ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/355191038

Creating spatially-detailed heterogeneous synthetic populations for agent-based microsimulation

Article *in* Computers Environment and Urban Systems · October 2021

DOI: 10.1016/j.compenvurbsys.2021.101717

CITATION
1

READS
117

4 authors:

Meng Zhou
Singapore-MIT Alliance for Research and Technology
13 PUBLICATIONS **396** CITATIONS
SEE PROFILE

Jason Li
Massachusetts Institute of Technology
1 PUBLICATION **1** CITATION
SEE PROFILE

Rounaq Basu
Massachusetts Institute of Technology
36 PUBLICATIONS **250** CITATIONS
SEE PROFILE

Joseph Ferreira
Massachusetts Institute of Technology
77 PUBLICATIONS **2,484** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Energy Transportation and Logistics View project

Informal and Public Transportation View project

All content following this page was uploaded by Rounaq Basu on 13 October 2021.

The user has requested enhancement of the downloaded file.

CREATING SPATIALLY-DETAILED HETEROGENEOUS SYNTHETIC POPULATIONS FOR AGENT-BASED MICROSIMULATION

A PREPRINT

Meng Zhou[1, 2, a], **Jason Li**[3], **Rounaq Basu**[3], and **Joseph Ferreira**[2, 3]

[1]School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen, Guangdong 518107, China
[2]Future Urban Mobility IRG, Singapore-MIT Alliance for Research and Technology, 138602, Singapore
[3]Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[a]Corresponding author; Email: zhoum89@mail.sysu.edu.cn

October 13, 2021

ABSTRACT

Agent-based models (ABMs) of urban systems have grown in popularity and complexity due to the widespread availability of high-performance computing resources and large data storage capabilities. Credible synthetic populations are crucial for the application of ABMs to understand urban phenomena. Although several (agent) population synthesis methods have been suggested over the years, the spatial dimension of synthetic populations has not received as much attention. This study addresses this myopic treatment of synthetic populations by creating two distinct components - *agents* and the *built environment* - that are integrated to form a 'full' spatially-detailed synthetic population. To generate agents, we used multiple Bayesian Networks (BN) to probabilistically draw pools from the microsample, followed by a Generalized Raking (GR) adjustment to match marginal controls. Using various measures, we demonstrate that our BN + GR framework performs better than more commonly used synthesis methods in both capturing the heterogeneity in the microsample and matching marginal controls. We also highlight the importance of accounting for heterogeneity by using separate type-specific models based on an explicitly defined household typology. For built environment synthesis, we generated various spatial entities such as buildings, housing units, establishments, and jobs at distinct spatial locations by fusing data from various spatial datasets. Their spatial distributions are found to effectively approximate the 'real' built environment in our study area. Our proposed framework can be used to generate a 'full' synthetic population for use in ABMs with more spatio-demographic heterogeneity than can otherwise be estimated using traditional methods.

Keywords Synthetic population · Built environment · Agent-based microsimulation · Bayesian Network · Land Use-Transport Interaction (LUTI) model

Citation: Zhou, M., Li, J., Basu, R. and Ferreira, J. (2021). Creating spatially-detailed heterogeneous synthetic populations for agent-based microsimulation. *Computers, Environment and Urban Systems*. doi: 10.1016/j.compenvurbsys.2021.101717

1 Introduction

With the availability of increased computing power, applications of agent-based microsimulations in the fields of transportation and urban studies have burgeoned in recent years (Fagnant and Kockelman, 2014; Waddell, 2002; Salvini and Miller, 2005). In particular, decision support systems such as land use-transport interaction (LUTI) models have increasingly delved deeper to portray the complex interrelationships between urban development and travel behavior with high spatio-temporal resolution through dynamic microsimulations (Acheampong and Silva, 2015; Basu and Ferreira, 2020a; Waddell, 2011). These microsimulation platforms, integrated with econometric models, depict

behaviors of various agents (e.g., households or individuals) related to mobility patterns at various spatio-temporal scales in addition to the interactions between agent behaviors and urban systems.

To that end, these agent-based models (ABMs) require disaggregate and comprehensive representation of systems they aim to simulate, a major component of which is the *synthetic population*. The purpose of a synthetic population in a microsimulation platform is to characterize the population, including households and individual members, with socio-demographic attributes in rich detail. The extent to which a synthetic population can replicate the 'real' (or actual) population has a significant impact on the credibility of the simulation that relies on it. Separate from the transportation domain, a related research stream has focused on constructing *spatial microsimulations* (Tanton et al., 2014; Ballas et al., 2005) that seek to create, analyze, and model individual-level data allocated to geographic zones (Lovelace et al., 2017). Although the two research communities use different terms, spatial microsimulations can be considered analogous to population synthesis as both aim to generate spatially-detailed microdata from samples.

While nearly complete information of the real population is collected by national censuses, such data are largely inaccessible even for research purposes due to valid concerns over privacy and security. Instead, microsamples (often referred to as public use microdata samples or PUMS) collected from various types of surveys are often available along with marginal statistics of some key socio-demographic attributes. For example, in the U.S., the PUMS offer detailed data for every individual and household but the spatial resolution is purposely kept low (e.g., a large area with at least 100,000 residents) to deter reverse-engineering efforts. Alternatively, marginal distributions of socio-demographic attributes (e.g., number of children in the household) are available at high spatial resolution (e.g., usually up to the block group level). The major challenge in creating a synthetic population for agent-based microsimulations lies in combining agent-based information at coarse spatial resolution with aggregate summary information at high spatial resolution.

Despite the growing interest in population synthesis, the spatial dimension of synthetic populations has received limited attention. Most existing approaches assign aggregated zonal information to the synthetic agents and fail to go further in terms of spatial resolution. This may be because the use of ABMs in the LUTI realm has been largely dominated by transportation researchers, who are satisfied with the spatial resolution of aggregated zones (e.g., Traffic Analysis Zones or TAZs) that are adequate for their aim of simulating medium or short-term activity-travel patterns. However, aggregated zones are insufficient for the disaggregate modeling of long-term urban decisions, such as residential and workplace location choices (Zhu et al., 2018). For example, if we are to construct an ABM for exploring housing market dynamics, we would want households to bid on specific housing units in specific buildings at precise locations (not aggregated zones). Thus, we argue that the term '*synthetic population*' has received myopic treatment in the literature and should be extended to include not just agents within the population but also detailed representation of the built environment (e.g., spatial entities such as housing units, buildings, schools, and establishments) that may enable spatially disaggregate allocation of the population. This is in keeping with the rising importance of 'digital twins' that seek to include increasingly large and accurate building information models.

In this study, we apply state-of-the-art methods to generate a 'full' synthetic population accounting for the heterogeneity in household and individual characteristics as well as the marginal controls of key socio-demographics. Additionally, and more importantly, we augment the agent population synthesis by incorporating the construction of the city-wide building population and detailed inventories of housing units and establishments. The integration of these two components, *agents* (i.e., households and individuals) and the *built environment*, results in a spatially disaggregate 'full' synthetic population that replicates the urban system at a high spatial resolution. We demonstrate this framework through an application to the city-state of Singapore for the base year of 2016.

The remainder of the paper is organized as follows. The next section reviews relevant literature and discusses key gaps and contributions. We briefly introduce the study area and data used for our Singapore application before presenting our framework for the 'full' synthetic population generation in Section 3. Section 4 presents the results of our population synthesis framework with comparisons to other popular methods. The paper concludes with discussions and remarks on research extensions in Section 5.

2 Literature Review

In this section, we first discuss existing population synthesis methods from the transportation literature. Then, we draw on the spatial microsimulation literature to summarize efforts in incorporating spatial detail into synthetic populations. Finally, we reflect on the research gaps within existing literature and propose a few contributions that this study hopes to make.

2.1 Population synthesis methods

Perhaps the most popular method of population synthesis is the classical Iterative Proportional Fitting (IPF), which was originally introduced as a technique to adjust contingency tables (Deming and Stephan, 1940) and later widely applied in urban studies and transportation research for population synthesis (Arentze et al., 2007; Beckman et al., 1996; Guo and Bhat, 2007; Zhu and Ferreira Jr, 2014). The IPF method fits a multivariate contingency table initialized from microsample data to the target marginal control distributions in an iterative manner. Despite its conceptual simplicity and popularity, the IPF algorithm has some notable limitations. Its performance depends heavily on the quality of microsample data, which are often inadequate due to the recruitment of niche samples or inconsistencies in the sampling methodology. Additionally, the microsample is likely to reflect only a limited number of attribute combinations, which limits the heterogeneity of the constructed synthetic population (Sun and Erath, 2015). Trying to obtain unobserved (or limitedly observed) attribute combinations using IPF results in what is commonly referred to as the 'zero-cell problem' (Farooq et al., 2013; Guo and Bhat, 2007). The IPF method also suffers from scalability issues whereby inclusion of a large number of attributes, especially those with multiple categories, can impose heavy computational burdens (Farooq et al., 2013; Sun and Erath, 2015). While the IPF usually matches distributions only at one demographic level (i.e., either household or individual), a more recent variant known as the Iterative Proportional Updating (IPU) algorithm has been proposed to allow for matching both household-level and individual-level distributions (Ye et al., 2009). This algorithm has been implemented in PopGen, an open-source synthetic population generator (Konduri et al., 2016).

Another often-used technique - combinatorial optimization (CO) - attempts to reach an optimized solution of population synthesis by randomly drawing from the microsample while minimizing differences in marginals with algorithms such as Simulated Annealing (Abraham et al., 2012; Voas and Williamson, 2000). CO-based approaches resemble IPF in that they also replicate existing agents from the microsample (Sun and Erath, 2015). There are other variants of IPF or CO such as fitness-based methods (Ma and Srinivasan, 2015) that follow the process of microsample replication. However, as mentioned earlier, over-dependence on microsample replication can result in several conceptual and empirical challenges.

More recently, researchers have adopted a probabilistic paradigm instead of the deterministic approach of the conventional IPF-based methods (Farooq et al., 2013; Ilahi and Axhausen, 2019; Saadi et al., 2016; Sun and Erath, 2015; Zhang et al., 2019). These studies break down the synthesis process into two steps: (a) characterization of the joint distribution of agent attributes, and (b) sampling from the learned joint probability distribution. Thus, the synthesized agents generated through this approach are *not* replicas of the microsample, and consequently reflect a greater heterogeneity of agent attributes (Sun et al., 2018).

While some studies opted to use Markov Chains for probabilistic population synthesis (Farooq et al., 2013; Saadi et al., 2016), others adopted data-driven inferential methods such as Bayesian Networks (Sun and Erath, 2015; Zhang et al., 2019). Markov Chains capture correlations among variables sequentially, which can be challenging to model when the sequence (or ordering) of variables is unknown and complex interdependencies exist among attributes. In the two studies using Markov Chains, we observed that the sequence of variables was exogenously pre-determined instead of being conceptually driven or learnt from the data. Bayesian Networks are comparatively better at inferring the multivariate probabilistic relationships among attributes, as the joint distributions are determined through graphical representation.

A few studies have tried to adopt the best of both worlds by combining statistical learning techniques and fitting adjustments to generate synthetic populations that are representative of attribute interrelationships and consistent with marginal controls. Casati et al. (2015), for example, used MCMC and generalized raking (similar to an augmented IPF) to synthesize the population. Saadi et al. (2018) combined a Hidden Markov Model (HMM) with IPF and reported quasi-perfect marginal distributions and relatively accurate multivariate distributions. Ilahi and Axhausen (2019) applied generalized raking to adjust the BN-based synthesis and reported a good fit to the marginal controls.

Over the last couple of years, the popularity of machine learning has motivated the use of deep learning methods in population synthesis. Methods such as Variational Auto-Encoder (VAE) and Wasserstein Generative Adversarial Network (WGAN) have been found to work well for high-dimensional cases (Borysov et al., 2019; Garrido et al., 2020). While deep learning methods provide promising approaches for population synthesis, long-standing issues of machine learning like the lack of interpretability (i.e., the black-box nature of machine learning models) and the tendency to overfit the training data remain viable concerns (Basu and Ferreira, 2020c).

