The Journal of Transport and Land Use http://jtlu.org
Vol. 11 No. 1 [2018] pp. 1231–1256

JTLU

An agent- and GIS-based virtual city creator: A case study of Beijing, China

Chengxiang Zhuge
University of Cambridge
cz293@cam.ac.uk

Chunfu Shao
Beijing Jiaotong University
cfshao@bjtu.edu.cn

Shuling Wang
Beijing Transport Institute
wangshuling@bjjtw.gov.cn

Ying Hu
Beijing Transport Institute
xixing1@sina.co

Abstract: Many agent-based integrated urban models have been developed to investigate urban issues, considering the dynamics and feedbacks in complex urban systems. The lack of disaggregate data, however, has become one of the main barriers to the application of these models, though a number of data synthesis methods have been applied. To generate a complete dataset that contains full disaggregate input data for model initialization, this paper develops a virtual city creator as a key component of an agent-based land-use and transport model, SelfSim. The creator is a set of disaggregate data synthesis methods, including a genetic algorithm (GA)-based population synthesizer, a transport facility synthesizer, an activity facility synthesizer and a daily plan generator, which use the household travel survey data as the main input. Finally, the capital of China, Beijing, was used as a case study. The creator was applied to generate an agent- and Geographic Information System (GIS)-based virtual Beijing containing individuals, households, transport and activity facilities, as well as their attributes and linkages.

Article history:
Received: July 29, 2017
Received in revised form:
March 4, 2018
Accepted: July 16, 2018
Available online: December 5, 2018

1 Introduction

Agent-based modelling is a general approach to modelling dynamic and adaptive complex systems that contain autonomous agents and their interactions (Macal & North, 2010). Compared with equation-based models that are another general approaches to modelling complex systems, agent-based modelling has several strengths below (Huang, Parker, Filatova, & Sun, 2014; Twomey & Cadman, 2002): (1) Natural representations. The model is much easier to understand, as it represents the target system more straightforwardly; (2) Heterogeneity. The model is able to deal with the heterogeneous attributes and actions of agents; (3) Bounded/Unbounded rationality. The agent-based model is able to consider the rationality to be either bounded or unbounded; (4) Communication and social networking. Agents can communicate and share information easily through the social network, which is difficult to incorporate into traditional approaches; (5) Maintenance and refinement. It is easy to add, remove or modify agent types, attributes and behavioral rules of agents. Therefore, many agent-based integrated urban models

http://dx.doi.org/10.5198/jtlu.2018.1270
ISSN: 1938-7849 | Licensed under the Creative Commons Attribution – Noncommercial License 4.0

The *Journal of Transport and Land Use* is the official journal of the World Society for Transport and Land Use (WSTLUR) and is published and sponsored by the University of Minnesota Center for Transportation Studies.

(Zhuge, Shao, Gao, Meng, & Ji, 2014), such as UrbanSim (Waddell, 2002) and ILUTE (Salvini & Miller, 2005), have been developed to investigate urban issues at the micro scale, considering the dynamics and feedbacks in urban systems. However, such integrated urban models heavily rely on micro and disaggregate data in general, which is not available or accessible in some cases due to a high survey cost or privacy issues. The lack of disaggregate data has become one of the main barriers to the application of these models. As a result, some of the agent-based urban models were only tested in numerical experiments, rather than in real-world scenarios (Ettema, 2011; Filatova, Parker, & Van der Veen, 2009, 2007; Magliocca, Safirova, McConnell, & Walls, 2011; Parker & Filatova, 2008). In general, greater access to such disaggregate data is expected to be helpful for more easily developing the models with fewer constraints on the data availability and also for better testing the models.

In response to this, a number of methods and algorithms have been developed to generate the disaggregate inputs for agent-based urban models, using a small portion of micro sample data and some relevant macro-level constraints (Müller & Axhausen, 2011; Ye, Konduri, Pendyala, Sana, & Waddell, 2009) or even only using macro-level data (Barthelemy & Toint, 2013; Ge, Meng, Cao, Qiu, & Huang, 2014). However, most of them were used to generate synthetic populations, paying significantly less attention to the disaggregate activity and transport facilities, which are another two key inputs for agent-based urban models. Furthermore, disaggregate populations and facilities are highly correlated: on one hand, each household needs to be allocated with a residential facility (Li, Zhuge, Zhang, Gao, & Zhang, 2014); on the other hand, individuals perform their daily activities (e.g., shopping and leisure) based at different facilities. Therefore, an integrated approach, which is called virtual city creator in this paper, is needed to in sequence generate a synthetic population and physical environment containing activity and transport facilities, and then to link individuals and households in the population to the facilities, resulting in an agent- and GIS-based virtual city. This could be viewed as a process of initializing the integrated urban models. In addition, the virtual city creator has been developed as a component of an agent-based integrated land use and transport model, SelfSim, which has been applied in a Chinese medium-sized city, Baoding (Zhuge, Shao, Gao, Dong, & Zhang, 2016).

