

Supplemental Material

1. Synthetic Population Literature

There is an increasing need for disaggregated population data for use in modelling urban systems (Leao et al., 2017). These data can be used to address current planning challenges including the provision of sustainable and active transportation. While disaggregated data may be collected by governments from population censuses, census data are typically released in aggregate form because of the need to preserve anonymity. Conducting additional surveys for the entire population being studied at city or national levels is often too costly (Barthélemy and Toint, 2013, Grapperon, 2016, Huynh et al., 2016).

A feasible alternative to surveys is to create a synthetic population for each small area within a study region, such as for the UK Office of National Statistics' Output Area, or the Australian Bureau of Statistics' (ABS) Statistical Area Level 1 (SA1). This is needed because released census data contains population-level information, whereas datasets that contain extra information, such as public transport smart card data, are at finer scales. Without a synthetic population, these datasets cannot be easily joined together with statistical rigour. This leaves important planning applications, such as quantifying commuter physical activity, near impossible to calculate across large cities.

A synthetic population is comprised of synthetic individuals and synthetic households, with the distribution of individuals and households in the synthetic population resembling as closely as possible the distribution of the true population as represented in the aggregate census tables (Huynh et al., 2013). Synthetic individuals typically contain attributes such as age, sex, relationship within a household (such as couple, single parent, or child), income, and employment location, while attributes for synthetic households include household type (such as family household or lone person household) and household size (Huynh et al., 2016, Moreno and Moeckel, 2018, Zhu et al., 2018).

There are two main approaches for creating a synthetic population: synthetic reconstruction and reweighting. Both use aggregate data to create synthetic individuals and households (Huang and Williamson, 2001).

Iterative Proportional Fitting (IPF), as illustrated in the work of Deming and Stephan (1940) and Beckman et al. (1996), is the main method for synthetic reconstruction (Choupani and Mamdoohi, 2016). For each small area within a study, a synthetic dataset at the household and individual scales is derived from various census tables, which serves as a microdata set (Tanton, 2014). Microdata refers to individual or household-level data and consists of additional information not released in the aggregated census. Aggregated census data contains only marginal distributions, which are the row and column sums of the relevant table (Barthélemy and Toint, 2013, Williamson et al., 1998). Examples of microdata include

the United States Census Bureau Public Use Microdata Sample (PUMS) and the ABS Confidentialised Unit Record Files (CURFs). An advantage of synthetic reconstruction is that synthetic microdata is created as a result of the microsimulation process rather than requiring a microdata source as an input for the microsimulation method (Tanton, 2014).

IPF has two main steps: fitting and allocation (Choupani and Mamdoohi, 2016). Within the fitting stage, for a sample seed with k characteristics of the population, the k -way distribution is estimated such that it fits the marginal distributions of the aggregated dataset (Barthélemy and Toint, 2013, Choupani and Mamdoohi, 2016). This procedure produces cross tabulations which can then be used in the allocation stage (Choupani and Mamdoohi, 2016).

Within the allocation stage, conditional probabilities are generated based on the expected number of individuals in the synthetic population for each group of characteristics (Barthélemy and Toint, 2013). The values in the cross tabulations may then undergo 'integerisation' of weights to reflect whole individuals or households, rather than having a situation where non-integer weights reflect only partial individuals or households (Lovelace and Ballas, 2013, Choupani and Mamdoohi, 2016). Individuals are then randomly selected from the sample based on the weights and the matching characteristics are assigned to a new synthetic individual (Barthélemy and Toint, 2013). This selection is repeated until the expected number of individuals in the synthetic population are created (Barthélemy and Toint, 2013).

Reweighting techniques for synthetic population generation arose from the work of Williamson et al. (1998). Several reweighting methods can be split into two groups. The first is methods that adjust the initial weights of microdata such as IPF reweighting (Ballas et al., 2005) and the generalised regression method (Tanton et al., 2011). The second is methods that select individuals such that their characteristics represent that of the small area for which the synthetic population is being generated (Tanton, 2014, Harding et al., 2004), such as combinatorial optimisation reweighting (Williamson et al., 1998).

The heuristic algorithm presented by Huynh et al. (2016) is an example of a sample-free approach. It involves the creation of a pool of synthetic individuals and a pool of synthetic households from aggregated census tables. The individuals are then assigned a household relationship and household type based on these tables, with the first set of individuals being assigned so that the minimum requirement of individuals for a certain household type is met. Following this, any individuals that have not been assigned, such as a household's second or third child, are then allocated to households to minimise the Root Mean Square Error (RMSE) between the synthetic population and the census tables. This approach overcomes a lack of survey data and the financial costs of microdata.

2. Algorithm to Generate the Synthetic Population

The algorithm was implemented in Python 2.7.13 for the study area of Sydney, Australia. It used the Census tables *Relationship in Household by Age by Sex* (ABS, 2017a), *Family Composition* (ABS, 2017b), *Household Composition by Number of Persons Usually Resident* (ABS, 2017c), and built upon the algorithm used by Huynh et al. (2016) by using *Occupation by Age by Sex* (ABS, 2017d) and *Method of Travel to Work by Sex* (ABS, 2017e).

There are sixteen types of family households recorded by the Census (Table S1). These are distinguished by the family relationship types of individuals within each household, e.g. couple, lone parent, and children under 15 years old. These family households are designated by the labels, HF1 to HF16 in the synthetic population. There are two types of non-family households, lone person and group households. In the synthetic population, these are referred to as H Lone and H Group respectively.

Each household type has a minimum number of residents that it can be home to (Table S2).

Table S1: Summary of household types, adapted from Table 2 in Huynh et al. (2016). The category descriptions are terms used by the Census (ABS 2017) with extra description for H Lone and H Group.

Category Description in Census	Label in Synthetic Population
Couple family with no children	HF1
Couple family with children under 15 and dependent students and non-dependent children	HF2
Couple family with children under 15 and dependent students and no non-dependent children	HF3
Couple family with children under 15 and no dependent students and with non-dependent children	HF4
Couple family with children under 15 and no dependent students and no non-dependent children	HF5
Couple family with no children under 15 and with dependent students and non-dependent children	HF6
Couple family with no children under 15 and with dependent students and no non-dependent children	HF7
Couple family with no children under 15 and no dependent students and with non-dependent children	HF8
One parent family with children under 15 and dependent students and non-dependent children	HF9
One parent family with children under 15 and dependent students and no non-dependent children	HF10
One parent family with children under 15 and no dependent students and with non-dependent children	HF11
One parent family with children under 15 and no dependent students and no non-dependent children	HF12
One parent family with no children under 15 and with dependent students and non-dependent children	HF13
One parent family with no children under 15 and with dependent students and no non-dependent children	HF14
One parent family with no children under 15 and no dependent students and with non-dependent children	HF15
Other family	HF16

Category Description in Census	Label in Synthetic Population
Lone person household (non-family household with one person usually resident)	HLone
Group household (non-family household with more than one person usually resident)	HGroup

Table S2: Minimum number of residents for household types according to household relationship, adapted from Table 3 in Huynh et al. (2016).

Relationship	HF1	HF2	HF3	HF4	HF5	HF6	HF7	HF8	HF9
Couple	2	2	2	2	2	2	2	2	N/A
LoneParent	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	1
ChildU15	N/A	1	1	1	1	N/A	N/A	N/A	1
Student	N/A	1	1	N/A	N/A	1	1	N/A	1
NDChild	N/A	1	N/A	1	N/A	1	N/A	1	1
OtherRelated	0	0	0	0	0	0	0	0	0
Unrelated	0	0	0	0	0	0	0	0	0
LonePerson	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
GroupMember	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Minimum residents	2	5	4	4	3	4	3	3	4

Relationship	HF10	HF11	HF12	HF13	HF14	HF15	HF16	HLone	HGroup
Couple	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
LoneParent	1	1	1	1	1	1	N/A	N/A	N/A
ChildU15	1	1	1	N/A	N/A	N/A	N/A	N/A	N/A
Student	1	N/A	N/A	1	1	N/A	N/A	N/A	N/A
NDChild	N/A	1	N/A	1	N/A	1	N/A	N/A	N/A
OtherRelated	0	0	0	0	0	0	2	N/A	N/A
Unrelated	0	0	0	0	0	0	N/A	N/A	N/A
LonePerson	N/A	N/A	N/A	N/A	N/A	N/A	N/A	1	N/A
GroupMember	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	2
Minimum residents	3	3	2	3	2	2	2	1	2

Three custom tables, which provide greater detail on Method of Travel to Work in the Census were downloaded from the ABS TableBuilder website¹. These were *SA1_byDZN_POW* (hereafter OxD, referring to origin by destination) (ABS, 2018b), *SA1_byMethodTraveltoWork* (hereafter OxM, referring to origin by mode) (ABS, 2018c), and *DZN_POW_by_MTWP_MethodTravelWork* (hereafter DxM, referring to destination by mode) (ABS, 2018a).

¹ <https://auth.censusdata.abs.gov.au/webapi/jsf/login.xhtml>

2.1. Create individual and household pools

Firstly, two pools were created: the individual pool and the household pool. From these, synthetic individuals and synthetic households were selected to form the synthetic population. The synthetic individuals were created according to the *Relationship in Household by Age by Sex* table, with ages randomly assigned from each age group category. The synthetic households were created according to the *Family Composition* and *Household Composition by Number of Persons Usually Resident* tables (Huynh et al., 2016). If there was a discrepancy in the Census between the number of residents per household and the number of households, e.g. for an SA1 with 19 people in a certain household type, but zero households of that type, then x number of households were created. x represents the number of people in the household type for the SA1 divided by the minimum number of residents required for that household type. This was rounded down to the nearest whole number. If the household type required a minimum of five residents, then the number of households that were created was $19/5$, equalling three households. If the number of residents for a household type in an SA1 was smaller than the minimum required, then one household of that household type was created, with the potential to create extra individuals in later steps.

2.2. Assign couples to couple households

Each couple household (HF1 – HF8) was assigned a desired age gap between the parents which was derived from a normal distribution. For the couple households with children, a minimum age for the female or younger parent was ascribed to reflect the assumption that the age of consent is 16. As the minimum age for dependent students and non-dependent children is 15, the household types, HF2 – HF4 and HF6 – HF8, have minimum ages for females or younger parents as 31. Couples were assigned based on the minimum age and desired age gap where possible. Where the minimum age of the parent could not be satisfied, the minimum age was ignored. If a single individual was left in the pool and there were still households that were yet to have couple individuals assigned to it, an individual was generated according to the desired age of the household.

2.3. Assign lone parents to lone parent households

Lone parent households had minimum age requirements similar to the couple households. For each lone parent household, an individual that met the minimum age was selected and assigned. As with the couple households, if there were only individuals below the minimum age, the minimum age was ignored. If there were no lone parents left in the individual pool, but lone parent households remained, new lone parent individuals were created according to a normal distribution of the sex and ages recorded in the Census.

2.4. Assign one non-dependent child to each household that requires it

This component assigned non-dependent children to households that required these individuals. For couples, the female or younger parent was selected and combined into the pool with lone parents. This pool, along with the pool of non-dependent children were both

sorted by age in descending order. For each parent, the eldest child that satisfied the age constraint was assigned to the parent's household. If there were no children that met the age constraint, the child that was closest to this age was selected. If there were no non-dependent children left in the individual pool, but households that required non-dependent children remained, new individuals were created according to a normal distribution of the sex and ages recorded in the Census.

2.5. Assign one dependent student and/or one child under 15 to each household that requires it

The steps to assign dependent students and children under 15 were similar to that of assigning non-dependent children.

2.6. Assign two other related individuals to each household that requires it

Other related individuals were assigned to households of type HF16. If there were no more other related individuals in the individual pool, but households that required them, new individuals were created. For simplicity, if new individuals were created, they were allocated with other related individuals of the same sex.

2.7. Assign lone persons to lone person households

Lone persons were assigned to households of type HLon. If there were no lone persons in the individual pool, yet lone person households remained, new individuals were created according to a normal distribution of the sex and ages recorded in the Census.

