Constructing a Synthetic Population of Establishments for the Simmobility Microsimulation Platform

Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 19 (2016) 81 – 93

Transportation Research Procedia

www.elsevier.com/locate/procedia

International Scientific Conference on Mobility and Transport Transforming Urban Mobility, mobil.TUM 2016, 6-7 June 2016, Munich, Germany

Constructing a synthetic population of establishments for the SimMobility microsimulation platform

Diem-Trinh Le [a]*, Giulia Cernicchiaro [a], Chris Zegras [b], Joseph Ferreira Jr.[b]

[a] *Singapore-MIT Alliance for Research and Technology, Singapore*
[b] *Massachusetts Institute of Technology, USA*

Abstract

This paper presents a method for building a synthetic population of establishments in Singapore for incorporation into SimMobility platform, an integrated microsimulation model that represents households' and firms' short-, medium-, and long-term decisions, ranging from lane changing behavior, to daily activity pattern choices, to location choices. The synthetic population includes data on (1) establishments' locations, (2) establishments' industry type, (3) establishments' employment size, and (4) their occupied floor area.

Peer-review under responsibility of the organizing committee of mobil.TUM 2016.

Keywords: population synthesis; microsimulation; iterative proportional fitting, SimMobility

1. Introduction

Along with households, firms are key drivers of urban dynamics. As such, understanding firmography – or the growth, failure, and migration of new or existing firms – is critical in urban research and modeling. Firms' activities require accessibility to capital and labour, to other firms, and to customers; all of which influence their location

* *Corresponding author. Tel.: +65-6601-1548; fax: +65-6684-2118.*
E-mail address: diem@smart.mit.edu

Peer-review under responsibility of the organizing committee of mobil.TUM 2016.
doi:10.1016/j.trpro.2016.12.070

choices to some degrees. The recent wave of land use-transportation modeling tools reflects a move towards *micro*-simulation e.g. ILUTE, ILUMASS, Ramlas, and UrbanSim, aiming to represent all agents in the model system and their relevant decisions. Such models require a reasonable representation of these agents – the population of interest. However, in most cases, micro-data on firms are considered sensitive and thus not available or not accessible for research. As information for individual firms are needed for micro-simulation models, a synthetic population is often generated for this purpose.

There have been several studies on methods for synthesizing household population (e.g. Zhu & Ferreira 2014; Rich & Mulalic 2012; Beckman et al. 1996). Nevertheless, the literature on firms' synthetic population is comparably sparse. Generating a baseline synthetic population of firms or establishments is necessary to simulate firmography processes and incorporate into land-use transport models. A common method for household population synthesis is Iterative Proportional Fitting technique with Monte Carlo simulation. The same method has also been applied for firms in several studies (e.g. Khan et al. 2002; Maoh & Kanaroglou 2005; Moeckel 2009). In most cases, an establishment is described by its consumption and production, age, required floor space, type of industry, and location. However, how close the synthetic population resembled the reality was not discussed. Cernichiaro and Ferreria (2015) presented a method to build a synthetic population of establishments using directory websites. Firm characteristics simulated include employment size, floor area, and industry type. However, the method was specifically applied to the service sector and manufacturing firms were excluded.

This paper demonstrates a method for creating a synthetic population of establishments in Singapore using limited available data. The population synthesis aims to reflect the reality of firms in Singapore in 2012 for incorporation into SimMobility model (Ben-Akiva 2010; Lu et al. 2015), an integrated microsimulation model that represents households' and firms' short-, medium-, and long-term decisions, ranging from lane changing behavior, to daily activity pattern choices, to location choices. For the synthetic firm population, data to be generated include (1) establishments' locations, (2) establishments' industry type, (3) establishments' employment size, and (4) their occupied floor area.

The paper is organized as follows. First, we will briefly introduce the SimMobility framework. Following an explanation of the simulation methods, we will present the generated synthetic population. The paper concludes with a summary of the methods and some implications for future work.

