as dealing with spatially situated agents is presented in [Moreira et al. 2009] where Amazonian land change is simulated using two different geographical aggregations. Agent integration related discussion of this work (discussed under spatial coupler) is similar to linking different agent resolutions in multi-resolution modelling.

Another aspect is integrating conceptually similar agents that can only exist in one component simulation at any given time. For example, [Rieck et al. 2010] couple two traffic simulations that capture two distinct geographical areas in Frankfurt, Germany, where a vehicle agent must not exist in two geographical locations at the same time. Another example is work of [Wall et al. 2015] in which they integrate a freight terminal operation simulation (Arena) and VISSIM, the transport simulation. In both these instances, when an agent instance move from one simulation's domain to the other a new agent instance is spawned. This requires maintaining data of agent instances to be able to map the corresponding agent instances. Another instance of mapping different agents is integrating MATSim with UrbanSim in [Nicolai 2013]. In this particular case UrbanSim human agents are represented in MATSim as vehicles. Significant difference in this work is that MATSim takes UrbanSim population as input after UrbanSim has finished executing in a time step and output of MATSim is used as input of UrbanSim in the next time step, where as in previous two cases both simulations were run simultaneously in the same time step. Another slightly different integration technique is presented using GAMA framework where agent instance morphs from one agent species to another when moving from one simulation domain to the other [Grignard et al. 2013]. GAMA facilitates this by allowing to program agents using a high level abstract agent class, which has to be implemented by specialised agent species.

The main focus of above discussed work is runtime integrations of agent populations. All above examples use custom built models and it gives them the freedom of choosing the data and parameters that will be included in the models. This significantly simplifies initialisation and synchronisation of agent populations in integrated models. However, this is not the case when coupling already developed and verified models, which is the objective of this work. Independently developed models are often tuned to use their own set of data, parameters and initialisation sequences for agent populations. Because of that, similarities observed in agent populations in above discussed work cannot be observed in initial agent populations of models used in our work. We cannot allow component simulations to use different initial agent populations because it would make synchronisation extremely difficult and error prone.

We argue that, the initial agent populations used in component simulations need to be consistent through out the integrated simulation. Our view of agent population consistency, relates to conceptually same agent instances having consistent states and their social structures being consistent in every component simulation. Tweaking all component simulation to produce a consistent agent populations is very difficult and not scalable. Thus, we propose constructing the initial agent population of the integrated simulation by merging agent populations of component simulations into one population and using the merged population as the initial population of all the component simulations.

2.1.5 Summary

Building models by reusing existing ABMs is relatively an open area compared to other more mature modelling and simulation techniques. Though there had been several attempts for building integrated ABMs, they are far from being comprehensive integration approaches. These techniques range from naive dismantle and rebuild approaches [Dam and Chappin 2010] to coupling models by data exchange; for example, coupling MATSim and UrbanSim in a feedback loop [Nicolai 2013] and coupling two spatially situated agent based models in a layered approach [Bandini et al. 2002]. Integrating ABMs in a purely feedback loop limits integrated ABM's ability to model concurrent behaviours of agents in the two models, like simulating agents competing for a shared resource. Feedback loop based architecture would let the first simulation use the shared resource freely and give the agents in the second simulation an already updated view of the shared resource, which would always constrain second simulation agents' behaviours based on first simulation. [Bandini et al. 2002]'s work is focused on synchronising spatially situated agents in component models. They treat component models as different layers of the same geographical space and conceptually same agents are kept in sync by communicating state changes among them. However, this work assumes that component models are architecturally homogeneous. Agent synchronisation issues are also explored in work on distributing existing ABMs to enhance performance [Wang et al. 2004, Minson and Theodoropoulos 2008]. However, this work is different from integrating existing ABMs.

