Heuristics based population synthesis using Australian census data

Bhagya. N. Wickramasinghe

bhagya.wickramasinghe@rmit.edu.au
RMIT University, Australia

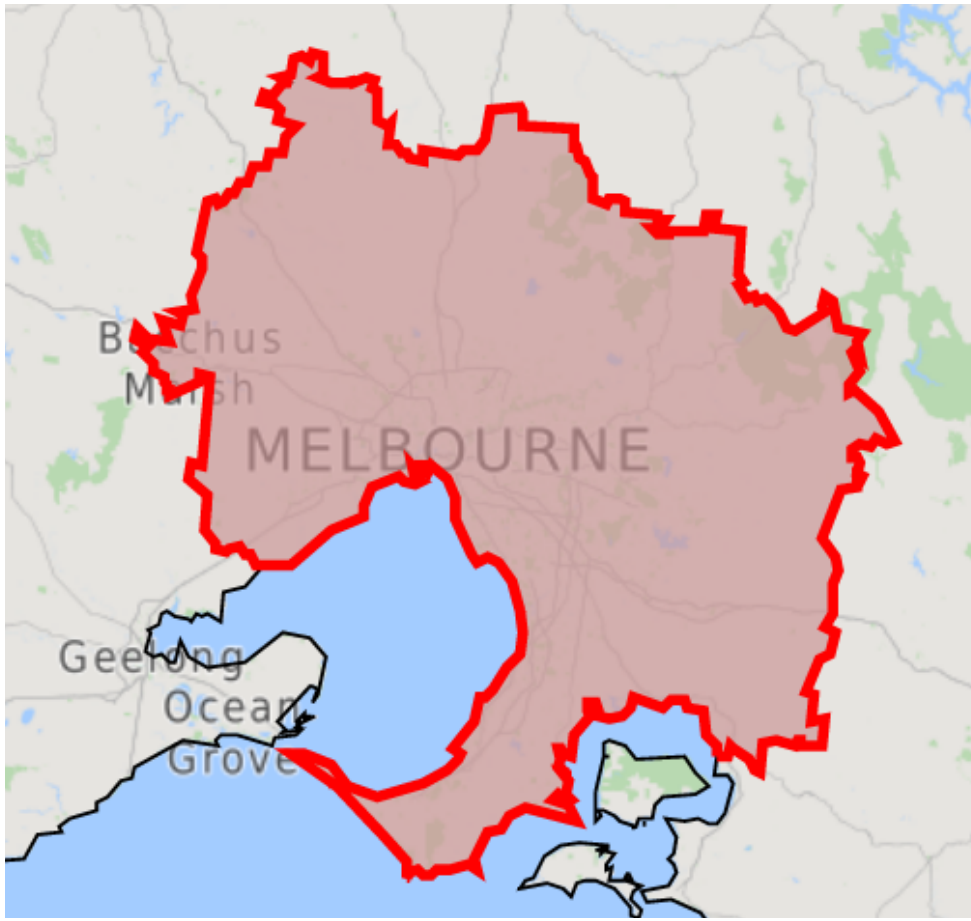
In collaboration with Dhirendra Singh, Lin Padgham and Shane Culpepper

RMIT UNIVERSITY

The Population - Greater Melbourne

Constructing a synthetic agent population, with realistic family household structures, that is consistent with person and household level aggregated census data distributions.

- Australian Statistical Geography Standard (ASGS) specifies 5 granularity levels.
- Statistical Area 2 (SA2) is the 3^{rd} smallest
- Greater Melbourne
  - 309 SA2s
  - 4,529,496 persons
  - 419,693 households



Greater Melbourne Area - www.abs.gov.au

Persons Distribution

Number of persons by
Relationship status $\times$ Sex $\times$ Age range

Relationship status

- Married
- Lone parent
- Dependent under 15 child (U15 Child)
- Dependent student (Student)
- Non-dependent over 15 child (O15 Child)
- Relative
- Group household
- Lone person

Sex

- Male
- Sex

Age range

- 0 – 14
- 15 – 24
- 25 – 39
- 40 – 54
- 55 – 69
- 70 – 84
- 85 – 99
- 100 or over

e.g. `(married, male, 25 - 39) : 430 persons`
`(student, female, 15 - 24) : 687 persons`

Relationship status categories

Custom relationship category	Original ABS categories
Married	Husband, Wife or Partner in de facto marriage, female same-sex couple Husband, Wife or Partner in de facto marriage, male same-sex couple Husband, Wife or Partner in de facto marriage, opposite-sex couple Husband, Wife or Partner in a registered marriage
Lone Parent	Lone parent
Dependent under 15 child	Foster child under 15 Natural or adopted child under 15 Otherwise related child under 15 Step child under 15 Unrelated child under 15 Grandchild under15
Dependent Student	Natural or adopted dependent student Dependent student step child Dependent student foster child
Non-dependent over 15 child	Non-dependent foster child Non-dependent step child Non-dependent natural, or adopted child
Relative	Brother/Sister Cousin Father/mother Grandfather/grandmother Nephew/niece Non-dependent grandchild Other related individual (nec) Uncle/aunt Unrelated individual living in family household
Group household	Group household member
Lone person	Lone person
Ignored	Visitor(from within Australia) Not applicable Other non-classifiable relationship