2.2 Spatial microsimulation methods

A related and often overlapping stream of studies beyond the transportation domain is referred to as *spatial microsimulation* or, more broadly, small area estimation (Pfeffermann, 2002; Tanton and Edwards, 2012; Tanton et al., 2014). In essence, spatial microsimulation models seek to simulate the population at spatially disaggregated scales. Such

models have been developed for several decades and applied in various domains. Spatial microsimulation extends the traditional agent-based microsimulation methods to incorporate a spatial dimension (Farrell et al., 2012) and is often considered to be analogous to population synthesis, or a broader approach of which population synthesis is a crucial part (Lovelace et al., 2017). Similar to the previously discussed population synthesis approaches, spatial microsimulation models mainly leverage deterministic or probabilistic sample reweighting techniques - IPF and its variants (Edwards and Clarke, 2012; Panori et al., 2017), CO (Voas and Williamson, 2000; Farrell et al., 2012), and generalized regression (Ballas et al., 2007; Tanton et al., 2011; Vidyattama et al., 2013) - to generate synthetic population microdata and assign them to geographic zones (Lovelace et al., 2017). These models often go beyond the 'mere' synthesis of agents and derive estimates of certain key indicators like income and its inequalities (Vidyattama et al., 2013; Panori et al., 2017) and obesity (Edwards et al., 2011; Edwards and Clarke, 2012) or model population dynamics over time (Rephann and Holm, 2004; Kavroudakis et al., 2012; Birkin et al., 2017). Several spatial simulation models have been operationalized for policy analysis in various areas such as demography (Ballas et al., 2005; Birkin et al., 2017), healthcare (Edwards et al., 2011; Edwards and Clarke, 2012), and economics (Campbell and Ballas, 2013; Kavroudakis et al., 2012).

2.3 Research contributions of this study

Although the population synthesis literature has continued to evolve in methodological rigor, the methods largely fail to consider spatial information of the synthetic agents or adequately represent the built environment. Detailed spatial information is of great value to ABMs seeking to model spatially disaggregate agent behaviors, e.g., housing market dynamics, evacuation behaviors, pandemic spreads. On the other hand, spatial microsimulation models account for the spatial dimension but usually assign agents to aggregated geographic zones (Tanton et al., 2014; Lovelace et al., 2017). Peters et al. (2014) is an exception where housing units are considered but the study has limited explicit representation of spatial entities. Additionally, most methods utilize reweighting techniques (IPF, CO etc.) that replicate microsamples, which limits the heterogeneity of the synthetic microdata.

The 'digital twin' approach that is recently gaining attention aims to provide a digital replication of living as well as non-living entities that can facilitate the means to monitor, understand, and optimize the functions of all physical entities and for humans to provide continuous feedback to improve quality of life and well-being (El Saddik, 2018). Translated to a more ABM-friendly language, this provides an impetus for greater attention to modeling the 'non-living' entities within urban systems (e.g., the built environment comprising housing units, buildings, jobs, schools, and establishments) by using 'full' synthetic populations.

This study aims to contribute to the population synthesis and spatial microsimulation literatures on several counts. First, we propose a combined Bayesian Network and Generalized Raking framework for agent synthesis that can incorporate microdata heterogeneity and match marginal controls better than more traditional and popular methods such as IPF. Second, we construct the built environment at a more spatially detailed resolution than in previous studies (e.g., housing units, buildings, and establishments). Third, we assign synthetic agents to specific housing units and jobs, not just aggregated zones, that enable us to simulate detailed residential and job location dynamics (although we do not show these simulation results here). Fourth, by virtue of using a probabilistic sampling design, our agents are truly synthetic and cannot be traced back to the observations in the microdata, thereby lending an additional layer of privacy to the original data.

3 Research Methods

In this section, we first describe the study area of Singapore which we use as a case study to demonstrate the application of our framework. Second, we provide an overview of the various data sources that are used to construct the 'full' synthetic population for Singapore, i.e., both agents and the built environment. Finally, we outline our proposed frameworks for agent synthesis and built environment synthesis.

3.1 Study Area

Singapore is a city-state that covers a total area of 719 square-kilometers. Located south of Peninsular Malaysia, it has a total population of around 5.61 million (as of 2016), among which 3.93 million are local residents (i.e., Citizens and Permanent Residents) belonging to 1.26 million resident households.[1] The land area of Singapore is divided into six planning regions and subsequently 55 planning districts (or planning areas), as shown in Figure 1. As of 2016, there are 1,422 TAZs and around 126,000 postcodes in use. Unlike the more conventional definition of postcodes (or ZIP codes) that most readers may be used to, postcodes in Singapore usually refer to a specific building in most areas (or a block in

[1]These statistics are sourced from the Singapore Department of Statistics (commonly referred to as SingStat), available at `https://www.tablebuilder.singstat.gov.sg/publicfacing/mainMenu.action`.

less dense and undeveloped areas). Thus, Singaporean postcodes are point features, not polygon features, lending high spatial resolution to the representation of urban systems (which we will subsequently use for population synthesis).

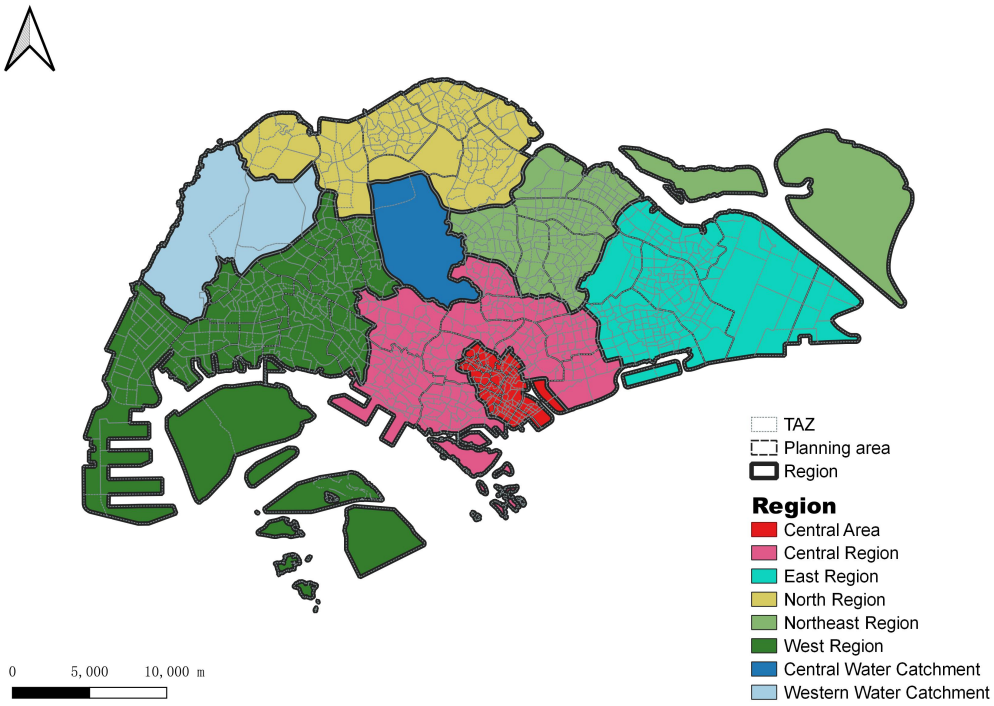


Figure 1: Geographical layout of Singapore [2]

Unlike most of the U.S., land use and housing policies in Singapore prioritize public housing and mixed land use. Various 'New Towns' have been developed that are designed to be self-sufficient in terms of providing everyday facilities within close proximity. The Housing Development Board (HDB) of Singapore is responsible for public housing policies and has overseen the implementation of various housing schemes that provide public housing to over 80% of Singaporean households (Singapore Housing & Development Board, 2019). Public housing flats (or HDB flats, as they are more commonly referred to) can be sold by current owners under certain conditions on the occupancy period. In addition, there are several types of private housing in Singapore including condominiums, apartments, and landed properties (i.e., where the property deed includes the land as well as the built structure).

Our proposed framework for population synthesis can be generalized to any other metropolitan region with similar data sources (which we intend to demonstrate through future research). We chose Singapore as the study area in this paper because of the availability of a rich variety of data sources (some of which are proprietary), and our need for a synthetic Singapore population to initialize the ABM we have developed for a project to simulate urban futures.

3.2 Data

In this study, we use the 2016 Household Interview Travel Survey (HITS) as the detailed microsample. The HITS is a 10% representative sample of Singaporean households that contain at least one resident (i.e., Citizen or Permanent Resident). This dataset is proprietary and was provided by the Land Transport Authority (LTA) of Singapore that conducts these travel surveys periodically every four years. That being said, any other representative microsample, such as a Public Use Microdata Sample (PUMS), a housing survey, or a consumer expenditure survey, would be just as

[2]The Central Water Catchment is part of the Central Region and the West Water Catchment is part of the West Region. Both are designated as natural reserves that have restrictions on residential, commercial, and industrial uses.

viable as long as it contains detailed information on a rich set of variables that are of significance and interest to the phenomena modelers seek to explore through ABMs.

IPF-like algorithms also use a set of marginal controls that specify the total number (or proportion) of households or individuals that belong to certain categories across one or more variables. Although most studies in the literature used only socio-demographic marginals, we used marginals that controlled for *both* socio-demographic and spatial variations. These datasets were obtained from the 2015 General Household Survey (GHS), which is available on the open data portal provided by the Singapore government.[3] The mid-decade GHS provides comprehensive data on Singapore's population and households in between the Population Censuses (which are conducted every ten years at the turn of the decade).

We used a variety of datasets in this study for built environment synthesis. These data were obtained through collaborative projects from local agency partners or sourced from open sources and include a wide range of information regarding the urban space. Proprietary data such as building addresses (postcodes) and building footprints were provided by the Singapore Land Authority (SLA) for 2016. We also used the open-source land use layer from the 2014 Master Plan created by the Urban Redevelopment Authority (URA). Likewise, public housing building information is openly available on the HDB website. Other open-source third-party data[4] that provide information on building types, building heights (number of stories), and construction times were also utilized. Additionally, data on zone-to-zone travel times at the TAZ level (i.e., travel skim matrices) were provided by LTA and used for the assignment of jobs to synthetic worker agents.

3.3 Full population synthesis

Although we present our proposed framework for synthesis of both agents and the built environment in Figure 2, the two components of the framework along with their integration are discussed separately.

3.3.1 Agent synthesis

In this subsection, we will focus only on the framework for agent synthesis (i.e., the left component of Figure 2). Our synthesis approach for agents, i.e., households and individuals, emulates the two-step process discussed earlier, consisting of sampling from a probabilistic model followed by adjustment to marginal controls using IPF or a similar technique. We chose the Bayesian Network (BN) as our probabilistic model because it has the necessary flexibility to capture heterogeneous multivariate joint distributions, and Generalized Raking (GR) for matching multivariate marginal controls at both the household and individual levels because of its higher efficiency compared to the traditional IPF algorithm.

The BN is a graphical model that can learn and represent complex relationships between a large set of variables. It is commonly depicted as a directed acyclic graph[5] where nodes correspond to variables and edges indicate correlation between variables - this is known as the "structure" of the BN. Each node also possesses a conditional probability distribution that defines the probability of observing certain values, given the values of its parent node(s) - these form the "parameters" of the BN. Using its structure and parameters, a BN can model a complex joint distribution that contains a wide range of variable correlations.

