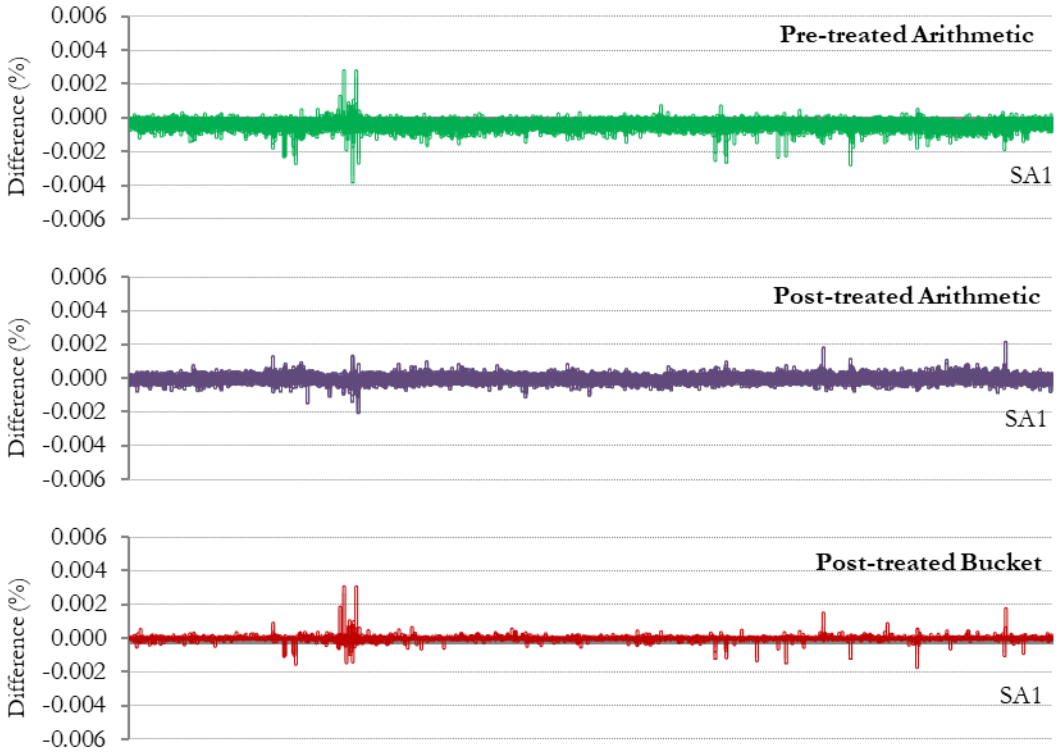
arithmetic estimates. The post-treated bucket estimates performed better in distributions for gender and labour force status but slightly worse off in distribution for age.

Table 18 Distributions of Actual and Adjusted Synthesised Person Estimates, Greater Sydney 2011

	Actual Distribution (%)	**% Difference between Synthetic and Actual Data**		
		Pre-treatment	Post- treatment	
		Arithmetic	Arithmetic	Bucket
Gender				
Male	49.09	-0.06	-0.06	-0.03
Female	50.91	0.06	0.06	0.03
Age				
0-4 years	6.93	0.16	0.16	0.26
5-9 years	6.38	0.09	0.21	0.35
10-14 years	6.22	0.12	0.22	0.39
15-19 years	6.31	0.05	0.17	0.01
20-24 years	7.05	-0.23	-0.01	-0.06
25–29 years	7.84	-0.17	-0.14	-0.10
30–34 years	7.72	0.00	-0.02	0.00
35–39 years	7.66	0.00	0.09	0.10
40–44 years	7.31	0.01	0.09	0.15
45–49 years	7.05	0.02	0.03	-0.08
50–54 years	6.6	-0.04	-0.04	-0.18
55–59 years	5.76	0.04	-0.10	-0.12
60–64 years	5.19	0.01	-0.16	-0.18
65–69 years	3.84	0.01	-0.13	-0.13
70–74 years	2.91	-0.02	-0.13	-0.13
75–79 years	2.21	0.00	-0.09	-0.11
80–84 years	1.69	0.00	-0.07	-0.09
85 years and over	1.31	-0.02	-0.07	-0.08
Labour Force Status				
Employed	47.21	0.64	0.68	0.18
Unemployed	2.85	-0.40	-0.38	-0.36
Not in the labour force	25.39	0.12	-0.06	-0.23
Not stated	24.56	-0.38	-0.24	0.40

Figure 22 show differences in percentages between the number of actual and synthetic persons at SA1 level for Greater Sydney. The percent discrepancies in the pre-treated arithmetic estimates are the highest. These discrepancies progressively became smaller or distributed closer to zeros in the post-treated arithmetic estimates and even smaller in the post-treated bucket or stochastic estimates.