Composability research covers both technical and semantic integration of simulation models [Petty and Weisel 2003]. Their work discuss whether models can be integrated meaningfully in addition to just technical feasibility. They define technical feasibility (or *interoperability*) as the ability to exchange data and services at runtime. Furthermore, interoperability of two models is required but not sufficient for them to be deemed as composable. According to [Teo and Szabo 2008], for two simulations to be semantically compatible, data passed between them have to be meaningful to both parties and exchanged data should not prevent simulations from producing valid outputs. Several groups including [Xiang et al. 2004] have worked on composability of simulation models over the years, however they have not discussed ABM specific issues in particular.

2.2 Synthetic Population Generation

Merging agent populations of different MASs is a relatively less researched area. The work on MAS in the literature that require consistent agent populations often have inherently consistent agent populations or assume that initial populations are already consistent. For example, spatially situated agents in Amazonian land use models are inherently consistent because they represent the same geographical units [Moreira et al. 2009], MATSim and UrbanSim integration for Puget Sound region, Washington, USA opt to use UrbanSim population as it is in MATSim [Nicolai 2013] and Muti-Resolution Modelling establish links between similar agent instances at runtime capitalising on existing similarities of agents [Navarro et al. 2011]. To the best of our knowledge there are no examples for merging different agent populations for ABM integration; however, there is a vast body of literature on synthetic population generation (or spatial microsimulation), which is very similar to what we try to achieve. An important distinction between that work and the application of this project is that synthetic population generation is usually applied on data of the same human population but taken from different sources, as opposed to constructing a consistent agent population based on data from different agent populations. In the latter case, there is no guarantee that there will be a common merged population that satisfies all the qualities of input agent populations.

Synthetic population generation generally estimates the population based on known aggregated distributions of agent attributes and a microdata sample (disaggregated data) taken from the target population. Microdata sample can be fine grained information collected on individuals during a survey; or, in a MAS, it can be detailed attributes of a randomly selected set of agents. The two main variants of synthetic population generation techniques are synthetic reconstruction and combinatorial optimisation. The basic approach of synthetic reconstruction is to first produce joint distributions (contingency table) of target agent attributes and then construct the population based on the joint distributions. The most common approach for producing joint distributions is employing Iterative Proportional Fitting (IPF) procedure [Deming and Stephan 1940]. We will discuss further on IPF based approaches and its variants in next subsection. Most of the combinatorial optimisation based methods utilise available household level microdata to generate the synthetic population. Combinatorial optimisation methods first produces an initial estimate of the population by replicating a subset of households from household level microdata. Then the initial estimate is improved by swapping random households with households from rest of the household microdata based on a suitable error minimisation function. Literature on combinatorial optimisation is discussed in detail in subsection 2.2.2. Apart from the most common approaches there are few novel techniques based on K-Nearest Neighbour method and copula theory, which will be discussed in subsection 2.2.3.

2.2.1 Iterative Proportional Fitting (IPF) based methods

IPF is one of the earliest techniques proposed to overcome limitations of census data. Usually census data comes as tables that give the distribution of a certain attribute, for example, distribution of number of persons by age categories or distribution of number of persons by income categories. To generate a population that capture persons by both age and income, we need to construct a joint distribution to find how many persons are in a given age-income range. IPF procedure [Deming and Stephan 1940] is one of the most tested solutions used for this problem in synthetic population generation [Fienberg 1970, Beckman et al. 1996, Auld and Mohammadian 2010, Lovelace et al. 2015].

The process for merging two data sets using IPF starts with an empty two way contingency table, two aggregated distributions (income and age distributions) and a seed, i.e. data of a sample of persons in the population (disaggregated data or microdata, in technical terms). The sample must have the values of attributes used in aggregated distributions recorded, for example, each person in sample data must have their income and age recorded. IPF uses the two aggregated distributions as row and column marginal totals of the contingency table and microdata summary to populate the table cells. In this case microdata act as an initial estimate of the joint distribution. The final step is applying IPF on the contingency table which re-weights the cell values to match marginal totals. The final cell values are the joint distribution of selected two individual distributions, i.e. it gives the number of persons by age and income. Though above example is only limited to two marginals, IPF procedure can be used with multiple marginal distributions [Pritchard and Miller 2012, Namazi-Rad et al. 2014a]. In addition to that "mipf" R statistical package provides functionalities needed for multi-dimensional IPF[2].