2. SimMoblity Framework

SimMoblity is a system of mobility sensitive behavioural models integrated in a multi-scale simulation platform that considers land-use, transportation, and communication interactions. It focuses on the impacts on transportation networks, intelligent transportation services, and vehicular emissions, thereby enabling the simulation of a portfolio of technology, policy, and investment options under alternative future scenarios (Adnan et al. 2016). The framework consists of three different sub-models:

- Short-term (ST) simulator is a traffic micro-simulator, extended with a communications simulator as well as pedestrians and public transport. The time step can be a fraction of a second and agent decisions include lane changing, braking, accelerating, gap acceptance, but also route choice.
- Mid-term (MT) simulator is a mesoscopic simulator, designed for activity-based modelling, with explicit pre-day and within-day behaviour including re-routing and re-scheduling, and multiple transport modes. The time step is in the range of seconds to minutes and agent decisions include route choice, mode choice, activity pattern and its (re)scheduling, departure time choice.
- Long-term (LT) simulator is a land-use and transport (LUT) simulator, with a market transaction bidding model. The time step is in the range of days to months to years, and agent decisions include house location choice, job location choice, land development, and car ownership.

The LT simulator is responsible for the generation and updating of a population of agents and their corresponding demographic and locational attributes. In the beginning, a two-stage data synthesis methodology is employed for construction of a synthetic population of households and firm establishments at building scale. The approach is designed to accommodate the need for spatially disaggregated details in a manner that can be readily adjusted and rerun to incorporate new data sources, changed time frames, and updated relationships and hierarchies across overlapping datasets. Long-term behaviours of agents and their effects on urban form, markets and other

agents are implemented by a group of behavioural models that are connected in a sequential/event-based framework. Zhu and Ferreira (2014) described the methods for generating the population synthesis of households and individuals for SimMobility. In this paper, how firms and their characteristics were created is described.

3. Methods

3.1. Data collection

Much information is needed to construct a synthetic population of establishments in Singapore. Ideally, the synthetic population should include establishments' characteristics that are relevant for modeling firms' behavior, such as locations, business types, number of employees, occupied floor areas, number of vehicles, operating receipts, and so on. However, the level of details can be generated depends on the available data. For the purpose of this study, establishments to be simulated need to have information on the location, business type, occupied floor area, and employment size. Accordingly, a sample of establishments with these characteristics is required.

3.1.1. Aggregate data

Several governmental agencies in Singapore provide data on the aggregated level. Department of Statistics Singapore (SingStats) and Ministry of Manpower (MOM) provide information on the total number of jobs by industry types and number of establishments by floor type, which are accessible from their websites. The national statistics are important and useful as marginal controls for the synthetic population. Data on regional and planning areas in Singapore are managed by Urban Redevelopment Authority (URA). The five geographical regions in Singapore (Central, East, West, North, and Northeast) are further divided into 55 planning areas, with an average area of 13km^2. URA's statistics on floor space by floor type in each planning area can be used to estimate the number of jobs by planning area. Singapore's Land Authority (SLA) provides data on parcels and building footprints in the country, which are needed for estimating building sizes. Furthermore, the Accounting and Corporate Regulatory Authority (ACRA) manages the registration of business entities in Singapore. Full addresses of establishments are recorded at ACRA, providing a geographical distribution of establishments. No information on employment size and floor area however can be found at the firm or establishment level.

3.1.2. The study sample

A list of actual entities of selected industries registered at ACRA in February 2015 was obtained for the research purpose. The aim is to synthesize a population of firms in Singapore in December 2012, therefore entities registered after 31/12/2012 were removed. The dataset (282,907 observations) includes information on establishments' name, address, and industry type. However, establishments' floor space and jobs by occupation type were not recorded. As there is no restriction on the number of establishments that can be registered in one address, there were much more establishments compared to the number of unique addresses in the dataset, with an average of two establishments per unit (Table 1). To avoid overestimating employment size and floor area in one location, a reduced sample was extracted, in which only one establishment was retained for a unique address. The selected establishment is one that has the latest registered date and its floor type matched with the building type. Table 1 shows the differences between the complete dataset and the extracted study sample. There are 142,642 establishments in the sample, which belong to 134,894 firms. Most firms are single entity establishments (130,485). In this paper, establishments are treated as independent individuals and the synthetic population is simulated with no relationship between establishments as headquarter and branches were observed at this stage. The sampled establishments are located throughout Singapore, covering 37,566 postcodes, where in most cases, a postcode is associated with one building.

Table 1: Study sample description

Description	Complete list	Extracted sample
Total number of establishments	282,907	142,642
Unique Entity Number (UEN)	269,796	134,894
Unique SSIC	858	826
Unique address	142,642	142,642
Unique postcode	37,566	37,566

Establishments in the sample cover 826 detailed categories indicated by Singapore Standard Industrial Code (SSIC). For practical reasons, establishments are differentiated only by SSIC sections (the highest level of industrial type aggregation), which can be aggregated into four floor types as shown in Table 2. It is noted that SSIC has 25 sections yet not all were included in the acquired dataset. Data for organizations in the governmental sectors for example were excluded.