Figure 22 Comparison of Percent Difference between Adjusted and Unadjusted Synthesised Person Estimates at SA1, Greater Sydney 2011



Detailed validation results for Greater Melbourne and Brisbane can be found in Lim (2019). The results are consistent with Greater Sydney. Overall, validation results show that the synthesised estimates have improved at all level after the data treatment. At city level, the discrepancy between the number of actual and synthesised persons for each city has been reduced to nearly zero. The enumeration of the synthetic population in each city is now at least 99.99 percent complete.

At variable level, the distribution of the only household variable (Nprd) affected by the algorithm remain almost the same for the pre-treated and post-treated arithmetic estimates and has improved in matching the actual aggregate total and distribution by Nprd for post-treated bucket or stochastic estimates. As all person control variables are affected by implementing the proposed algorithm, synthesised distributions of these variables were re-examined. Generally, the percent differences between synthesised and actual aggregate totals have been reduced in the post-treated estimates, especial the post-treated bucket or stochastics estimate. That means, after data treatment, the aggregate totals by person control variables are now even closer to the actual aggregate totals. In terms of distributions, there are some minor gains and loses between the pre-treated arithmetic, post-treated arithmetic and post-treated bucket/stochastic estimates. However, these movements in percentages are mostly within than ± 0.01 percent. Overall, the distributions by control variables after data treatments remained closely matched to the actual distributions.

At geographical level, the synthesised distributions after data treatments have clearly reduced the percent differences in the number of synthesised and actual persons. All three cities have shown prominent improvement in the distributions by SA1 level from pre-treated arithmetic to post-treated estimates.

The proposed algorithm has shown promising results for improving the completeness of synthesised estimates generated from the IPU algorithm. The correction of the synthesised persons is important in any subsequent model simulation. Errors carried forward from the synthetic population often remerge if not magnified in subsequent model outcomes. The heuristic algorithm proposed in this section is a relatively straight forward and time efficient method to improve the representation of an under-synthesised population.

Conclusion

The requirement of spatial microdata has continued to be a barrier to the development of a fully operational microsimulation activity-based travel demand model in Australia. Overall, the restricted access of detailed geocoded micro data and the demand of specialised computing skills for building a synthetic population often hinders the progress of further model development in activity-based travel demand modelling.

Numerous population synthesis techniques have been proposed as an alternate approach to supplement the inadequacy in readily available microdata for microsimulation analysis. In practice, the lack of reusability of existing population synthesisers often impose the need to develop a synthesis routine from scratch for a new research project that required comprehensive microdata. Many existing population procedures are often concealed in computer codes and shrouded by inaccessible language. Some are overcomplicated or, lack implementation details or transparency in validations. There is also a fine balance between the ever-increasing complexity and resources required to create the synthetic data. Enhanced complexity often increases both costs to build and the time required to implement with no clear certainty of better performance results (O'Donoghue 2018). The challenge is to develop population synthesis techniques that are user friendly and at the same time retains sufficient complexity to produce a well-represented synthetic population.

This research study contributes in setting up a replicable population synthesis routine that can be included into a standard methodological toolbox for transport researchers and mainstream social scientists to produce synthetic population using Australian census data. This research intends to alleviate the cumbersome and costly process of building synthetic microdata by presenting a practical pathway to building synthetic populations for Australian cities

The synthetic populations generated in this research study for the three major Australian cities have been extensively validated. The performance results consistently displayed excellent fit with high level of confidence in matching the synthesised to actual data. Two heuristic procedures were formulated to ease the data handling process, specifically for Australian data. The procedure proposed for data treatment before the synthesis routine ensures the consistency of the input data and the procedure proposed for data treatment after the synthesis routine extends under-synthesised estimates to a complete synthetic population. Multiple experiments have also been conducted to test the efficacy and reliability of the IPU algorithm. The treated post-synthesised estimates have been revalidated and proven to further increase the accuracy of the synthesised estimates. The approach serves as a practical solution to building the necessary synthetic individuals and households in the context of activity based microsimulation modelling in Australia.

