How to build a synthetic population for the Service Sector using directory websites.

Giulia Cernicchiaro and Joseph Ferreira

Abstract

Study of firm (re)location behavior occupies an important place in urban dynamics literature. However, spatially detailed data about firm establishments are often unavailable or hard to collect. Directory websites provide a voluminous source of spatially-detailed data that could be especially helpful in quantifying the characteristics of non-work destinations at people-scales and in formulating firm (re)location models at building and establishment scales.

This paper builds a synthetic population of firm establishments, using mostly data that are available online. Official statistics of different governmental entities are used as marginal controls. Basemaps of parcels and building footprints provided by the Singapore Land Authority (SLA) and a set of real estate transaction data for commercial units provided by the Urban Redevelopment Authority (URA) are also utilized.

G. Cernicchiaro (Corresponding author)
IRG Future Urban Mobility, Singapore-MIT Alliance for Research and Technology, 138602 Singapore, Singapore
Email: giulia@smart.mit.edu

J. Ferreira
Department of Urban Information Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, US
Email: jf@mit.edu

1. Introduction

Firm (re)location behavior is an important part of understanding urban dynamics. Not only do their choices impact urban land use, but firms also locate jobs and activities that impact individual mobility. Moreover, commodity movement, within firms and from firms to final consumers, generates freight traffic impacting urban mobility. To cite a few, Bodenmann (2011), Ciari (2011) and DeBok (2011) studied this problem.

Spatially detailed data about firm establishments are often unavailable or hard to collect. Due of this lack of data, firm locations are often modeled at aggregate scales in the literature (Anas and Liu (2007)). Even when a micro econometric approach is adopted, location choice is not considered at the building level but aggregated at a zonal level (Bodenmann (2011)) or the study focus on a restricted set of firms (Ciari (2011), Nguyen et al. (2013)).

However, a great amount of information, especially for firms operating in the service sector, is provided by business directory websites. Such websites are designed to provide users with the location of points of interest (POI) for undertaking a broad range of urban activities. Although such data are loosely structured with uneven reliability, they provide a voluminous source of spatially-detailed micro data that could be especially helpful in quantifying the characteristics of non-work destinations at people-scales and the freight volume, and in formulating firm (re)location and individual workplace choice models at building and establishment scales.

An important issue to face when using this kinds of data is the highly disaggregated business activity classifications provided by the website. This classification mostly does not match the national codification of industry (or business activity). Jiang et al. (2014) highlight the problem and propose a machine learning method to solve it.

If one could observe a sample of establishments through a directory website, then one could use that information to estimate characteristics of the whole population while matching official aggregated statistics on firm characteristics. Methods for estimating such a synthetic population can be found in the literature on residential location, and Iterative Proportional Fitting (IPF) methods (Deming and Stephan (1940), Zhu and Ferreira (2013)).

This paper presents a machine learning procedure to build a sample of establishments with information on their location, industry and size, using web data. A method is then developed to synthesize the full establishment population based on this sample. To illustrate the methodology, the synthesis method is applied to 2008 data for Singapore.

Following SingStat definition, an establishment is an entity (a mono-branch firm or a branch of a multi-branch firm) operating in a unique location and characterized by the industry sector of its principal activity. Each of these entities is composed of multiple jobs and occupies a floor area.

Translating for the firms' case what proposed for households in Zhu and Ferreira (2013), we can treat establishments, jobs, and floor area as three related populations to be simultaneously estimated. Such an approach improves the estimation's precision. Moreover, if one wants to link this population to synthetic workers and buildings, independently available totals for these two sub populations have to be respected. These marginal could not correctly match if the number of jobs and occupied floor area were estimated separately for each establishment.

What is estimated here is the number of establishments, jobs and occupied square meters, at the building level by industry. Jobs are not packed into individual establishments.

The paper is structured as follows. First, we identify available data and the work needed to transform them into an appropriate establishment sample completes of all information for which marginal totals are observed. In order to build the needed initial sample, an automatic procedure is developed to clean and format data from directory websites. A machine learning method is defined to deduce a standardized industry code based on the disaggregated information provided by the website. Space occupied by each establishment is estimated through a two-steps method: a regression model is calibrated to estimate the floor area, when possible, and an approximation of the latter, based on available space within postal code (when known), is attached to the remaining establishments.

In Singapore statistics on establishments, jobs and floor area are provided by different governmental entities. As a consequence, agency distributions are available only for part of the needed characteristics for each population. A set of initial hypotheses about the links among agency datasets are then defined in order to estimate the full set of marginal totals.

