CHAPTER 4

Heuristics Based Population Synthesis Using Different Data Sources

In this section we explore constructing a synthetic population by merging aggregated data from 2016 Australian census. Australian Statistical Geography Standard (ASGS) developed by Australian Bureau Statistics defines a hierarchical system that divides the country into smaller geographical areas[1]. Statistical Area 2 (SA2) is the third smallest area defined in the system and in most cases they correspond to officially gazetted state suburbs and localities. To construct the population we collected individual level and household level marginal distributions of SA2s that fall under Greater Melbourne area, which consists of 309 SA2s. The population consists of 4,485,211 persons in 1,832,043 households[2]. To represent the population in these areas we need to infer the structure of households, explaining the persons that live in them and the relationships among them using the available marginal distributions. The majority of the literature on this matter relies on microdata samples to infer family/household compositions [Beckman et al. 1996, Williamson et al. 1998, Ye et al. 2009]. However, in this case privacy related restrictions are a major deterrent for using them. Because of that, the proposed sample free technique is more desirable when constructing the synthetic population. We describe first the data we use, then the algorithm followed by an analysis of the goodness of the resulting synthetic population.

4.1 Data

4.1.1 Person Level Data

Person level information collected under each SA2 includes joint distributions for the number of persons by age, gender and the person's relationship in a household. There are eight age categories, eight relationship status categories and two gender categories. The table 4.1 gives the list

[1] www.abs.gov.au/ausstats/abs@.nsf/Lookup/2901.0Chapter23102011
[2] www.censusdata.abs.gov.au/census_services/getproduct/census/2016/quickstat/2GMEL

of custom categories used in this exercise under each person level characteristic. A person type is represented by a combination of an age category, a gender category and a relationship status, for example (`Male`, `Married`, `age 25-39`). The person level joint distribution gives the number of persons under each of such 128 ($8 \times 8 \times 2$) person types. The actual relationship categories used by Australian Bureau of Statistics is more detailed than the what is used here. The table 4.2 shows how original categories were aggregated for the purpose of this exercise. A complete description of the special terms, person and household level categories is available in Australian Bureau of Statics website[3].

Table 4.1: Individual level characteristics and categories

Characteristic	**Categories**							
Sex	Male	Female						
Relationship status	Married	Lone parent	Dependent under 15 children	Dependent students	Non-dependent over 15 children	Group household	Lone person	Relative
Age	0-14	15-24	25-39	40-55	56-69	70-84	85-99	100+

In reference to the relationship status categories used here, the married category includes persons in a registered marriage or in a de facto partnership. Though census data includes all the hetero and homo sexual marital partnerships, here all the marital relationships are assumed to be heterosexual partnerships for simplicity. The children in the population are categorised into three categories as dependent under 15 children, dependent students and non-dependent over 15 children. All the children that are 14 years old or younger are categorised as dependent under 15 children. Dependent students are the persons aged between 15 and 24, lives with their parents and enrolled as a full-time student. The children aged 15-24 but not a full-time student and children aged 25 or more are categorised as Non-dependent over 15 children. We use following short forms to refer to children types respectively: `U15 Child`, `Student` and `O15 Child`. `Relative` is short for other related individuals, which encompasses individuals who live in a family household but not part of the family nucleus. `Relative` may also include persons that form family nuclei because of relationships other than marital and parental relationships; for example brothers and sisters living in the same household. `Group Household` persons are individuals living together with other non-related individuals, for example tenants in a shared house, and a `Lone Person` is a person living in alone in a house.

Following directives on determining familial relationships of persons can be derived based on descriptions available in census data. When a person's relationship status is determined, the marital relationship is given a higher priority over parental relationships. For instance, consider person m_1 living with wife w_1 and his mother w_0. Though m_1 has both marital and parental relationships, m_1 is categorised as a `Married` person. Furthermore, m_1 and w_1 are treated as the family nucleus, thus the family is categorised as a couple family and w_0 is categorised as a `Relative` living with the couple because w_0 does not form a family on her own. If m_1 and

[3] www.abs.gov.au/ausstats/abs@.nsf/Previousproducts/2901.0Main%20Features702011

Table 4.2: Custom relationship status categories based on classifications used by Australian Bureau of Statistics

Custom category	**Original 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 Grandchild under15 Natural or adopted child under 15 Otherwise related child under 15 Step child under 15 Unrelated child under 15
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

w_1 have a child (w_2), they are still categorised as `Married` persons, however their family type becomes `Couple family with children`. w_2 can be in either of above three children categories depending on the two parent's age. If the household only consists w_0, m_1 and w_2; where w_0 is the parent of m_1 and m_1 is the parent of w_2, the youngest child's parental relationship is given the priority. Thus w_2 is categorised as a child and m_1 is categorised as a `Lone Parent`, forming a `One parent` family nucleus, and w_0 is again categorised as a `Relative` living with the family.