Family households distribution

Number of households by
Household size $\times$ Family household composition

Household sizes

- 1 person
- 2 persons
- 3 persons
- ...
- ...
- 8 or more persons

Family household compositions

- 1 family: Couple family with no children
- 1 family: Couple family with children
- 1 family: One parent family
- 1 family: Other family
- ...
- 3 family: Couple family with no children
- 3 family: Couple family with children
- 3 family: One parent family
- 3 family: Other family
- Group household
- Lone person household

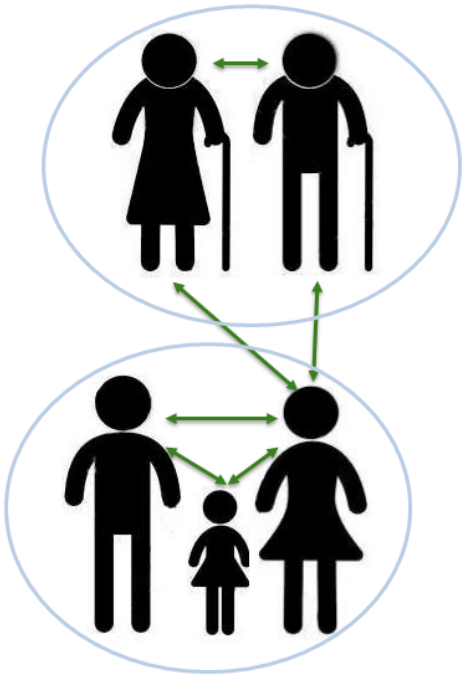
e.g. `(2 persons, 1 family: one parent family) : 30 households`
`(5 persons, 2 family: couple family with children) : 11 households`

Family Types

- Family composition

Couple family with no children	:	Married Male + Married Female + 1..* $\times$ Relative
Couple family with children	:	Married Male + Married Female + 1..* $\times$ child + 1..* $\times$ Relative
Lone parent	:	Lone parent + 1..* $\times$ child + 1..* $\times$ Relative
Other family	:	2..* $\times$ Relative
Group household	:	2..* $\times$ Group household person
Lone person household	:	Lone person

- Priority order of deciding the primary family
  1. The Couple family with children unit in the household
  2. The One parent family in the household
  3. Couple with no children and Other families get similar priority
- Secondary and tertiary families are not explained in data



Data Cleaning

1. Constructs the population taking household distribution as the reference.
2. Set all unrealistic values to 0, e.g. (`male, married, age 0-15`) and (`3 person, lone person household`).
3. If the household and person level `group household` persons are different update person level distribution while preserving age and sex distribution
4. Proportionally update `lone persons` to match `lone person` households.
5. Proportionally update `married males` and/or `females` to match the number of primary family units consisting married couples

Data Cleaning

6. If the number of `married males` and `females` are different, no changes are made as this discrepancy is handled by the algorithm.

7. If the number of `lone parents` is less than the number of `one parent` primary family units, increase the `lone parents` proportionally

8. If there are not enough `children` to form all primary families that have children increase `children` categories proportionally.