Next, we identify additional estimation issues and decompose the task into sub problems associated with different geographical level. A multi-steps approach to solve this decomposed problem is then suggested.

In conclusion, after summarizing the key points, possible improvements in the method will be suggested and future work will be discussed.

2. Data preparation

The objective here is to build a synthetic population of establishments that were operating in Singapore in 2008. That is, we want to estimate the total number of establishments, jobs and occupied floor area by industry for each existing building. In order to perform such a task a minimum of information is needed on existing establishments, as well as on the distribution of total population across the island.

Singapore is composed of 5 regions. Each of these regions is composed of sub-zones, called planning area. Singapore accounts for 55 planning areas. About 163k buildings are distributed over these areas.

2.1 Establishments Sample

An important issue that we have to face here is the absence of satisfactory micro data. Although each firm operating in Singapore has to provide the Accounting and Corporate Regulatory Authority (ACRA) with its head office address, only some business entities are asked to declare the physical address of their branches.

Only Singapore's Department of Statistics (DOS) interviews establishments, collecting mostly financial and no geographical information. Principal activity and total number of workers are collected, but cannot be linked with their location. Some information on POI[1] is provided by Singapore Land Authority (SLA). However, available data are limited and only 5% of the total establishment population is observed.

Moreover, during this study a pilot survey has been performed to test retailers' willingness to answer some question on location and workforce. The results suggest that these kinds of information are considered sensitive and hence hard to collect. A possible solution is provided by directory websites. In Singapore POI data are provided online by multiple websites, such as StDirectory[2] and StreetDirectory to cite only a few.

[1] Here we will use points of interest (POI) and establishments as synonyms.

[2] StDirectory was the most extensive online business information site that we accessed. StDirectory is provided by Singapore Press Holdings Ltd. Co. (http://www.sph.com), whose online and new media initiatives include ST701 (http://www.st701.com/) which is dedicated to providing the most comprehensive repository and platform to meet advertisers' & consumers' needs. StDirectory (http://directory.stclassifieds.sg/) is part of this consumer-to-consumer online marketplace.

Such websites' goal is to provide users with POI's location.
In order to increase user's probability to find what he's looking for, each observed establishment is attached to a list of classes describing the whole set of its activities. Information on establishment's activities is hence provided at a highly detailed level. As observed in Jiang et al. (2014), this level of detail is deeper than the one adopted in National Industry Classification.
In Singapore this classification is defined by Singapore Standard Industry Classification (SSIC) and is structured on six levels of details. Most of the time macro statistics are provided at an even more aggregated level than the first SSIC level. The SSIC's first level consists in 24 classes, reported in Table 1[3], versus about 6k classes used in the web data that we examine.
Moreover, such a website doesn't provide information on occupied floor area and/or employment size.
A preparation work is hence needed in order to build a sample of establishments with location (postal code), activity type (SSIC in Table 1) and occupied floor space[4] (square meters). A modular automatic procedure, here presented, is then developed to perform this task.

Available web data are structured as follow. Each record is composed of the establishment's name and complete address. Other information that will be not of use here are also available, such as provided services and products, firm's structure (mono or multi-branch) and the establishment's role (head office or branch).
The address, formatted according to Singapore's standard, is a string of the form:

```
#floor(s)-unit(s) & Number & Street & Singapore & Postcode
```

An automated module is used to parse the provided address in order to detect the postal code, the position within the building (underground, low floor, high floor), and number of occupied units. This procedure also creates two variables to register building number and street name.

[3] Only classes representing activity types here considered are reported.

[4] Workforce size is estimated as function of floor space. Readers can refer to next section for a presentation of this task.

Table 1. SSIC with related Floor Type and Aggregated Class

SSIC Class	Floor Type	Aggregated Class
Accommodation	H	AF
Food and beverage service activities	R	AF
Wholesale trade	R	RW
Retail trade	R	RW
Education	R	CSP
Health and social services	R	CSP
Arts, entertainment And recreation	R	CSP
Other service activities	R	CSP
Professional, scientific and technical activities	O	RB
Real estate activities	O	RB
Administrative and support service activities	O	RB
Financial and Insurance activities	O	FN
Information and communications	O	IC
Transportation and storage	W	TP
Manufacturing	I	MAN

. Reported aggregated classes are those used in publishing statistics by Singapore Department of Statistics.

. H: Hotel; R: Retail; O: Office; W: Warehouse; I: Industrial.