Abbreviations

ABS	Australian Bureau of Statistics
ABSHID	Dwelling Record Identifier
ABSPID	Person Record Identifier
ASGS	Australian Statistical Geography structure
ASU	Arizona State University
AUS	Australia
Bg	Block group
BITRE	Bureau of Infrastructure, Transport and Regional Economics
CO	Combinatorial Optimisation
CSF	Census Sample File
CURF	Confidentialised Unit Record File
GCCSAs	Greater Capital Cities Statistical Areas
IPF	Iteration Proportional Fitting
IPU	Iteration Proportional Update
PUMS	Public Use Microdata Sample
PUMA	Public Use Microdata Area
Pumano	Identifier for the Public Use Microdata Area (PUMA) of the corresponding geography
SAS	Statistical Analysis System
Simtravel	**Sim***ulator of* **T***ransport,* **R***outes,* **A***ctivities,* **E***missions, and* **L***and* model
SPSS	Statistical Package for the Social Sciences
SR	Statistical Region
S/T	State and Territory
TAZ	Traffic Analysis Zone
TRANSIMS	TRansportation ANalysis SIMulation System (Los Alamos National Laboratory 2005).

References

Abraham, J, Stefan, K & Hunt, J 2012, 'Population Synthesis Using Combinatorial Optimization at Multiple Levels', *Transportation Research Board 91st Annual Meeting, 2012*, pp. 17, viewed on 15 December 2018, < https://trid.trb.org/view.aspx?id=1130260>.

Australia Bureau of Statistics (ABS) 2011, *Australian Statistical Geography Standard (ASGS): Volume 1 - Main Structure and Greater Capital City Statistical Areas July 2011*, No 1270.0.55.001, Commonwealth of Australia, Canberra, ACT.

Auld, J & Mohammadian, A 2010, 'Efficient methodology for generating synthetic populations with multiple control levels', *Transportation Research Record: Journal of the Transportation Research Board*, 2175(1), pp.138-147.

Australia Bureau of Statistics (ABS) 2011, *Australian Statistical Geography Standard (ASGS): Volume 1 - Main Structure and Greater Capital City Statistical Areas July 2011,* No 1270.0.55.001, Commonwealth of Australia, Canberra, ACT.

Ballas D & Clarke GP 2001, 'Modelling the local impacts of national social policies: A spatial microsimulation approach', *Environment and Planning C: Government and Policy,* vol. 19, no. 4, pp. 587-606.

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

Birkin M 2013, 'Challenges for spatial dynamic microsimulation modelling', in Tanton R., Edwards K.L. (Eds.), *Spatial microsimulation: A reference guide for users*, Springer, Dordrecht, pp. 223-245.

Bishop. Y 1967, 'Multidimensional Contingency Tables. Cell Estimates', PhD thesis, Harvard University.

Bishop, M 1969, 'Calculating smoothed contingency tables', in JP Bunker (ed.), *The National Halothane Study: A Study of the Possible Association Between Halothane Anesthesia and Postoperative Hepatic Necrosis*, Bethesda, MD: National Institutes of General Medical Sciences.

Bishop Y, Fienberg S & Holland P 1975, *Discrete Multivariate Analysis: Theory and Practice,* MIT Press; Cambridge, Massachusetts.

Bhat, C, Guo, J, Srinivasan, S. & Sivakumar, A 2003, *Guidebook on activity-based Travel Demand Modelling for Planners*. Product 4080-P3, prepared for the Texas Department of Transportation, October 2003.

Birkin, M & Clarke, M 1988, 'SYNTHESIS - a synthetic spatial information system for urban and regional analysis: methods and examples', *Environment and Planning A*, vol. 20, pp. 1645-1671.

Bowman, JL 2004, 'A comparison of population synthesizers used in microsimulation models of activity and travel demand', Unpublished working paper.

Brown, DT 1959, 'A note on approximations to discrete probability distributions', *Information and Control*, *2*(4), 386–392.

Brown, MB 1976, 'Screening effects in multidimensional contingency tables', *Applied Statistics*, *25*(1), 37–46.

Bureau of Infrastructure, Transport and Regional Economics (BITRE) 2015, *Traffic and congestion cost trends for Australian capital cities*. Information Sheet 74. Bureau of Infrastructure, Transport and Regional Economics, Canberra, ACT.

Caldwell, S, & Keister, L 1996, 'Wealth in America: Family stock ownership and accumulation, 1960–1995', in GP Clarke (ed.), *Microsimulation for urban and regional policy analysis,* pp. 88–116, London: Pion.

Chin, FS, Haring, A & Bill, A 2006, *Regional Dimensions: Preparation of 1988-99 HES for Reweighting to Small-area Benchmarks*, Technical Paper no. 34, NATSEM, University of Canberra.