2 Literature review

Many population synthesizers have been developed to generate synthetic populations comprising households and persons, as well as their attributes (e.g., car ownership and age). Iterative Proportional Fitting (IPF) (Beckman, Baggerly, & McKay, 1996) and Combinatorial Optimization (CO) method are two traditional approaches to population synthesis (Abraham, Stefan, & Hunt, 2012). In general, these two synthesizers use the household travel survey data as the key input, together with some macro-level constraints. Apart from the two traditional methods and their variants, the other synthesizers can be broadly categorized into the following four groups (Zhuge, Li, Ku, Gao, & Zhang, 2016): (1) Simulation-based models (Farooq, Bierlaire, Hurtubia, & Flötteröd, 2013) that, for example, used Markov Chain Monte Carlo method to generate a synthetic population; (2) Data Constriction models (Gargiulo, Ternes, Huet, & Deffuant, 2010) that were particularly for those cases where the general input data above was not available; (3) Entropy Optimization Models (Bar-Gera, Konduri, Sana, Ye, & Pendyala, 2009) that were capable of searching for a most-likely set of household weights for population synthesis; (4) Fitness-based Models (Ma & Srinivasan, 2015) that "generate a list of households, directly matching several multilevel controls". A detailed review on population synthesis can be found in the work of Müller and Axhausen (2010).

The physical environment in this paper refers to the GIS-based disaggregate data on activity facilities and transport infrastructures. Open online sources, such as OpenStreetMap (See http://www.openstreetmap.org/), are good datasets for preparing such disaggregate data, as they can provide relatively

fine-scale transport networks (Gao, Zhao, Zhuge, & Zhang, 2012), as well as the locations of activity and transport facilities (Liu & Long, 2016; Zilske, Neumann, & Nagel, 2011). However, the capacities and prices of the facilities, which could influence several forms of individual behavior (e.g., residential location choice), are difficult to access or may even not be available online. Furthermore, these open online sources generally can only provide the latest information, but have no historical data available which are needed in some studies investigating past issues. Compared with population synthesizers, the methods to create the physical environment have received significantly less attention. Ge et al. (2014) proposed an algorithm to generate the physical environment that was composed of six facility categories, "houses, educational institutions, workplaces, consumption locations, entertainment locations and medical institutions." Given a synthetic population and physical environment, individuals in the population need to be further linked to activity facilities, such as houses, workplaces, and educational institutes, which can be viewed as the generation of initial travel demand, resulting in initial daily plans for each agent. This kind of generation can be done with activity-based models (Chu, Cheng, & Chen, 2012; Horni, Nagel, & Axhausen, 2016; Pinjari & Bhat, 2011; Zhuge, Shao, Gao, Meng, & Xu, 2014; Zhuge, Shao, Wang, & Hu, 2017).

In summary, the synthesis methods for transport and activity facilities have received significantly less attention than population synthesizers. In response to this, both activity and transport facility synthesizers will be developed to generate activity and transport facilities, respectively, which make up a GIS-based physical environment. Furthermore, this paper will develop a so-called virtual city creator by integrating several synthesis methods, including population and facility synthesizers, to generate and link individuals and the physical environment at the micro level, resulting in an agent- and GIS-based virtual city which can be used as the input for agent-based urban micro-simulations.