The data collection process resulted in:

- The number of jobs by industry in entire Singapore
- The number of establishments by floor type in entire Singapore
- The number of floor area by floor type in each planning area
- The list of buildings and building footprints by planning area
- A sample of establishments with information on their location and business type.

Table 2: Establishments' industrial and floor type

SSIC	**Section**	**Floor type**
C	Manufacturing	Industrial
H	Transportation and Storage	Warehouse
J	Information and Communications	Office
K	Financial and Insurance Activities	
L	Real Estate Activities	
M	Professional, Scientific and Technical Activities	
N	Administrative and Support Service Activities	
G	Wholesale and Retail Trade	Retail
I2	Food Service Activities	
P	Education	
Q	Health and Social Services	
R	Arts, Entertainment and Recreation	
S	Other Service Activities	

3.2. Estimating establishments' size

The extracted list from ACRA dataset has information on establishments' location, industrial type (and floor type), yet knowing establishments' occupied floor area and employment size is essential to create a complete sample. Given that no statistics in the ACRA sample can be used to estimate establishments' size, an indirect method was adopted. Data used for this task is provided by URA. Specifically, property transactions in Singapore

since 1995 have been recorded on URA's REALIS database. The transaction history includes information on property's floor area, type, and location, but not on buyers and sellers. Whenever two or more units were jointly sold, only the total floor areas were provided. However, based on the unit address, information can be deduced to obtain the number of units and average floor area.

Assuming that a unit's floor area is determined by its location, floor type, and building, a regression model can be calibrated for units' floor area. Unit's characteristics such as building type and location with regards to shopping mall and public transit stations were identified by matching the transaction data with building data (Zhu & Ferreira 2015). Given their distinctive characteristics, commercial and industrial properties were treated separately. Table 3 explains the variable coding and Table 4 shows the results of the regressions on the two property types.

Table 3: Variable coding

Variable name	**Explanation**
pt_retail	Property type retail
pt_office	Property type office
pt_factory	Property type factory
freehold	Property located in a freehold building
floor_space	Estimated building area
distance_mrt	Distance to the closest metro station
distance_bus	Distance to the closest bus stop
distance_express	Distance to the closest express way
distance_pms30	Distance to the closest primary school
distance_cbd	Distance to the central business district
distance_mall	Distance to the closest mall
west	Property located in the west region
northeast	Property located in the northeast region
north	Property located in the north region
east	Property located in the east region
underground	Property located at the underground level of the building
low_floor	Property located at the low levels of the building (level 5 and below)
high_floor	Property located at the high levels of the building (level 6 and above)
whole_building	Property occupied the whole building
bt_commercial	Property located in commercial building type
bt_residential	Property located in residential building type
bt_mixedresidential	Property located in mixed residential building type

Table 4: Regression results

REALIS commercial property transactions 1995-Oct 2015
16348 observations
Dependent variable: floor area

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.733e+00	9.689e-02	38.531	< 2e-16	***
pt_retail	4.555e-01	6.499e-02	7.010	2.48e-12	***
pt_office	9.348e-01	6.477e-02	14.433	< 2e-16	***
freehold	-8.376e-02	2.164e-02	-3.871	0.000109	***
floor_space	-2.098e-07	1.178e-07	-1.781	0.074918	.
distance_mrt	4.997e-02	1.498e-02	3.336	0.000851	***
distance_bus	4.478e-02	1.363e-01	0.328	0.742562	
distance_express	-4.150e-02	1.653e-02	-2.510	0.012080	*
distance_pms30	1.043e-01	1.286e-02	8.112	5.31e-16	***
distance_cbd	-1.471e-02	4.591e-03	-3.204	0.001357	**
distance_mall	-1.028e-01	1.866e-02	-5.510	3.64e-08	***
west	3.648e-01	7.418e-02	4.918	8.84e-07	***
northeast	-1.435e-01	1.166e-01	-1.230	0.218612	
north	3.503e-01	1.226e-01	2.858	0.004269	**
east	1.772e-01	4.611e-02	3.842	0.000122	***
underground	-1.303e-01	2.880e-02	-4.523	6.14e-06	***
low_floor	-7.766e-02	1.933e-02	-4.017	5.91e-05	***
high_floor	4.709e-01	2.905e-02	16.211	< 2e-16	***
whole_building	1.787e+00	6.410e-02	27.886	< 2e-16	***
bt_commercial	-6.598e-01	6.436e-02	-10.253	< 2e-16	***
bt_residential	-3.544e-01	7.250e-02	-4.888	1.03e-06	***
bt_mixedresidential	-6.795e-01	6.351e-02	-10.700	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7837 on 16348 degrees of freedom
Multiple R-squared: 0.3912, Adjusted R-squared: 0.3904
F-statistic: 500.2 on 21 and 16348 DF, p-value: < 2.2e-16