Cho, S, Bellemans, T, Knapen, L, Janssents, D, Wets, G 2014, 'Synthetic Population Techniques in activity-based research, Data Science and Simulation', in *Transportation Research*, page 23, chapter 3.

Choupani, A & Mamdoohi, AR 2016, 'Population Synthesis Using Iterative Proportional Fitting (IPF): A Review and Future Research', *Transportation Research Procedia*, vol. 17, 2016, pp. 223-233.

Csiszar, I 1975, 'I-Divergence Geometry of Probability Distributions and Minimization Problems', *The Annals of Probability,* 3. 146-158.

Deming, WE & Stephan, FF 1940, 'On a least squares adjustment of a sampled frequency table when the expected marginal totals are known', *Annals of Mathematical Statistics*, 11, pp. 57-66.

Edwards, KL & Clarke, GP 2009, 'The design and validation of a spatial microsimulation model of obesogenic environments for children in Leeds, UK: SimObesity'. *Social science & medicine*, vol. 69, no. 7, pp. 1127–1134.

Edwards KL & Tanton R 2013, 'Validation of spatial microsimulation models'; in Tanton R., Edwards K.L. (Eds.), Spatial microsimulation: A reference guide for users, Springer, Dordrecht, pp. 249-258.

Evans, VP 1999, 'Strategies for detecting outliers in regression analysis: An introductory primer', in B. Thompson (ed.), *Advances in social science methodology*, Stamford, CT: JAI Press, vol. 5, pp. 213-233.

Farooq, B, Bierlaire, M, Hurtubia, R. & Flötteröd, G 2013, 'Simulation based population synthesis. *Transportation Research Part B: Methodological*, 58, pp.243-263.

Fienberg, SE 1970a, 'An Iterative Procedure for Estimation in Contingency Tables', *The Annals of Mathematical Statistics,* Vol. 41, No. 3 (Jun 1970), pp. 907-917.

Fienberg, SE 1970b, 'Quasi-Independence and Maximum Likelihood Estimation in Incomplete Contingency Tables', *Journal of the American Statistical Association*, Vol. 65 No 332, pp. 1610-1616.

Fienberg, SE 1968, 'The geometry of an r x c contingency table', *Ann. Math. Statist. 39*, 1186-1190.

Fienberg, SE & Gilbert, JP 1970, 'The geometry of a 2 x 2 contingency table', *Journal of the American Statistical Association,* Vol. 65, No. 330, pp. 694-701.

Fienberg, SE & Meyer, MM 1981, 'Iterative Proportional Fitting', Technical Report No 270, Department of Statistics, Carnegie-Mellon University.
Guo, J & Bhat, C 2007, 'Population Synthesis for Microsimulating Travel Behavior'. *Transportation Research Record: Journal of the Transportation Research Board*, 2014(1), pp.92-101.

Giryes, R, Eldar, YC, Bronstein, AM & Sapiro, G 2016, 'Tradeoffs Between Convergence Speed and Reconstruction Accuracy in Inverse Problems', *IEEE Transactions on Signal Processing*, pp 99.

Guo, J & Bhat, C 2007, 'Population Synthesis for Microsimulating Travel Behavior'. *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2014, no. 1, pp.92-101.

Haberman, Sj 1984, 'Adjustment by minimum discriminant information', *Ann. Statist. 12*, 971-988.

Haberman, SJ 1974, *The Analysis of Frequency Data,* Chicago: The University of Chicago Press.

Harland, K, Heppenstall, A, Smith, D, & Birkin, M 2012, 'Creating realistic synthetic populations at varying spatial scales: a comparative critique of population synthesis techniques', *Journal of Artificial Societies and Social Simulation*, vol. 15, no. 1, pp. 1, view on 13 December 2018, <http://jasss.soc.surrey.ac.uk/15/1/1.html>.

Huang, Z & Williamson, P 2001, 'A Comparison of Synthetic Reconstruction and Combinatorial Optimisation Approaches to the Creation of Small-Area Microdata', Working paper. Liverpool, United Kingdom: University of Liverpool, Department of Geography.

Ireland, CT & Kullback, S 1968, 'Contingency tables with given marginals', *Biometrika*, **55**, 179–188.

Jain, S, Ronald, N & Winter, S 2015, 'Creating a Synthetic Population: A Comparison of Tools', 3rd Conference of Transportation Research Group of India (3rd CTRG).

Knight J, Wells S, Marshall R, Exeter D, Jackson R 2017, 'Developing a synthetic national population to investigate the impact of different cardiovascular disease risk management strategies: A derivation and validation study'. *PLoS One*. 2017;12(4): e0173170. Published in doi: 10.1371/journal.pone.0173170.