3 Methodology

3.1 An agent-based integrated urban model—SelfSim

As shown in Figure 1, the virtual city creator is developed as a key component of SelfSim, an agent-based integrated urban model, and is used to initialize the model prior to a simulation. SelfSim comprises several agent-based spatial sub-models, which make up an annual loop to simulate how urban systems evolve over time at the micro level. A brief introduction to each sub-model is given as follows (Zhuge & Shao, 2018a, 2018b; Zhuge, Shao, Gao, et al., 2016):

- Demographic Evolution Model: is used to simulate the demographic transitions, such as births and deaths, at the individual level (Zhuge & Shao, 2018b);
- Activity-based Travel Demand Model: is based on MATSim (Multi-Agent Transport Simulation) and is used to simulate how agents perform their daily activities and travel from one activity location to the next throughout a whole day (Horni et al., 2016);
- Micro-Simulation-based Accessibility Model: calculates the accessibility for different transport and activity facilities based on the activity-based simulation (or the MATSim simulation);
- Joint Model of Residential Location Choice & Real Estate Price: is used to simulate the decision-making of household agents on their residential locations and to predict the real estate price of residential facilities (Zhuge & Shao, 2018b);
- Transport Facility Development Model: simulates the development of transport facilities, including parking lots and refueling stations (Zhuge & Shao, 2018a);
- Activity Facility Location Model: is used to simulate the development of activity facilities, such as shops. Note that this sub-model has not been implemented yet. Therefore, the data on activity facilities needs to be input before the model is run.

One may have found that such an agent-based urban model requires various disaggregate input data. A virtual city creator could be very useful here for preparing data to initialize the model prior to a simulation.

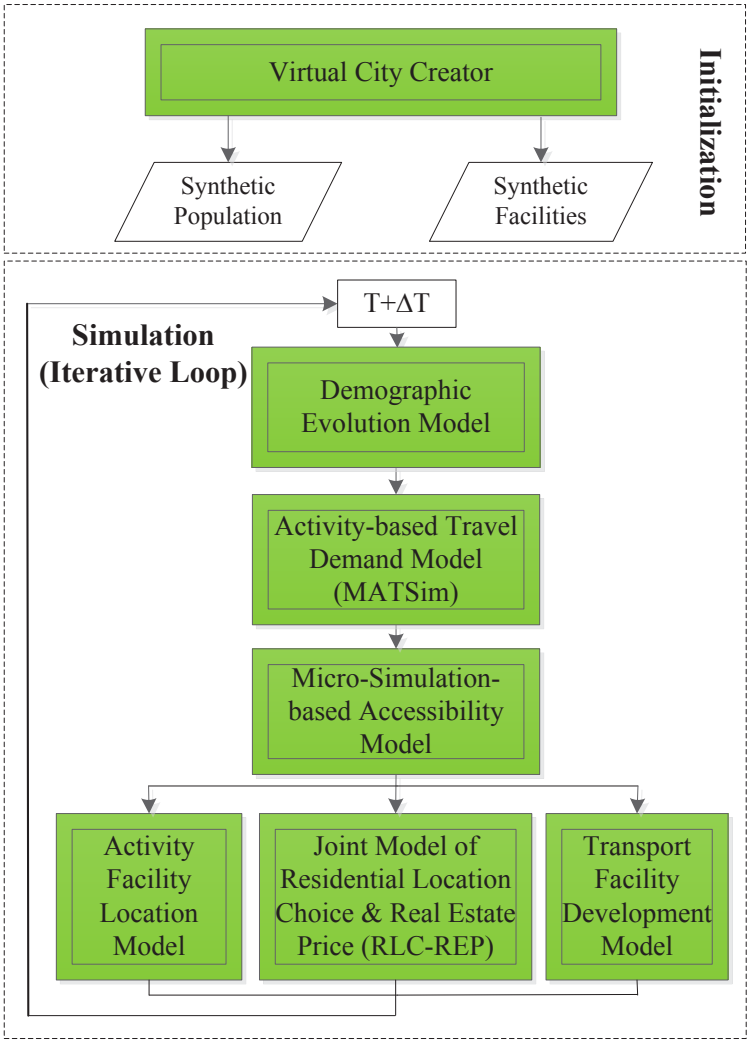


Figure 1: Framework of SelfSim (Zhuge, Shao, Gao, et al., 2016)

Figure 2 shows the sketch of an agent- and GIS-based virtual city generated by the creator, which contains persons, households and facilities, as well as their attributes and linkages. SelfSim assumes that all individuals and facilities are centered on road nodes (or interactions) for the purpose of model simplification, meaning that transport and activity facilities are located on nodes, and individuals are resident and perform their daily activities in node-based facilities. It is worth noting that both individuals and facilities have their own attributes attached, such as sex, age and facility capacity.