Departing from previous literature, we train different BNs based on explicitly defined household types. As households comprise a diverse set of family structures and living arrangements, we expect different household types (or structures) to exhibit different joint distributions representing unique relationships between variables. To this end, the microsample data are partitioned into six sub-samples (Table 1), each corresponding to a household type: *single-member households* (SM), *multi-generational households* (MG), *married couples without co-residing children* (MC), *single parents with children* (SP), *nuclear households* (NH), and *others* (OT). These types mirror the household structures defined by the Ministry of Social and Family Development of Singapore (Singapore Ministry of Social and Family Development, 2017). The first 5 types (excluding '*others*') encompass over 90% of Singaporean households, which makes us reasonably confident that this typology should be able to satisfactorily represent a large amount of the variation within

[3] `https://data.gov.sg/`

[4] For example, Streetdirectory (`https://www.streetdirectory.com/`) and EMPORIS (`https://www.emporis.com/`)

[5] A directed acyclic graph (DAG) consists of vertices and edges (or arcs), with each edge directed from one vertex to another, such that following those directions will never form a closed loop.

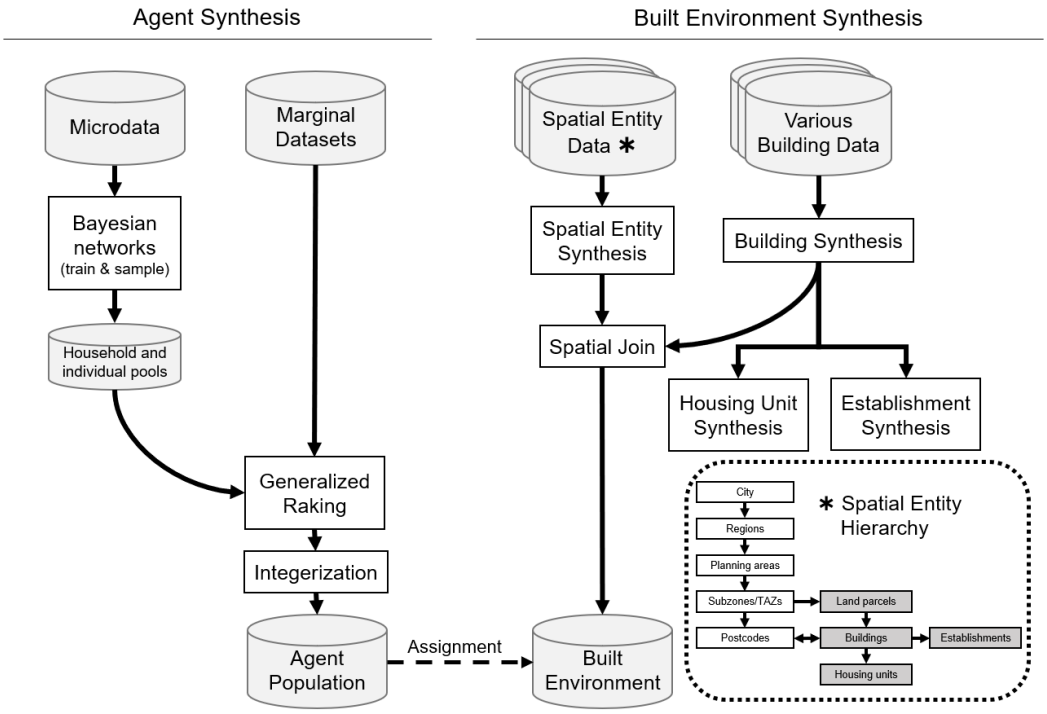


Figure 2: Integrated framework for agent and built environment population synthesis

Table 1: Household typology for agent synthesis

Type	Definition	Count	Share (%)
SM	Single-member (*contains exactly 1 individual*)	156,950	12.5%
MG	Multi-generational (*contains individuals of at least 3 age ranges, each separated by at least 15 years*)	118,850	9.5%
MC	Married couple without co-residing children (*contains 2 individuals whose age groups are not more than 15 years apart*)	186,800	14.8%
SP	Single parent with children (*contains 2 or more individuals, exactly 1 of which is at least 15 years older than the others*)	87,500	7.0%
NH	Nuclear household (*contains 3 or more individuals, exactly 2 of which are at least 15 years older than the others*)	598,950	47.7%
OT	Other households (*any household that does not fulfill the above criteria, e.g., multiple siblings living together*)	108,550	8.6%

Singaporean households. We note here that any other classification scheme that is a reasonable representation of the socio-cultural context of a particular study area will likely be just as appropriate. In addition to modeling heterogeneous interrelationships, explicitly separating household types also facilitates expert-guided verification and, if necessary, modification of the BN structure and parameters discovered by the algorithm.

BNs can be specified directly based on expert knowledge. In addition, there exist a wide variety of data-driven algorithms for learning both the structure and parameters of BNs. A popular subset of these are "score-based" algorithms, which search for a network structure that results in an optimal score measure, such as the Bayesian Information Criterion (BIC) or Akaike Information Criterion (AIC). In keeping with previous BN-based population synthesis methods, we chose an algorithm that selects a parsimonious model with good fit based on the AIC measure. A household-level BN was constructed for each of the six microsample data sub-samples corresponding to a household type. We used the greedy hill-climbing algorithm provided by the R package *bnlearn* for learning the structure of the BN, followed by

maximum likelihood estimation for learning the parameters of the BN (Scutari, 2010). This method is also robust to missing data, as we can use multiple imputation to obtain several BNs and average them into a single model (not unlike ensemble-based models in machine learning such as random forests).

To create a pool of synthetic households for Singapore, the appropriate number of households of each type (Table 1) were drawn from the corresponding BN using forward sampling, producing a total of 1.26 million households. A similar, but slightly modified, process was used to create a pool of synthetic individuals. First, the separate household- and individual-level microsample tables were joined together to create a combined dataset containing all variables of interest. Modelers may choose to ignore this step for applications where a combined dataset for households and individuals is directly available. Second, this dataset was also partitioned into the six aforementioned household types, since differences in variable relationships and distributions at the household level are expected to percolate down to the individuals within the households as well. Third, individual-level BNs (which contain both household- and individual-level variables) were learned from the sub-samples. In training the individual-level BNs, we initially used the entire set of household-level variables used for the household-level BNs and consequently trimmed edges that were found to be inconsequential or have very weak strength. The resultant individual-level BNs are parsimonious representations which prohibit edges to household-level variables that do not play a part in the following individual sampling step.

Building on the previous step of sampling households, we drew as many individuals as determined by the sampled household size variable from the corresponding individual-level BN (that is specific to the household type), by forward sampling conditional on the sampled household variables. This ensures that the characteristics of household members correspond to the characteristics of the household they belong to. A pool of roughly 4 million synthetic individuals was constructed from this individual sampling process.

Although the pools of synthetic households and individuals sampled from the BNs can reproduce the microsample's joint distributions fairly well, they do not match reported aggregate marginal distributions of certain control variables. This is because the BNs approximate the joint distributions observed in the microsample, which is an imperfect representation of the overall population (usually represented through the Census) despite best efforts to obtain representativeness. Several issues could occur that distort the representativeness of the microsample (e.g., sampling bias, attrition, non-response), but those are usually beyond the purview of the modeler seeking to construct synthetic populations from a given microsample and are outside the scope of this paper.

For ABMs with a spatial dimension (e.g., location choice simulations), it is imperative to ensure that spatial distributions of people, housing, and jobs are appropriately represented. Failure to consider the spatial dimension in the population synthesis approach can affect the veracity of the ABM and may reduce the effectiveness of scenario explorations. In order to adjust the BN-generated samples spatially, we used the Generalized Raking (GR) procedure from the R package *MultiLevelIPF* for the household and individual pools to simultaneously fit them to a set of selected multivariate marginal controls (Mueller, 2018). We chose the planning area as our spatial unit of analysis (recall from Figure 1 that a planning area in Singapore is equivalent to a neighborhood), as our ABM focuses on location choices that are pertinent to this level of detail. Thus, the marginal controls we chose to match our synthetic population with are: (a) Planning Area × Dwelling Type, Planning Area × Household Income, and Dwelling Type × Number of Workers at the household level, and (b) Planning Area × Age and Planning Area × Employment Status at the individual level. Choosing any other spatial unit (e.g., the commonly used TAZ for traffic simulations) is just as acceptable; we suggest that the modeler choose the scale of spatial detail for marginals based on the granularity with which they wish to model spatial processes.

After convergence, GR produces fractional weights, separately for each household and for each individual. However, in order to avoid disarranging the grouping of individuals in households, using only the household weights should be sufficient. To create a synthetic population, an integerization procedure must be performed on the fractional household weights to convert them to integers. The truncate, replicate, sample (TRS) method is chosen for this purpose (Lovelace and Ballas, 2013). TRS first truncates weights to their integer part, then randomly samples weights to increment by one, weighted by the fractional part, until the original total weight is restored. We then replicated each household according to its computed integer weight, replicating the constituent individuals along with it. Since GR and TRS preserve the total size of the population, we finish the resident synthetic population generation with 1.26 million households and 3.96 million individuals.

Finally, since Singapore has a large non-resident population (consisting of 1.67 million individuals in 2016) that factors significantly into ABMs of mobility choices, we add these households and individuals as a post-processing step. Most non-residents have residential and job locations that are largely dictated by government and employer policies. For example, construction workers and single-individual household work permit holders live in assigned dormitories and work at assigned locations. Therefore, a certain number of individuals holding each work visa type (e.g., Employment Pass, Student Pass, Construction Work Permit) as determined by statistics from the Ministry of Manpower are inserted into the synthetic population (Singapore Ministry of Manpower, 2020). Their characteristics are determined by a

combination of expert knowledge about the different foreigner demographics and observed variable distributions in the HITS microsample. Adding non-residents to the synthetic population allows the transportation modeling component of ABMs to model the full set of daily trips. However, these non-resident households and individuals are not included in the results and discussion that follow in this paper, since the reference data (i.e., HITS microsample and GHS marginal controls) that we use for comparison do not include non-residents.

3.3.2 Built environment synthesis

We represent the urban built environment through a series of spatial entities in a hierarchical structure at different spatial scales (see the right side of Figure 2). These levels of spatial aggregation (e.g., planning regions, planning areas, etc.) may vary by the study area but the general strategy of adopting a hierarchical structure of spatial entities is expected to serve the modeler well for all cases. The shaded boxes in the sub-figure are spatial entities that we specifically constructed to record the residences and workplaces for households and workers. Their size and location are estimated using sources independent from the socio-demographic data (on the left side of Figure 2). They are linked to the demographic data through elements in the spatial hierarchy, usually the postcode.

The ontology-based approach is well-suited for the synthesis of the built environment in this study given the variety of data utilized. Datasets from various sources containing different aspects of the built environment are integrated based on the ontology that links semantic features that characterize the entities in the built environment. Based on the created ontology, the integration process then constructs the full list of entities, retrieves attribute values based on the relationships in the ontology, and imputes missing values with similar entities. Spatial entities are created either in sequence (in the cases of buildings, housing units, and establishments) or in parallel (in the case of land parcels). We refer readers to Zhu and Ferreira (2015) for further details of this ontology-based approach.

The primary element of built environment synthesis is the creation of buildings, as they form the basis for synthesizing housing units and establishments. This process includes cleaning building geometry data and inferring various building attributes. We obtained spatial locations and building geometries directly from building footprint data with relatively minimal processing (such as merging multiple postcodes that point to the same building). Next, we inferred building attributes such as building type, height, and space. Building types (e.g., residential, commercial, industrial, etc.) were inferred first based on the footprint data and third-party datasets matched through postcodes. For residential buildings, specific types (e.g., public, private, and landed) and subtypes (e.g., condos, apartments, terrace houses, etc.) were also identified. For most cases, we were able to retrieve the number of stories within each building directly from available data. For cases with missing data, we used the building heights to estimate the number of stories. The building space for different types was estimated using the number of stories and the area directly retrieved from the building footprints. We also estimated the age of the buildings using data on construction and commencement dates, when available.