Konduri, K, You, D, Garikapati, V & Pendyala, R 2016, 'Enhanced Synthetic Population Generator That Accommodates Control Variables at Multiple Geographic Resolutions', *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2563, no. 1, pp. 40-50.

Lambert, S, Percival, R & Schofield, D & Paul, S. 1994, *An introduction to STINMOD: A static microsimulation model*, Series National Centre for Social and Economic Modelling (NATSEM) Technical Papers, Issue 1994/1, University of Canberra.

Lee, KS, Eom, JK & Moon, DS 2014, 'Applications of TRANSIM in transportation: A literature review', *Procedia Computer Science 32*, 769 – 773.

Lenormand, M & Deffuant, G 2013, 'Generating a synthetic population of individuals in households: Sample-free vs sample-based methods', *Journal of Artificial Societies and Social Simulation*, SimSoc Consortium, 16 (4), pp.12.

Lim, PP 2019, 'Population Synthesis for Travel Demand Modelling in Australian Capital Cities' (Submitted), PhD thesis, University of Queensland, Brisbane.

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, DOI: 10.1080/00330124.2015.1099449.

Los Alamos National Laboratory 2005, *TRANSIMS: Transportation analysis and simulation system.* http://www.ccs.lanl.gov/transims/index.shtml, Accessed on December 2, 2005.

Lovelace, R, Birkin, M, Ballas, D & van Leeuwen, E 2015, 'Evaluating the Performance of Iterative Proportional Fitting for Spatial Microsimulation: New Tests for an Established Technique', *Journal of Artificial Societies and Social Simulation*, vol. 18, no. 2, pp. 21.

Ma, L & Srinivasan, S 2015, 'Synthetic Population Generation with Multilevel Controls: A Fitness-Based Synthesis Approach and Validations', *Computer-Aided Civil and Infrastructure Engineering*, vol. 30, no. 2, pp. 135-150.

McBride, E, Davis, AW, Lee, JH & Goulais, KG 2016, *Spatial transferability using synthetic population generation methods*, University of California Transportation Centre, University of California, Santa Barbara.

McNally, MG, Rindt, C 2007, 'The Four-Step Model', Chapter 3 in Hensher, DA & Button, KJ (eds), Handbook of Transport Modelling, Pergamon [2nd Edition], pp. 35-53.

McNally, MG and Recker, WW 1987, 'On the Formation of Household Travel-Activity Patterns: A Simulation Approach', Final Report, *USDOT*, Washington, DC.

McNally, MG, Rindt, C 2007, 'The activity-based approach', Chapter 4 in Hensher, DA & Button, KJ (eds), *Handbook of Transport Modelling*, Pergamon [2nd Edition], pp. 55-73.

Mladenovic, M & Trifunović, A 2014, 'The Shortcomings of the Conventional Four Step Travel Demand Forecasting Process', *Journal of Road and Traffic Engineering,* Vol. 60, No. 1, Pages 5-12.

Morrissey, K & O'Donoghue, C 2013, 'Validation Issues and Spatial Patterns of Household Income', Chapter 5 in O'Donoghue, C, Ballas, D, Clarke, G, Hynes, S & Morrissey, K (eds), Spatial Microsimulation for Rural Policy Analysis, Springer, pp. 87-102.

Mosteller, F 1968, 'Association and estimation in contingency tables', *J. Amer. Statist. Assoc. 63*, 1-28.

Melhuish, T, Blake, M, & Day, S 2002, 'An evaluation of synthetic household populations for census collection districts created using optimisation techniques', *Australasian Journal of Regional Studies*, *8*(3), pp 369–387.

Müller, K & Axhausen KW 2011, 'Population synthesis for microsimulation: State of the art', paper presented at the *90th Annual Meeting of the Transportation Research Board*, Washington, D.C.

Müller, K 2017, 'A Generalized Approach to Population Synthesis', Doctoral thesis, ETH Zürich, Zürich.

Namazi-Rad, M, Mokhtarian, P &Perez, P, 2014, 'Generating a Dynamic Synthetic Population – Using an Age-Structured Two-Sex Model for Household Dynamics', *PLoS ONE*, 9(4): e94761, doi:10.1371/journal.pone.0094761.

O'Donoghue C 2018, 'Increasing the Impact on Dynamic Microsimulation Modelling', *International Journal of Microsimulation*, 11(1) 61-96.

Osborne, JW & Overbay, A 2004, 'The power of outliers (and why researchers should ALWAYS check for them', *Practical Assessment, Research and Evaluation*, vol. 9, no. 6, ISSN 1531-7714.