REALIS factory/warehouse property transactions 1995-Oct 2015
23689 observations
Dependent variable: floor area

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	5.413e+00	7.653e-02	70.732	< 2e-16	***
pt_factory	-6.045e-03	1.704e-02	-0.355	0.72277	
freehold	-6.573e-01	3.783e-01	-1.737	0.08237	.
floor_space	1.685e-06	7.711e-08	21.853	< 2e-16	***
distance_mrt	-8.385e-02	1.298e-02	-6.457	1.09e-10	***
distance_bus	-4.357e-01	4.313e-02	-10.104	< 2e-16	***
distance_express	-9.425e-03	7.540e-03	-1.250	0.21132	
distance_pms30	-6.739e-03	5.536e-03	-1.217	0.22356	
distance_cbd	1.054e-02	2.697e-03	3.906	9.39e-05	***
distance_mall	8.847e-02	1.029e-02	8.598	< 2e-16	***
west	8.259e-02	2.704e-02	3.055	0.00225	**
northeast	-2.028e-01	2.879e-02	-7.044	1.92e-12	***
north	1.698e-01	3.157e-02	5.379	7.58e-08	***
east	1.284e-01	1.831e-02	7.012	2.42e-12	***
underground	5.578e-02	8.018e-02	0.696	0.48663	
low_floor	-3.310e-01	1.475e-02	-22.438	< 2e-16	***
high_floor	-5.291e-01	1.555e-02	-34.022	< 2e-16	***
whole_building	2.161e+00	1.977e-02	109.309	< 2e-16	***
bt_industrial	-5.836e-02	7.322e-02	-0.797	0.42540	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6548 on 23689 degrees of freedom
Multiple R-squared: 0.6534, Adjusted R-squared: 0.6531
F-statistic: 2481 on 18 and 23689 DF, p-value: < 2.2e-16

The regression results were applied to the two subsets of firms. In particular, results from the regression model on commercial transactions was applied to services and retail establishments; and floor areas of manufactory establishments were calculated based on the model on factory and warehouse floor type. The floor area of establishments in the sample was estimated as the number of units occupied times the unit's floor area.

To predict establishments' employment size, establishments' floor areas were assumed to be correlated with number of employees. Average floor space per employee for different floor type α_{ft} was then used as a conversion factor between floor size and employment size.

$$\alpha_{ft} = \text{total floor space for floor type } ft \text{ (URA)/ total number of workers in floor type } ft \text{ (MOM)} \quad (1)$$

Specifically, the number of employees j^e in an establishment e of floor type ft is defined as:

$$j^e=(1/\alpha_{ft})\, f^e \quad (2)$$

Using the average number of space per worker to estimate an establishment's job size is a simplification. Even though the converting factors distinguish between four floor types, the differences between industry types, which are typically influential on floor space, were not captured. Furthermore, the method did not account for the distinction between large and small firms as large firms tend to have less space per worker. However, as no data were available for better classification, the simple converting factor of average floor space was used. The expected mismatch between the estimated numbers and the real total numbers can then be adjusted using the IPF method in the next step.

3.3. IPF

At the end of the previous step, we have a complete sample of establishments in Singapore with information on location, business type, floor area occupied, and number of employees. However, as these estimated numbers did not add up to the real total number of jobs in Singapore, IPF technique was adopted to adjust the numbers (Zhu & Ferreira 2014). To improve the accuracy of the job locations, the adjustment of jobs by industry was done at the planning area level. The sample of establishments registered at ACRA with estimated number of employees was used as the seed matrix. The numbers of jobs by floor type estimated by available floor space by planning area

provided by URA were used as row marginals, whereas MOM's statistics on employment by industry were column marginals. The procedure resulted in the adjusted number of jobs by industry type for each planning area in Singapore.

3.4. Distributing establishments and jobs

The total number of jobs adjusted by IPF was distributed among buildings within each planning area and floor type proportionally to the approximated buildings' occupied floor area. Buildings' commercial space were estimated based on building footprints, number of floors, and other characteristics as described in Zhu and Ferreira (2015).