StDirectory provide an activity classes' index enabling a direct access to the whole list of referenced establishments attached to each class. Data are hence crawled from the web sequentially for each class in the website's activity classification. That is, for each observed establishment we will crawl as many records as the number of classes describing its full set of possible activities. We have hence to solve the problem of merging multiple records for the same establishment. A possible solution is to aggregate records by name. However, different branches of a same firm can be registered under the same name in a directory database. Moreover, a preliminary semi-automatic analysis suggested the existence of establishments registered twice, where one record seems to have partially incorrect location information.

A second module performs such aggregation by distinguishing three cases.

- When two or more identical records are observed under more than one class only one record is conserved and the list of related classes is registered;
- When part of the address is different for two records with the same name, within a same class or not, only one record is conserved with its list of classes (when needed) and corrected address. The correction task is performed by matching observed data with a complete list of existing Singapore's address with related postal code;
- When two records with the same name have different information both are conserved.

About 384k records are initially observed. The module performs records' aggregation sequentially. First, only case 1 is treated. This produce 95.5k aggregated records, 97% of which a unique address is systematically declared for records registered under the same establishment's name. Tasks 2 and 3 are hence performed for the remaining 3%, producing 94.6k unique records with corrected address info. Of these records, less than 0.2% has no observed postal code.

This module's output is an initial set of 94399 records with location and list of activities. The next module receives these data and deduces an aggregated activity code based on the disaggregated information provided in input.

As in Jiang et al. (2014), we dispose of a training sample (46k observations) of web data matched with industrial classification down to the first SSIC level (Table 1), containing less information than the one obtained from StDirectory that we want to use. However, here the disaggregated industry classification definition used in the available training sample is not the same that in our sample. Moreover, trying to join these two classifications we obtain less than 10% classes successful matching. Because of this discrepancy, a method clustering training classes in first SSIC level could not be applied here, but we observe that single words can be easily matched. Even a simple algorithm matching only words identical or of a different form (singular and plural) allows finding more than 70% of the 3.7k words composing the 5.7k classes in our observed sample.

Another issue highlighted in Jiang et al. (2014) is the time needed to run methods they use. Here we need a faster method that can be easily integrated in a platform running simultaneously multiple algorithms.

An alternative machine learning method, with supervised learning, is hence proposed. This method aims to classify observed establishments down to first SSIC level, by choosing the SSIC code that maximizes probability that 1-gram elements composing its classes' labels be linked to the code in the training sample. Firstly, a measure of the probability that a given SSIC be attached to a 1-gram element is computed for each word encountered at least once in the training sample and linked to matched words in our observed sample. An equivalent measure is computed for each class in the observed sample by aggregation of this measure over words composing its label, and the same is repeated for each establishment over classes describing its activity. The algorithm hence picks the SSIC that maximize for each observed establishment this final aggregated measure, of the form $E_{classes\in establishment}\{E_{words\in class}[m(word, SSIC)]\}$.

The measure of probability that a word be found attached to a given SSIC in the training sample is defined following the here presented logic. We want this measure increase proportionally to the repartition by SICC of the word in training sample ($p(word, SSIC)$). In order to make this measure comparable over 1-gram elements we rank SSIC codes by repartition (from most to less represented) and use this rank ($r(word, SSIC)$) to normalize it. Finally we weight this measure by the number of times the word appears divided number of SSIC found attached to the word ($w(word)$). This is because we want to attach a lower weight to those words that has more sparse definition.

We hence have $(word, SSIC) = \frac{p(word,SSIC)*w(word)}{r(word,SSIC)}$.

The choice made for the aggregated measure for classes here is simply the average of previous measure $m(word, SSIC)$, over words composing each class label. This aggregated measure is weighted for each class by number of words composing its label possibly attached to a SSIC divided total number of possible match word-SSIC. Similarly to previous step, we normalize this repartition measure using the corresponding rank.

The same is done at the last step for observed establishments, where the average is computed over observed activity classes. Using such a measure we found less than 0.3% cases without a unique maximum.

Testing the method on the training sample we obtain 61% correctly classified establishments. This is 20% less than what obtained in Jiang et al. (2014). However, looking at this percentage per SSIC class (Table 2) we can see that the method works better for personal services (i.e. food and beverage, education, ...) than for business services and manufacturing sector.

Moreover, the method needs less than 5 minutes to run and attach a SSIC code to 99% input records.

We need hence to improve classification quality. An idea is to plug the obtained establishment population in 2008's building population in order to fix discrepancy between building type and the initial SSIC code, output of the previous module. Moreover, the preliminary analysis also records instances registered with a same location but under different names. These are to be identified and aggregated in a unique record for each corresponding establishment.