IPF based household construction with individual level IPF joint distributions requires household level microdata to obtain information on household compositions. Household microdata are usually available in census data (e.g. CURFs in Australia and PUMA in the USA & Canada). The general approach is to select households from microdata and add them to the synthetic population based on a weighting function. This approach was first proposed by [Beckman et al. 1996]. Weighted probability of a household type being in the synthetic population is calculated based on IPF weights assigned to each agent type in the household. Synthetic population is constructed by sampling households from microdata according to the weighted probabilities of the households [Auld and Mohammadian 2010]. The main problem of this approach is sampling from microdata may limit the heterogeneity in the population because same households gets replicated in every selection.

Though IPF is a widely used method, it is not without limitations. One of them is called zero cell problem. The problem is caused by the seed used in the contingency table not being a comprehensive representation of the underlying population thus setting some cells to zero although corresponding marginal values contradicts that. Final result of the cell in this case is zero, an erroneous value. The solutions proposed for this problem include: assigning very small non-zero values [Beckman et al. 1996], adjusting attribute categorisation to avoid zero cell values [Guo and Bhat 2007] and calculating the values that should be there based on other available data [Lovelace et al. 2015]. An interesting observation made in latter work is that accuracy of the seed has very little influence on final synthetic population. Integerisation (rounding), on the other hand, has a significant negative influence on final result, especially when it comes to smaller categories [Lovelace et al. 2015].

Dependency on microdata is a major limitation in IPF procedure based synthetic population generation approaches. Microdata are often expensive, because of needing door to door surveying, or hard to obtain due to privacy concerns. An example for a method that does not rely on microdata is given in [Barthelemy and Toint 2013]. They proposed producing joint individual distributions using IPF procedure (presumably using 1s and 0s instead of microdata sample) and household distribution using a method based on entropy maximisation and Tabu search. Then households

[2] https://cran.r-project.org/web/packages/mipfp/ index.html

are selected based on an error minimisation function and individuals are grouped into them based on heuristics on household composition. Another important problem in producing synthetic populations with IPF is its inability to match household and individual level distributions at the same time. Work discussed in [Ye et al. 2009] proposes new Iterative Proportional Update (IPU), which matches both household level and individual level joint distributions at the same time.

Dependency on microdata is one of the main obstacles that IPF based population merging has to overcome when integrating MASs. Though modellers may have access to agent populations of component simulations, each population is different in composition and in attributes. So, a sample from one simulation will not be a correct representation of the other simulations. The second problem is constructing social structures. This is not just forming groups; it should construct realistic relationship links among agents in households.

2.2.2 Combinatorial Optimisation (CO) based methods

CO based methods are mainly used to construct small area populations. CO starts with an initial set of households/individuals chosen randomly from household/individual microdata. Then it iteratively swaps already selected households with ones from microdata if the swap improves the accuracy. This process continues until the desired level of fitness is achieved or maximum number of tries are exhausted. Williamson et. al. showed that simulated annealing works better than hill climbing and genetic algorithms in one of the earliest applications of CO for synthetic population generation [Williamson et al. 1998]. The reason being that latter two having a tendency to get stuck in a local optima. The fitness measure they used for CO is Total Absolute Error (TAE), which is calculated by summing errors of all households/individual types. Since its inception CO based population synthesis has been applied in various contexts, including in Flexible Modelling Framework [Harland 2013] and producing synthetic population for Wollongong, NSW in Australia [Namazi-Rad et al. 2014a].