Based on the synthetic buildings and their use types (e.g., residential, commercial, industrial), we then created housing units of different types and establishments and firms in various industry sectors. Counts of these entities were retrieved directly from available data or estimated based on building space for the specific use type. We estimated other entity attributes (e.g., sizes, ages, etc.) based on the relevant characteristics of the buildings. Additionally, we synthesized land parcels using open-source land use data from URA. We also used spatial data on different amenities (such as public transit facilities, top schools, shopping malls, and expressway access points) to compute 'local' accessibility measures for each building (and postcode) along the road network.

3.3.3 Assigning agents to the built environment

Synthetic household agents need to be assigned to housing units, and worker agents (individuals) need to be assigned to jobs to produce a spatially detailed synthetic population. In this study, we matched housing units to synthetic households based on their planning areas (neighborhoods) and dwelling types, which are known from the HITS microsample data. A rule-based matching heuristic was implemented for this assignment. First, a predefined percentage of units was reserved in each zone-dwelling type bin based on expert knowledge and historical trends, reflecting the vacancy rate. Then, within each bin, housing units were randomly assigned to households, with larger units more likely to be assigned to larger and wealthier households. If we ran out of units before all households in that bin were matched, such households were matched to either a unit of similar dwelling type in the same neighborhood, or, if still unavailable, to a unit of same dwelling type in a nearby neighborhood.

We assigned jobs to worker agents using a destination choice model estimated using the HITS microsample. The destination choice set of each worker comprises a set of 30 TAZs that contain at least one job pertaining to their industry sector. The explanatory variables include the number of jobs in that sector, the log-transformed commuting cost (travel time) between the worker's home and destination (workplace), and interactions between commuting cost and individual-level attributes of the worker (e.g., age, gender, income, etc.). The estimated model was used to predict

the probabilities of choosing the 30 TAZs for each worker, following which we used a probability-weighted random assignment to match a worker with a job within the chosen TAZ.

4 Results and Discussion

We first evaluate the performance of our framework against more commonly used population synthesis methods using a variety of metrics. Then, we describe the synthetic agents we generated by focusing on the different BNs we learned at both the household- and individual-levels. Finally, we conclude this section with a discussion of the spatial entities that are derived from the built environment synthesis.

4.1 Model performance

Since the built environment synthesis is largely a data integration and fusion process sans the use of any statistical modeling approaches, we will focus exclusively on the agent synthesis component of our synthetic population generation framework to evaluate model performance. As a reminder, our synthetic agent population consists of 1.26 million households and 3.96 million individuals, accurately reflecting the resident population of Singapore in 2016. The agent population synthesis process, including data imputation and processing, BN learning and sampling, GR, and integerization, takes about 35 minutes on a x64 laptop PC with 16 GB of RAM and a 1.90GHz Intel Core i7-8650U CPU.

The population synthesis framework used in this study is a combination of Bayesian Networks (BN) and Generalized Raking (GR), which we call the '*BN + GR*' method for simplicity. We compared this framework with the BN-sampled pools generated before the application of GR (i.e., *BN only*), as well as iterative proportional updating (IPU), a multilevel IPF algorithm that is used as part of the popular open-source software PopGen. We used the IPU implementation from the R package *ipfr*, with the same marginal controls listed previously to ensure a fair comparison (Ward, 2020). The synthetic populations generated by these three methods (i.e., BN + GR, BN only, and IPU) are evaluated using two criteria: (a) similarity to the joint distribution of the weighted microsample, and (b) similarity to the marginal control distributions.

4.1.1 Similarity to the joint distribution of the weighted microsample

We evaluate the similarity of the generated synthetic populations to the joint distribution of the weighted microsample using two methods. First, we use an objective error measure to quantify the extent of the similarity, whereby a lower error value indicates a greater similarity. Second, we use a graphical method to understand the performance of each method in greater detail. It is worth noting here that we used the unweighted microsample to generate our synthetic populations. Sampling weights are usually calculated and provided by the agency conducting the survey in order to account for stratified sampling or other known deviations from a purely random sample. In our case, there is no need for sampling weights since the BN sampling process generates a full synthetic population based on the multivariate correlations observed in the travel survey.

As an objective measure of differences between each of our three generated synthetic populations and the one generated using the HITS microsample with sampling weights, we use the standard root mean square error (SRMSE) as defined by Sun and Erath (2015):

$$SMRSE = \sqrt{\sum_{m_1=1}^{M_1} \dots \sum_{m_n=1}^{M_n} (f_{m_1,\dots,m_n} - \hat{f}_{m_1,\dots,m_n})^2 \times \prod_{i=1}^{n} M_i} \tag{1}$$

where n is the total number of variables upon which the joint distribution is defined; $f_{m_1,\dots,m_n}$ and $\hat{f}_{m_1,\dots,m_n}$ are the relative frequencies of a particular variable combination in the weighted microsample and synthetic population, respectively; and M_i is the total number of categories of the ith variable. AA SRMSE value of zero represents a perfect match between the two joint distributions under comparison, while higher values represent greater mismatch. However, we neither expect nor desire a zero SRMSE, since the HITS, even with sampling weights, is an imperfect representation of the population (as evidenced by HITS' discrepancies with the marginal controls). SMRSE values for the three different methods at both the household- and individual-levels are reported in Table 2.

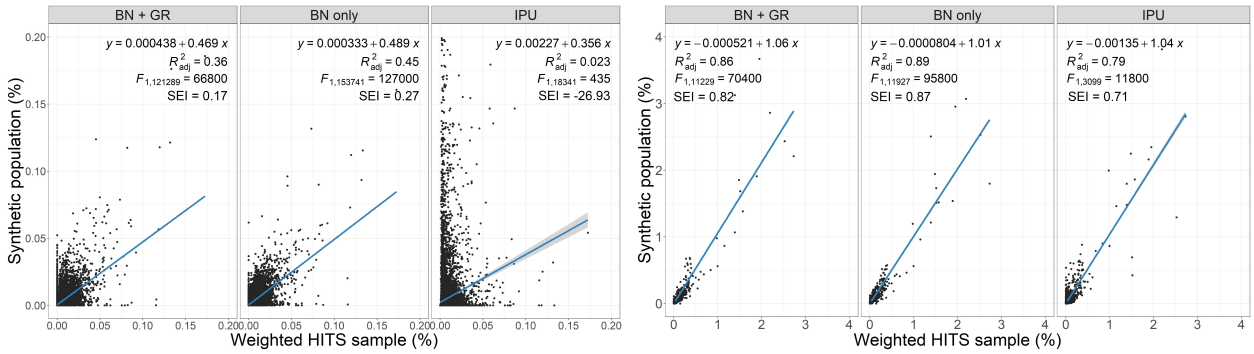
We expect the BN-only method to have the lowest SRMSEs, because the BNs directly attempt to model the joint distribution of the microsample. When GR is applied to match marginal controls, the joint distribution is altered slightly, which leads to a small increase in SRMSE observed for the BN + GR method. IPU yields the greatest error of all;

Table 2: SRMSE values for different agent population synthesis methods

SynPop method	**SRMSE** (***household-level***)	**SRMSE** (***individual-level***)
BN + GR (this study)	23.66	7.90
BN only	22.34	6.58
IPU	102.86	9.67

this is also unsurprising since it tries to replicate the microsample, which constitutes a significantly smaller sample size than the BN-sampled pools. IPF-like methods tend to achieve greater accuracy with larger sample sizes that can capture greater heterogeneity (Wong, 1992), but microsamples are expensive to collect and 'real-world' surveys rarely go beyond a 10% sampling rate (as is the case with HITS in Singapore). Finally, we note that the larger magnitudes of the household-level SRMSEs are not due to any particular biases of the methods in balancing household-level versus individual-level fits, but are simply because we define nine variables of interest at the household-level but only six at the individual-level.

The second comparative approach visualizes the fit of the synthetic population to the weighted microsample through a frequency plot, where the frequencies of every unique variable combination in the two datasets are plotted against each other. Each point in the frequency plot represents a unique variable combination. A perfect match is represented by a line of best fit with zero intercept, unit slope, and an R^2 value of one. Additionally, for each plot, we report the Standard Error around Identity (SEI), which resembles R^2 in that higher values are better but instead measures the extent of dispersion from the perfect line of best fit (Tanton et al., 2011). Thus, unlike R^2, the SEI measure can account for systematic biases in the synthetic population. We present frequency plots at both the household- and individual-levels in Figure 3.



(a) Household-level comparison

(b) Individual-level comparison

Figure 3: Frequency plots to assess joint distribution match between synthetic populations and weighted microsample

Three important observations emerge from this visual analysis. First, the synthetic population generated by the BN-only method achieves the best fit to the joint distribution, as evidenced by the slope and adjusted R^2 values being closer to one and the SEI being higher than the rest. Second, the IPU method performs the worst among the three, possibly because it cannot generate as many variable combinations (there are only 22,494 degrees of freedom for IPU at the household-level as compared to over 190,000 d.o.f. for BN-based methods). Third, the individual-level distribution is much easier to fit to than the household-level distribution, as the latter contains more variables (each with more categories) that leads to increased complexity. In summary, this visual analysis reinforces our observations from the SMRSE comparisons and shows that BN + GR performs reasonably well at approximating the joint distribution, but not as well as BN-only. We will explain in the following subsection why BN + GR is a better choice, although it may seem counterintuitive at this point.

4.1.2 Similarity to the marginal control distributions

We also evaluated the similarity of the synthetic populations generated by the three methods (i.e., BN + GR, BN-only, and IPU) to the reported marginal control distributions. As a reminder, the marginal controls we chose to match to our

synthetic agent populations are: (a) Planning Area × Dwelling Type, Planning Area × Household Income, and Dwelling Type × Number of Workers at the household level, and (b) Planning Area × Age and Planning Area × Employment Status at the individual level. We explore the similarities to the marginal controls through two methods. First, we look at bar plots that highlight how the distributions match up for the socio-demographic variables in the aforementioned set of marginal controls. Second, we look at maps that highlight the spatial variation in the differences (or errors) between the distributions of marginal controls. To maintain a parsimonious representation of the large set of comparisons that can be generated (given the number of marginal controls we use), we show bar plots for household income and number of workers (at the household-level) and age (at the individual-level), and maps for selected dwelling types (at the household-level) and employment status (at the individual-level).

The three bar plots for household income, number of workers, and individual age are shown in Figure 4. In general, we find that the synthetic populations generated by the BN + GR and IPU methods are able to match the marginal control distributions almost perfectly. This is unsurprising as these methods include an explicit IPF-like process that aims to match the reported marginals. On the other hand, the microsample with sampling weights (or weighted HITS) and the synthetic population generated by the BN-only method have skewed distributions that are often very different from the marginal controls (see, for example, the case of household income where higher-income households are under-represented in the microsample). Using sampling weights or only the BN creates an overreliance on the microsample, which usually falls short of being representative of the population despite best design efforts, and does not align the microsample with the reported marginal distributions. Thus, it seems clear that both BN + GR and IPU are equally adept at matching marginal controls and perform much better at this objective than the BN-only method. This, in combination with our findings regarding the similarity to the joint distribution of the weighted microsample, leads us to conclude that the BN + GR method ('our' method) achieves a fine balance between the two objectives, which the other two methods cannot since they perform well on only one of the two objectives.