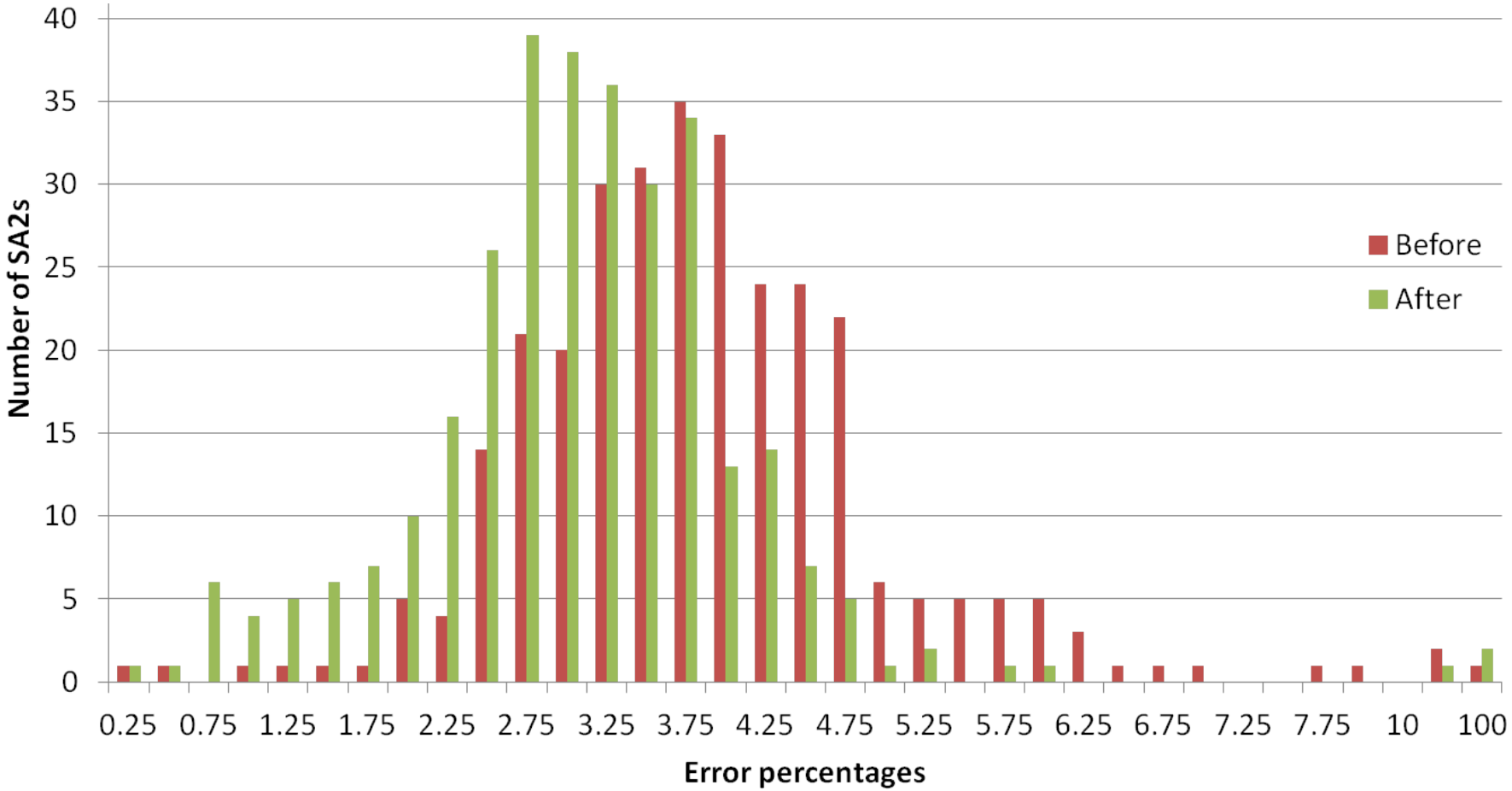
9. If the number of `relatives` is not sufficient to form all primary `other family` units and `1 family: couple with no children` (size > 2) households, increase `relatives` proportionally.

10. If the total number of persons and the number of persons required by households are still different, the discrepancy is handled by the algorithm.

Error percentage histograms before and after cleaning

The error percentage is the difference between the no. of persons in households and the no. of persons in persons distribution as a percentage of the persons in households.

The overall absolute percentage error decreases after cleaning the data.



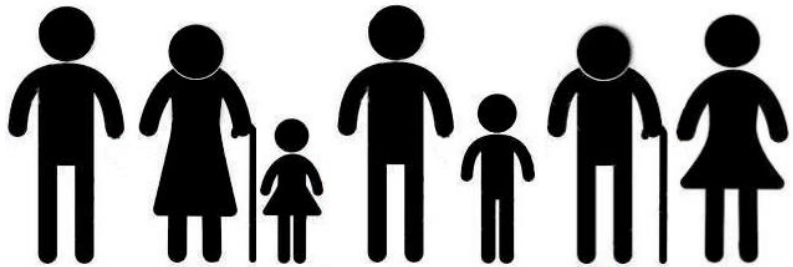
Algorithm - Overview

1. Instantiate all the persons with their characteristics.
2. Form the basic structures for all the inferable families.
3. Instantiate all family households with the primary family, and `group` and `lone person` households with respective person instances.
4. Heuristically add non-primary families to the incomplete family households.
5. Complete households by adding remaining `children` and `relatives`.

Algorithm - Instantiate all persons

1. Instantiate all persons with properties
2. If there are extra persons in household level data than person level data, create a matching number persons without any characteristics.

   Extras = persons in households distribution - persons in persons distribution



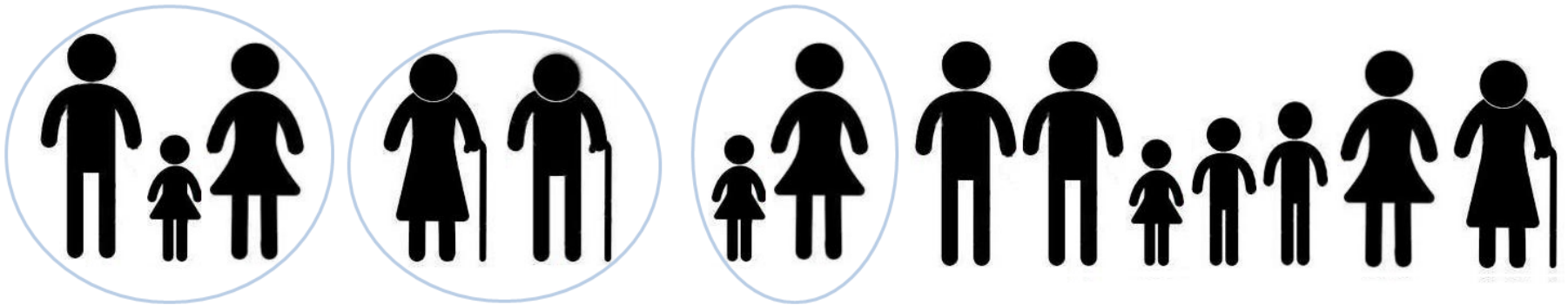
Algorithm - Form all inferable basic family structures