The building dataset provides information on building type, estimated total space, and estimated floor area for each of the included 109,709 buildings in Singapore. For buildings that have non-residential floor space (49,728), occupied floor area for each for the four floor types were estimated as:

Occupied area of floor type *ft* = Assigned floor area of floor type *ft* * (1-vacancy rate of floor type *ft*) (3)

In buildings where assigned floor area was not known, occupied floor area were estimated based on the building types as shown in Table 5, and is defined as:

Occupied area of floor type *ft* = Building space * % of floor type *ft* * (1-vacancy rate of floor type *ft*) (4)

Table 5: Floor area assignment based on building type

Building type	**Floor type**	**% of floor type**
Office	Office	100%
Retail	Retail	100%
Mixed office and retail	Office	50%
	Retail	50%
Mixed residential and retail	Residential	50%
	Retail	50%
Residential	Residential	100%
Industrial	Industrial	100%
Warehouse	Warehouse	100%

Establishments and jobs were distributed to buildings according to several constraints. Specifically, the establishments *e*, in the resulting synthetic population are characterized by the establishment's building number *i*, their respective industry type *k*, the number of employees *j* and their occupied floor area *f*. The number of establishments, jobs, and occupied square meters in building *i* of industry type *k*, which are designated as $n^e_{i,k}$, $n^j_{i,k} = \sum_{e \in i,k} j_e$ and $n^f_e = \sum_{e \in ik} f_e$, respectively, are controlled by the marginal totals as follows:

$$\begin{cases} \sum_{i \in pa} n^j_{i,k} = N^j_{pa,k} \\ \sum_{e \in i,\, ft} n^f_e = N^f_{i,\, ft} \\ \sum_{i,k \in ft} n^e_{i,k} = N^e_{ft} \end{cases}$$

That is, in the generated synthetic population:

(1) the numbers of jobs in all buildings *i* of a planning area *pa* for a particular industry *k* add up to the total number of jobs in that industry in that planning area resulted from the IPF step.

(2) the assigned space to establishments in a building for a particular floor type adds up to the total number of floor space estimated for that floor type in that building.

(3) the number of establishments for a particular floor type *ft* in all buildings equals to the given total number of establishments for that floor type given by MOM.

4. Results and validation

4.1. Summary statistics

Table 6 compares the numbers of synthesized establishments and jobs by floor type with the real numbers provided by SingStats. In 2012, 160,371 establishments in Singapore employed 3,332,900 workers (SingStats, 2012). The distribution of jobs across four floor types in the synthetic population is very close to the real numbers. There are some slight deviations in the establishment numbers in each floor type. However, the total number still corresponds to the actual number.

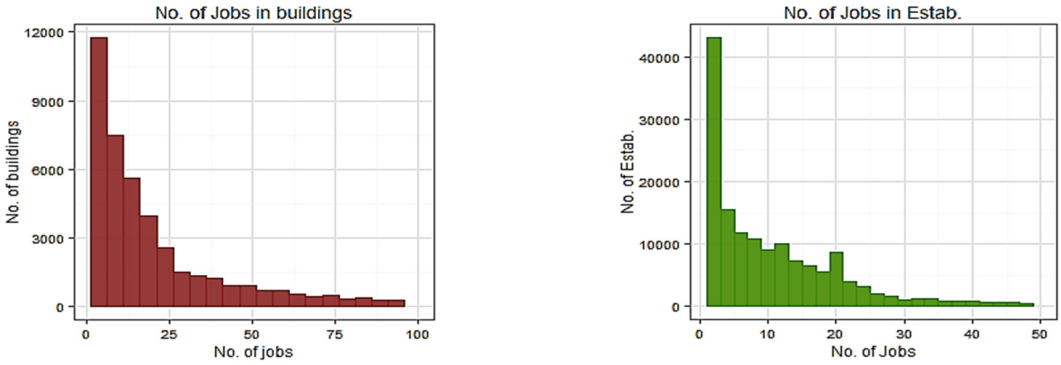
Table 6: National statistics vs. synthetic population

Floor type	**Establishments**		**Jobs**	
	Pop. Syn	*SingStats*	*Pop. Syn.*	*SingStats*
Office	58,732	139,718	1,170,157	2,580,200
Retail	76,132		1,340,522	
Manufacturing	15,217	9,577	533,620	535,000
Warehouse	9918	11,076	215,105	217,700
Total	160,000	160,371	3,259,404	3,332,900

4.2. Distribution of jobs in establishments and buildings

Jobs were distributed to buildings according to the available floor space as explained above. Figure 1a shows that the number of buildings decreases as buildings' job size increases i.e., more buildings have smaller number of jobs. On average, there are 70 jobs per building. However, the majority of buildings (67%) have fewer than 25 jobs, whereas there are only a small number (9%) of large buildings with at least 100 jobs. This reflects the reality that in Singapore there are a few large commercial buildings in the central and industrial buildings in the north and west of the city. Most often jobs and establishments are located in smaller-size buildings. It is also common that in many residential buildings, the first levels are used for commercial purposes, and the jobs assigned to these buildings are accordingly small.