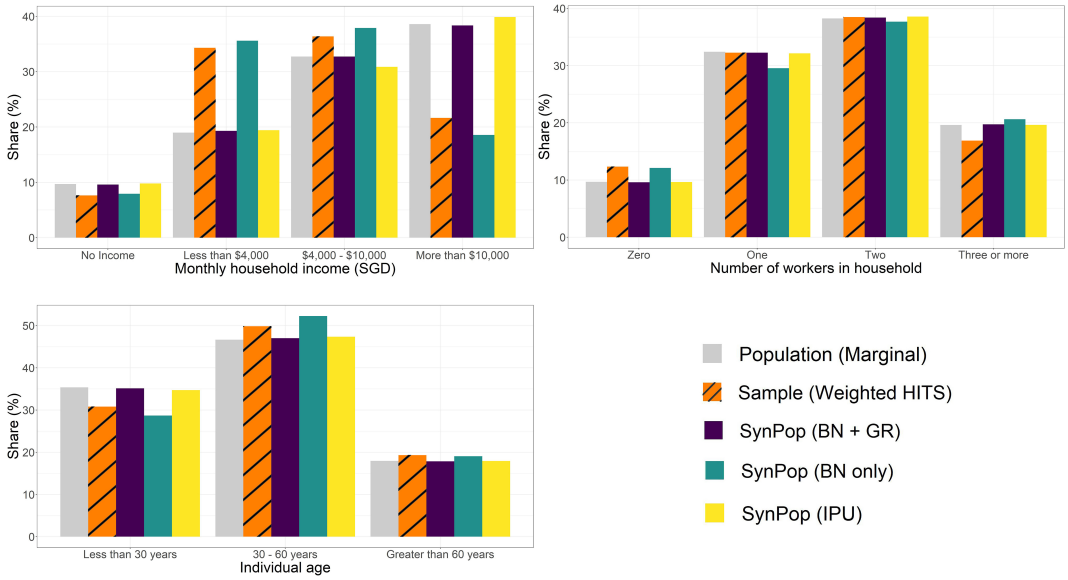
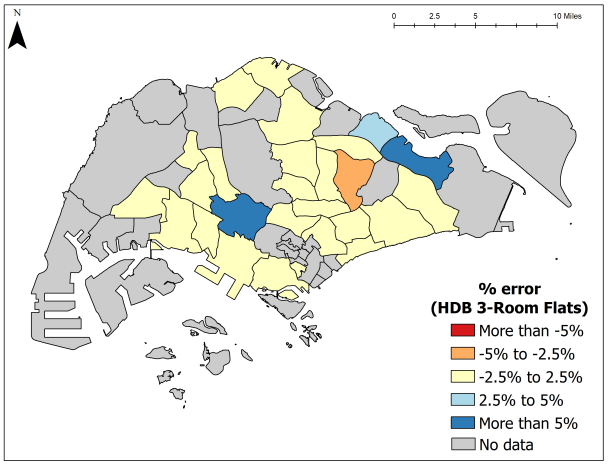


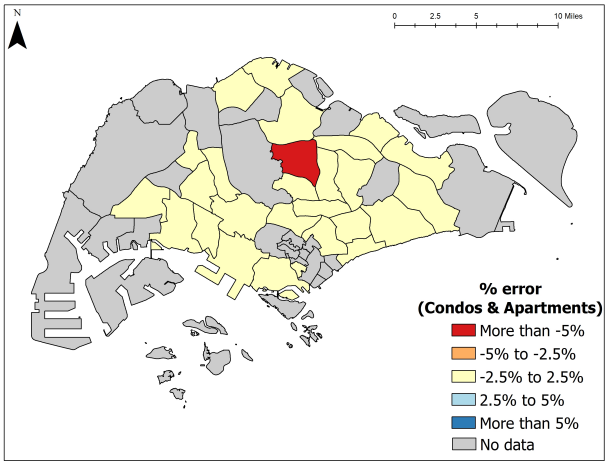
Figure 4: Bar plots to assess distribution match between synthetic populations and socio-demographic marginal controls

In addition to exploring the match with socio-demographic marginal controls, we also explored the match with spatial marginal controls. We present the differences (or errors) between the spatial distributions of the synthetic population and the marginal controls for selected dwelling types and employment statuses (for brevity) in Figure 5. Instead of a comparative analysis as earlier, we only show error maps for the synthetic population generated by the BN + GR method to understand the extent to which systematic spatial biases might be generated by our method, if any. Looking at the most popular dwelling types for public and private housing (i.e., HDB flats with 3 rooms, and condominiums and apartments respectively), we do not find any observable non-random patterns of spatial errors. In particular, we find that almost all planning areas have dwelling type distribution errors between -2.5% to 2.5%, which are quite reasonable. Errors greater than 5% are infrequent and occur in different planning areas across the maps.

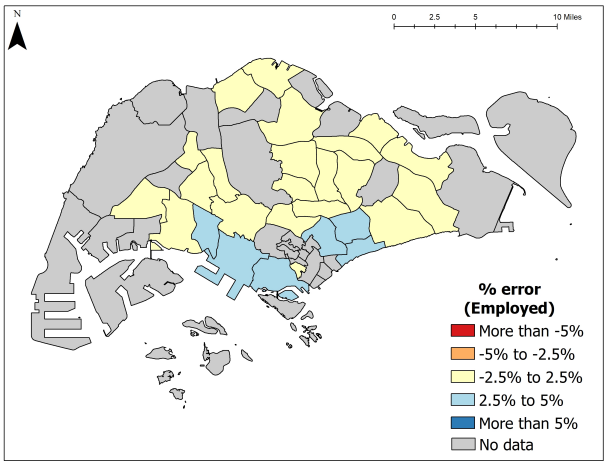
We obtain similar observations from the spatial error distribution of the employment status of individuals. In particular, we find that we are able to predict inactive individual counts (i.e., individuals who are not in the labor force) within a 2.5% error margin. Our predictions for employed individuals are within the 2.5% error margin for the majority of



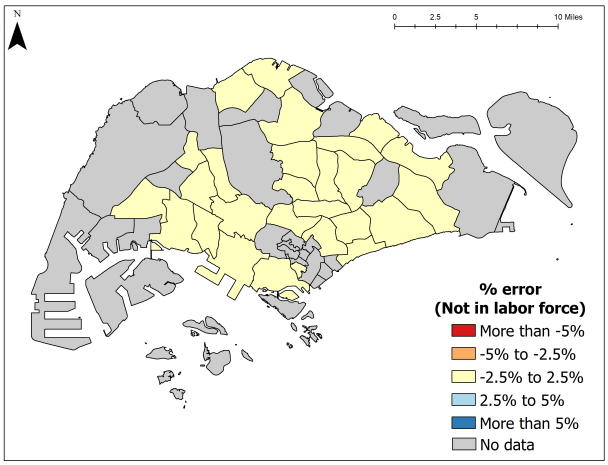
(a) Households living in HDB 3-room flats



(b) Households living in condos and apartments



(c) Employed individuals



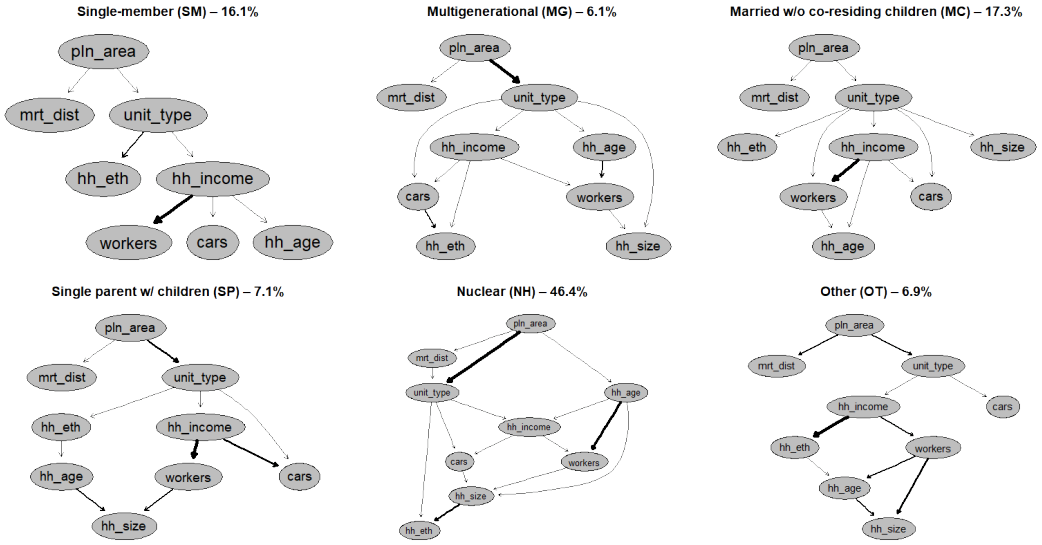
(d) Individuals not in labor force

Figure 5: Maps to assess spatial distribution match between BN + GR synthetic population and selected socio-demographic marginal controls

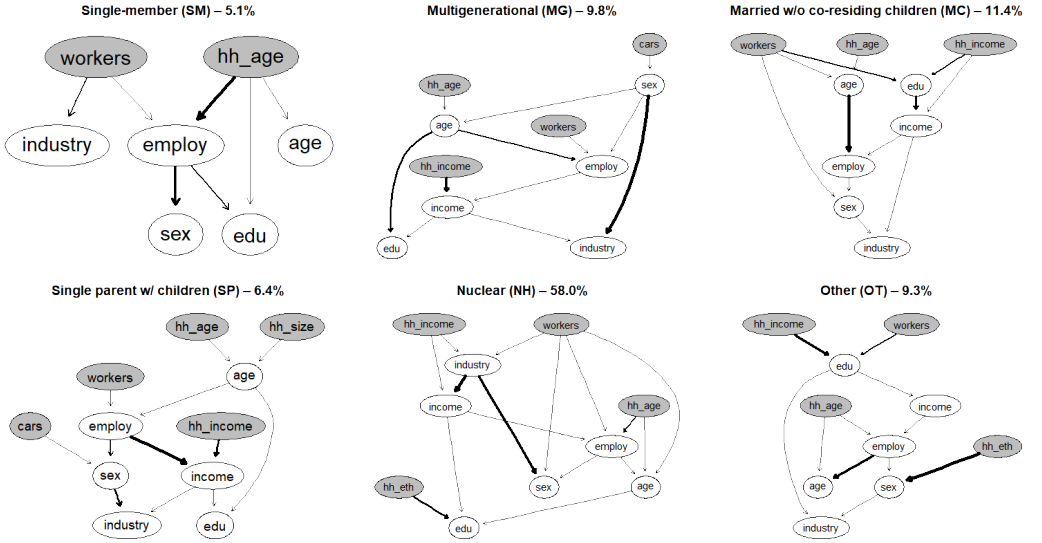
planning areas, with the exception of a few cases in the Central Region where the error margin is slightly higher but still less than 5%. In general, we do not find any reason to suspect that our method may have introduced systematic spatial biases within the synthetic agent population. Additionally, upon detailed examination of our results, we found that our predictions after the GR step are quite accurate with less than 1% difference. The errors increase (but not by much, as demonstrated through the maps) due to the integerization process. An improved integer programming algorithm might find a better solution, but finding one with adequate performance on our scale of problem is beyond the scope of this paper.

4.2 Synthetic agents: Households and individuals

In this subsection, we briefly discuss the Bayesian Networks we obtained for households and individuals at the end of the BN training step in our BN + GR framework. Recall that we had defined a typology of six household categories based on which we trained six BNs for households and an additional six BNs for individuals. We represent these BNs in Figure 6, where the nodes are variables and the edges have varying thickness that are directly related to the strength of the relationship between the nodes they connect. The shaded nodes in the figure are household-level variables, while the unshaded nodes are individual-level variables. Each variable is treated as a categorical (i.e., ordinal or nominal) variable with multiple levels (or categories), which are listed in detail in the appendix.