3. Form basic couples by pairing `married males` and `females` in age descending order.

4. Form basic one parent families by pairing each `lone parent` with a `child` from a younger age category.

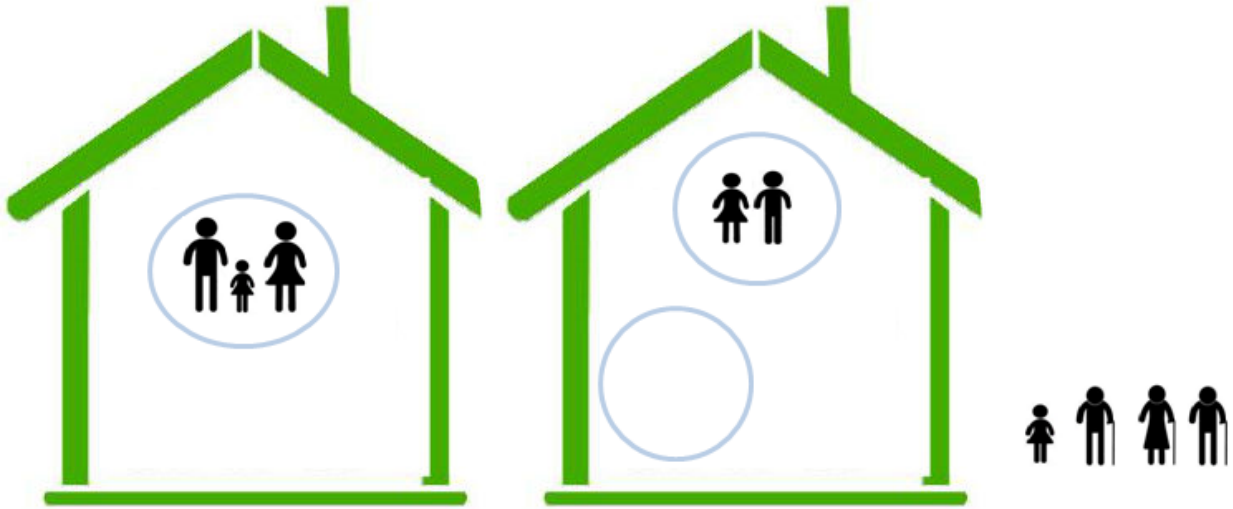
5. Form the primary `couple with children` families shown in household distribution by pairing a random couple with a `child` from a younger age category.

6. Form the primary `other family` units shown in the household distribution by pairing two `relatives`



Algorithm - Instantiate all households

7. Instantiate empty household instances with the expected properties.
8. Add `lone person` instance to `lone person` households.
9. Add `group household` person instances to `group households`.
10. Add previously formed basic family units to suitable household instances as primary families, as per households' primary family type descriptions.

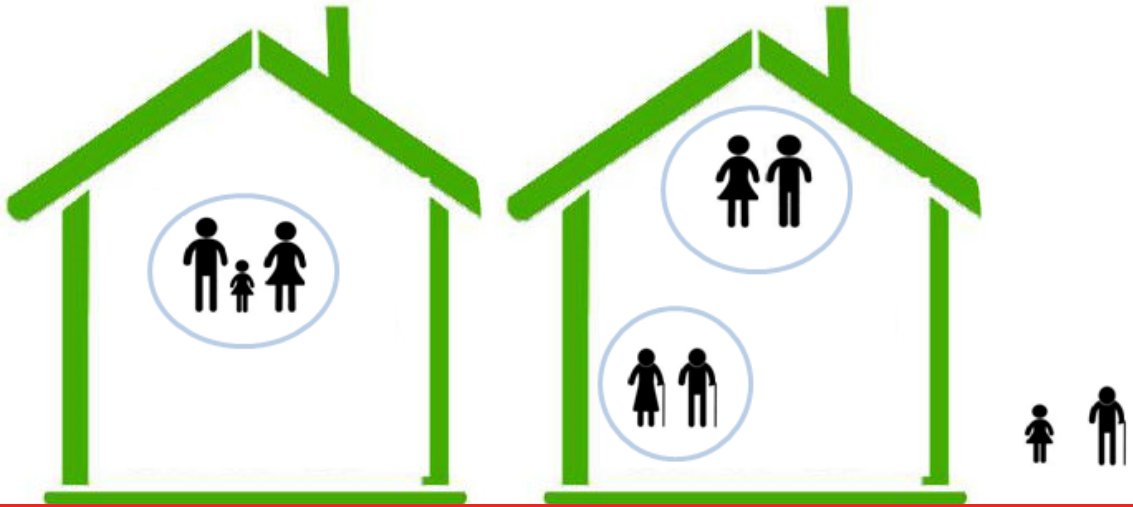


Algorithm - Assign non-primary families to households

11. Add remaining `one parent` basic families to eligible households as non-primary families.

12. Add remaining `couples` to households either as a `couple family with no children` or a `couple family with children` units.
    - The exact family type is determined based on the probability of observing a non-primary `couple family with children` unit in a household.
    - The probability is assumed to be 0.2 (configurable).