Table 2. Classification Method's Test Results

SSIC Class	N	C
Food and beverage service activities	4809	96%
Education	2331	84%
Other service activities	5747	77%
Accommodation	300	70%
Transportation and storage	3667	65%
Real estate activities	1802	62%
Health and social services	2544	62%
Retail trade	11086	62%
Financial and insurance activities	4218	60%
Professional, scientific and technical activities	8068	59%
Wholesale trade	16946	51%
Administrative and support service activities	6113	48%
Information and communications	3683	43%
Manufacturing	11216	43%
Arts, entertainment and recreation	777	26%

. N: number of establishments in the training sample.

. C: percentage of correctly classified establishments.

A fourth module performs these tasks by comparing names and industries of records for problematic case. The code first matches data with a 2008's building population, preliminarily prepared by deducing characteristics from multiple web sources for the whole list of building existing in Singapore5. It then identifies if records correspond to the same establish-

[5] The whole set of existing building, complete of location and base area, has been provided by Singapore Land Authority (SLA).

ment as well as industry of principal activity, on the basis of name similarity and industry compatibility both within records and with building type. This a semi-automatic method that use rules defined by direct observation of the sample. Through this observation those words systematically attached to a wrong SSIC within a given building type are highlighted in order to enable the procedure to correct the classification.

This produces 76.9k cleaned establishments with coherent activity and building type. Only 0.2% has now not found SSIC. Moreover about 3k records are classified under industries not treated here because of absence of macro statistics[6]. We dispose here of 73.5k establishments classified under SSIC in Table 1. The repartition of these establishments by aggregated class is reported in Table 3[7].

Table 3. Repartition by Aggregated Class, Sample vs Known Totals

	MARGINAL	SAMPLE	StD/DOS
AF	6258 4%	8079 11%	1.3
WR	55854 36%	21432 29%	0.4
CSP	25500 16%	18428 25%	0.7
RB	35036 22%	18201 25%	0.5
FN	9533 6%	1432 2%	0.2
IC	6767 4%	1749 2%	0.3
TP	9653 6%	897 1%	0.1
MAN	8640 5%	3319 5%	0.4
TOT	**157241**	**7537**	**0.5**

We can see that these correspond to half of the existing population. Some discrepancy with the observed population can be highlighted. Food and Beverage and Personal Services come out to be overrepresented, where Business Services are underrepresented. This is not surprising be-

[6] Agriculture And Fishing; Construction; Electricity, Gas And Air-Conditioning Supply; Mining And Quarrying; Public Administration And Defence; Water Supply; Sewerage, Waste Management And Remediation Activities.

[7] Manufacturing sector is deducted here mostly from additional data.

cause of our data source. That is, Business Directory Websites aim principally to provide final consumers with information on points of interest more than business. It's hence more common to found the location of a retailer or a restaurant on such kind of website. A possible solution would be to sample establishments within this set in order to have a more representative initial population. Here we continue, however, with the whole sample as the rate of observed establishments on known totals seems to be quite stable.

In order to have a complete sample we need now to estimate occupied floor space for each establishment. Data available for this task are provided by the Urban Redevelopment Authority (URA). These data consist of a set of observed transactions of commercial and industrial units since 1995, complete with the units' floor area and location. However, no information is provided about the units' buyer and/or occupier. Moreover units are often jointly sold. That is, a same agent can buy two or more contiguous (or at least close) units. When this happens, URA provides only the total floor area of the joint units. Applying the first module to each address string we can also deduce information about the location within the building and the number of occupied units.

Let suppose that a unit's floor area is a function of building and environment characteristics, as well as the unit's floor type and its location within the building. We can hence calibrate a regression model for a units' floor area, using URA data matched with 2008's building population. Two regression models have been calibrated, one for Service Sector and other for Manufacturing. Results for Services are reported in Figure 1.