(a) Household-level Bayesian Networks
(Variable dictionary - *pln_area*: Planning Area, *mrt_dist*: Distance to nearest MRT station, *unit_type*: Dwelling type, *hh_size*: Household size, *hh_income*: Household income, *hh_eth*: Ethnicity of head of household, *hh_age*: Age of head of household, *workers*: Number of workers, *cars*: Number of cars)



(b) Individual-level Bayesian Networks
(Variable dictionary - *age*: Age, *sex*: Gender, *income*: Income of individual, *industry*: Industry sector of job, *edu*: Highest educational qualification, *employ*: Employment status)

Figure 6: Trained Bayesian Networks for the six household categories at both household and individual levels

The household-level BNs shown in Figure 6a share the same 'root' relationships, whereby the dwelling type and distance to the nearest MRT station are both conditional on the planning area where the household resides. The importance of considering the spatial dimension in population synthesis is underscored by the fact that the planning area is the root node for all six BNs, implying that socio-demographic correlations are largely dependent on (and vary by) spatial locations. We observe further commonalities as we proceed 'deeper' down the BNs, e.g., the number of workers and the number of cars owned by the household are conditional on the household income, which in turn is conditional on the dwelling type. However, the nature of these relationships vary by household type. In single-member households, single-parent households with children, and married households without co-residing children, the strongest link is

unsurprisingly between the number of workers and household income as these households are less likely to earn income through non-work means (e.g., pension or government benefits). The strongest relationship for multigenerational households and nuclear households is between dwelling type and planning area, as these choices are likely to be driven by similar preferences (e.g., proximity to primary schools, community centers, or parks). For nuclear households (who comprise almost 50% of the Singaporean population), we also notice a strong relationship between the number of workers and the age of the household head. This is because nuclear households with a younger household head are likely to have more workers (and vice versa).

Individual-level BNs comprise both individual-level variables (unshaded nodes) and household-level variables (shaded nodes), as shown in Figure 6b. Among all household types, we find common relationships between age and employment status, employment status and income, and gender and industry sector of job. However, these relationships vary by strength across the household types. For example, for individuals in nuclear households (forming almost 50% of the population), there are very strong relationships between the job industry sector and the individual's gender and income. Additionally, the ethnicity of the household head can influence the highest educational qualification of individuals within the household (which may reflect ethnic disparities in access to education resources and/or opportunities). For married households without co-residing children, the age of the individual has a strong impact on their employment status. For single parents without children, their employment status strongly influences their income. Among single-member households, the age of the household head (who is also the only individual in the household) relates strongly with employment status. This perhaps reflects how younger individuals (likely students) are less likely to be employed.

Despite all BNs exhibiting a common and 'expected' set of relationships in general, there are important differences between the BNs for the different types of households. For instance, the household-level BN for nuclear households is unique in that the number of cars owned is related to household size, which likely reflects the likelihood of car ownership (and consequently the number of cars owned) being directly proportional to the number of children in nuclear households. Another example is the individual-level BN for single parents with children, which is unique in that household size determines individual age. This lines up with our intuition because the age distribution in a single-parent household depends strongly on the number of children, which is simply one less than household size. Finally, but importantly, we note that many significant differences between the BNs for different household types are hidden in the parameters (i.e., node probability distributions) in addition to those observed from the graph structures. For example, although number of cars owned is determined by household income in both single-member and nuclear households, single-member households rarely, if ever, own multiple cars, whereas this is not unexpected for nuclear households. These numerous differences in both structures and parameters between the BNs justify our choice of using a household typology to learn type-specific BN models.

4.3 Synthetic built environment: Spatial entities and zones

We generated the synthetic built environment comprising various spatial entities in Singapore for the year 2016. First, our building synthesis process resulted in the creation of 116,415 buildings. We present the spatial distribution of buildings in Singapore by use type in Figure 7, where we find residential buildings distributed across the island, commercial buildings mostly located within the central area, and industrial buildings situated mainly in the suburban areas (particularly along the south-west shore of the island). This spatial distribution lines up with our first-hand knowledge of Singapore and external data sources (e.g., official land use maps and geospatial services). These buildings are home to the housing units and establishments of different industries that provide housing and jobs to the household and individual agents respectively.

Using the residential buildings generated through the building synthesis process, we created around 1.66 million housing units. We purposely create more units than households to allow for a reasonable vacancy rate that can mimic the 'real' housing market. We also included dormitories reserved for foreign migrant workers and units occupied by foreigner-headed households (whom we had to explicitly include in our agent population as a post-processing step due to data unavailability in the microsample). The spatial distributions of different types of housing units are shown in Figure 8. Public HDB housing units are located within HDB buildings that are located across the island in HDB estates and New Towns. Private units (i.e., condominiums and apartments), on the other hand, are more clustered within the Central Region. Landed properties are scattered around the island and other types of units, such as dormitories and shophouses (i.e., mixed-use landed houses with the ground floor being commercially used while upper levels are used for residential purposes), are located more in the suburban areas.

Similar to the generation of housing units, we generated synthetic establishments using the generated commercial and industrial buildings to accommodate the employment opportunities for synthetic individuals. 194,044 establishments are created with around 3.42 million jobs, which is slightly higher than the 3.12 million employed individuals obtained from the synthetic agent population (to allow for a 'vacancy' rate in the job market, similar to the housing market). The spatial distribution of establishments proportional to the number of jobs they contain is presented in Figure 9. The

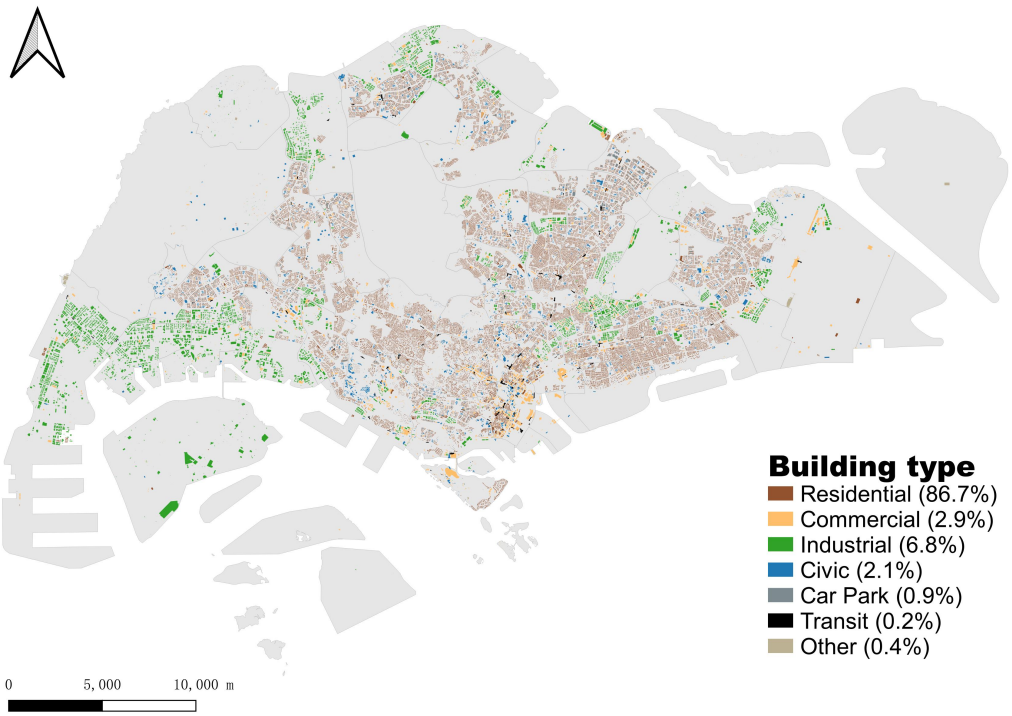


Figure 7: Spatial distribution of synthetic buildings (N = 116,415)

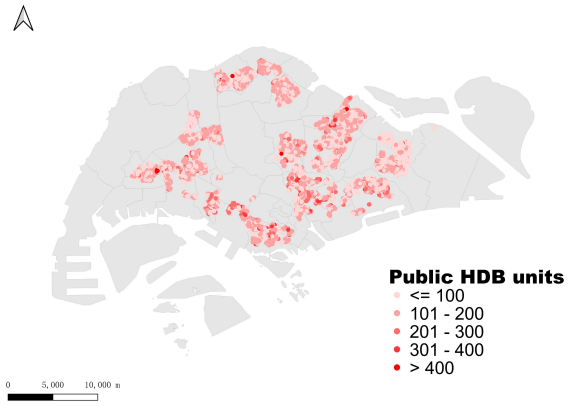
distributions of synthetic establishments of different industries are consistent with the distribution of synthetic buildings, with manufacturing jobs concentrated in suburban areas mostly in industrial buildings and jobs in the finance and real estate sectors mostly located within the city center in commercial buildings.

4.4 Full synthetic population: Linking agents to locations

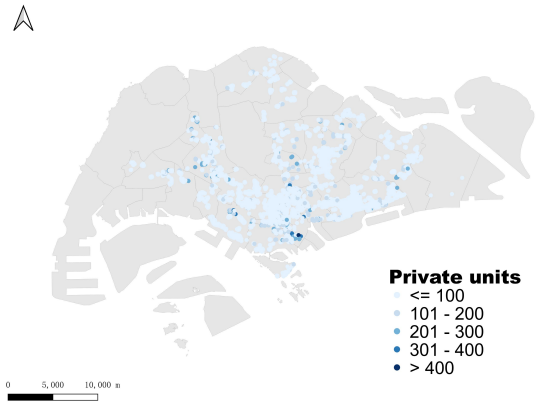
The final step of the population synthesis is to match the synthesized agents with the synthesized spatial entities of the built environment by assigning housing units to households and jobs to workers. We find that our rule-based heuristic of matching household agents with housing units works remarkably well. Assuming a vacancy rate of 2%, we were able to assign housing units to over 99% of the households within their preferred planning area (neighborhood) and dwelling type. Only 0.8% of households required an adjustment for neighborhood or dwelling type. The spatial distribution of residential locations of all 1.26 million households is shown in Figure 10a.

After assigning housing units to households, we were able to assign a job to each of the 3.12 million workers in their preferred industry sector. As a recap, we estimated a destination choice model on the HITS microsample whereby assignment likelihood ratios were (for each industry sector) directly proportional to the number of jobs in the destination TAZ and inversely proportional to the commute distance. The spatial distribution of job locations of employed individuals is shown in Figure 10b. Additionally, as a measure of the accuracy of our job assignment, we compared the job-housing distances for workers in our synthetic population with those in the HITS microsample. We found that the distributions look quite similar, although we tend to slightly overestimate the commute distances (see Figure A1 in the appendix). The median commute distance for our synthetic workers is 8.44 kilometers, as compared to 7.93 kilometers for the HITS sample.

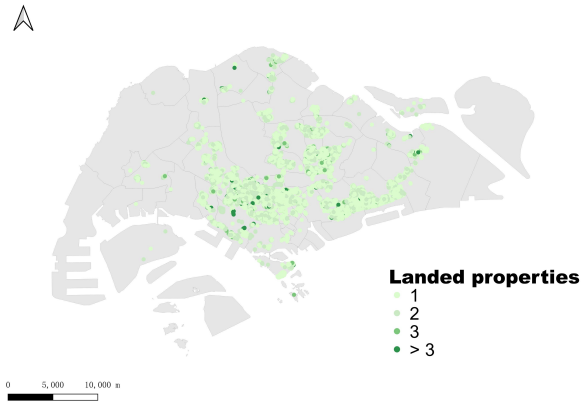
After both matching procedures are complete, the spatially assigned agent population is ready for use in large-scale agent-based microsimulations. These initial assignments can be further adjusted to match predictions of behavioral models (such as residential and job location choice models) by performing a 'burn-in' simulation using the ABM (Basu and Ferreira, 2020b).