13. Add remaining `lone parents` and `couples` to the *extras* pool by nullifying the relationship category.



Algorithm - Assign non-primary families to households

14. Perform next 4 steps if there are households that need more family units.
15. Calculate the probability distribution of `couple`(with and without children), `lone parent` and `other` family units in primary families.
16. Determine potential secondary and/or tertiary family types of a household considering heuristic family priority order explained earlier.
17. Select one of the family types probabilistically based on the distribution calculated in step 15 and add to the household.
18. If the selected family type is `couple` decide whether to add a `couple family with no children` or a `couple family with children` unit as described in step 12.
19. If certain person types have depleted, use the *extras*.

Algorithm - Complete households by adding children and relatives

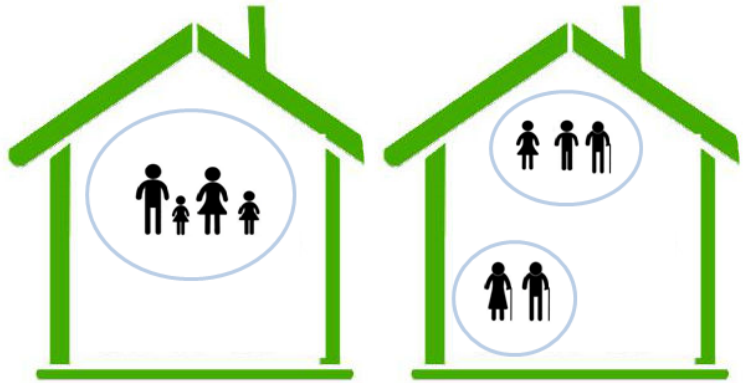
20. At this stage all the households have the required family units but the number of persons may not be complete.

21. Add remaining `children` to eligible families considering parents' ages.

22. Add all remaining unassigned person instances, except for `relatives`, to *extras* by nullifying their relationship status.

23. Probabilistically add all persons in *extras* to families as `children` or `relatives` based on the input persons distribution.

24. Add `relatives` to primary families in remaining incomplete households.



Assigning exact age

Assign the exact ages to person instances based on the percentage of persons in each yearly age category in an SA2.
e.g.

- Brunswick SA2 synthesised population age distribution

Age category	0–14	15–24	..	100+
Persons count	2063	866	..	0

- Brunswick SA2 exact age distribution

Age (year)	0	1	2	3	..	115
Persons percentage	1.12%	0.98%	0.87%	0.86%	..	0%

SA1 level household distribution

The synthesised households are assigned to SA1s based on the SA1 level household types distribution obtained from census data.