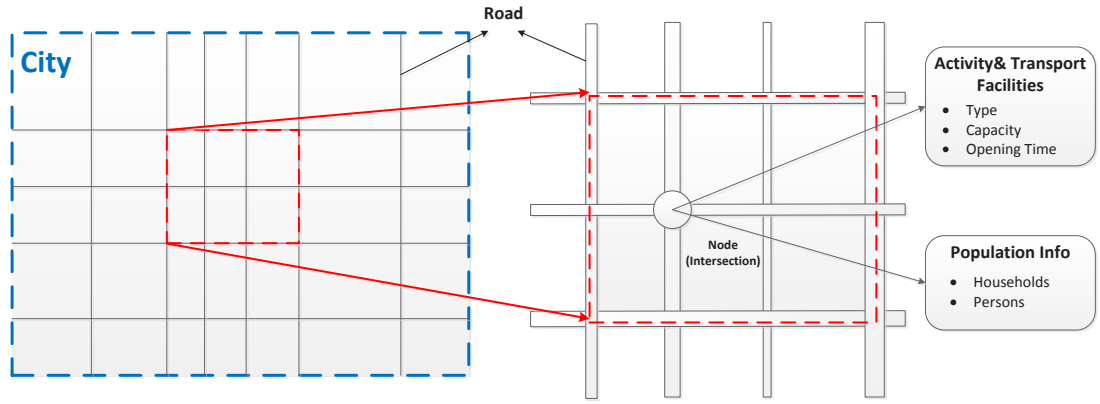


Figure 2: Sketch of an agent- and GIS-based virtual city

3.2 Procedure of data synthesis for the creator

Figure 3 demonstrates how to create and link individuals and physical environment at the micro level, given the household travel survey data, macro-level constraints (e.g., total number of households) and a transport network. Among the inputs above, the household travel survey data, which contains rich spatial and temporal information on travel and activities of respondents, is the key input. The procedure of data synthesis is composed of three steps below: Steps 1 and 2 are to generate a synthetic population and physical environment comprising activity and transport facilities, respectively. Step 3 is to link individuals in the population to the physical environment. As a result, an agent- and GIS-based virtual city containing persons, households and facilities, as well as their attributes and relationships can be obtained. The virtual city can be viewed as the base-year input of SelfSim. It is worth noting that Steps 2 and 3 are interdependent, meaning that the outputs of one component are used as the inputs of the other. For instance, the transport facility synthesizer in Step 2 uses the daily plans generated in Step 3 as inputs to generate transport facilities, including parking lots and refueling stations. More details on the interdependency will be given below where relevant.

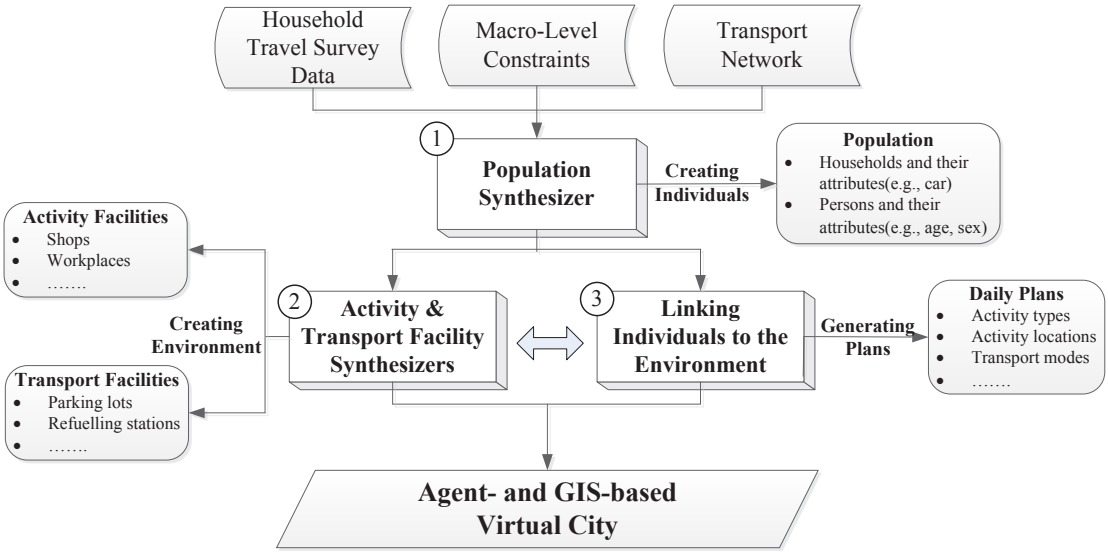


Figure 3: Procedure of data synthesis

3.3 Creating individuals: Genetic algorithm (GA)-based population synthesis

Population synthesis can be mathematically viewed as an optimization problem that tries to minimize the differences between the generated and target frequencies of control variables by searching for an optimal set of household weights, which can be formulated as Equation (1).