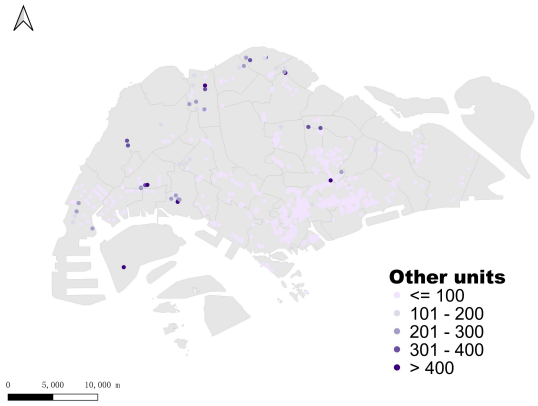
(a) HDB units (74.5%)



(b) Private units (17.4%)



(c) Landed properties (6.2%)



(d) Other units (1.9%)

Figure 8: Spatial distributions of synthetic housing units (N = 1,661,284)

5 Conclusion

Agent-based models (ABMs) of urban systems have been in use for several decades. In recent times, ABMs have grown in popularity due to the availability of high-performance computing resources and large data storage capabilities. ABMs also continue to grow in complexity by attempting to model urban systems in increasing spatio-temporal detail. Perhaps the most crucial component of ABMs is the population they seek to model, thus requiring the creation of a synthetic population. Data availability challenges affect the resolution at which synthetic populations can be created, whereby agent-based information at coarse spatial resolution needs to be combined with aggregate summary information at high spatial resolution. Even though data may be available across agencies, variable definitions and data collection periods differ, confidentiality issues persist, and considerable time and funding are needed to piece together the elements. We think it would be worthwhile to invest in periodic construction of detailed synthetic populations that can be used for many modeling purposes. The construction could be timed to coincide with the periodic travel surveys that many metropolitan areas conduct every 4-10 years. Several population synthesis methods have been suggested over the years, starting from iteratively updating weights in a relatively simple manner to complex deep learning models. Despite the growing research interest in population synthesis, the spatial dimension of synthetic populations has remained largely neglected. Most existing approaches assign aggregated zonal information to the synthetic agents and fail to go further in terms of spatial granularity.

In this study, we addressed this myopic treatment of the synthetic population by creating two distinct components - *agents* and the *built environment* - that could be integrated to form what we call the 'full' synthetic population. In terms of creating the built environment, we generated synthetic spatial entities such as buildings, housing units, establishments, and jobs at various spatial scales (e.g., postcodes, land use parcels, planning areas, planning regions, etc.). We employed

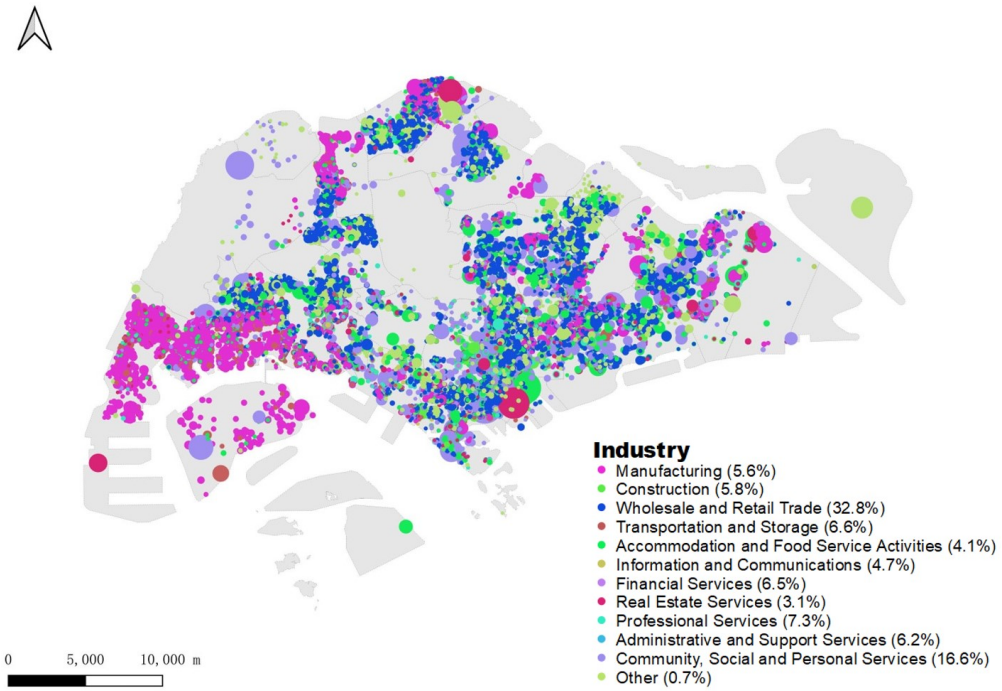
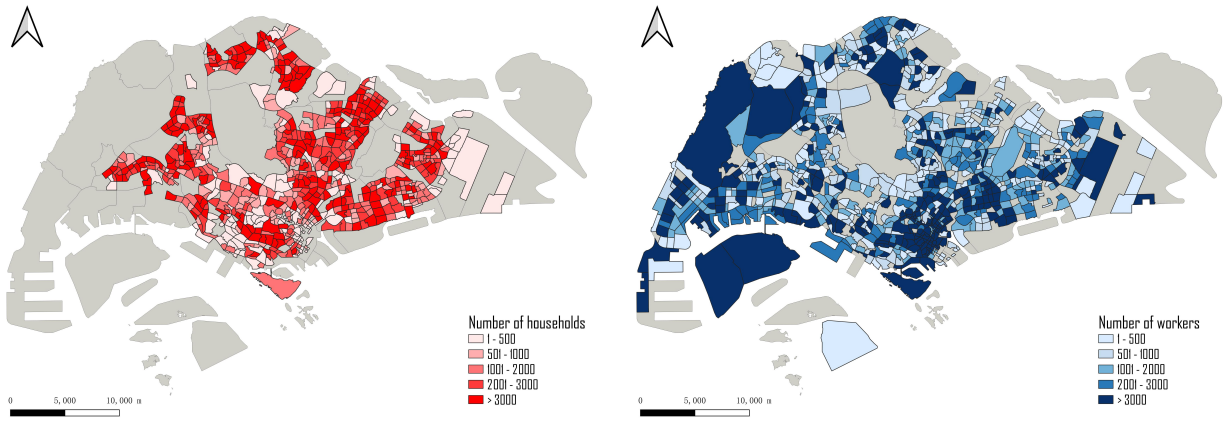


Figure 9: Spatial distribution of synthetic establishments (N = 194,044)



(a) Residential locations

(b) Job locations

Figure 10: Spatial distributions of residential locations of households and job locations of workers

a two-stage framework to probabilistically sample households and individuals from the microsample and subsequently adjust these pools to match distributions of marginal control variables. Using various measures, we demonstrated that our BN + GR framework (combining Bayesian Networks and Generalized Raking) performed better than more commonly used methods (such as IPU and BN only) in both capturing the heterogeneity in the microsample and matching marginal controls. We also highlighted the importance of accounting for heterogeneity by using separate type-specific models based on an explicitly defined household typology. Using data fusion techniques on multiple spatial datasets, we generated various disaggregate spatial entities and found their spatial distributions to match the 'real' built environment in our study area. Thus, we highlighted how our proposed framework can be used to generate a 'full' synthetic population for use in ABMs of any study area of choice.

The research presented in this paper can be extended in various ways. One area of future research is the development of better and faster algorithms for population synthesis. We found that probabilistic models such as the BN can replicate the microsample well, but an additional proportional update step (using the GR or another IPF-variant) is necessary to match the marginal controls. The second step of matching reduces the goodness-of-fit of the first step of sampling, as we highlighted through various error measures. A combined and simultaneous framework (e.g., a multi-objective optimization routine) could address this issue that arises during sequential adjustment. Additionally, we found that the integerization process introduced about 2.5-5% errors in the spatial distribution of our synthetic population. Perhaps a better (and faster) integerization algorithm might be able to reduce these admittedly small errors even further.

The exploration of better conceptual frameworks for synthetic populations (and their subsequent use in ABMs) is another promising research area. While we went further than most in generating disaggregate spatial entities such as buildings, it is possible to go even further in the pursuit of creating digital twins and synthesizing even the interiors of buildings. Such detailed synthesis can enable the use of ABMs to model building evacuation techniques in case of emergencies, building energy use, and airflow within populated buildings (to name but a few applications). The reader might also wonder if a separate household typology could have resulted in a more 'accurate' synthetic population. We simply chose our typology because it best represented the population in our study area, which is what should guide modelers. However, it is certainly feasible to explore alternative typologies with different categories or different numbers of categories, or even sidestep these explicit definitions by deriving the typology from the data (through, e.g., latent class analysis). Finally, we note that every population synthesis paper that we reviewed has demonstrated their proposed framework in only one study area (which we are equally guilty of). It would behoove the ABM community to begin thinking about extending their population synthesis frameworks to other study areas or to demonstrate the use of a generalizable framework in multiple study areas.

In closing, we hope that we are able to convince readers and the ABM community at large to pay more attention to standardizing easily repeatable methods for creating synthetic populations in greater spatial detail and with adequate representation of heterogeneity. We anticipate that 'full' synthetic populations (comprising both agents and the built environment) can enable the exploration of hitherto unanswered research questions about urban processes with high spatio-temporal granularity.

Acknowledgements

This research was funded in part by the Singapore National Research Foundation through the Future Urban Mobility group at the Singapore-MIT Alliance for Research and Technology Center. We appreciate the support of our agency partners in Singapore for sharing relevant data and information.

References

J. E. Abraham, K. J. Stefan, and J. Hunt. Population synthesis using combinatorial optimization at multiple levels. Technical report, 2012.

R. A. Acheampong and E. A. Silva. Land use–transport interaction modeling: A review of the literature and future research directions. *Journal of Transport and Land use*, 8(3):11–38, 2015.

T. Arentze, H. Timmermans, and F. Hofman. Creating synthetic household populations: Problems and approach. *Transportation Research Record*, 2014(1):85–91, 2007.

D. Ballas, G. P. Clarke, and E. Wiemers. Building a dynamic spatial microsimulation model for ireland. *Population, Space and Place*, 11(3):157–172, 2005.

D. Ballas, G. Clarke, D. Dorling, and D. Rossiter. Using simbritain to model the geographical impact of national government policies. *Geographical Analysis*, 39(1):44–77, 2007.

R. Basu and J. Ferreira. A LUTI microsimulation framework to evaluate long-term impacts of automated mobility on the choice of housing-mobility bundles. *Environment and Planning B: Urban Analytics and City Science*, 47(8): 1397–1417, 2020a. doi: 10.1177/2399808320925278.

R. Basu and J. Ferreira. Planning car-lite neighborhoods: Examining long-term impacts of accessibility boosts on vehicle ownership. *Transportation research part D: transport and environment*, 86:102394, 2020b. doi: 10.1016/j.trd.2020.102394.

R. Basu and J. Ferreira. Understanding household vehicle ownership in singapore through a comparison of econometric and machine learning models. *Transportation Research Procedia*, 48:1674–1693, 2020c. doi: 10.1016/j.trpro.2020.08.207.

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

M. Birkin, B. Wu, and P. Rees. Moses: dynamic spatial microsimulation with demographic interactions. In *New frontiers in microsimulation modelling*, pages 53–77. Routledge, 2017.

S. S. Borysov, J. Rich, and F. C. Pereira. How to generate micro-agents? a deep generative modeling approach to population synthesis. *Transportation Research Part C: Emerging Technologies*, 106:73–97, 2019.

M. Campbell and D. Ballas. A spatial microsimulation approach to economic policy analysis in s cotland. *Regional Science Policy & Practice*, 5(3):263–288, 2013.

D. Casati, K. Müller, P. J. Fourie, A. Erath, and K. W. Axhausen. Synthetic population generation by combining a hierarchical, simulation-based approach with reweighting by generalized raking. *Transportation Research Record*, 2493(1):107–116, 2015.

W. E. Deming and F. F. Stephan. 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, 1940.

K. L. Edwards and G. Clarke. Simobesity: combinatorial optimisation (deterministic) model. In *Spatial Microsimulation: A reference guide for users*, pages 69–85. Springer, 2012.

K. L. Edwards, G. P. Clarke, J. Thomas, and D. Forman. Internal and external validation of spatial microsimulation models: small area estimates of adult obesity. *Applied Spatial Analysis and Policy*, 4(4):281–300, 2011.