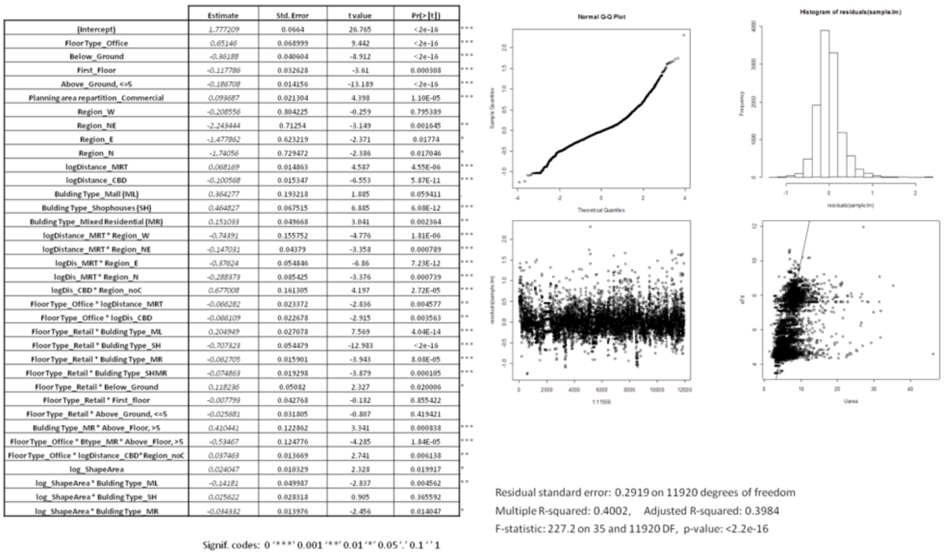
A fifth module uses this regression function to simulate units' floor area in our sample, when possible, and estimate establishment floor area as the simulated unit's floor area times the number of occupied units.

For remaining establishments, a last module approximates floor area based on available space (i.e. a whole building, or a whole floor within a building) when known.

Figure 2 shows distributions of establishments by industry and floor area by geographical location in the sample obtained in the output of the modular procedure. We can compare the obtained distribution by aggregated industry class with the one provided by DOS, for the Service Sector, and Singapore Economic Development Board (EDB), for the Manufacturing Sector. We can see that the two are quite closes.

Floor area repartition is compared to data that URA provides. We can see that commercial floor area (Office and Retail) seems to be underrepresented in the Central Region. Industrial floor area distribution seems quite far from the observed one.

Figure 1. Regression Specification and Results for Service Sector

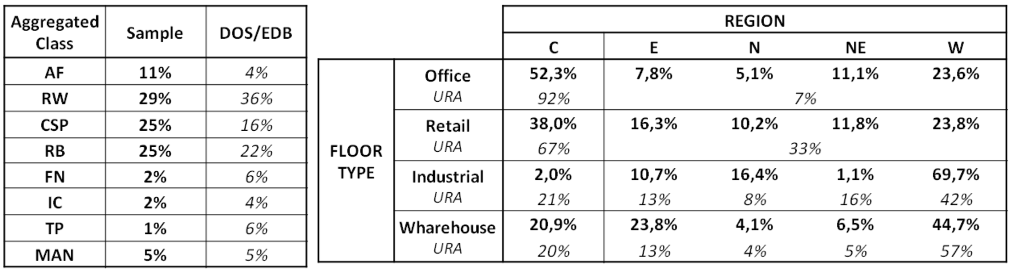


	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.777209	0.0664	26.765	<2e-16	***
FloorType_Office	0.65146	0.068999	9.442	<2e-16	***
Below_Ground	-0.36188	0.040604	-8.912	<2e-16	***
First_Floor	-0.117786	0.032628	-3.61	0.000308	***
Above_Ground, <=5	-0.186708	0.014156	-13.189	<2e-16	***
Planning area repartition_Commercial	0.093687	0.021304	4.398	1.10E-05	***
Region_W	-0.208556	0.804225	-0.259	0.795389	
Region_NE	-2.243444	0.71254	-3.149	0.001645	**
Region_E	-1.477862	0.623219	-2.371	0.01774	*
Region_N	-1.74056	0.729472	-2.386	0.017046	*
logDistance_MRT	0.068169	0.014863	4.587	4.55E-06	***
logDistance_CBD	-0.100568	0.015347	-6.553	5.87E-11	***
Bulding Type_Mall (ML)	0.364277	0.193218	1.885	0.059411	.
Bulding Type_Shophouses (SH)	0.464827	0.067515	6.885	6.08E-12	***
Bulding Type_Mixed Residential (MR)	0.151033	0.049668	3.041	0.002364	**
logDistance_MRT * Region_W	-0.74391	0.155752	-4.776	1.81E-06	***
logDistance_MRT * Region_NE	-0.147031	0.04379	-3.358	0.000789	***
logDis_MRT * Region_E	-0.37624	0.054846	-6.86	7.23E-12	***
logDis_MRT * Region_N	-0.288373	0.085425	-3.376	0.000739	***
logDis_CBD * Region_noC	0.677008	0.161305	4.197	2.72E-05	***
Floor Type_Office * logDistance_MRT	-0.066282	0.023372	-2.836	0.004577	**
Floor Type_Office * logDis_CBD	-0.066109	0.022678	-2.915	0.003563	**
Floor Type_Retail * Bulding Type_ML	0.204949	0.027078	7.569	4.04E-14	***
Floor Type_Retail * Bulding Type_SH	-0.707323	0.054479	-12.983	<2e-16	***
Floor Type_Retail * Bulding Type_MR	-0.062705	0.015901	-3.943	8.08E-05	***
Floor Type_Retail * Bulding Type_SHMR	-0.074863	0.019298	-3.879	0.000105	***
Floor Type_Retail * Below_Ground	0.118236	0.05082	2.327	0.020006	*
Floor Type_Retail * First_floor	-0.007793	0.042768	-0.182	0.855422	
Floor Type_Retail * Above_Ground, <=5	-0.025681	0.031805	-0.807	0.419421	
Bulding Type_MR * Above_Floor, >5	0.410441	0.122862	3.341	0.000838	***
Floor Type_Office * Btype_MR * Above_Floor, >5	-0.53467	0.124776	-4.285	1.84E-05	***
Floor Type_Office * logDistance_CBD* Region_noC	0.037463	0.013669	2.741	0.006138	**
log_ShapeArea	0.024047	0.010329	2.328	0.019917	*
log_ShapeArea * Bulding Type_ML	-0.14181	0.049987	-2.837	0.004562	**
log_ShapeArea * Bulding Type_SH	0.025622	0.028318	0.905	0.365592	
log_ShapeArea * Bulding Type_MR	-0.034332	0.013976	-2.456	0.014047	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2919 on 11920 degrees of freedom
Multiple R-squared: 0.4002, Adjusted R-squared: 0.3984
F-statistic: 227.2 on 35 and 11920 DF, p-value: <2.2e-16