Pritchard, DR & Miller, EJ 2009, 'Advances in agent population synthesis and application in an integrated land use and transportation model', paper presented at the 88th Annual Meeting of the Transportation Research Board, Washington, DC.

Pritchard, DR & Miller, EJ 2012, 'Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously'. *Transportation*, vol. 39, no. 3, pp. 685-704.

Recker, WW, McNally, MG, Root, GS 1986, 'A model of complex travel behavior: Part I—Theoretical development', Transportation Research Part A: General Volume 20, Issue 4, July 1986, Pages 307-318.

Rose, AN & Nagle, NN 2016, 'Validation of spatiodemographic estimates produced through data fusion of small area census records and household microdata', *Computers, Environment and Urban Systems,* vol. 63, pp. 38-49.

Ruther M, Maclaurin G, Leyk S, Buttenfield B, Nagle N 2013, 'Validation of spatially allocated small area estimates for 1880 census demography', *Demographic Research*, vol. 29, pp. 579-616.

Rüschendorf, L 1995. 'Convergence of the iterative proportional fitting procedure', *The Annals of Statistics*, Vol. 23, No. 4, 1160-1174.

Ryan, J, Maoh, H & Kanaroglou, P 2007, 'Population synthesis: Comparing the major techniques using a small, complete population of firms', *Geographical Analysis*, 41 (2) 181–203.

Saadi, I, Mustafa, A, Teller, J, Cools,M., 2016, 'An integrated framework for forecasting travel behavior using markov chain monte carlo simulation and profile hidden markov models', in: Proceedings of the *95th Annual Meeting of the Transportation Research Board*, Transportation Research Board of the National Academies, Washington, DC.

Sachs, L 1982, *Applied statistics: A handbook of techniques (2nd ed)*, New York: Springer-Verlag.

Srinivasan, S, Ma, L & Yathindra, K 2008, 'Procedure for forecasting household characteristics for input to travel-demand models', Final Report, TRC-FDOT-64011-2008, Transportation Research Centre, University of Florida.

Sun, L & Erath, A, 2015, 'A bayesian network approach for population synthesis' Transportation Research Part C: Emerging Technologies, Vol.16, pp. 49-62, DOI: 10.1016/j.trc.2015.10.010. search Part C: Emerging Technologies 61, 49–62. doi: http://dx.doi.org/10.1016/j.trc.

Tanton, R 2014, 'A Review of Spatial Microsimulation Methods', *International Journal of Microsimulation*, 7(1) 4-25, International Microsimulation Association.

Voas, D, Williamson, P 2001, 'Evaluating Goodness-of-Fit Measures for Synthetic Microdata', *Geographical and Environmental Modelling* 5: 177–200.