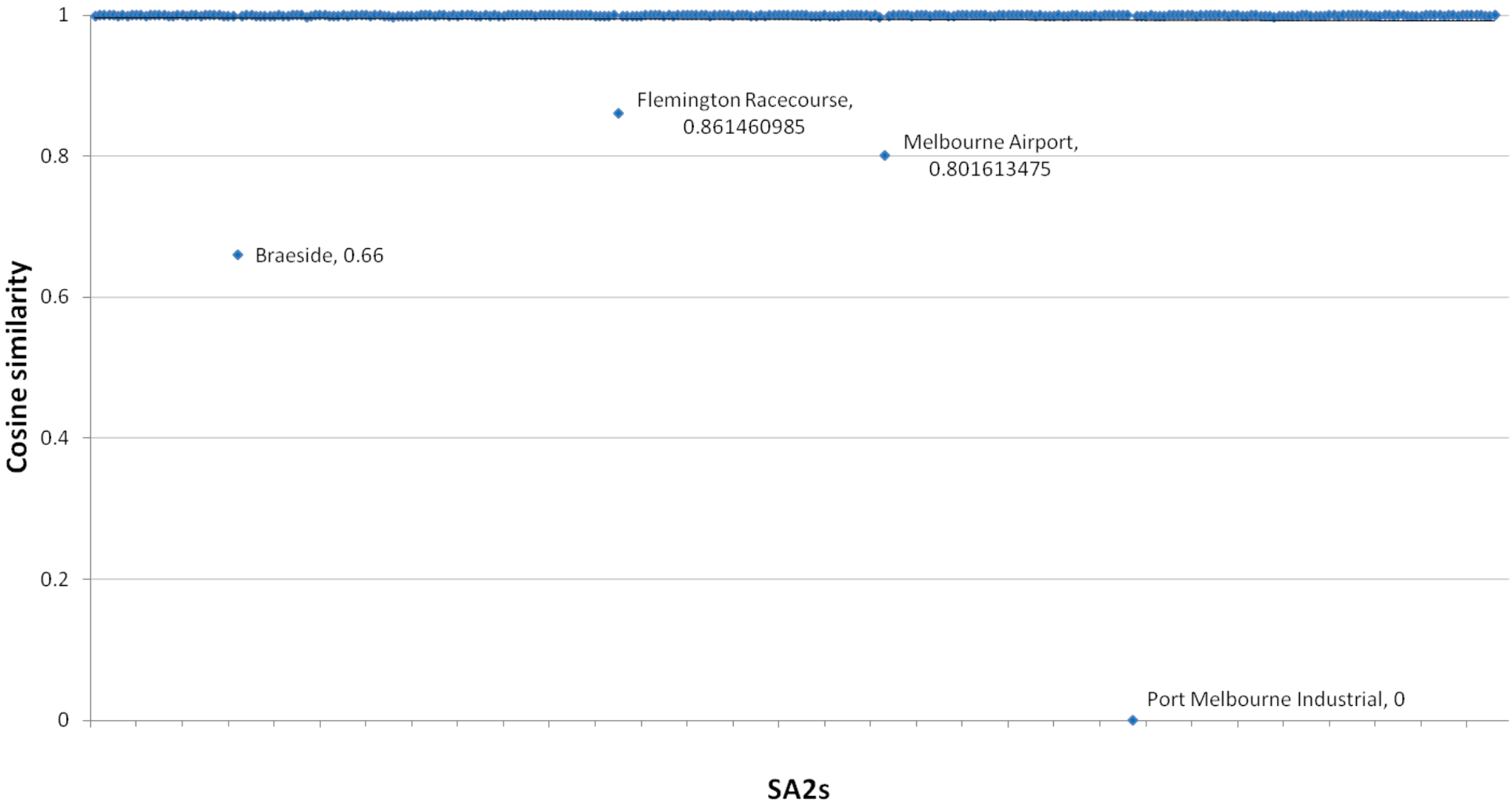
- The households in synthetic population matches the SA1 level household distribution.
- But the persons distribution may not match the SA1 level persons distribution.

Assigning households to addresses

1. The address shape files with geographical locations are obtained from Vicmaps (`www.data.vic.gov.au`)
2. The mesh block polygon shape files are obtained from ABS (`www.abs.gov.au`)
3. Find the ABS mesh block and the SA1 of each address using GIS tools.
4. Assign synthesised households in an SA1 to corresponding addresses.

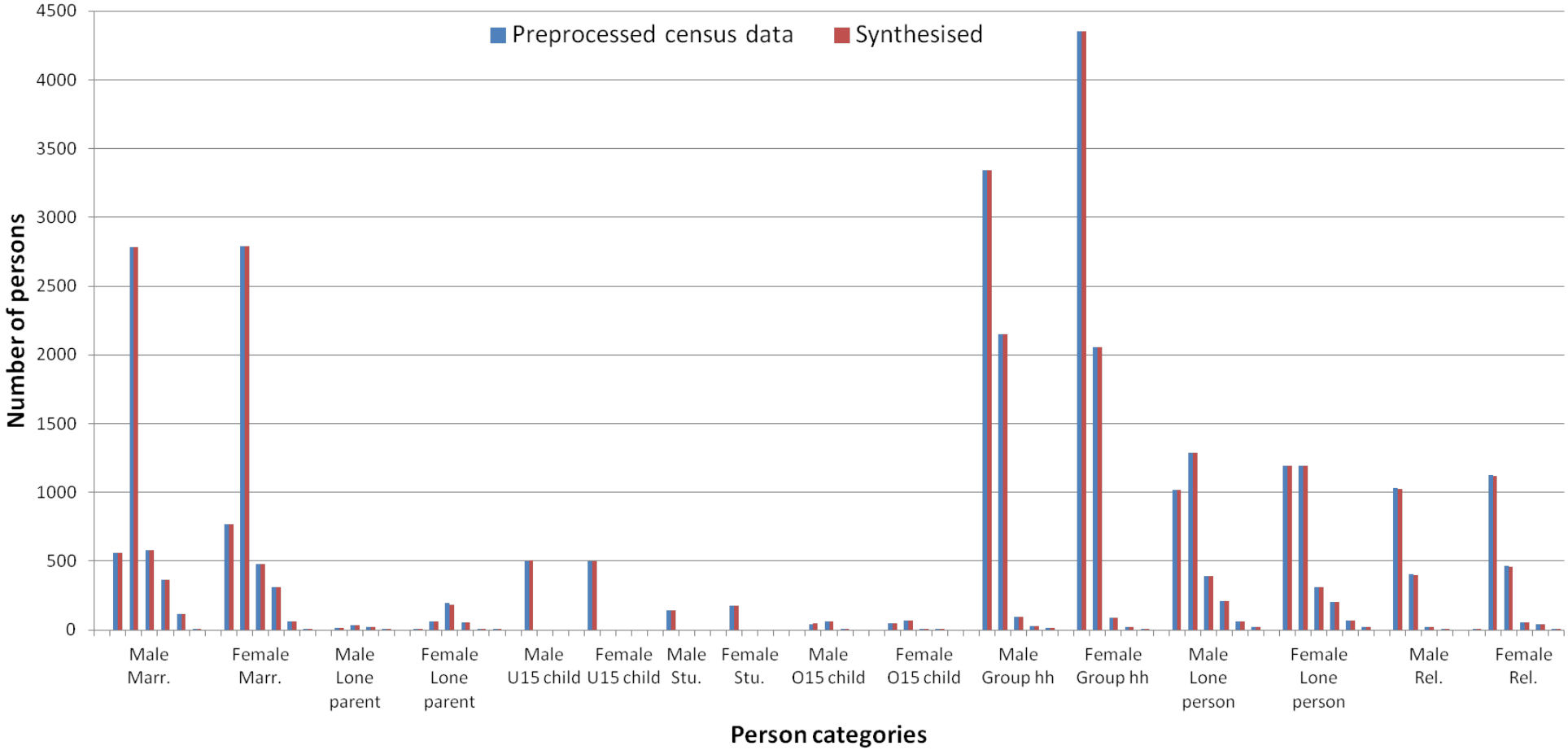
Cosine similarity test results preprocessed census data vs. synthesised