Establishments' size was determined by the number of establishments and jobs assigned to the buildings. Similar to the distribution of jobs in buildings, establishments' size has an inverse relationship with the number of establishments (Figure 1b). The average number of employment in an establishment is 20 yet a large number of establishments (38%) has 5 employers or below. Most establishments (99%) in the synthetic population are small and medium enterprises (not more than 200 employers), which is the same with the reality in Singapore.



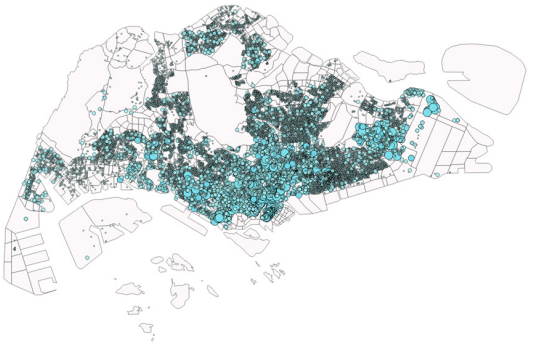
(a) Job distribution in buildings (b) Job distribution in establishments

Figure 1: Job distribution in buildings and establishments in the synthetic population

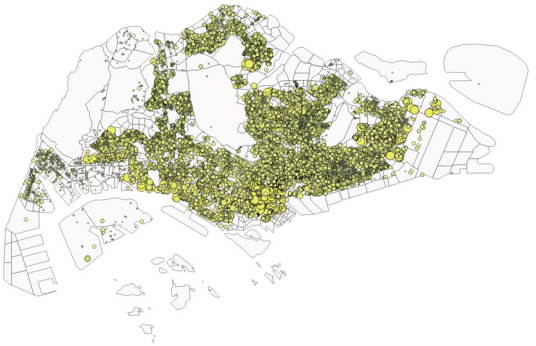
4.3. Establishments' locations

Singapore is an island state and has a land area of around 719km^2 and a population of approximately 5.3 million in 2012. The country is divided into five regions, which are further classified into 55 planning areas and 1169 zones. Figure 2 shows the locations of the synthetic establishments in Singapore by employment size. Retail and office establishments have similar distributions. These types of establishments spread across the entire island with a high concentration in the central area. Manufacturing and warehouse establishments share some similarities in their location patterns. They mostly distribute in areas far from the city center, particularly in the west, north and east, at the edge of the island. Figure 3 describes the distribution of establishments by showing the concentration of establishments at the zonal level (darker colors represent larger numbers). A high number of office establishments are located in the central business district and nearby areas (Figure 3a). Retail establishments are dense in the central and Orchard areas (Figure 3b). Most manufacturing establishments are in the west, north, and east regions (Figure 3c). There is a high concentration of warehouse establishment in the west and east of Singapore.

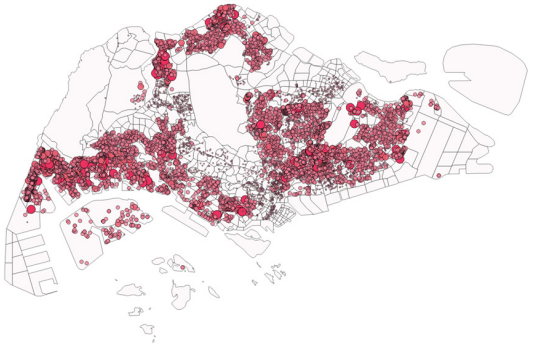
The locations of the synthetic establishments reflect the land allocation in 2012 for industrial and commercial property by SLA. It also captures the location distribution of establishments in the study sample i.e. the ACRA dataset.