2.8. Assign group members to group households

Group members were assigned to households of type HGroup. The table, *Household Composition by Number of Persons Usually Resident*, was used to estimate the number of residents for each group household. If there were households that still needed to have group members assigned, yet there were no more group members in the individual pool, new individuals were created according to a normal distribution of the sex and ages recorded in the Census. If there were still group members left in the individual pool and all households had been assigned their allocated number, these individuals were randomly assigned to households that were designated as having six or more individuals.

2.9. Assign remaining non-dependent children, dependent students, children under 15, other related individuals, and unrelated individuals to family households

The remaining individuals in the individual pool were assigned to HF2 – HF16 households. The individuals were assigned to minimise the Root Mean Square Error (RMSE) for the “distribution of family households by household size” and the “distribution of individuals by family household type” in the synthetic population compared to the Census (Huynh et al., 2016, p. 10).

The RMSE formulae in Huynh et al. (2016) was used as follows. Equation S1 illustrates the RMSE for the distribution of individuals.

(S1)

$$RMSE_{IndCount} = \sqrt{\frac{1}{n_{HFType}} \sum_{i=1}^{n_{HFType}} (PIC_i - PIS_i)^2}$$

where

$$PIS_i = \begin{cases} \frac{IS_i}{1 + \sum_{j=1}^{n_{HFType}} IS_j}, & i \neq k \\ \frac{1 + IS_i}{1 + \sum_{j=1}^{n_{HFType}} IS_j}, & i = k \end{cases}$$

and

$$PIC_i = \frac{IC_i}{\sum_{j=1}^{n_{HFType}} IC_j}.$$

As in Huynh et al. (2016), IC represents the array of individual counts in the Census data according to the type of household, while IS represents the array of counts in the synthetic population prior to a new individual being assigned to a household. n_{HFType} represents the number of categories of family households (i.e. HF1 – HF16, of which there are 16) and k refers to the index of IS for the household type for which a new individual may be assigned (e.g. HF2 would be represented by $k = 2$ etc.).

Equation S2 illustrates the RMSE for the distribution of households.

(S2)

$$RMSE_{HholdCount} = \sqrt{\frac{1}{n_{HFSize}} \sum_{i=1}^{n_{HFSize}} (PHC_i - PHS_i)^2}$$

where

$$PHS_i = \begin{cases} \frac{HS_i}{1 + \sum_{j=1}^{n_{HFSize}} HS_j}, & i \neq k \\ \frac{1 + HS_i}{1 + \sum_{j=1}^{n_{HFSize}} HS_j}, & i = k \end{cases}$$

and

$$PHC_i = \frac{HC_i}{\sum_{j=1}^{n_{HFSize}} HC_j}.$$

As in Huynh et al. (2016), HC represents the array of counts of family households according to size in the Census data, while HS represents the array of counts in the synthetic population prior to a new individual being assigned to a household. n_{HFSize} is 5, representing the number of household size types (i.e. two person households up to six or more person households) and k refers to the index of HS for the household size for which a

new individual may be assigned (e.g. a two person household would be represented by $k = 1$ etc.).

For each individual to be assigned, the $RMSE_{IndCount}$ for each HFTYPE and $RMS_{HholdCount}$ of each HFSIZE combinations were summed. The smallest value was considered the best HFTYPE-HFSIZE option. If there were no households that matched the best option for HFTYPE and HFSIZE, then any household with a feasible HFTYPE could be selected, provided it matched the HFSIZE. For a household to match the HFSIZE, it must have one person fewer than the ideal HFSIZE, so that when the new person is added to the household, the size will equal the desired HFSIZE. If there were no feasible HFTYPES for households that would match the HFSIZE, then any household with a feasible HFTYPE could be chosen, regardless of the HFSIZE. If any children could not be assigned, this was because there were no couple or lone parent households available in the SA1 that had a feasible HFTYPE to accommodate the child.

2.10. Assign working and commuter statuses to synthetic adults

The working and commuter statuses of adults in the synthetic population were then assigned. These were derived from the Census tables, *Occupation by Age by Sex* and *Method of Travel to Work by Sex*.

2.11. Assign modes and destinations

The OxD, OXM, and DxM tables were used to calculate, for each origin, the probability that a person took each possible mode-destination combination (Equation S3).

(S3)

$$P(\text{Mode} \cap \text{Destination} \mid \text{Origin})$$

Workplace addresses, derived from geocoded address data, were then assigned using random selection based on the chosen destination zone (DZN). Firstly, address points that were located in non-residential MBs were considered. If there were only address points in residential MBs, then these were randomly selected from. If there were no address points available, an approximate midpoint, or one that aligned with imagery available through Google Maps (2019) was chosen as the address point. The work MB that the address fell within was noted.

2.12. Assign home MBs and home addresses

MBs were assigned for each synthetic household and their subsequent synthetic individuals. The smallest RMSE combination between the distribution of individuals in each MB in the Census and in the synthetic population and the distribution of households in each MB in the Census and in the synthetic population were calculated.

Equations S4 and S5 were adapted from Huynh et al. (2016) but were used to calculate the RMSE for the distribution of individuals in each MB and the RMSE for the distribution of households in each MB respectively.

(S4)

$$RMSE_{IndCount} = \sqrt{\frac{1}{n_{MB}} \sum_{i=1}^{n_{MB}} (PIC_i - PIS_i)^2}$$

where

$$PIS_i = \begin{cases} \frac{IS_i}{size + \sum_{j=1}^{n_{MB}} IS_j}, & i \neq k \\ \frac{size + IS_i}{size + \sum_{j=1}^{n_{MB}} IS_j}, & i = k \end{cases}$$

and

$$PIC_i = \frac{IC_i}{\sum_{j=1}^{n_{MB}} IC_j}.$$

IC represents the array of individual counts in the Census data per MB, while IS represents the array of counts in the synthetic population prior to new individual(s) being assigned to a MB. k refers to the index of IS for the MB for which new individual(s) may be assigned (Huynh et al., 2016). n_{MB} refers to the number of MBs in an SA1, while $size$ indicates the number of residents per household for the household that is being assigned.

(S5)

$$RMSE_{HholdCount} = \sqrt{\frac{1}{n_{MB}} \sum_{i=1}^{n_{MB}} (PHC_i - PHS_i)^2}$$

where

$$PHS_i = \begin{cases} \frac{HS_i}{1 + \sum_{j=1}^{n_{MB}} HS_j}, & i \neq k \\ \frac{1 + HS_i}{1 + \sum_{j=1}^{n_{MB}} HS_j}, & i = k \end{cases}$$

and

$$PHC_i = \frac{HC_i}{\sum_{j=1}^{n_{MB}} HC_j}.$$

HC represents the array of counts of households per MB in the Census data, whereas HS represents the array of counts in the synthetic population prior to a household being assigned to a MB. k refers to the index of HS for the MB for which a household may be assigned (Huynh et al., 2016).

Home addresses were assigned similarly to work addresses, except that addresses located in residential MBs were first considered before addresses located in non-residential MBs.

3. Synthetic Population Characteristics

The Freeman-Tukey statistic (Equation S6) was used to measure the fit of the distribution of the synthetic population relative to the Census data for eight demographic attributes for each SA1 (Figure S1). T represents the distribution of the Census data and T' represents the distribution of the synthetic population (Huynh et al., 2016). A large p-value (>0.05) indicates that there is no evidence that the synthetic population is not drawn from the Census data.

(S6)

$$FT(T, T') = 4 \sum_i \left(\sqrt{T_i} - \sqrt{T'_i} \right)^2$$

The attributes that had the best match in the synthetic population with the Census data were family by household type and non-family by size (Figure S1). For both attributes, the distribution of the synthetic population matched the distribution of Census data for 100% of SA1s as Census distributions were used to create the household pool (Huynh et al., 2016). Although Census distributions were also used to create the individual pool, the attributes male by relationship and female by relationship averaged 81.1% and 77.5% respectively as new individuals could be created under certain circumstances where random adjustments had been introduced into the Census data (Supplementary Material s2.2-2.8). For example, if there was an odd number of individuals in a couple relationship for an SA1, or there were lone parent households but no lone parents in the individual pool, then new synthetic individuals would be created.

The attribute with the lowest match was family by size, with a mean of 58.1% of SA1s recording a match across each of the replications. This was followed by persons per MB (66.0%).

During the assignment of remaining children, other related and unrelated individuals, the distribution of family by size may have been affected if there were no households that matched the best HFTYPE-HFSIZE option. In cases where there were no feasible HFTYPE for households that matched the HFSIZE, then any household with a feasible HFTYPE may have been chosen, regardless of the HFSIZE. This would cause the distribution of family by size of the synthetic population to be more dissimilar to the Census data. Additionally, some children were not able to be assigned if there were no couple or lone parent households available in the SA1 that had a feasible HFTYPE to accommodate the child. These children were removed from the synthetic population, which may have also influenced the distributions of the synthetic population.

Another contributing factor to the results of both family by size and persons per MB is the fact that different Census tables had different total populations (Table S3), with the largest

representation being that derived from the *MB population* (4,320,224 persons). This was a discrepancy of 600,842 persons from the *Relationship in Household by Age by Sex* table, which had only 3,719,382 persons. For the *Relationship in Household by Age by Sex* table, random errors and the exclusion of the 'Visitor (from within Australia)' relationship category contributed to the smaller figure of persons. Using the totals column and including the 'Visitor (from within Australia)' relationship category results in an extra 283,979 individuals than when individual cell counts are used and the 'Visitor (from within Australia)' relationship category is excluded. The remaining unaccounted individuals in the *Relationship in Household by Age by Sex* table are also expected to be from 'Visitors only' households and 'Other non-classifiable' households which are excluded by the ABS in this table.

Despite the discrepancies inherent in the Census data, the algorithm performed well overall. The mean of the proportion of p-values that exceed 0.05 across the eight attributes was 82.0% (Figure S1).

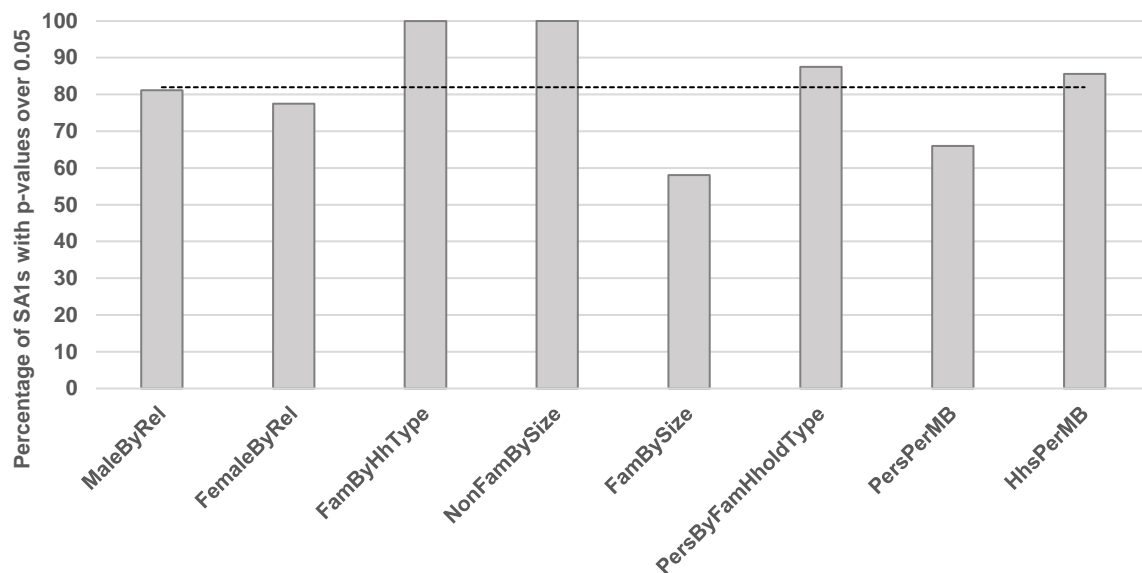


Figure S1: Percentage of SA1s for which the synthetic population matches the Census data for eight demographic attributes. Data for the mean across ten replications. Based on ABS data.

Table S3: Differences for total population of the study area in Census tables and compared with the synthetic population. Based on ABS data.