4.1.2 Household Level Data

The household level information extracted for each SA2 includes the joint distribution of the number of households by household size and family-household composition. Structurally, a household

consists of one or more families and a family consists of at least two persons. A person can only belong to one family. `Lone Person` households and `Group households` are exceptions to this as they do not consist of families.

Here we represent the household types using two characteristics: the number of persons living in a residential dwelling (household size) and family household composition, for example (`4 persons`,
sectionSimilarity AnalysistexttttOne family household: `One parent family`) is a household type. The categories used for household size characteristic range from one to eight persons. The family household composition characteristic includes the number of family units in the household and the type of the primary family unit. For example, `Two family household: Couple family with no children` refers to households with two family units where the primary family unit is a couple family that has no children. The census data used here recognises up to three family units in a household and the type of the primary family. However, it does not identify the types of the secondary and tertiary family units. Thus their family types need to be inferred based on available data and heuristics. The households distribution joining these two characteristics gives the number of households that fall under each of the 112 household types (8 household sizes $\times$ 14 family household compositions) per SA2. Table 4.3 gives the list of family household composition types.

Table 4.3: Household level characteristics and categories

Family household composition categories
One family household: Couple family with no children
One family household: Couple family with children
One family household: One parent family
One family household: Other family
Two family household: Couple family with no children
Two family household: Couple family with children
Two family household: One parent family
Two family household: Other family
Three or more family household: Couple family with no children
Three or more family household: Couple family with children
Three or more family household: One parent family
Three or more family household: Other family
Lone person household
Group household

The family nucleus determining process follows the priority order of relationship categories as described in the above section 5.4.1. If two persons have a marital relationship they are always considered a family unit. For example, if a household consists of two married persons (m_1, w_1) and the married parents (m_0, w_0) of m_1, the household is considered a two family household; the younger couple as one family and the older couple as another family. Here, m_1 is not considered as a member of the older couple's family. In absence of marital relationships, parental relationships

are given prominence, for instance, `One Parent` families. If the family only consists of two single brothers, they are identified as `Relative` persons, thus their family type is categorised as `Other Family`.

The primary family of a household is mainly determined based on the number of children in the families. If there is a family of a couple with children in a household, it is always selected as the primary family. If there are multiple couple families with children, then the family unit with the highest number of children is selected as the primary family. If the number of children is the same, the family with the youngest child is treated as the primary family. The second priority is given to `One Parent` family units and ties are broken in the same way. If there are no children in a multi-family household either a `Couple family with no children` or a `Other Family` can become the primary family depending on the order of the details entered in the census form. When there are `Relatives` that do not form a family of type `Other Family` they are always treated as members of the primary family by the family classification system used by Australian Bureau of Statistics.

4.2 Population Construction

Before constructing the synthetic populations with census data we need to perform several data cleaning routines to minimise the effect of observed data discrepancies. These errors are caused either because of limitations in data collection process or deliberately introduced errors to protect privacy. The data cleaning performed here is based on knowledge of the population and intended to improve the consistency of the individual level data distribution to the household level data distribution. Following is the list of adjustments we made to the population to minimise the errors. However, it is important to note that the data set is not descriptive enough to solve the inconsistencies completely. One such example is inability to know exact number of required married persons due to lack of information on secondary and tertiary family units in multifamily households.

1. Set all unrealistic values to 0, e.g. (`Male`, `Married`, `age 0-15`) and (`3 person`, `Lone person household`).

2. If the number of `Group Household` persons in person level data does not match with the number of persons expected according to household level data, update the person level `Group Household`s distribution proportionally while preserving sex and age distribution.

3. Proportionally update the number of `Lone Person`s in person level distribution to match persons required to form `Lone person` households, if they are different.

4. If there are not enough `Married Male`s or `Female`s to form all primary family units that contain married couples, proportionally increase the number of persons in the relevant categories in the person level distribution to match the required number.