Figure 2. Output Sample's Distribution



Aggregated Class	Sample	DOS/EDB
AF	11%	4%
RW	29%	36%
CSP	25%	16%
RB	25%	22%
FN	2%	6%
IC	2%	4%
TP	1%	6%
MAN	5%	5%

		REGION				
		C	E	N	NE	W
FLOOR TYPE	Office	52,3%	7,8%	5,1%	11,1%	23,6%
	URA	92%	7%			
	Retail	38,0%	16,3%	10,2%	11,8%	23,8%
	URA	67%	33%			
	Industrial	2,0%	10,7%	16,4%	1,1%	69,7%
	URA	21%	13%	8%	16%	42%
	Wharehouse	20,9%	23,8%	4,1%	6,5%	44,7%
	URA	20%	13%	4%	5%	57%

2.2 Marginal Totals

The whole list of available marginal distributions is reported in Table 4. We can see that no official statistics on establishments' geographical distribution is available. DOS collects a set of data on establishments operating in the Service Sector, interviewing a sample of firm registered with ACRA. However, no information on establishment's location is included here or in the published macro statistics. EDB publishes equivalent statistics for Manufacturing Sector. Here too, no geographical statistic is estimated.

Table 4. Available Macro Statistics

Establishments' Distribution by Aggregated Industry and Workforce Class Size	
Service Sector	DOS
Manufacturing	EDB
Jobs' Distribution by Industry (include vacancy)	MOM
SQMs' Distribution by Planning Area (2011)	URA
SQMs' Distribution by Region	URA

The geographical distribution of jobs can be observed from the 2008 Household Interview Transport Survey (HITs), where individuals are asked to provide Land Transport Authority (LTA) with their workplace's location. However, in 2008 only worker's occupation and not industry sector of his employer was collected. These data are hence not useful here. Only floor area's geographical distribution is provided.

As reported in Table 4, distribution by planning area is provided only for 2011. We observe, however, distribution at region level for 2008. Moreover, both are provided by floor type. We can hence estimate a repartition at the planning area by floor type in 2008 by application of 2011's distribution within the region.

We have now:

- Number of establishments by industry and class of workforce size;
- Number of jobs by industry and class of establishment's workforce;
- Number of occupied square meters by planning area and floor type.

3. Population Synthesis

Our objective here is to build a synthetic population of establishments e with information on their location ($e \in i$), industry ($e \in k$) and size (workforce j_e8 and occupied floor area f_e). That is, we look for number of establishments (n_{ik}^{e}), jobs ($n_{ik}^{j} = \sum_{e \in ik} j_e$) and occupied square meters ($n_{ik}^{f} = \sum_{e \in ik} f_e$) by building by floor type (i, ft) and industry by size k, for which

[8] Total number of jobs, including both vacant and occupied jobs.

$$\begin{cases} \sum_i n_{ik}^c = N_k^c \\ \sum_k n_{ik}^c = N_{i,ft}^c \end{cases}, \quad \forall\, ft, k \in ft, c \in \{e, j, f\} \tag{3.1}$$

Zhu and Ferreira (2013) proposed a method to solve simultaneously these problems for households and individuals ($c \in \{h, i\}$) when building a synthetic household's population with residential location. However, that proposed method cannot be applied in the establishment and job case. First, the number of individuals within a household is generally smaller than the number of jobs per establishment and has a known upper limit. Moreover, marginal totals are provided by classes of workforce size that are too large to apply the Zhu and Ferreira (2013) method. Accordingly, we need to define an alternative solution method.