	Population	Notes
Synthetic Population	3,823,920	-Mean across ten replications. -Excludes the 'Visitor (from within Australia)' category.
Relationship in Household by Age by Sex (excluding visitors)	3,719,382	-Excludes the 'Visitor (from within Australia)' category. -"Excludes persons in 'Visitors only' and 'Other non-classifiable' households" (ABS, 2017a).
Household Composition by Number of Persons Usually	3,967,799	-Assumes that the number of people in six or more households equals 6.

	Population	Notes
Resident (assuming six or more = 6)		-“Excludes family members temporarily absent on Census Night. Excludes overseas visitors” (ABS, 2017b).
Household Composition by Number of Persons Usually Resident (assuming six or more = 7)	4,034,928	-Assumes that the number of people in six or more households equals 7. -“Excludes family members temporarily absent on Census Night. Excludes overseas visitors” (ABS, 2017b).
MB population	4,320,224	-

The distributions of the Census data and synthetic population were similar across attributes (Figure S2, Figure S3, Figure S4). The results across the ten replications showed very little variation, with standard deviations of the proportion of population allocated to each subgroup being too small to plot (<0.00004 in Figure S2, <0.00007 in Figure S3 and <0.00008 in Figure S4).

When the mean absolute percentage error (MAPE) of total population between the Census data and synthetic population at the SA1 level was calculated, the errors were typically small (Figure S5). 80% of the data had errors $<6.0\%$, whereas 50% of the data had errors $<3.1\%$. There was also no meaningful spatial structure in the accuracy surface when an Anselin Local Moran’s I was executed (Figure S6). It should be noted that the *Relationship in Household by Age by Sex* table (excluding visitors) was used, whereas the algorithm to create the synthetic population also used tables such as *Household Composition by Number of Persons Usually Resident*.

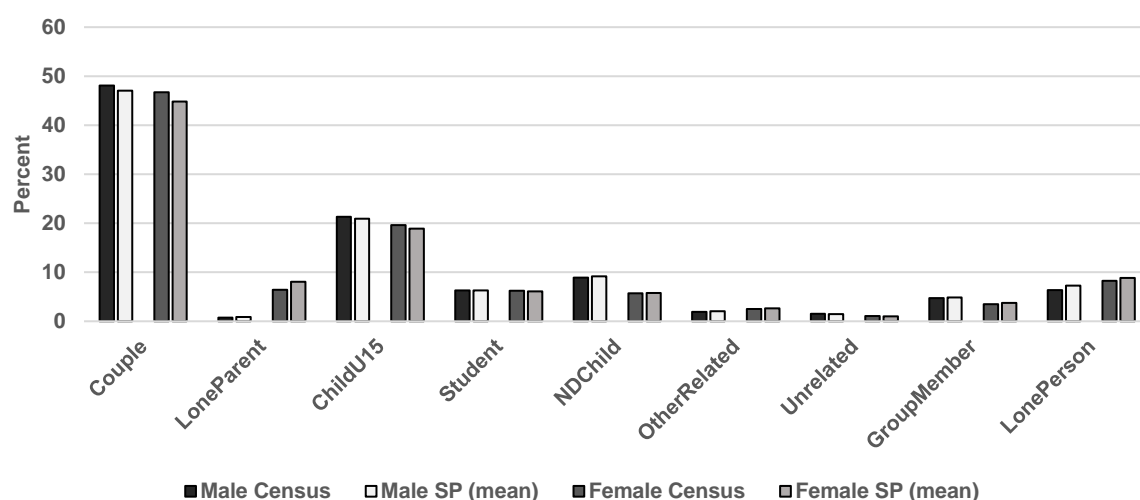


Figure S2: Mean distribution of males and females by household relationship for ten replications. Based on ABS data.

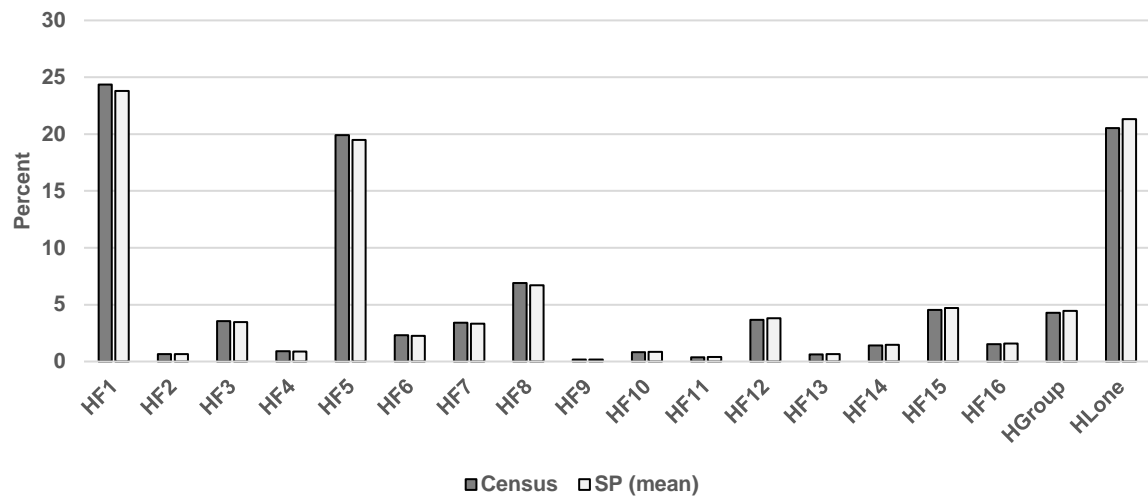


Figure S3: Mean distribution of family and non-family households by household type for ten replications. See Table S1 for category descriptions. Based on ABS data.

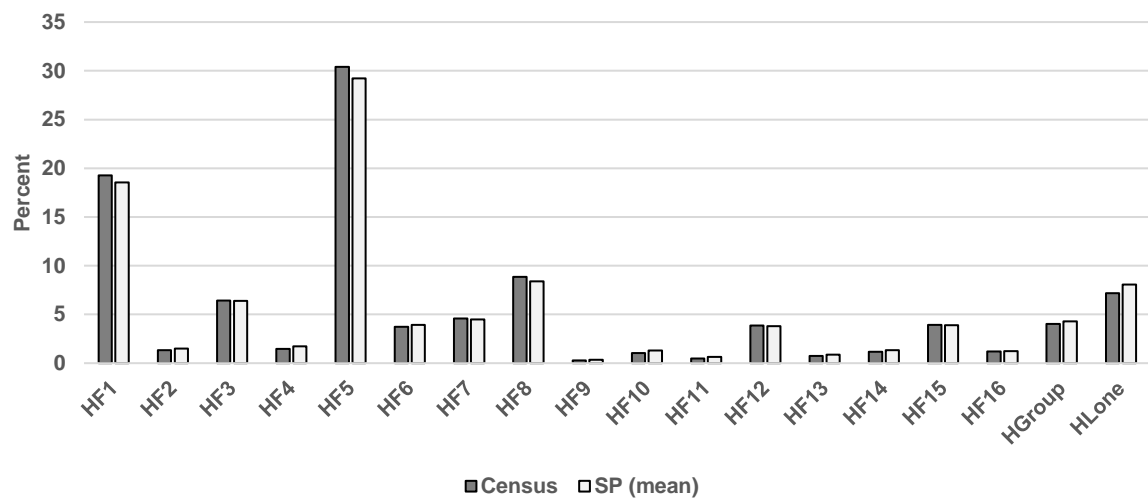


Figure S4: Mean distribution of persons by household type for ten replications. See Table S1 for category descriptions. Based on ABS data.

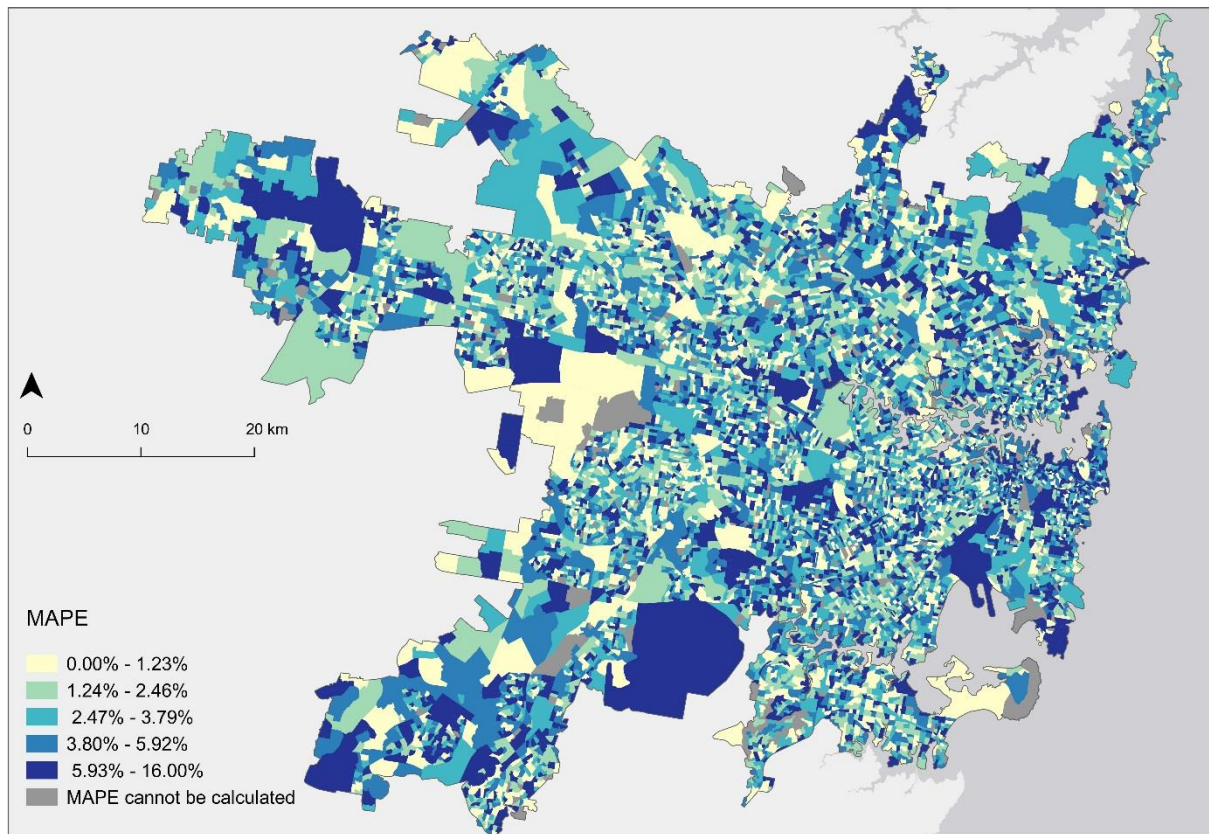


Figure S5: MAPE between total populations of Census and synthetic population by SA1. Data is displayed in quintiles (Based on data from: ABS, 2016a, ABS, 2016b).