CO based methods use microdata to produce realistic new samples during the iterative optimisation process. For example, in household construction, households drawn from microdata are actual compositions observed in the real population thus researchers do not have to worry about accuracy of household compositions. However, this reliance on microdata is a hindrance in agent population merging because of previously explained differences in source populations. Some work in the literature propose alternatives that use heuristics for constructing household structures [Huynh et al. 2013; 2016]. However, their heuristics are application specific and will not work with different data sets.

2.2.3 Alternative Methods

Generalised Regression: GREGWT is a generalise regression based deterministic re-weighting technique developed by Australian Bureau of Statistics. The algorithm was first published in [Singh and Mohl 1996] and used for small area synthetic population estimations. It also starts with a sample of microdata with initial weight estimates assigned. Algorithm iteratively (Newton-Raphson method) updates the weights to minimise chi-square distance (X^2 distance) between the benchmark tables (marginal distributions) and weight estimates [Tanton et al. 2011].

Markov Chain Monte Carlo (MCMC) based approach: This approach takes the view that joint marginal distribution of individuals/households is the cumulative result (*equilibrium distribution*) of a series of stateless conditional probabilistic selections (a Markov Chain) of individuals/households. In other words, joint marginal distribution is the equilibrium distribution of a Markov Chain. Based on that, if we are given the joint marginal distribution we can construct the population by finding its Markov Chain. There are various random (Monte Carlo) sampling techniques that facilitate simulating random draws from distributions, i.e. reconstructing a Markov Chain that leads to a equilibrium distribution. Gibbs sampler is one of such MCMC based approach used in synthetic population generation [Farooq et al. 2013]. The main advantage of this approach is not loosing heterogeneity of the synthetic population due to microdata cloning approaches used by IPF and CO.

Latest work on population synthesis by various groups have proposed several alternative methods other than the ones described above. Bayesian network based approach proposed by [Sun and Erath 2015] produces better results than IPF and MCMC based approaches, specially when microdata sample is small. It also has lesser data demand than MCMC. Another approach is copula based population synthesis which is similar to IPF in performance but without the limitations related to zero cell or convergence in IPF. Future work related to copula based approaches may focus on sample free techniques [Jeong et al. 2016]. Another alternative to IPF and IPU is K-Nearest Neighbour crossover kernel based population synthesis program proposed in [Hamada et al. 2015], which, unlike IPF based methods, does not loose the heterogeneity of the population. Though these new approaches are promising alternatives, IPF still has the upper hand of being the most tested and applied methodology for constructing synthetic populations.

2.2.4 Goodness of Fit Measures

Goodness of fit measures are used to validate the synthetic population against known distributions of the actual population. The general objective of a statistical test is evaluating whether observed characteristics are statistically significant or not. The most used goodness of fit measures in synthetic population construction are Pearson's Chi-squared test (e.g. [Ye et al. 2009, Lenormand and Deffuant 2013]), Freeman-Tukey (e.g. [Auld and Mohammadian 2010, Barthelemy and Toint 2013, Huynh et al. 2016]), Total Absolute Error (e.g. [Williamson et al. 1998, Smith et al. 2009]) and Z-statistic (e.g. [Williamson et al. 1998, Voas and Williamson 2001, Lovelace et al. 2015]).

Though the ideal goodness of fit measure depends on the application, few guidelines can be drawn based on published work. Z-statistic and Total Absolute Error (TAE) are influenced by application's total population size, while a Standardised Absolute Error (SAE) provides the ability to cross compare different applications [Lovelace et al. 2015]. TAE and SAE also have the advantage of being simple methodologies. However, Z-statistic is a more favourable goodness of fit measure considering its wide community acceptance, ability to produces results corresponding to our intuitive sense of fit, ease of calculation and ability to perform comparisons among different data tables [Voas and Williamson 2001]. They also claim the standard Pearson Chi-squared test as a close alternative.

CHAPTER 3

Merging agent populations from different models

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