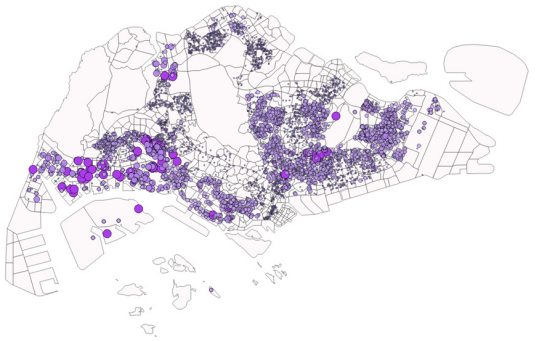
(a) Location of synthetic office establishments



(b) Location of synthetic retail establishments



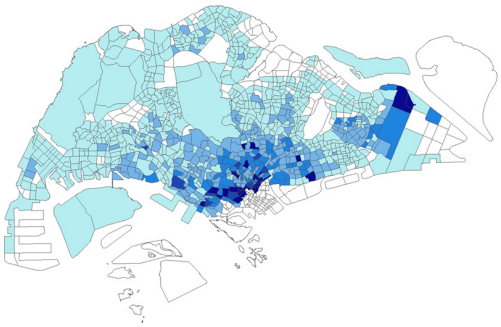
(c) Location of synthetic manufacturing establishments



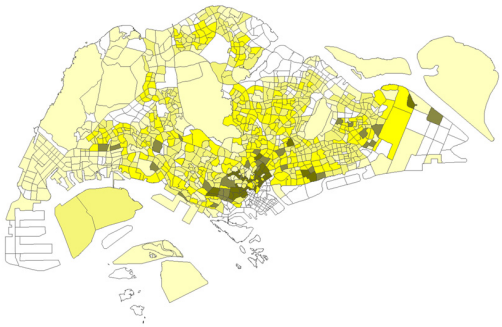
(d) Location of synthetic warehouse establishments

Under 10 employees | 10-99 employees | 100-199 employees | 200 employees and more

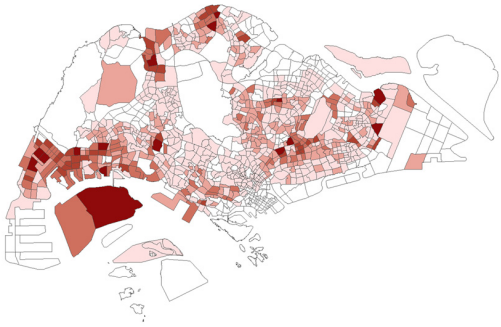
Figure 2: Location of all establishments in the synthetic population



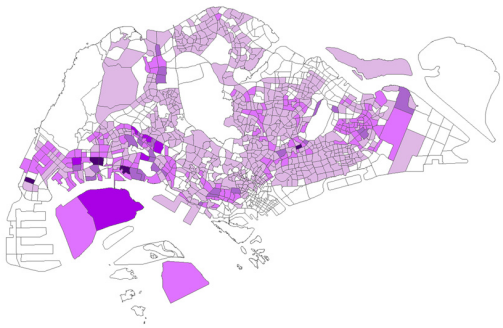
(e) Location of office establishments at zonal level



(f) Location of retail establishments at zonal level



(g) Location of manufacturing establishments at zonal level



(h) Location of warehouse establishments at zonal level

Figure 3: Location of establishments by floor type at zonal level

5. Conclusions

This paper has presented a viable method for generating a synthetic population of establishments for a metropolitan area using relatively sparse data, together with a description of the synthetic population produced through this method for Singapore in 2012. The population synthesis includes information on the number of establishments and the floor area by industry sector, and the number of jobs by occupation type for each building in Singapore. The procedure involves four basic steps. First, to estimate establishments' floor area, we ran linear regression models for units' floor area using official property transactions data from URA. A unit's floor size is believed to be a function of its floor type, location, and building characteristics. The results were then applied to the two subsets of firms. In particular, results from the regression model on commercial transactions was applied to services establishments; and floor areas of manufactory establishments were estimated from the model on factory and warehouse floor type. Next, we assumed establishments' floor areas are correlated with their employment size and thus estimated an establishment's number of workers based on its floor area. Average floor space per employee for different floor type was used as a conversion factor between floor size and employment size. Iterative Proportional Fitting was used to adjust the number of jobs; in the IPF, the control total for number of jobs is set at the planning area level. Official statistics from different authorities were used as marginal controls and data from ACRA contributed to create the seed table. Finally, the total number of jobs is distributed to individual buildings

within each planning area and floor type proportionally to the approximated buildings' occupied floor area. Buildings' volumes were estimated by fusing data from multiple sources with footprint layer obtained from SLA.