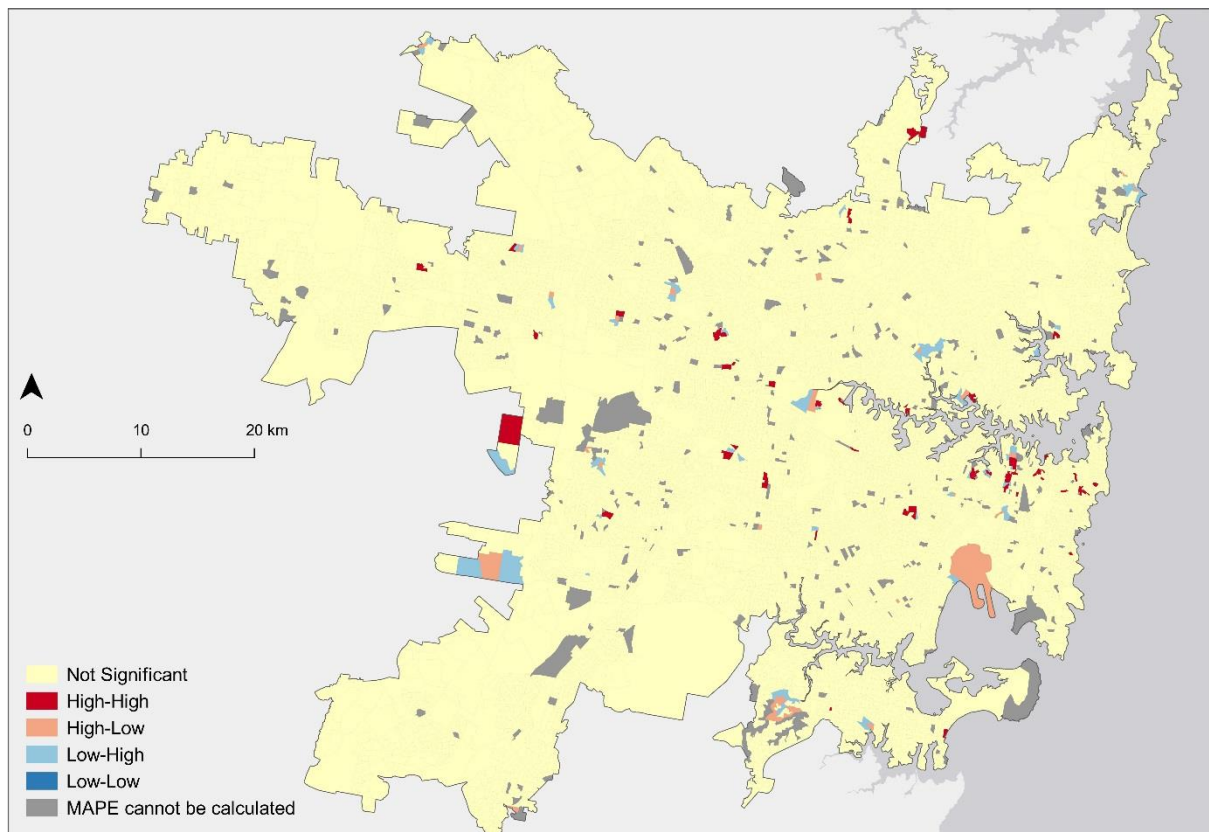


Figure S6: Anselin Local Moran's I analysis of the MAPE between total populations of Census and synthetic population by SA1, using Queen's case adjacency (Based on data from: ABS, 2016a, ABS, 2016b).

4. Assigning Opal journeys – Detailed Explanation

Commuters typically travel to work in the morning (Alsger et al., 2015, Zhang et al., 2016). For each Opal journey that had the first tap-on (entry to the first public transport mode or station/wharf used in a journey) in the AM Peak or Inter Peak period, the first tap-on was linked with home MBs (based on spatial location). The last tap-off (exiting from the last public transport mode or station/wharf used in a journey) was linked with work MBs. If the first tap-on for an Opal journey was in the Other period, but the last tap-off was in the AM Peak, the journey was considered as belonging to the AM Peak. Commuters typically travel back to their home in the evening (Alsger et al., 2015, Zhang et al., 2016). For each Opal journey with the first tap-on in the PM Peak or Other period, the first tap-on was linked with work MBs (based on spatial location) and the last tap-off was linked with home MBs. If the first tap-on of an Opal journey was in the Inter Peak period, but the last tap-off was in the PM Peak, the journey was considered as belonging to the PM Peak.

Network spatial weights matrices were created using ArcGIS for each MB using pedestrian networks (see Streets data in Table 3). This was to identify neighbouring MBs within 0.5km, 1km, and 1.5km distances. It was assumed that commuters could walk up to 1.5km between their home and public transport, depending on the mode used. Daniels and Mulley (2013) showed that train and bus users can walk up to 2km to catch public transport in Sydney, with a mean of 805m for train users and a mean of 461m for bus users. Burke and Brown (2007) also found that people were more likely to walk shorter distances (<500m) to bus stops rather than train stations. This was due to the larger density of bus stops compared to train stations. The mean for walking to bus stops was 600m and the 85th percentile was 1.07km. For both train and ferry, the mean distance was ~1km and the 85th percentile was slightly larger than 1.5km. The values assumed for light rail were the same as for bus (Table 4) as light rail stops are typically spaced more closely than train stations. Since distances walked from public transport to destinations are often less than access distances (Burke and Brown, 2007), this was reflected in the smaller maximum search radii around work MB centroids (Table 4).

For commuters that used public-with-private transport, car network spatial weights matrices were created. This was to identify neighbouring MBs within 1km, 2km, and 5km of each other. Private modes of transport increase commuters' access distance to public transport, especially in low-density areas where people are more likely to drive instead of walk to public transport (Vijayakumar et al., 2011). As such, search distances larger than for walking were chosen. The larger search distance thresholds for train-with-private transport and ferry-with-private transport compared to bus-with-private modes and light rail-with-private modes (Table 4) were due to the sparser coverage of train stations and ferry wharves compared to bus stops and light rail stops. In this research, private transport modes included car, taxi, truck, motorcycles and bicycles and any other modes that were not public modes (bus, ferry, light rail, train/metro) or walking.

For any synthetic commuters not assigned journey to work Opal data, their work MB may have been reassigned based on actual travel patterns observed in the Opal data rather than relying on the origin x destination (OxD) Census data (ABS, 2018b). The summed cell values in this table only accounted for 55.3% of data compared to the totals provided by the ABS (Table S4). This difference can be explained by the ABS introducing errors to prevent the identification of individuals, with certain OxD combinations likely to be excluded from reporting if their cell counts were low. Approximately 873,166 commuting individuals were not represented in the OxD table. This likely affected the ability of OxDxM probability calculations to accurately predict OxD combinations, especially for combinations with small counts.

Table S4: Comparison of data in cells compared to totals provided by the ABS in the custom tables OxD (ABS, 2018c), OxD (ABS, 2018b), and DxM (ABS, 2018a).

	Sum of Individual Cell Counts	Totals Provided	% of Data in Cells
OxD	3,465,026	3,515,747	98.5%
OxD	1,078,704	1,951,870	55.3%
DxM	2,194,819	2,209,296	99.3%

If a commuter lived near two modes of transport, they could likely have chosen to take either of those modes. If no matches were found in the Opal data, it was assumed that only private modes were used. There is evidence that different modes can be used between the journey to work and the return journey. For example, Burke and Brown (2007) found that commuters were more likely to use private modes to access public transport for the journey to work than the return journey. In the absence of data for the return journey mode in the Census, the following assumptions were made based on the distance travelled and influence of private modes on the journey to work:

- If the distance from the centroid of the last stop's MB to the centroid of the home MB was within 0.5km, it was assumed that only public transport modes were used for the return journey.
- If the distance from the centroid of the last stop's MB to the centroid of the home MB was between 0.5-1km, it was assumed that only public transport modes were used for the return journey if the journey to work used public transport modes only. Conversely, if public-with-private modes were used for the journey to work, it was assumed that private modes were also used for the return journey.
- If the distance from the centroid of the last stop's MB to the centroid of the home MB was between 1-2km, it was assumed that public-with-private modes were used for the return journey.

5. Validation and Calculation of Door-to-door Travel Time – Example, Assumptions and Limitations

For example, when the coordinates of home (origin) and work (destination) were entered into the Trip Planner API (Figure S7), valid options that a synthetic commuter who used a bus could take were:

- boarding a bus from Stop B and alighting from Stop D on the green route; and
- boarding a bus from Stop C and alighting from Stop D on the blue route.

A journey in which a commuter departed from home and walked past Stop B and boarded the same bus at Stop A would not be validated for the commuter's journey to work. The method did not validate other tap-offs and tap-ons that comprise multiple-segment journeys (for example those extra tap-ons and tap-offs that occur by catching two or more buses).

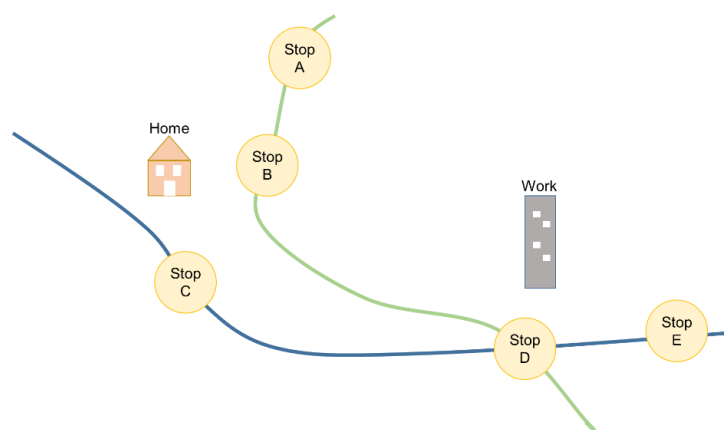


Figure S7: Validation using the TfNSW Open Data Hub and Developer Portal Trip Planner APIs. This process validated whether the journey's first tap-on location and last tap-off location were appropriate based on the home location, work location, and mode used by the synthetic commuter. Valid options that a synthetic commuter who used a bus to work could take were: (1) boarding a bus from Stop B and alighting from Stop D on the green route; and (2) boarding a bus from Stop C and alighting from Stop D on the blue route.

Assumptions and limitations of the validation process and calculation of door-to-door travel time include:

- The time for queries to the Trip Planner API for journeys to work was set to start at 8:00 am. For return journeys from work, the start time was set to 5:00 pm. Commuters may have started their commutes before or after these times, though these times reflect the middle of the morning/afternoon peak periods as determined by TfNSW.
- The mode set for queries to the API was the single mode that a synthetic commuter used (except for synthetic commuters who used multiple public transport options). For example, if the commuter used a bus for their journey to work, then only journeys that used a bus or school bus were included in the output from the API for validating the journey to work.

- The date set for queries to the API was Tuesday 13 October 2020, since the API did not provide access to historical data. This means that journeys returned by the API reflect public transport infrastructure and timetables that may not have been in operation at the time of the 2016 Census. For example, the Sydney Metro Northwest was not operational in 2016 but was opened in May 2019 and the CBD and South East Light Rail was opened from December 2019 (see Figure S12). This means that there may have been synthetic commuters for which the journeys could have been validated if the process had been conducted in 2016 rather than 2020.
- Even though commuters were not able to use the metro in 2016, the North-West Rail Link train line between Epping and Chatswood was operational in 2016 (see Figure S12). Hence, for synthetic train users, metro was also included as a valid mode in addition to train. However, reported travel time may have been slightly shorter if the metro was used compared to regular train services.
- When several options were returned by the API for a single commuter, only the first three options were considered. The shortest of these options' travel time was selected as the travel time for that commuter since it was possible in theory to travel within that time.
- It was assumed that the services operated according to the timetable, as timetabling would incorporate longer-term averages.
- Only public transport journeys were validated. It was beyond the scope of the Trip Planner API to account for private modes of transport for accessing or egressing from public transport stops.
- The process did not validate whether the home or workplace was correct based on satellite imagery, which can have issues when using linear regression (Chapuis et al., 2018), or fine-scale land-use data that incorporates population density per building footprint. MB data for homes were the finest scale at which reasonable population estimates were available and the DZNs for workplaces were constructed by the ABS to analyse the Census' Place of Work data.

6. Additional Results for the Journey to Work

6.1. Public transport commuters

On average, 442,085 synthetic commuters used public transport to work modelled across the 10 realisations (Figure S8). Train was the most common mode type and was double that of bus.