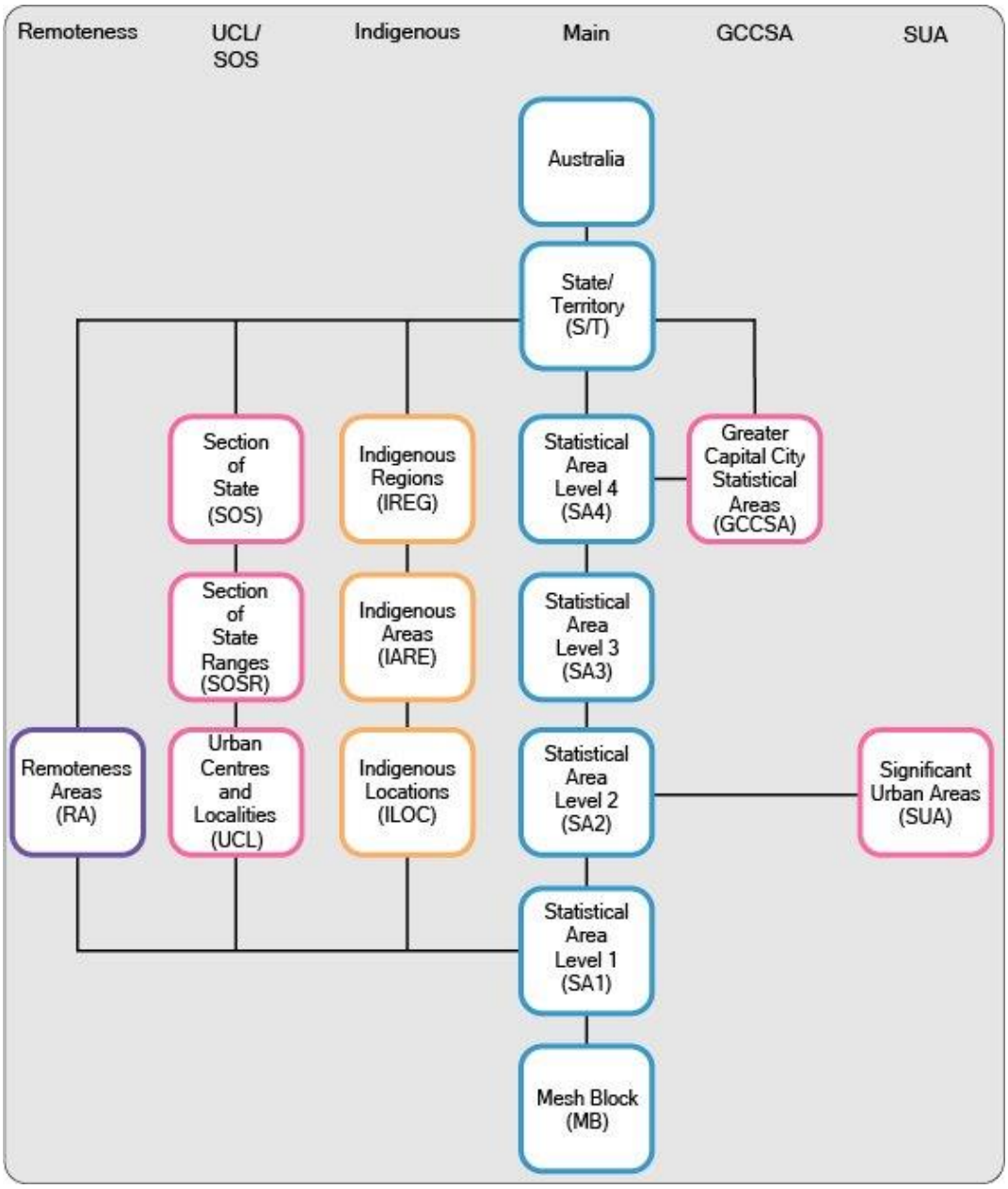
Voas, D & Williamson, P 2000, 'An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata', International Journal of Population Geography, 6(5):349-366.

Williamson P, Birkin, M, Rees, PH 1998, 'The Estimation of Population Microdata by Using Data from Small Area Statistics and Samples of Anonymised records', *Environment and Planning A* 30: 785–816.

Ye, X, Karthik K, Pendyala, R, Sana, B, & Waddell, P 2009, 'A Methodology to Match Distributions of Both Household and Person Attributes in the Generation of Synthetic Populations', presented at the *88th Annual Meetings of the Transportation Research Board.*

Zhuge, C, Li, X, Ku, C, Gao, J & Zhang, H 2017, 'A Heuristic-Based Population Synthesis Method for Micro-Simulation in Transportation', KSCE Journal of Civil Engineering, 21(6):2373-2383, DOI 10.1007/s12205-016-0704-1.

Appendix 1 ABS ASGS Structures



Source: ABS 2011

Appendix 2 Geographical Correspondence between Population Census and CURF 2011 for Greater Sydney, Greater Melbourne and Greater Brisbane

GCCSA Name	ASGS Statistical Region code (Based on SA4)	ASGS Statistical Region Name	CURF Area code
Greater Sydney	102	Central Coast	02
	115, 126	Sydney - Baulkham Hills and Hawkesbury, Sydney - Ryde	08
	116	Sydney - Blacktown	09
	117	Sydney - City and Inner South	10
	118	Sydney - Eastern Suburbs	11
	119	Sydney - Inner South West	12
	120	Sydney - Inner West	13
	121, 122	Sydney - North Sydney and Hornsby, Sydney - Northern Beaches	14
	123, 128	Sydney - Outer South West, Sydney - Sutherland	15
	124	Sydney - Outer West and Blue Mountains	16
	125	Sydney - Parramatta	17
	127	Sydney - South West	18
Greater Melbourne	206	Melbourne - Inner	22
	207	Melbourne - Inner East	23
	208	Melbourne - Inner South	24
	209	Melbourne - North East	25
	210	Melbourne - North West	26
	211	Melbourne - Outer East	27
	212	Melbourne - South East	28
	213	Melbourne - West	29
	214	Mornington Peninsula	30
Greater Brisbane	301, 302	Brisbane - East, Brisbane - North	32
	303	Brisbane - South	33
	304, 305	Brisbane - West, Brisbane Inner City	34
	310, 317	Ipswich, Toowoomba	38
	311	Logan - Beaudesert	39
	313, 314	Moreton Bay - North, Moreton Bay - South	41