Cosine similarity test results raw census data vs. synthesised



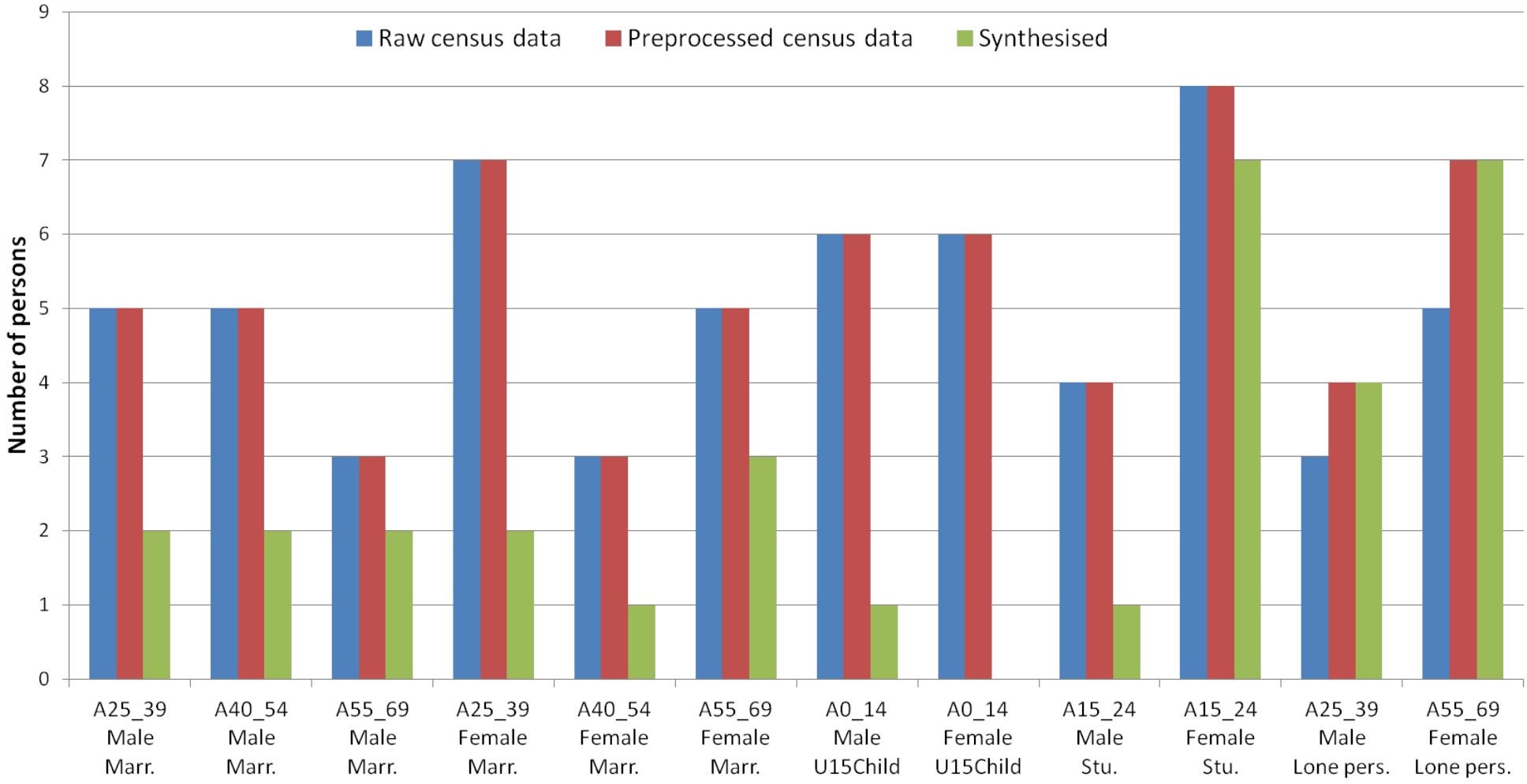
Melbourne SA2 persons distribution - synthesised vs. preprocessed census data