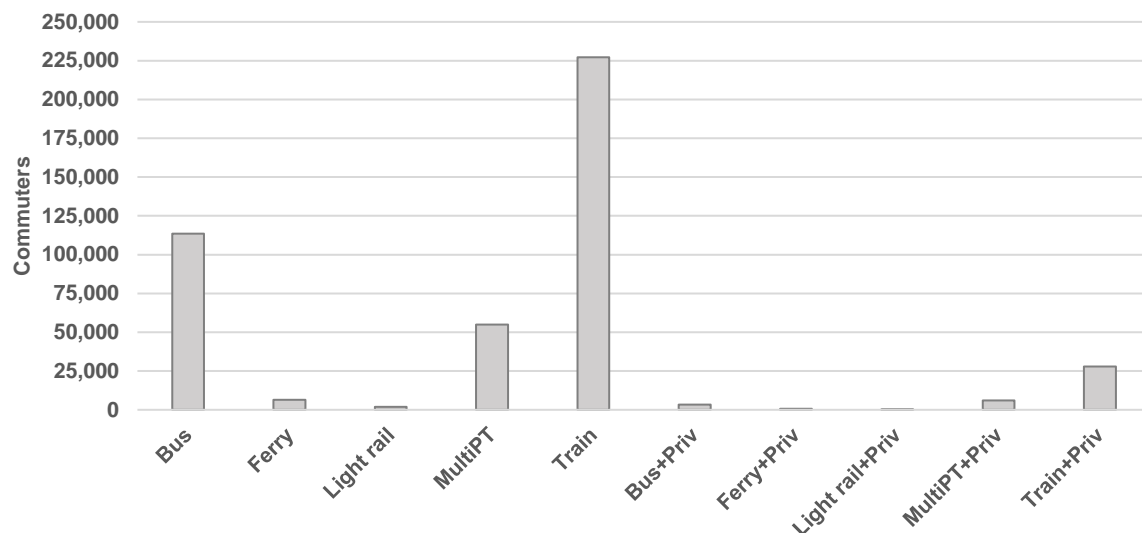


Figure S8: Breakdown of commuters who use public transport travelling to work by mode, showing the mean of ten replications. Commuters that were unable to be assigned with Opal data are included in the totals. Based on ABS data.

6.2. Opal data assignment

For commuters travelling to work, there were ten times as many commuters only using public transport than public-with-private commuters. 40.3% of commuters only using public transport modes could be matched with Opal card data (Figure S9). A further 31.3% could be matched if their work MB was reassigned. For public-with-private commuters, 35.7% were matched with Opal card data while a further 25.4% were matched if their work MB was reassigned.

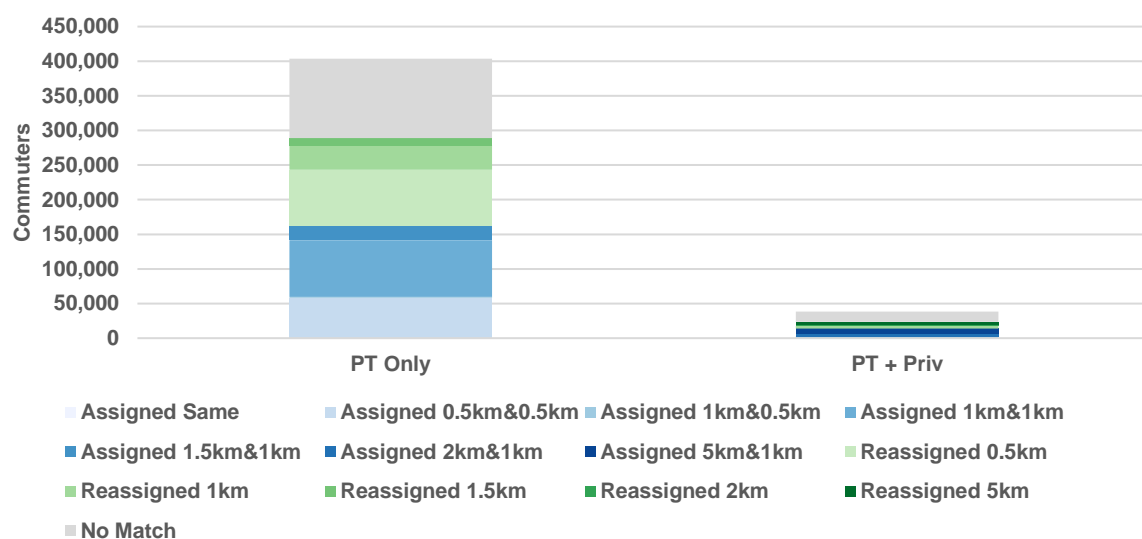


Figure S9: Summary of Opal data assignment for synthetic workers travelling to work for Replication 5. Based on ABS data and TfNSW data.

Proportionally, train was the best mode for which the most Opal data assignments were made without needing to reassign work MBs (Figure S10). When work MB reassignments were taken into consideration, multiple public transport options performed the best. Opal data was assigned for 69.2% of synthetic train commuters when work MB reassignments were considered.

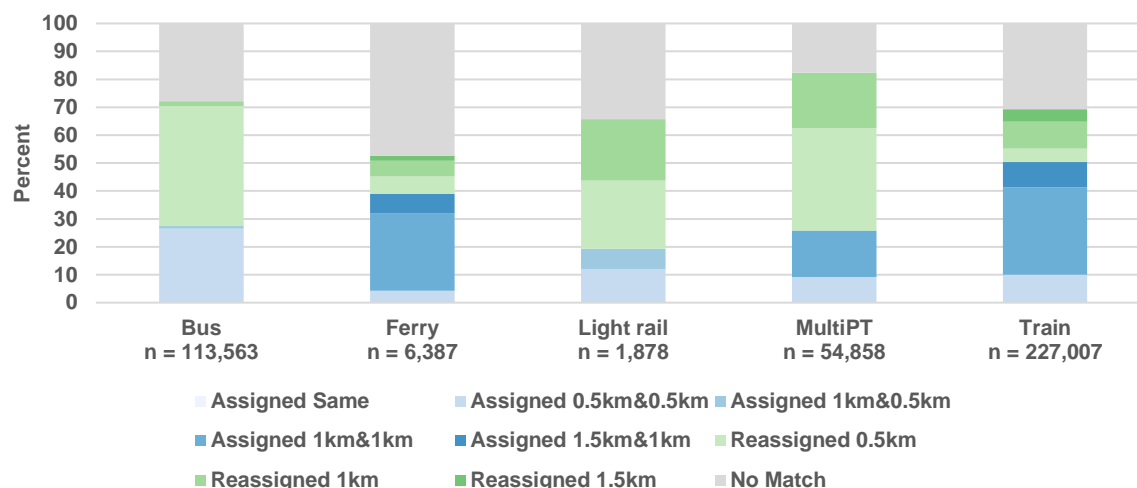


Figure S10: Opal data assignment by mode for synthetic workers travelling to work who use public transport only. Figure denotes Replication 5. Based on ABS data and TfNSW data.

The model is better able to match Opal data with commuters that live near train lines than commuters living further away. For the journey to work, 25.4% of SA1s have over 90% of public transport commuters paired with Opal data across the 10 replicates (Figure S11). SA1s with these larger percentages of commuters being paired (>90%) tended to follow the rail network.

Train was the dominant travel mode around train lines, especially in the central-eastern portion of the study area (Figure S12). As the distance from the train line increased, train-with-private modes and multiple public transport options were more likely to dominate. Multiple public transport options were likely used rather than train-with-private modes in cases where there was less car ownership or there was limited car parking around train stations. Bus was the dominant mode in South-east Sydney, the Northern Beaches, The Hills District, and parts of the Inner West and Lower North Shore that were isolated from train services.

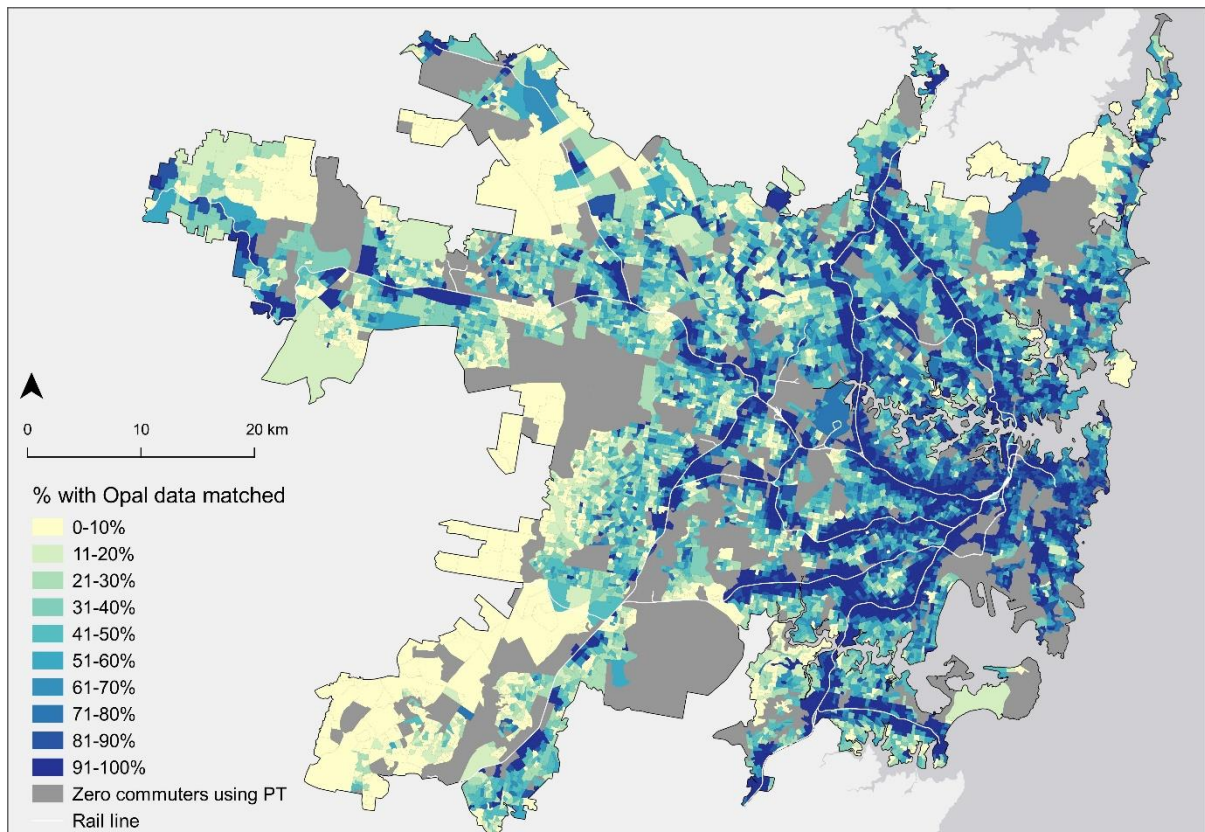


Figure S11: Percentage of commuters using public transport matched with Opal data for the journey to work per SA1
(Based on data from: TfNSW, ABS, 2016a, ABS, 2016b, HERE, 2017). Data for mean across 10 replications.