As suggested in Zhu and Ferreira (2013), this problem can be decomposed into two subsequent problems at two different levels of geographical aggregation. Let $\{z\}$ be a partition of the whole considered geographical zone9 at a more aggregated level than our set of locations $\{i\}$. We have hence $\sum_{i \in z} N_{i,ft}^c = N_{z,ft}^c \ \ \forall\, c \in \{e, j, f\}$ and we can write the problem as two subsequent constrained problems

$$\forall\, ft, k \in ft, c \in \{e, j, f\} \ : \ N_{zk}^c \ u.c. \begin{cases} \sum_i N_{zk}^c = N_k^c \\ \sum_k N_{zk}^c = N_{z,ft}^c \end{cases} \tag{3.2}$$

$$\rightarrow \quad n_{ik}^c \ u.c. \begin{cases} \sum_{i \in z} n_{ik}^c = N_{zk}^c \\ \sum_k n_{ik}^c = N_{i,ft}^c \end{cases}$$

where $\bar{N}_{zk}^c$ and $\bar{n}_{ik}^c$ are observed total, computed from our initial establishments' sample.

We also know that the number of establishments and jobs are linked by the establishment's size, $n_{ik}^j / n_{ik}^e = \bar{n}_{ik}^j \in \left[n_k^{j-};\, n_k^{j+}\right] \ \forall\, i, \forall\, k$.

[9] Here planning area, Singapore's partition.

Moreover, each establishment's employee occupies a physical space. One can hence suppose that an establishment's floor area f_e is function of its workforce j_e, $f_e = g(j_e)$, $\forall\ e$

Data available for our study case are limited and mostly obtained from web sources. We begin with a sample of establishments complete with their location, industry and occupied floor area. Also, we apply our method to the whole population of existing buildings with estimated floor area by floor type. Finally, we know marginal totals by industry and workforce size for establishments and jobs, and floor area by geographically aggregated location (planning area) and floor type.

A first issue in estimating the number of jobs is that we don't observe workforce in our sample, so we could hence not build a seed table for number of jobs. As previously said, we can treat jobs and floor area as linked through a function $g()$. We can hence estimate establishments' workforce in the sample as $j_e = g^{-1}(f_e)$.

We will adopt here the simplified hypothesis that the same space is assigned to each job occupying the same floor type. That is, we estimate average floor area per job by floor type $a_{ft} = \left(\sum_{k\in ft} N_k^f\right)/\left(\sum_{k\in ft} N_k^j\right) = a_k\ \forall\ k \in ft$ and hence $j_e = \frac{1}{a_{k_e}} f_e$, $\forall\ e$. As total number of jobs (N_k^j) includes vacant jobs, j_e here estimated is total number of jobs, vacant and occupied, per establishment.

Also, we never observe both N_k^c and $N_{z,ft}^c$ for a same c. However, under previous hypothesis we can estimate $N_{z,ft}^j$ as $\frac{1}{a_{ft}} N_{z,ft}^f$. We then know both N_k^j and $N_{z,ft}^j$ for jobs.

Another obstacle to face is that no marginal is available at building level. While we observe building size10, we don't know the vacancy rate within the building. However, using building type, which is available, we can estimate the distribution by floor type within each building. A possible solution is hence to approximate the occupied building's floor area of each floor type as

$$\text{total floor area } * \text{ percentage of floor type within building } * \\ (1 - \text{building vacancy rate for floor type}) ,$$

using observed vacancy rate by floor type and region.

Finally, with available data we cannot solve the problem for the total number of establishments (c = e). However, as previously saw, we know that number of establishments and jobs within a building are related

[10] Computed as base area time number of floors.

through average establishments' workforce, $n_{ik}^{j}/n_{ik}^{e} = \bar{n}_{ik}^{j}$. We can hence estimate average workforce by industry and workforce's class per building from our sample as $f(x_i, x_e, x_{ie})$, where x_i is a vector of building's characteristics (including its environment), $x_e = k$ are establishment's industry and workforce size class and $x_{ie} = x_u$ are characteristics of unit u occupied by establishment e within building i.