5. If the number of `Married Males` or `Females` is higher than the other, no changes are made as this discrepancy is handled by the algorithm.

6. If the number of `Lone Parent` persons is less than the number of `One parent` family units in households, increase the `Lone Parent` persons proportionally to match the required number of persons.

7. There must be enough children to form `Couple family with children` and `One parent` primary family units with at least one child in them. If not increase the number of persons in children categories proportionally.

8. If there are not enough `Relative` persons to form all primary `Other family` units, increase the number of `Relative` persons proportionally.

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

At a high level the algorithm presented here constructs the population in five stages. The first is to instantiate all the persons instances with their characteristics. Then form the basic structures for all the inferable families with the minimum persons required as per the family type. We can infer that all the `Lone Parents` must form basic `one parent` families with a child because `Lone Parents` cannot be in any other family type, and all the `Married Males` and `Females` must form couples. Further based on the households data distribution we can form basic structures needed for primary `Couple Family with Children` families by pairing a previously formed couple with a child and primary `Other Family` units by pairing two `Relatives`. The third stage is instantiating all the family households with the correct primary family using the basic family instances, `Group Households` with `Group Household` persons and `Lone Person` households with `Lone Person` instances. The forth stage is heuristically adding non-primary families to the incomplete family households. After this step all the family households have required families and some may even have all the required members. Finally, the remaining incomplete households are completed by adding remaining `Children` and `Relatives`.

When forming families we mainly follow two age related heuristics. For couples we assume that the male partner is in the same or one age category older than the female partner. The actual ages are assigned stochastically based on observed age distribution in census data after the population construction while taking care that persons are assigned realistic ages considering other family members. This allows some female partners be older than the male partner if they are in the same age category. For parental relationships, we assume that the age gap between the parent and a child is at least 15 years but not more then 60 years or the child comes from the three immediately younger age categories than the parent's. That is if the parent is in 70-84 age category, the children can be selected either from 15-24, 25-39, 40-55 or 56-69.

As input households distribution does not conclusively describe the secondary and tertiary families, the fourth stage of the households construction process needs to infer them based on

available limited information. Households distribution provides information on the number of persons in a household, the type of the primary family and the number of families in the household. In addition to that the minimum size of family can be inferred based on its family type, that is, a `Couple family with Children` must consist of three members and all other three family types must consist of at least two members. Since a basic family instance of the primary family is already assigned to the household, all the persons in non-primary families must not exceed the remaining number of slots in a household. For instance, a seven person three family household where the primary family type is `Couple family with children` has four remaining positions given basic family instance of the primary family type is already assigned. This indicates that remaining two families can have up to four members, thus the two non-primary families cannot be another `Couple family with children` instance as it requires three members, which would leave room for only one member of the third family. Because of that the two non-primary families have to be selected from `Couple family with no children`, `One parent` or `Other family` types. In this case the household completes with all the required members once the three families are added. Following is list of criteria that needs to be considered when selecting non-primary families.

(a) To add any family type there must be at least one free slot for a new non-primary family.

(b) To add a `Couple family with no children` or a `Other family` units there must be room for at least two more persons.

(c) To add a `One Parent` family there must be room for at least two persons and the primary family must be either `Couple Family with Children` or `One Parent` family.

(d) To add a `Couple Family with Children` unit there must be room for at least three persons and the primary family type must also be `Couple Family with Children`.

.

In situations where there are multiple possible family types for non-primary families, there needs to be a methodology to approximate the most suitable type. Though the distribution of primary family types is known, it is highly unlikely that the non-primary families would follow the same distribution. For example, based on general knowledge, though there is a large proportion of `Couple family with children` primary family units, it is highly unlikely that the same high proportion would be observed in non-primary families. Instead, we argue here that it is reasonable to assume that the distribution of the number of family units by the relationship type between the two main family members commonly represents all the families in the population regardless of the order of families in the households. Thus, the approximation mechanism used here first calculates the probability distribution of primary families by categorising them to *marital, parental* and *other* based on the relationship between the two main members of the family. The three categories respectively correspond to `Couple family with no children` and

`Couple family with children` families combined into one group, `One parent` families and `Other family` units. When there are multiple suitable family types for a non-primary family one of them is selected based on the calculated probability distribution. If the main relationship of a newly added non-primary family is determined to be a *marital* relationship and the primary family type of the household is `Couple Family with Children`, we can either add the new family as a `Couple family with no children` or a `Couple family with children` unit. In such cases we opt one of the two based on a configurable parameter giving the probability of observing a `Couple family with children` unit given the primary family of the household is of the same type, as per population heuristics. If the primary family type is something else the new family is added as a `Couple Family with no Children` unit.