- Total persons:32,046
- Wrongly categorised persons: 52



Person types distribution - Melbourne Airport SA2

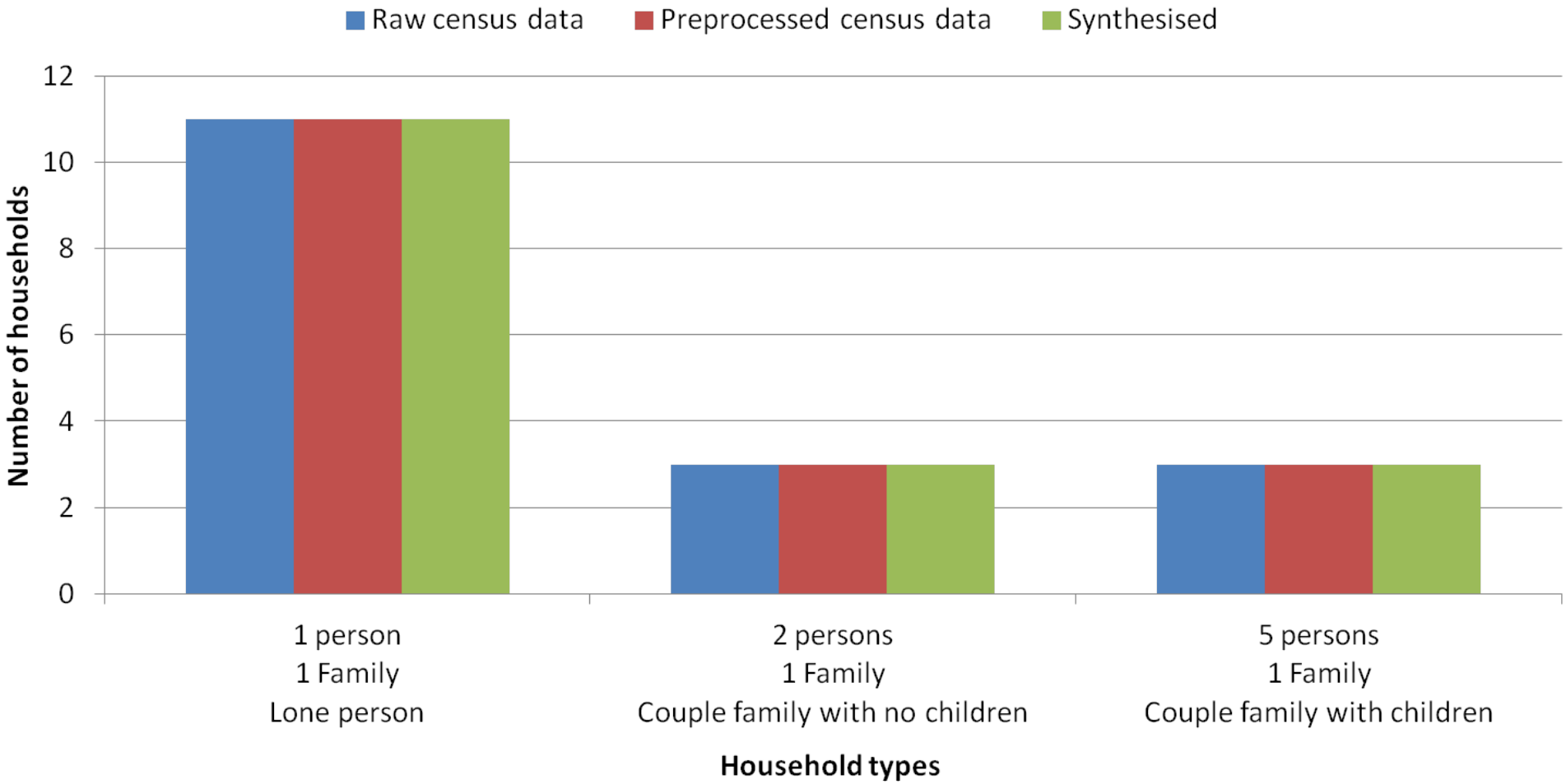
- Persons in input person level data: 63
- Persons in synthetic population: 32
- Absolute error: 31
- Persons in input household distribution: 32



Household types distribution - Melbourne Airport SA2

All households are formed though significant person level discrepancies are observed

- Total households: 17



Thank you and Questions?

Source code

`www.github.com/agentsoz/synthetic-population`