A. El Saddik. Digital twins: The convergence of multimedia technologies. *IEEE multimedia*, 25(2):87–92, 2018.

D. J. Fagnant and K. M. Kockelman. The travel and environmental implications of shared autonomous vehicles, using agent-based model scenarios. *Transportation Research Part C: Emerging Technologies*, 40:1–13, 2014.

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

N. Farrell, K. Morrissey, and C. O'Donoghue. Creating a spatial microsimulation model of the irish local economy. In *Spatial microsimulation: A reference guide for users*, pages 105–125. Springer, 2012.

S. Garrido, S. S. Borysov, F. C. Pereira, and J. Rich. Prediction of rare feature combinations in population synthesis: Application of deep generative modelling. *Transportation Research Part C: Emerging Technologies*, 120:102787, 2020.

J. Y. Guo and C. R. Bhat. Population synthesis for microsimulating travel behavior. *Transportation Research Record*, 2014(1):92–101, 2007.

A. Ilahi and K. W. Axhausen. Integrating bayesian network and generalized raking for population synthesis in greater jakarta. *Regional Studies, Regional Science*, 6(1):623–636, 2019.

D. Kavroudakis, D. Ballas, and M. Birkin. Simeducation: A dynamic spatial microsimulation model for understanding educational inequalities. In *Spatial microsimulation: A reference guide for users*, pages 209–222. Springer, 2012.

K. C. Konduri, D. You, V. M. Garikapati, and R. M. Pendyala. Application of an enhanced population synthesis model that accommodates controls at multiple geographic resolutions. In *Proceedings of the 95th Annual Meeting of the Transportation Research Board, Washington, DC, USA*, pages 10–14, 2016.

R. Lovelace and D. Ballas. 'truncate, replicate, sample': A method for creating integer weights for spatial microsimulation. *Computers, Environment and Urban Systems*, 41:1–11, 2013.

R. Lovelace, M. Dumont, R. Ellison, and M. Založnik. *Spatial microsimulation with R*. Chapman and Hall/CRC, 2017.

L. Ma and S. Srinivasan. Synthetic population generation with multilevel controls: A fitness-based synthesis approach and validations. *Computer-Aided Civil and Infrastructure Engineering*, 30(2):135–150, 2015.

K. Mueller. Multilevelipf: Implementation of algorithms that extend ipf to nested structures. *Available online from* `https://github.com/krlmlr/MultiLevelIPF`, 2018.

A. Panori, D. Ballas, and Y. Psycharis. Simathens: A spatial microsimulation approach to the estimation and analysis of small area income distributions and poverty rates in the city of athens, greece. *Computers, Environment and Urban Systems*, 63:15–25, 2017.

I. Peters et al. Constructing an urban microsimulation model to assess the influence of demographics on heat consumption. *International Journal of Microsimulation*, 7(1):127–157, 2014.

D. Pfeffermann. Small area estimation-new developments and directions. *International Statistical Review*, 70(1): 125–143, 2002.

T. J. Rephann and E. Holm. Economic-demographic effects of immigration: results from a dynamic spatial microsimulation model. *International Regional Science Review*, 27(4):379–410, 2004.

I. Saadi, A. Mustafa, J. Teller, B. Farooq, and M. Cools. Hidden markov model-based population synthesis. *Transportation Research Part B: Methodological*, 90:1–21, 2016.

I. Saadi, B. Farooq, A. Mustafa, J. Teller, and M. Cools. An efficient hierarchical model for multi-source information fusion. *Expert Systems with Applications*, 110:352–362, 2018.

P. Salvini and E. J. Miller. Ilute: An operational prototype of a comprehensive microsimulation model of urban systems. *Networks and spatial economics*, 5(2):217–234, 2005.

M. Scutari. Learning bayesian networks with the bnlearn r package. *Journal of Statistical Software, Articles*, 35(3): 1–22, 2010. doi: 10.18637/jss.v035.i03.

Singapore Housing & Development Board. Key Statistics, 2018/2019 Annual Report. Technical report, 2019. URL `https://services2.hdb.gov.sg/ebook/AR2019-keystats/html5/index.html?&locale=CHS&pn=9`.

Singapore Ministry of Manpower. Foreign workforce numbers. Technical report, 2020. URL `https://www.mom.gov.sg/documents-and-publications/foreign-workforce-numbers`.

Singapore Ministry of Social and Family Development. Families and Households in Singapore, 2000 - 2017. Technical report, 2017. URL `https://www.msf.gov.sg/research-and-data/Research-and-Data-Series/Documents/Families%20and%20Households%20in%20Singapore%20-%20Statistics%20Series%202019%20%282000%20-%202017%29.pdf`.

L. Sun and A. Erath. A bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies*, 61:49–62, 2015.

L. Sun, A. Erath, and M. Cai. A hierarchical mixture modeling framework for population synthesis. *Transportation Research Part B: Methodological*, 114:199–212, 2018.

R. Tanton and K. Edwards. *Spatial microsimulation: a reference guide for users*, volume 6. Springer Science & Business Media, 2012.

R. Tanton, Y. Vidyattama, B. Nepal, and J. McNamara. Small area estimation using a reweighting algorithm. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(4):931–951, 2011.

R. Tanton et al. A review of spatial microsimulation methods. *International Journal of Microsimulation*, 7(1):4–25, 2014.

Y. Vidyattama, R. Cassells, A. Harding, and J. Mcnamara. Rich or poor in retirement? a small area analysis of australian private superannuation savings in 2006 using spatial microsimulation. *Regional Studies*, 47(5):722–739, 2013.

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

P. Waddell. Urbansim: Modeling urban development for land use, transportation, and environmental planning. *Journal of the American planning association*, 68(3):297–314, 2002.

P. Waddell. Integrated land use and transportation planning and modelling: addressing challenges in research and practice. *Transport reviews*, 31(2):209–229, 2011.

K. Ward. ipfr: List Balancing for Reweighting and Population Synthesis. Technical report, 2020. URL `https://CRAN.R-project.org/package=ipfr`.

D. W. Wong. The reliability of using the iterative proportional fitting procedure. *The Professional Geographer*, 44(3): 340–348, 1992.

X. Ye, K. Konduri, R. M. Pendyala, B. Sana, and P. Waddell. A methodology to match distributions of both household and person attributes in the generation of synthetic populations. In *88th Annual Meeting of the Transportation Research Board, Washington, DC*, 2009.

D. Zhang, J. Cao, S. Feygin, D. Tang, Z.-J. M. Shen, and A. Pozdnoukhov. Connected population synthesis for transportation simulation. *Transportation research part C: emerging technologies*, 103:1–16, 2019.

Y. Zhu and J. Ferreira. Data integration to create large-scale spatially detailed synthetic populations. In *Planning support systems and smart cities*, pages 121–141. Springer, 2015.

Y. Zhu and J. Ferreira Jr. Synthetic population generation at disaggregated spatial scales for land use and transportation microsimulation. *Transportation Research Record*, 2429(1):168–177, 2014.

Y. Zhu, M. Diao, J. Ferreira, and P. C. Zegras. An integrated microsimulation approach to land-use and mobility modeling. *Journal of Transport and Land Use*, 11(1):633–659, 2018.

Appendix

Table A1: Household-level variables used in the BNs

Variable	Categories	Sample share (%)
Dwelling type (*unit_type*)	HDB 1- and 2-Room Flats	5.0%
	HDB 3-Room Flats	20.5%
	HDB 4-Room Flats	35.9%
	HDB 5-Room and Executive Flats	25.7%
	Condominiums and Apartments	6.6%
	Landed properties	6.2%
	Others	0.01%
Household size (*hh_size*)	One	3.8%
	Two	13.8%
	Three	19.6%
	Four	27.9%
	Five	19.2%
	Six or more	15.8%
Monthly household income (*hh_income*)	No Income	4.5%
	Less than $1,000	2.8%
	$1,000 to $2,000	7.6%
	$2,000 to $4,000	23.6%
	$4,000 to $6,000	21.0%
	$6,000 to $10,000	20.0%
	$10,000 to $15,000	12.6%
	$15,000 to $20,000	3.1%
	More than $20,000	5.0%
Number of workers (*workers*)	Zero	7.2%
	One	27.4%
	Two	41.7%
	Three or more	23.7%
Number of cars (*cars*)	Zero	61.6%
	One	33.6%
	Two	4.0%
	Three or more	0.7%
Age of head of household (*hh_age*)	15 to 30 years	1.7%
	30 to 60 years	67.2%
	More than 60 years	31.1%
Ethnicity of head of household (*hh_eth*)	Chinese	72.7%
	Indian	11.4%
	Malay	13.2%
	Others	2.7%
Distance to nearest MRT station (*mrt_dist*)	Less than 400 meters	16.6%
	400 to 800 meters	37.0%
	More than 800 meters	46.5%

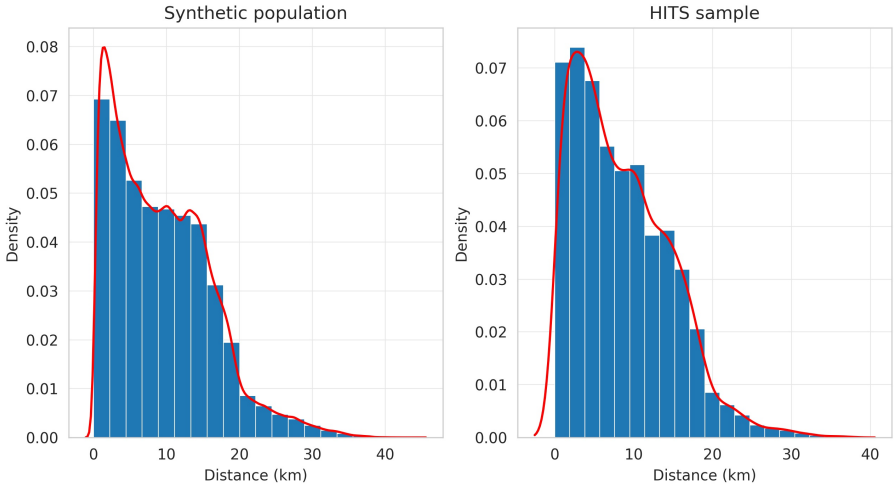


Figure A1: Distributions of job-housing distances for workers

Table A2: Individual-level variables used in the BNs

Variable	Categories	Sample share (%)
Gender (*sex*)	Male	47.0%
	Female	53.0%
Age (*age*)	Less than 15 years	6.6%
	15 to 30 years	51.6%
	30 to 60 years	20.4%
	More than 60 years	21.4%
Monthly individual income (*income*)	No Income	44.1%
	Less than $1,000	3.8%
	$1,000 to $2,000	9.9%
	$2,000 to $4,000	22.7%
	$4,000 to $6,000	11.2%
	$6,000 to $10,000	5.5%
	More than $10,000	2.9%
Industry (*industry*)	Accomodation and Food Services	3.6%
	Administrative and Support Services	6.0%
	Community, Social and Personal Services	10.3%
	Construction	3.4%
	Financial Services	4.1%
	Information and Communications	4.1%
	Manufacturing	6.0%
	Professional Driver	1.2%
	Professional Services	7.8%
	Real Estate Services	0.9%
	Transport and Storage	4.5%
	Wholesale and Retail Trade	4.4%
	Others	0.1%
	None (for those without a job)	44.4%
Employment status (*employ*)	Employed Full Time	46.5%
	Employed Part Time	5.4%
	Self-Employed	3.6%
	Full Time Student	18.0%
	Retired	9.2%
	Homemaker	13.4%
	Unemployed	2.7%
	Others	1.1%
Highest educational qualification (*edu*)	Primary	6.1%
	Secondary	18.4%
	Post-Secondary	5.6%
	Polytechnic	11.5%
	Bachelor's	17.1%
	Master's/Doctorate	5.4%
	Postgraduate certification	3.6%
	Professional degree	3.7%
	Others	19.2%
	None	9.4%

View publication stats