4.2.1 The Algorithm

This section describes the population construction steps in detail. When constructing the population the algorithm places persons, families and households in different pools depending on their characteristics and states. The table 4.4 is a guide to these different pools.

Forming Persons

1. Clean the input data distributions by removing assessable discrepancies based on known population properties as described above.

2. Form all person instances represented in the input person level distribution and set their properties according to the categories they belong to in the distribution. The formed instances are added to separate pools as: *married-males*, *married-females*, *lone parents*, *children*, *relatives*, *group-households* and *lone persons*, corresponding to person properties.

3. If there are more persons represented in household level distribution compared to the individual level distribution, they are instantiated without any characteristics and added to the *extras* pool. These extras are used later in population construction if any person type does not have enough instances to form the required households.

Forming All Known Basic Family Structures

4. Form basic structures of `One parent` families by pairing each person in the *lone parents* pool with a person from the *children* pool in age descending order. The child is always selected conforming to parent-child age heuristic described above. The constructed family structures are put in the *basic one parent* family units pool.

5. Form all the possible married *basic couples* by forming marital relationships between (`Male`, `Married`) and (`Female`, `Married`) persons. The couples are formed by pairing two persons, each taken from the *married-males* and the *married-females* in the age descending

Table 4.4: Person and family instances pools

Pool name	Description
Extras	Characteristics unknown person instances yet unassigned to a family or a household.
Married-males	`Married Male` person instances yet unassigned to a family.
Married-females	`Married Female` person instances yet unassigned to a family.
Lone-parents	`Lone Parent` person instances yet unassigned to a family.
Children	`U15 Child`, `Student` and `O15 Child` person instances yet to be assigned to a family combined into one pool
Relatives	`Relative` person instances yet unassigned to a family
Group households	`Group Household` person instances yet unassigned to a household
Lone-persons	`Lone Person` instances yet unassigned to a household
Basic couples	The pool of married couples, a (`Married`, `Male`) and a (`Married`, `Female`), yet unassigned to households
Basic one parent family units	The pool of family units consisting a `Lone Parent` and a child (`U15 Child`, `Student` or `O15 Child`), yet unassigned to a household.
Basic couple with child family units	The pool of three member family units consisting a married couple and a child, yet unassigned to a household
Basic other family units	The pool of family units consisting two persons of type `Relative`, yet unassigned to a household
Incomplete-households	The pool of partially completed households.
Completed-households	The pool of households completed with all the required family structures and persons instances.

order, until one or both genders deplete. All marital relationships are assumed to be heterosexual. Any remaining persons will be used in later stages.

6. Form the pool of *basic couple with child* family units by grouping a randomly selected unit from the *couples* and a child from the *children* pool, to match the number of households in the input household distribution where the primary family is of type `Couple Family with Child`. The child is selected considering the parent-child age heuristic restrictions on both parent's ages.

7. Form *basic other family* units by grouping two randomly selected `Relatives` to match the number of households where the primary family is of type `Other Family` according to the input household distribution.

Forming Households

8. Instantiate all the households and set their expected household characteristics: household size, the type of the primary family and the number of family units according to the household distribution and add them to the *incomplete-households* pool.

9. Add each individual in the *lone-persons* pool to each `Lone person` households in the input households distribution and add the households to the *completed-households*.

10. Complete the `Group households` by adding all the required person instances from the *group-households* pool. The constructed households are added to the *completed-households* pool.

11. Assign the primary family to each household in *incomplete-households* by selecting a family unit that matches the household's primary family type from one of the basic family unit pools created in steps 5, 6, 4 and 7. This step utilises all the units in the *basic other family* and the *basic couple with child* pools, as these pools only have families sufficient for primary families.

Assigning Non-primary Families

12. Assign remaining units in the *basic one parent families* pool to randomly selected eligible households in the *incomplete-households* pool until the *basic one parent* family units or the eligible households deplete. To be eligible, a household must meet above (a) and (c) conditions, and the same household may be selected more than once as long as it meets eligibility criteria. If a household completes during the process, it is added to the *completed-households* pool.

13. Disassemble any remaining units in the *basic one parent* pool and add the persons to the *children* and the *lone parent* pools accordingly.