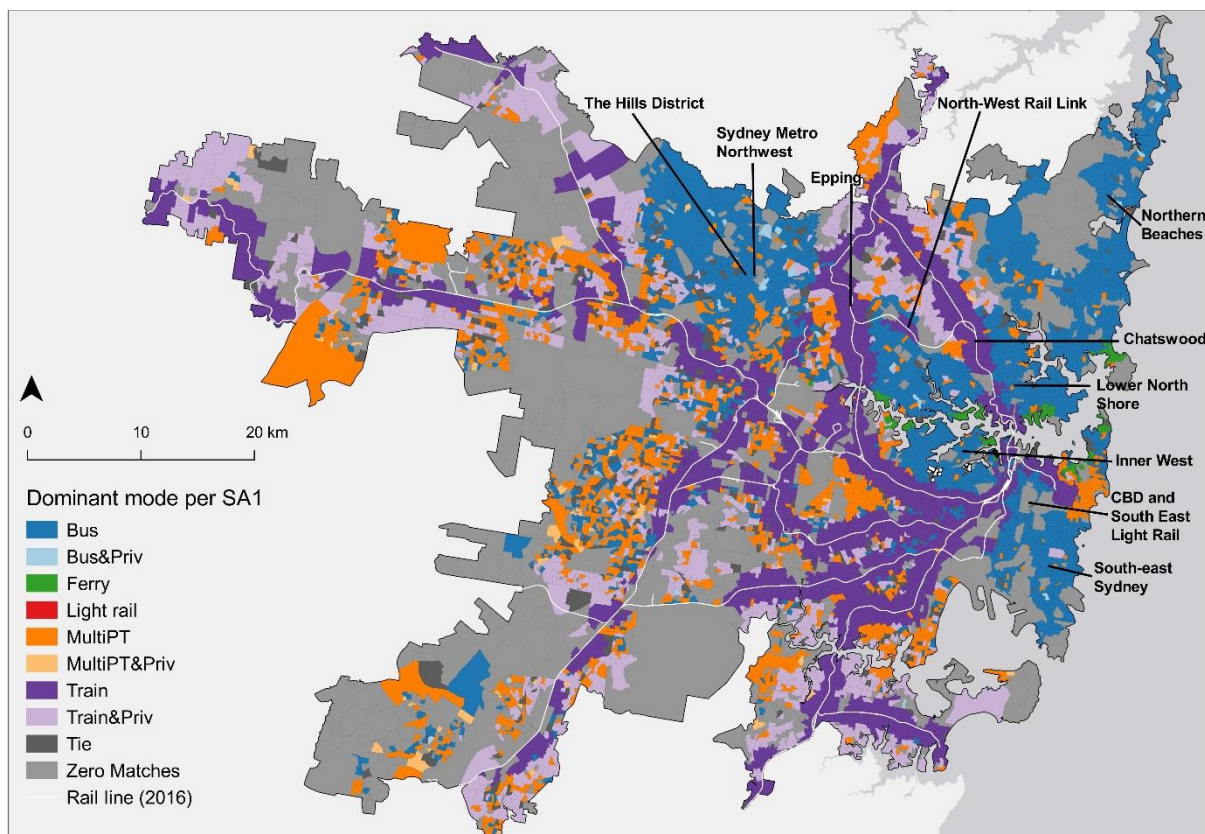


Figure S12: Dominant mode per SA1. Data for Replication 5 (Based on data from: TfNSW, ABS, 2016a, ABS, 2016b, HERE, 2017).

7. Results for the Return Journey

7.1. Public transport commuters

312,352 commuters used public transport for the return journey, on average (Figure S13). Train was the most common mode type, representing 50.1% of all modes used, followed by bus (25.5%).

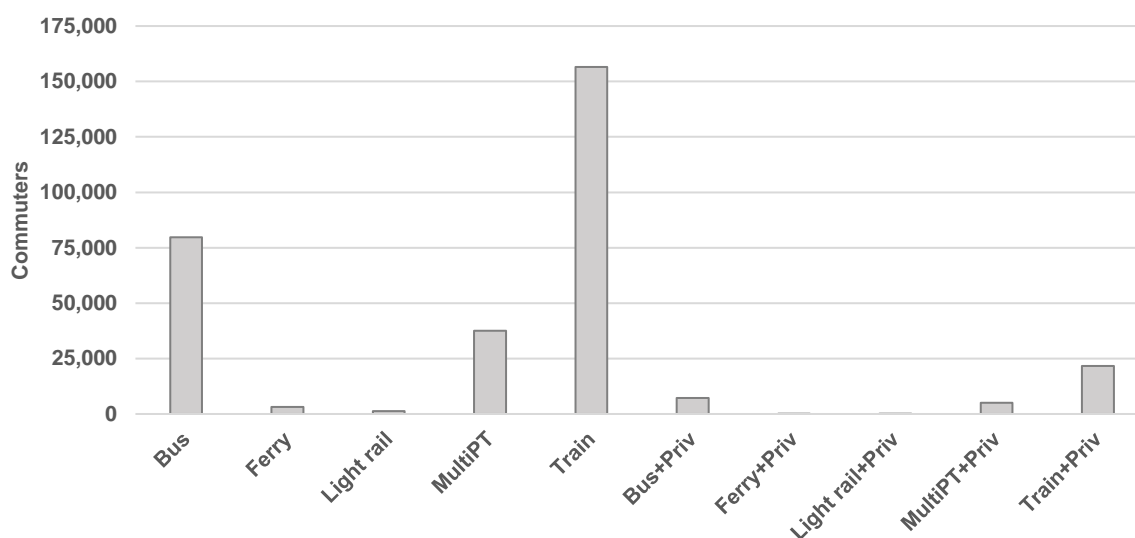


Figure S13: Breakdown of commuters who use public transport for the return journey by mode, showing the mean of ten replications. Commuters that were unable to be assigned with Opal data for the return journey are not included in the totals. Based on ABS data.

7.2. Opal data assignment

For the return journey, over 90% of commuters only using public transport could be reassigned without a change in mode (Figure S14). However, for public transport commuters who also used private modes, this figure was lower (60%), due to more modes being reassigned for such commuters.

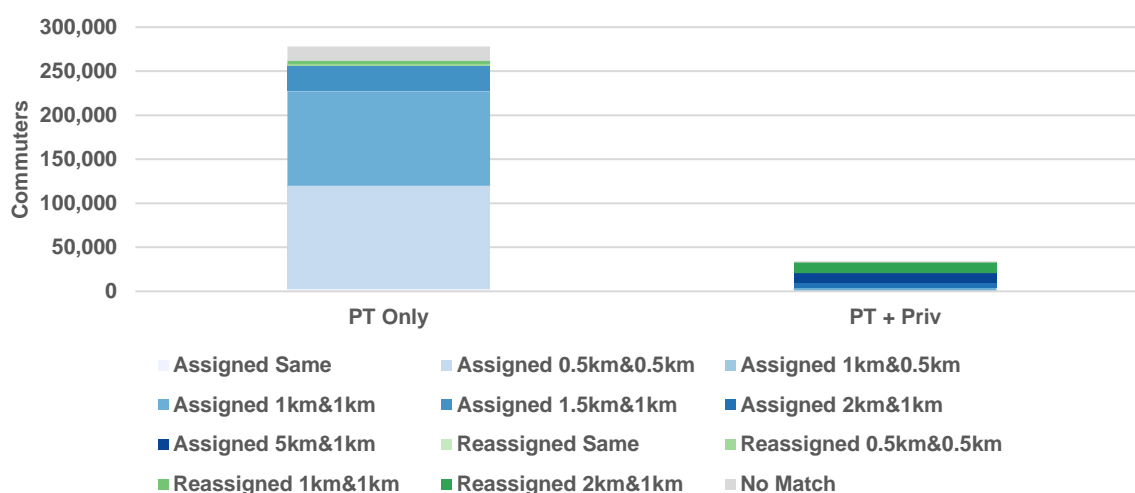


Figure S14: Summary of Opal data assignment for synthetic workers travelling on the return journey. Data for Replication 5. Based on ABS data and TfNSW data.

For public transport only commuters travelling on the return journey, proportionally, the best mode for which the most Opal data assignments were made (without needing to reassign work MBs) was train (Figure S15). When mode reassignments were taken into consideration, however, light rail matched best, closely followed by train.

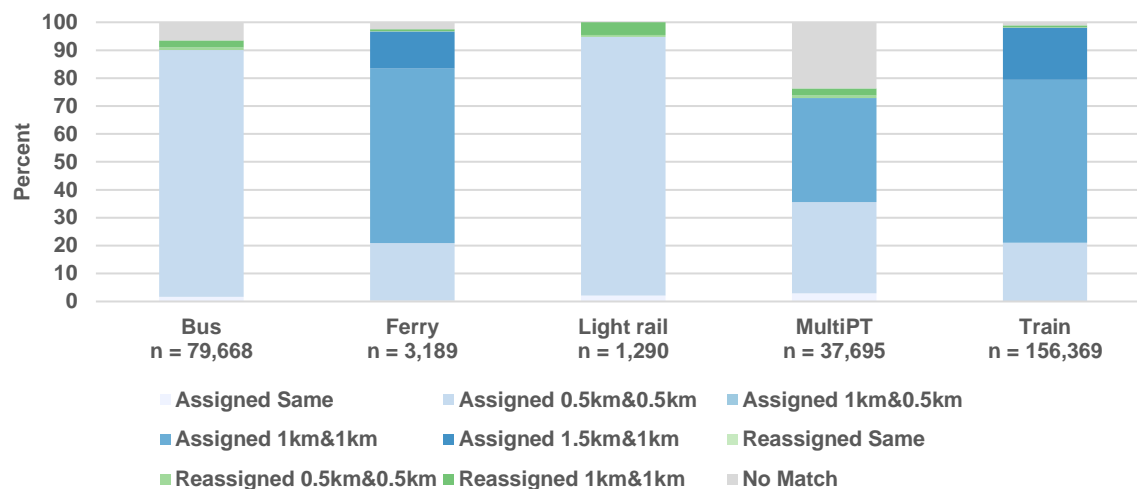


Figure S15: Opal data assignment by mode for synthetic workers travelling on the return journey who use public transport only. For workers whose mode could not be reassigned ("Not Reassigned"), the mode is that used for the journey to work. Data for Replication 5. Based on ABS data and TfNSW data.

For public transport commuters who also used private modes (Figure S16), proportionally, the best mode for which the most assignments were made (without needing to reassign work MBs) was train with private mode(s). When mode reassignments were taken into consideration, light rail with private mode(s) matched best.

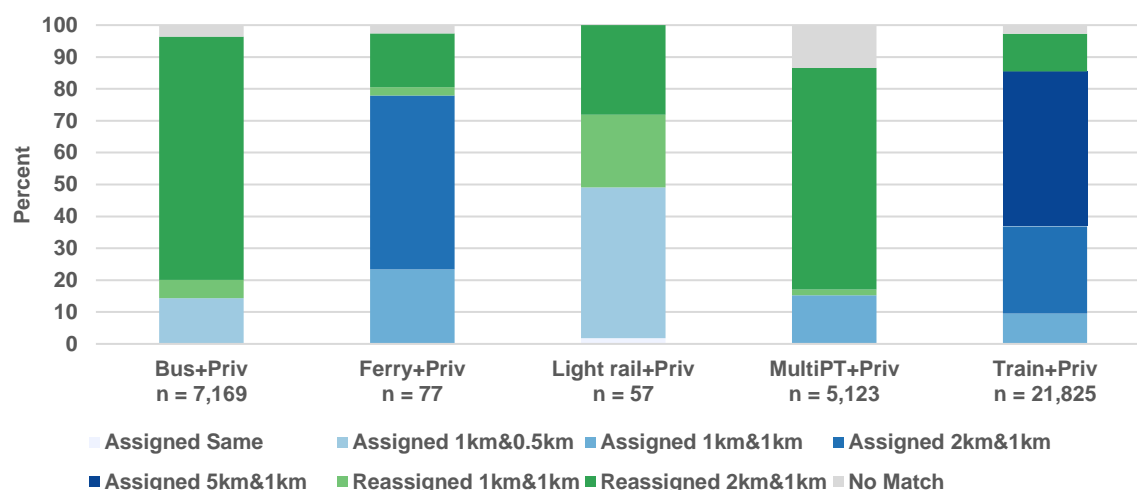


Figure S16: Opal data assignment by mode for synthetic workers travelling on the return journey who use public transport in combination with private transport. For workers whose mode could not be reassigned ("Not Reassigned"), the mode is that used for the journey to work. Data for Replication 5. Based on ABS data and TfNSW data.

For the return journey, 56.0% of SA1s have over 90% of public transport commuters paired with Opal data (Figure S17). These SA1s tended to follow the rail network, as well as in the Eastern Suburbs, Northern Beaches, and parts of The Hills District.