Appendix 3 Record Linkage for Household Variables between Population and Housing Census data and CURF 2011

Selected Control Variables at Household Level				
Household Composition (HHCD)				
CSF Code	1% Basic CSF Classification	Linked	Census Code	Census Classification
1	One family household	HHCD1	1	One family household
2	Two or more family household	HHCD2	2	Multiple family household
3	Lone person household	HHCD3	31	Lone person household
4	Group household		32	Group household
5	Visitors only	HHCD4	41	Visitors only
6	Other non-classifiable		42	Other non-classifiable
7	Not Applicable		@@@	Not Applicable
Note: HHCD3 = Non family household				
Dwelling Structure (STRD)				
1	Separate house	STRD1	11	Separate house
2	Semi-detached, row or terrace house, town house, etc.	STRD2	21,22	Semi-detached, row or terrace house, town house, etc.
3	Flat, unit or apartment	STRD3	31,32,33, 34	Flat, unit or apartment
4	Other dwelling	STRD4	91,93,94	Other dwelling
5	Not stated	STRD5	&&	Not stated
6	Not applicable		@@	Not applicable
Number of Persons Usually Resident in Dwelling (NPRD)				
1	One person	NPRD1	1	One person
2	Two persons	NPRD2	2	Two persons
3	Three persons	NPRD3	3	Three persons
4	Four persons	NPRD4	4	Four persons
5	Five persons	NPRD5	5	Five persons
6	Six or more	NPRD6	6	Six
7	Not applicable	NPRD7	7	Seven
			8	Eight or more
			@	Not applicable
Number of Motor Vehicles (ranges) VEHRD				
0	No motor vehicles	VEHRD0	0	No motor vehicles
1	1 motor vehicle	VEHRD1	1	1 motor vehicle
2	2 motor vehicles	VEHRD2	2	2 motor vehicles
3	3 motor vehicles	VEHRD3	3	3 motor vehicles
4	4 or more motor	VEHRD4	4	4 or more motor
5	Not stated	VEHRD5	&	Not stated
6	Not applicable		@	Not applicable

Appendix 4 Record Linkage for Person Variables between Population and Housing Census data and CURF 2011

Selected Control Variables at Person Level				
Age of person (AGEP)				
CSF Code	1% Basic CSF Classification	Linked	Census Code	Census Classification
0 to 24	0 to 24 year singly Regroup to match 5-year age groups in census			Used by 5-year age groups (AGE5P)
	0-4 years	AGEP1	0-4 years	0-4 years
	5-9 years	AGEP2	5-9 years	5-9 years
	10-14 years	AGEP3	10-14 years	10-14 years
	15-19 years	AGEP4	15-19 years	15-19 years
	20-24 years	AGEP5	20-24 years	20-24 years
25	25-29 years	AGEP6	25-29 years	25-29 years
26	30-34 years	AGEP7	30-34 years	30-34 years
27	35-39 years	AGEP8	35-39 years	35-39 years
28	40-44 years	AGEP9	40-44 years	40-44 years
29	45-49 years	AGEP10	45-49 years	45-49 years
30	50-54 years	AGEP11	50-54 years	50-54 years
31	55-59 years	AGEP12	55-59 years	55-59 years
32	60-64 years	AGEP13	60-64 years	60-64 years
33	65-69 years	AGEP14	65-69 years	65-69 years
34	70-74 years	AGEP15	70-74 years	70-74 years
35	75-79 years	AGEP16	75-79 years	75-79 years
36	80-84 years	AGEP17	80-84 years	80-84 years
37	85 years and over	AGEP18	85 - 115	85 - 115
Sex (SEXP)				
1	Male	SEXP1	1	Male
2	Female	SEXP2	2	Female
Labour Force Status (LSFP)				
1	Employed	LFSP1	1,2,3	Employed
2	Unemployed	LFSP2	4,5	Unemployed
3	Not in the labour force	LFSP3	6	Not in the labour force
4	Not stated	LFSP4	&	Not stated
5	Not applicable		@	Not applicable
6	Overseas visitor		V	Overseas visitor

ISBN: ???

May 2020

Use of the Coat of Arms

Contact us

Department of Infrastructure and Regional Development
Bureau of Infrastructure, Transport and Regional Economics (BITRE)
GPO Box 501, Canberra ACT 2601, Australia

Phone: (international) +61 2 6274 7210
Fax: (international) +61 2 6274 6855
Email: bitre@infrastructure.gov.au
Website: www.bitre.gov.au