The synthesis method is hence defined. Firstly the total numbers of jobs are estimated at the planning area level through Iterative Proportional Fitting method (Deming and Stephan (1940)), as the solution of a constrained minimum least squared problem

$$\forall\, ft, k \in ft \; : \; \min_{N_{zk}^{c}} \sum_{z,k} w_1 \left(N_{zk}^{f} - \bar{N}_{zk}^{f}\right)^2 \; u.c. \; \begin{cases} \sum_{i} N_{zk}^{f} = N_{k}^{f} \\ \sum_{k} N_{zk}^{f} = N_{z,ft}^{f} \end{cases} \qquad (3.2)$$

where $\bar{N}_{zk}^{c}$ are observed total, computed from our initial establishments' sample.

Totals are hence distributed among buildings within each planning area and floor type proportionally to the approximated occupied building's floor area. Corresponding total numbers of occupied square meters are estimated using the relation previously defined.

In conclusion, we have estimated, for each building in the 2008 Singapore population, the number of establishments and the floor area by industry sector, and the number of jobs by occupation type.

4. Conclusions

This paper presents a method to synthesize a business establishment population for 2008 in Singapore. The aim is to overcome the lack of spatially detailed data on firms' characteristics and behaviors. Such a population would enable the study of business establishment (re)location behavior as well as individual workplace choice.

We face here a significant limitation in data availability. The proposed solution is hence to observe directory websites' data in order to build a sample of establishments. Such sample enables us to estimate total numbers of establishments, jobs and occupied square meters at building level, by industry and workforce size categories.

A modular automatic procedure is proposed for directory website data preparation, establishment activity classification and occupied floor area estimation. The results are quite satisfactory for the Service Sector. As we could expect, however, the procedure seems to be not efficient for the Manufacturing Sector.

Although this procedure is written for specific Singapore data here observed, it can be applied to any other case. Directory website structure, in fact, is quite stable. Only the first module, performing data cleaning and formatting, would have to be adapted.

The proposed population synthesis method for firms is defined under constraints imposed by data availability. Such a method can be applied to other cases, as long as one can obtain (a) a sample of floor area (or job counts) data for establishments at building (or detailed postcode) geographical level, (b) distributions by industry and by more aggregated geographical location for floor area (or jobs), (c) an approximation of floor area per job by occupation category, (d) aggregate data about the business establishments distribution by industry, and (e) a building population with each building's floor area by general business type.

Both the automated procedure and synthetic population construction method can be improved along multiple directions. First, we have estimated the number of establishments, jobs and square meters within each building, but we still don't know each establishment's mixture of occupation types. Knowing this mixture will be helpful for modeling business establishment (re)location behavior. A packaging method is currently under development to estimate how jobs and floor area are distributed among establishments within a building.

As stated earlier, the sample quality can be still improved, especially for the industrial sector and business services. One could hence improve industrial classification quality by tighter integration of the modules in the automated procedure. Also, a more intensive use of building type information at this stage could help in increasing classification precision.

Acknowledgments

We acknowledge the Singapore Land Authority for providing building footprint base maps, the online business directory, StDirectory, provided by Singapore Press Holdings, and other public and private entities that have provided the datasets used in this research. We also acknowledge partial support of the Singapore National Research Foundation through the Future Urban Mobility program of the Singapore-MIT Alliance for Research and Technology.

References

Anas, A., & Liu, Y. (2007). A regional economy, land use, and transportation model (RELU-TRAN©): formulation, algorithm design, and testing. Journal of Regional Science, 47(3), 415-455.

Bodenmann, B. R. (2011, September). Modelling firm (re-) location choice in UrbanSim. In ERSA conference papers (No. ersa11p1091). European Regional Science Association. DeBok (2011).

Ciari, F. (2011). Modeling location decisions of retailers with an agent-based approach. Eidgenössische Technische Hochschule Zürich, IVT, Institut für Verkehrsplanung und Transportsysteme.

Deming, W. E., & Stephan, F. F. (1940). 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.

Nguyen, C. Y., Sano, K., Tran, T. V., and Doan, T. T. (2013). Firm relocation patterns incorporating spatial interactions. The Annals of Regional Science, 50(3), 685–703.

Rodrigues, F., Alves, A. O., Pereira, F. C., Jiang, S., & Ferreira, J. (2012, January). Automatic Classification of Points-of-Interest for Land-use Analysis. In GEOProcessing 2012, The Fourth International Conference on Advanced Geographic Information Systems, Applications, and Services, 41-49.

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(1), 168-177.