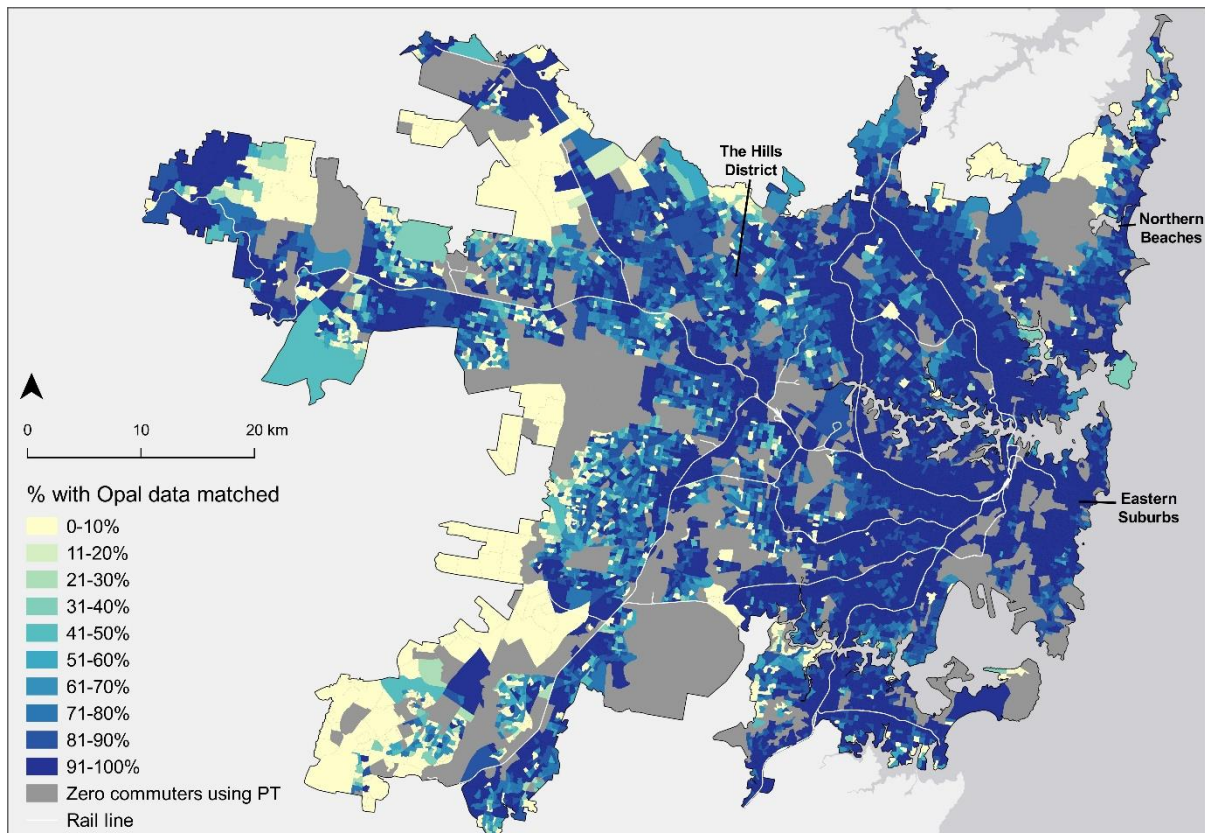


Figure S17: Percentage of commuters using public transport matched with Opal data for the return journey per SA1. Note that the denominator is the number of commuters using public transport matched with Opal data for the journey to work, rather than total commuters using public transport according to the Census (Based on data from: TfNSW, ABS, 2016a, ABS, 2016b, HERE, 2017). Data for mean across 10 replications.

It is to be noted, that for the return journey, a commuter may have alighted at a stop that would make them closer to their home, rather than the stop where they would have likely parked their car near for the journey to work. However, the presence of two or more adults in a household may mean that one commuter drove the car on their way to work, while another picked up the car later in the day.

7.3. Validation

For the return journey, 58.2% of public transport commuters had their first tap-on and last tap-off locations validated based on their home and work location and mode used from TfNSW's Trip Planner API. Specifically, for train commuters, this figure was 81.0% (Table S5).

Table S5: Validation of synthetic commuters who use public transport only for the return journey using TfNSW's Trip Planner API for Replication 5. Note validation for the return journey using the TripPlanner API is not contingent on validation for the journey to work.

Mode	Validated	Total	% Validated
Bus	16,434	74,501	22.1%
Ferry	2134	3111	68.6%
Light rail	489	1290	37.9%
Train	125,267	154,589	81.0%
Total Single Mode Types	144,324	233,491	61.8%
Multiple PT Mode Types	8410	28,756	29.2%
Total All Modes	152,734	262,247	58.2%

Journeys were not considered valid if:

- The final tap-off of a multiple-segment journey for the journey to work is in the same suburb as the home suburb.
- The distance between home and the first tap-on for the journey to work is larger than the distance between home and work.
- Transfer distances for the return journey exceed 1.6km. Since Opal data has a 20-minute threshold between tap-ons and the preceding tap-off before a new journey was created instead of a multiple-segment journey, it was assumed commuters could only travel up to 1.6km on foot within this time frame. This was calculated from a 5-minute walking distance equating to approximately 400m (Levinson, 2020).
- The calculation of travel time without first- and last-miles for the return journey exceeded the door-to-door travel time for the return journey.

It is to be noted that validation for the return journey using the TripPlanner API is not contingent on validation for the journey to work.

7.4. Door-to-door travel time

Of the 152,734 validated commuters using public transport only for the return journey from work, 33,808 recorded a shortest possible travel time of ≤ 30 minutes (22.1%). Of the 6301 SA1s with travel time recorded, 816 SA1s have mean shortest public transport time for the return journey from work that is ≤ 30 minutes (Figure S18). This represents 13.0% of SA1s when SA1s with zero public transport commuters or zero validated commuters are excluded.

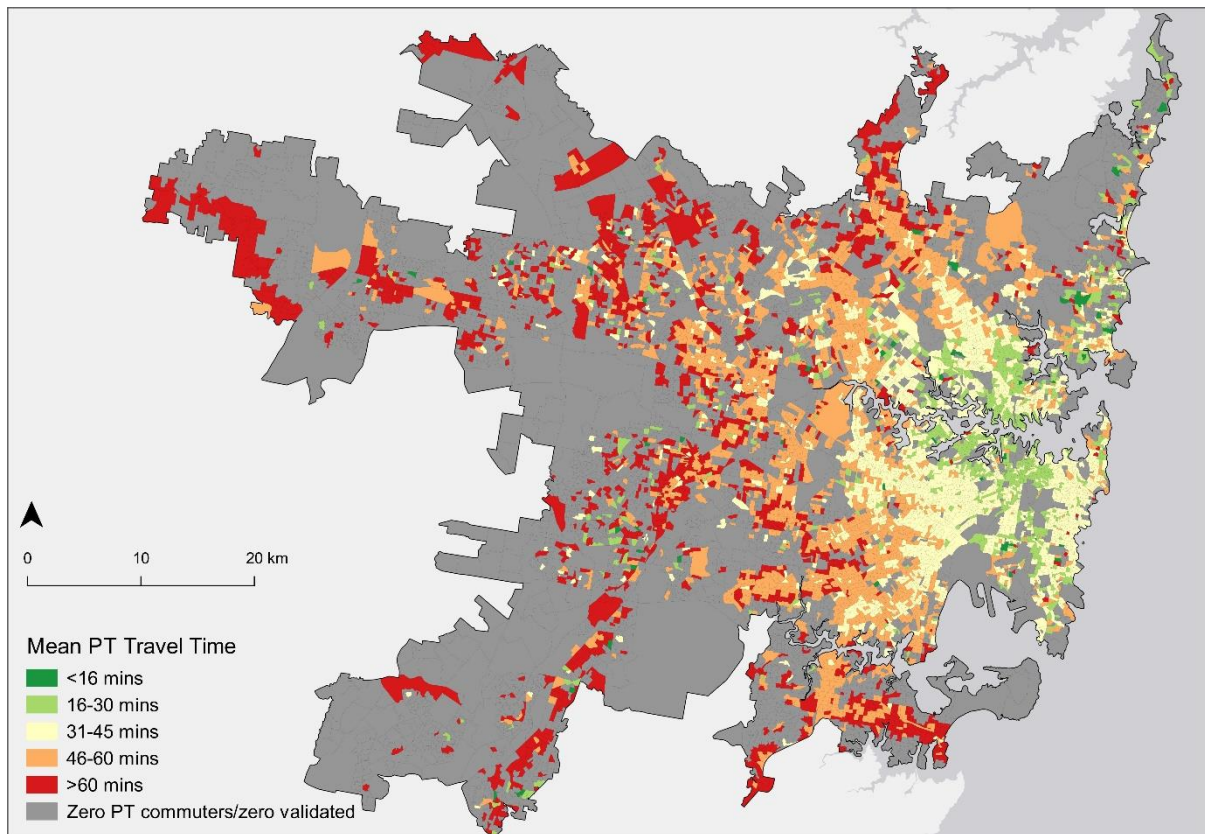


Figure S18: Mean shortest travel time per validated commuter using only public transport modes per SA1 (origin) for the return journey from work. Data for Replication 5 (Based on data from: TfNSW, ABS, 2016a, ABS, 2016b).

7.5. Comparison of door-to-door travel time with travel time without first- and last-miles

According to the analysis without first- and last-miles, 100,654 public transport only commuters recorded a shortest possible travel time of ≤ 30 minutes for the return journey. However, when door-to-door travel time was considered, 66,846 commuters were incorrectly classified as having a public transport commute of ≤ 30 minutes (Table S6), 95.3% of which were different by more than 10 minutes. The analysis without first- and last-miles also misclassified 2662 SA1s as having mean shortest travel time of ≤ 30 minutes (Table S6, Figure S19).

Table S6: Comparison of analysis without first- and last-miles with door-to-door public transport travel time for the return journey from work. Data for Replication 5.

	Analysis without first- and last-miles		Door-to-door analysis	
Commuters with shortest travel time ≤ 30 mins	100,654	65.9% of validated commuters	33,808	22.1% of validated commuters
SA1s with mean shortest travel time ≤ 30 mins	3478	35.2% of total SA1s	816	8.3% of total SA1s

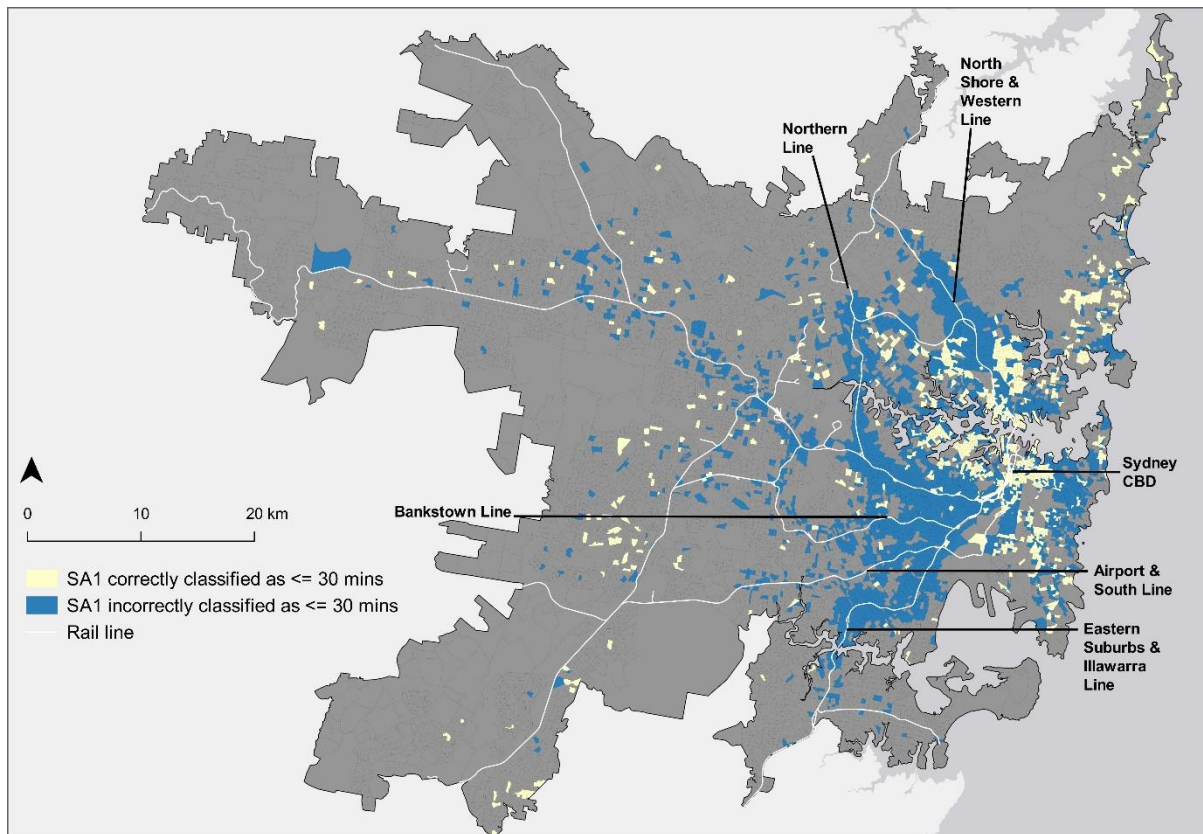


Figure S19: Classification of SA1s (origins) with mean shortest travel time ≤ 30 minutes using the analysis without first- and last-miles for the return journey from work. Data for Replication 5 (Based on data from: TfNSW, ABS, 2016a, ABS, 2016b, HERE, 2017).