The highlights of the resulted synthetic population are:

- The numbers of jobs in each industry type and in each planning area are close to real numbers.
- The numbers of buildings that have jobs are similar to the number of buildings in the ACRA dataset.
- The establishments are relatively well distributed among the buildings according to the availability of suitable floor space.

In sum, a synthetic population of establishments in Singapore was generated for the purpose of modeling firms' behavior in the SimMobility platform. The methods can be applied to any other cases provided the required (or similar) data are available: (1) a sample of establishments at the building geographical level, (2) aggregate data on employment distribution by industry and by area, (3) an approximation of floor area per employee, and (4) a building population with information on each building's floor area by general business type. Improvements however are possible in some areas such as establishment distribution among industry types, which currently have some deviations from the real number. In this population the jobs in one floor type are more evenly distributed among the industry types of that floor type than in reality. Currently only traditional types of jobs were included in the population, the methods may thus need to be modified to accommodate for future job types such as those with high flexibility in location and work hours. More detailed classification of industries will also be beneficial for the understanding firm behaviours, especially with regarding to the study of production and consumption of freight modelling.

Acknowledgements

This research is part of the SimMobility project funded by the Singapore National Research Foundation through the Singapore–MIT Alliance for Research and Technology Center for Future Mobility. The authors appreciate the support of the Accounting and Corporate Regulatory Authority, the Singapore Urban Redevelopment Authority, and Singapore Land Transport Authority on the collection of the ACRA datasets, the REALIS data set, and other helpful information.

References

Adnan, M. et al., 2016. SimMobility: A Multi-scale Integrated Agent-based Simulation Platform. In *Paper Presented at the 95th Annual Meeting of the Transportation Research Board Forthcoming in Transportation Research Record*.

Beckman, R.J., Baggerly, K.A. & McKay, M.D., 1996. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30(6), pp.415–429. Available at: http://www.sciencedirect.com/science/article/pii/0965856496000043 [Accessed January 27, 2016].

Ben-Akiva, M.., 2010. SMART – Future Urban Mobility. *Journeys*, (November). Available at: http://www.lta.gov.sg/ltaacademy/doc/J10Nov-p30Ben-Akiva_FutureUrbanMobility.pdf.

Cernicchiaro, G. & Ferreira, J., 2015. How to build a synthetic population for the service sector using directory websites. In *Paper presented at the 14th International Conference on Computers in Urban Planning and Urban Management*.

Department of Statistics, Singapore Statistics 2012. Available at: http://www.singstat.gov.sg/.

Khan, A.S., Abraham, J.E. & Hunt, J.D., 2002. Agent-based micro-simulation of business establishments. In *42nd ERSA Congress Dortmund*.

Lu, Y. et al., 2015. SimMobility Mid-Term Simulator: A State of the Art Integrated Agent Based Demand and Supply Model. In *In Transportation Research Board 94th Annual Meeting*.

Maoh, H.F. & Kanaroglou, P.S., 2005. Agent-Based Firmographic Models: A Simulation Framework for the City of Hamilton. In *PROCESSUS Second International Colloquium on the Behavioural Foundations of Integrated Land-use and Transportation Models: Frameworks, Models and Applications,*. Available at: http://www.civ.utoronto.ca/sect/traeng/ilute/processus2005/PaperSession/Paper08_Maoh-Kanaroglou_Agent-basedFirmographicModels_CD.pdf.

Moeckel, R., 2009. Simulation of Firms as a Planning Support System to Limit Urban Sprawl of Jobs. *Environment and Planning B: Planning and Design* , 36 (5), pp.883–905. Available at: http://epb.sagepub.com/content/36/5/883.abstract.

Rich, J. & Mulalic, I., 2012. Generating synthetic baseline populations from register data. *Transportation Research Part A: Policy and Practice*, 46(3), pp.467–479. Available at: http://www.sciencedirect.com/science/article/pii/S0965856411001716 [Accessed January 27, 2016].

Zhu, Y. & Ferreira, J., 2015. Data integration to create large-scale spatially detailed synthetic populations. In S. Geertman et al., eds. *Planning Support Systems and Smart Cities*. Heidelberg: Springer, pp. 121–141.

Zhu, Y. & Ferreira, J., 2014. Synthetic Population Generation at Disaggregated Spatial Scales for Land Use and Transportation Microsimulation. *Transportation Research Record: Journal of the Transportation Research Board*, 2429, pp.168–177. Available at: http://dx.doi.org/10.3141/2429-18.