14. Add remaining units in *basic couples* pool to randomly selected households in the *incomplete-households* that meet conditions (a), (b) and (d). If a household only meets conditions (a) and (b) the new family is added as a `Couple family with no children` unit. If the household meets conditions (b) and (d) the new family is either assigned as a `Couple Family with Children`, by adding a new child to it, or a `Couple Family with no Children` unit based on the afore mentioned user defined probability. The households that complete during the process are added to the *completed-households*.

15. Any unassigned families in the *basic couples* are disassembled and persons are re-added to the *married-males* and the *married-females* pools accordingly.

16. Calculate the probability distribution of primary family unit having either a *marital, parental* or *other* as the main relationship as described earlier.

17. For every household in the *incomplete-households* that meets condition (a), stochastically select the relationship between the main two persons in the new family using to the probability distribution calculated in step 16.

18. If the selected relationship type is *parental* and the household meets condition (c) determine `One Parent` as the new family type.

19. If the selected relationship type is *other* and the household meets condition (b) determine `Other Family` as the new family type.

20. If the selected relationship type is *marital* and the household meets condition (b) determine `Couple with no Children` as the new family type.

21. If the selected relationship type is *marital* and the household meets condition (d) the type of the new family is decided between `Couple Family with Children` and `Couple Family with no Children` based on the afore mentioned user defined probability.

22. Form a family with the basic structure for the determined family type according to the steps 4– 7. If there are not enough persons of any required relationship status, new persons are drawn from the *extras* pool. The age and sex properties of the new persons are set probabilistically based on the input persons distribution considering population heuristics.

23. Assign the new family to the household. If the household becomes complete add it to *completed-households*.

24. Repeat steps 17 to 23 until all the non-primary families are assigned to all the households.

Completing Households with Children and Relatives

25. Randomly select the households in the *incomplete-households* where the primary family type is `One Parent` family or `Couple Family with Children` and assign them persons from the *children* pool, while adhering to the parent-child age heuristic. The children are added only to the primary family to ensure it has more children than the others. Any household that fill all the persons is added to the *completed-households*.

26. If there are any remaining in the *children* pool, add them to `Couple Family with Children` and `One Parent` secondary and tertiary families considering the parent-child age heuristics without making the updated family larger than the preceding families. Any household that completes all the persons is added to the *completed-households*.

27. Convert all the remaining persons in the *married males*, *married-females*, *lone parents* and *children* to the *extras* pool by nullifying their relationship status but not age and sex categories.

28. Add the persons in the *extras* pool to the primary family of households in the *incomplete-households* until the household reaches its expected size. The properties of the new person is determined probabilistically based on the distribution of `U15 Child`, `Student`, `O15 Child` and `Relative` persons in the person input distribution and considering the population heuristics. If there are person instances in the *extras* that matches the decided age and sex categories use them by setting the correct relationship status, otherwise, all the properties are set to the expected values. Add the completed households to the *completed-households* pool.

29. Complete the population by assigning the `Relatives` to the remaining in the *incomplete-households* pool until each household reaches its expected household size. The relatives are only added to the primary family of the household.

After running above steps for all the SA2s in Greater Melbourne, the population is complete with proper person characteristics, family structures and households, except for exact age of persons. To assign age of the persons we obtain the proportional distribution of persons by their exact age (one year age categories) from census data for each SA2 and use it to determine a person's age stochastically by selecting an age within the persons already known age category. This also considers population heuristics when selecting potential age values for a person to avoid any unrealistic relationships between persons. For example, given a child in 0–14 age category, a parent in 25–39 age category and the population heuristic is the age gap between a parent and a child must be at least 15 years but not more than 45; we first determine potential ages for the parent allowing the child to have at least the youngest age within its age category. In this instance, the youngest age the child can have is 0 years (a new born) thus the parent can have any age between 25–39. Then probabilistically select an age for the parent within its age range, say 27 years. Next determine potential ages for the child based on parents age considering the heuristic and probabilistically assign one of them. Given a 27 years old parent, the oldest the child can be is 12 years, though the child's age category allows up to 14 years. The child is assigned an age between 0 and 12 probabilistically based on age distribution. Here, we only had to consider the upper bound for the child's age because the lower bound is already restricted to 0 years. If that is not the case the child's age must be selected from an interval that does not exceed the 45 years gap compared to the parent's age. This concludes the population synthesis process.