8. Household Travel Survey Report Sydney: 2012/13

The Bureau of Transport Statistics (BTS, 2014) conducted a Household Travel Survey (HTS) report for Sydney in 2012/13. Data from Table S7 and Table S8 come directly from the report. These two tables were combined to create Table S9 and Table S10.

Table S7: Number of trips by purpose (average weekday). Data from Table 4.2.1 in the HTS Report (BTS, 2014).

Purpose	2012/13
Social/recreation	4,057,000
Serve passenger	3,067,000
Shopping	2,670,000
Commute	2,528,000
Work related business	1,305,000
Education/childcare	1,554,000
Personal business	927,000
Other	561,000
Total	16,670,000

Table S8: Proportion of trips by mode and purpose (average weekday). Data from Table 4.3.3. in the HTS Report (BTS, 2014).

Purpose	Vehicle Driver	Vehicle Passenger	Train	Bus	Walk	Other	Total
Commute	63.4%	6.6%	14.1%	6.5%	6.4%	2.9%	100%
Work related business	80.5%	6.1%	3.3%	1.4%	6.9%	1.8%	100%
Education/ childcare	6.0%	48.8%	8.7%	18.5%	15.8%	2.2%	100%
Shopping	53.1%	13.1%	2.5%	4.1%	26.2%	1.0%	100%
Personal business	53.0%	19.2%	4.3%	4.8%	17.8%	0.9%	100%
Social/ recreation	36.4%	27.1%	2.8%	2.7%	28.1%	2.9%	100%
Serve passenger	58.5%	29.6%	0.9%	0.8%	9.9%	0.3%	100%

Table S9: Number of trips by mode and purpose. This table was created by multiplying the number of trips in Table 4.2.1 from the HTS Report (Table S7) by the percentages in Table 4.3.3 from the HTS Report (Table S8) for each respective category.

Purpose	Vehicle Driver	Vehicle Passenger	Train	Bus	Walk	Other	Total
Commute	1,602,752	166,848	356,448	164,320	161,792	73,312	2,525,472
Work related business	1,050,525	79,605	43,065	18,270	90,045	23,490	1,305,000
Education/ childcare	93,240	758,352	135,198	287,490	245,532	34,188	1,554,000
Shopping	1,417,770	349,770	66,750	109,470	699,540	26,700	2,670,000
Personal business	491,310	177,984	39,861	44,496	165,006	8,343	927,000
Social/ recreation	1,476,748	1,099,447	113,596	109,539	1,140,017	117,653	4,057,000
Serve passenger	1,794,195	907,832	27,603	24,536	303,633	9,201	3,067,000

Table S10: Number of train and bus trips by mode and purpose. This table adds the train and bus trips together by mode and purpose from Table S9.

Purpose	Train	Bus	Total	Percentage
Commute	356,448	164,320	520,768	33.8%
Work related business	43,065	18,270	61,335	4.0%
Education/childcare	135,198	287,490	422,688	27.4%
Shopping	66,750	109,470	176,220	11.4%
Personal business	39,861	44,496	84,357	5.5%
Social/recreation	113,596	109,539	223,135	14.5%
Serve passenger	27,603	24,536	52,139	3.4%
Total	782,521	758,121	1,540,642	100%

9. References

- ALSGER, A. A., MESBAH, M., FERREIRA, L. & SAFI, H. 2015. Use of Smart Card Fare Data to Estimate Public Transport Origin–Destination Matrix. *Transportation Research Record: Journal of the Transportation Research Board*, 2535, 88-96.
- AUSTRALIAN BUREAU OF STATISTICS 2016a. Australian Statistical Geography Standard (ASGS): Volume 1 - Main Structure and Greater Capital City Statistical Areas, July 2016, datacube: ESRI Shapefile, cat. no. 1270.0.55.001, viewed 27 September 2017, <<http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.001July%202016?OpenDocument>>.
- AUSTRALIAN BUREAU OF STATISTICS 2016b. Australian Statistical Geography Standard (ASGS): Volume 4 - Significant Urban Areas, Urban Centres and Localities, Section of State, July 2016, datacube: ESRI Shapefile, cat. no. 1270.0.55.004, viewed 14 November 2017, <<http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.004July%202016?OpenDocument>>.
- AUSTRALIAN BUREAU OF STATISTICS 2017a. Census of Population and Housing: General Community Profile, Australia, 2016, G23 Relationship in Household by Age by Sex, DataPacks: NSW, cat. no. 2001.0, viewed 22 May 2018, <<https://datapacks.censusdata.abs.gov.au/datapacks/>>.
- AUSTRALIAN BUREAU OF STATISTICS 2017b. Census of Population and Housing: General Community Profile, Australia, 2016, G25 Family Composition, DataPacks: NSW, cat. no. 2001.0, viewed 22 May 2018, <<https://datapacks.censusdata.abs.gov.au/datapacks/>>.
- AUSTRALIAN BUREAU OF STATISTICS 2017c. Census of Population and Housing: General Community Profile, Australia, 2016, G31 Household Composition by Number of Persons Usually Resident, DataPacks: NSW, cat. no. 2001.0, viewed 22 May 2018, <<https://datapacks.censusdata.abs.gov.au/datapacks/>>.
- AUSTRALIAN BUREAU OF STATISTICS 2017d. Census of Population and Housing: General Community Profile, Australia, 2016, G57 Occupation by Age by Sex, DataPacks: NSW, cat. no. 2001.0, viewed 22 May 2018, <<https://datapacks.censusdata.abs.gov.au/datapacks/>>.
- AUSTRALIAN BUREAU OF STATISTICS 2017e. Census of Population and Housing: General Community Profile, Australia, 2016, G59 Method of Travel to Work by Sex, DataPacks: NSW, cat. no. 2001.0, viewed 22 May 2018, <<https://datapacks.censusdata.abs.gov.au/datapacks/>>.
- AUSTRALIAN BUREAU OF STATISTICS 2018a. DZN_POW_by_MTWP_MethodTravelWork (August 2016), TableBuilder. Findings based on use of ABS TableBuilder data.
- AUSTRALIAN BUREAU OF STATISTICS 2018b. SA1_byDZN_POW (August 2016), TableBuilder. Findings based on use of ABS TableBuilder data.
- AUSTRALIAN BUREAU OF STATISTICS 2018c. SA1_byMethodTraveltoWork (August 2016), TableBuilder. Findings based on use of ABS TableBuilder data.
- BALLAS, D., CLARKE, G., DORLING, D., EYRE, H., THOMAS, B. & ROSSITER, D. 2005. SimBritain: a spatial microsimulation approach to population dynamics. *Population, Space and Place*, 11(1), 13-34.
- BARTHÉLEMY, J. & TOINT, P. L. 2013. Synthetic Population Generation Without a Sample. *Transportation Science*, 47(2), 266-279.

- BECKMAN, R. J., BAGGERLY, K. A. & MCKAY, M. D. 1996. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30(6), 415-429.
- BUREAU OF TRANSPORT STATISTICS 2014. Household Travel Survey Report: Sydney 2012/13. Haymarket.
- BURKE, M. & BROWN, A. L. 2007. Distances people walk for transport. *Road and Transport Research*, 16(3), 16-29.
- CHAPUIS, K., TAILLANDIER, P., RENAUD, M. & DROGOUL, A. 2018. Gen*: a generic toolkit to generate spatially explicit synthetic populations. *International Journal of Geographical Information Science*, 32(6), 1194-1210.
- CHOUPANI, A.-A. & MAMDOOHI, A. R. 2016. Population Synthesis Using Iterative Proportional Fitting (IPF): A Review and Future Research. *Transportation Research Procedia*, 17, 223-233.
- DANIELS, R. & MULLEY, C. 2013. Explaining walking distance to public transport: The dominance of public transport supply. *Journal of Transport and Land Use*, 6(2), 5-20.
- DEMING, W. E. & STEPHAN, F. F. 1940. On A Least Squares Adjustment Of A Sampled Frequency Table When The Expected Marginal Totals Are Known. *The Annals of Mathematical Statistics*, 11(4), 427-444.
- GRAPPERON, A. 2016. *Behavioural Approach to Estimation of Smart Card Holders Socio-Demographic Characteristics in a Public Transport System*. Université de Montréal.
- HARDING, A., LLOYD, R., BILL, A. & KING, A. 2004. *Assessing poverty and inequality at a detailed regional level: New advances in spatial microsimulation*. (No. 2004/26). Research Paper, UNU-WIDER, United Nations University (UNU).
- HERE 2017. RailRds, data: Esri Shapefile.
- HUANG, Z. & WILLIAMSON, P. 2001. A Comparison of Synthetic Reconstruction and Combinatorial Optimisation Approaches to the Creation of Small-Area Microdata. Liverpool: Department of Geography, University of Liverpool.
- HUYNH, N., BARTHÉLEMY, J. & PEREZ, P. 2016. A heuristic combinatorial optimisation approach to synthesising a population for agent-based modelling purposes. *Journal of Artificial Societies and Social Simulation*, 19(4): 11.
- HUYNH, N., NAMAZI-RAD, M., PEREZ, P., BERRYMAN, M. J., CHEN, Q. & BARTHÉLEMY, J. 2013. Generating a Synthetic Population in Support of Agent-Based Modeling of Transportation in Sydney. *20th International Congress on Modelling and Simulation*. Adelaide.
- LEAO, S. Z., HUYNH, N., TAYLOR, A., PETTIT, C. & PEREZ, P. 2017. Evolution of a synthetic population and its daily mobility patterns under spatial strategies for urban growth. In: GEERTMAN, S., ALLAN, A., PETTIT, C. & STILLWELL, J. (eds.) *Planning Support Science for Smarter Urban Futures*. Cham, Switzerland: Springer.
- LEVINSON, D. M. 2020. *The 30-Minute City: Designing for Access*, Network Design Lab.
- LOVELACE, R. & BALLAS, D. 2013. 'Truncate, replicate, sample': A method for creating integer weights for spatial microsimulation. *Computers, Environment and Urban Systems*, 41, 1-11.
- MORENO, A. & MOECKEL, R. 2018. Population Synthesis Handling Three Geographical Resolutions. *ISPRS International Journal of Geo-Information*, 7(5), 174-93.
- TANTON, R. 2014. A Review of Spatial Microsimulation Methods. *International Journal of Microsimulation*, 7(1), 4-25.

- TANTON, R., VIDYATTAMA, Y., NEPAL, B. & MCNAMARA, J. 2011. Small area estimation using a reweighting algorithm. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(4), 931-951.
- VIJAYAKUMAR, N., EL-GENEIDY, A. M. & PATTERSON, Z. 2011. Driving to Suburban Rail Stations. *Transportation Research Record: Journal of the Transportation Research Board*, 2219(1), 97-103.
- WILLIAMSON, P., BIRKIN, M. & REES, P. H. 1998. The estimation of population microdata by using data from small area statistics and samples of anonymised records. *Environment and Planning A*, 30(5), 785-816.
- ZHANG, F., ZHAO, J., TIAN, C., XU, C., LIU, X. & RAO, L. 2016. Spatiotemporal Segmentation of Metro Trips Using Smart Card Data. *IEEE Transactions on Vehicular Technology*, 65(3), 1137-1149.
- ZHU, Y., DIAO, M., FERREIRA JR., J. & ZEGRAS, C. 2018. An integrated microsimulation approach to land-use and mobility modeling. *Journal of Transport and Land Use*, 11(1), 633-659.