$$\min \sum_j \left| \sum_i d_{ij} w_i - c_j \right| \Big/ c_j \qquad (1)$$
$$s.t. \;\; w_i < \beta \cdot w_i^{'} \quad i = 1, 2, ...,$$

Where, i denotes a household category; j denotes a control variable; d_{ij} denotes the frequency of the control variable (household/person type) j in household category i ; w_i denotes the weight of household category i ; c_j denotes the constrain of the control variable j (Ye et al., 2009).

This paper develops a Genetic Algorithm (GA)-based population synthesizer (Zhuge, Shao, Li, Gao, & Zhang, 2016), which is able to deal with some particular cases in which the priority needs to be given to some more influential control variables, in order to make the resulting synthetic population better represent the real world in some particular aspects, for some specific research purposes. For example, in the studies of vehicle purchase, the priority may need to be given to the vehicle ownerships of each household when the population is synthesized, because the number of vehicles in each household, as well as the total number of vehicles, could significantly influence the research outcomes. To this end, the fitness function for the GA-based synthesizer shown by Equation (2) incorporates the weights of each variable (W_j) into Equation (1). Here, the weights are used to quantify the relative importance of the variables. It is expected that the resulting differences between target and generated frequencies of the variables that are given the priority could be smaller. Therefore, the synthetic population could be more realistic in terms of these important variables.

$$\min \sum_j W_j \cdot \left| \sum_i d_{ij} w_i - c_j \right| \Big/ c_j \qquad (2)$$

3.4 Creating physical environment: Generating activity facilities and transport infrastructures

In SelfSim, a physical environment is composed of activity facilities and transport infrastructures, as well as the prices associated with the facilities, such as house prices. For activity facilities, they provide space for agents to perform their daily activities, primarily including shopping, leisure, studying (at school), resting (at home) and work. The transport infrastructures, including transport network, parking lots and refueling stations, are the basic facilities for agents to travel from one place to another. The prices associated with activity facilities particularly refer to the real estate prices, including selling prices and rents.

(1) Synthesis method for activity facilities

The activity facility synthesis is to create activity facilities mainly using the household travel survey data and relevant macro-level constraints, such as the total numbers of houses in the study area. The basic assumption for the synthesizer is that the frequency of performing activities at a place that can be extracted from the household travel survey data is directly proportional to the number of activity facilities located in that place. Figure 4 shows the flowchart of generating GIS-based activity facilities with the capacity information attached. Taking shopping facility for example, the method to create and locate shopping facilities is introduced as follows:

Step 1: Extract Activity Location Information from the Household Travel Survey Data. In general, the survey data contains detailed activity and travel information of each participant, including activity type and location. For shopping facilities, the locations where participants go shopping can be extracted and counted at the zone level. As a result, the frequencies of shopping (or the number of shopping activities, F_i) for each zone (i) can be obtained.

Step 2: Calculate the Number of Activity Facilities at the Zone Level. According to the basic assumption of the facility synthesizer, the number of facilities can be calculated by Equations (3) and (4). Specifically, given the total number of shopping facilities (AF_{total}), Equation (3) is to calculate the ratio ($R_{AF/P}$) of AF_{total} to the sum frequency of performing shopping (F_{sum}) for the whole study area. Then Equation (4) is used to calculate the total number of shopping facilities for each zone (AF_i) by multiplying the ratio $R_{FA/F}$ by the frequency of performing shopping at the zone i (F_i).

$$R_{AF/F} = \frac{AF_{Total}}{F_{Sum}} = \frac{AF_{Total}}{\sum_{i=1}^{I} F_i} \tag{3}$$

$$AF_i = R_{AF/F} \cdot F_i \tag{4}$$

Step 3: Randomly Allocate the Facilities to Road Nodes. As all facilities in SelfSim are assumed to be centered on road nodes, the generated facilities for each zone need to be further allocated to a specific road node in the zone. For instance, given that the allocation is random, then each road node in a zone will have the same probability of being allocated with the activity facilities. After the allocation, the exact locations and capacities of each node-based facility can be obtained.

Likewise, the other activity facility types can be synthesized in the same way described above. As a result, a GIS-based layout of activity facilities with detailed information on locations and capacities can be obtained.

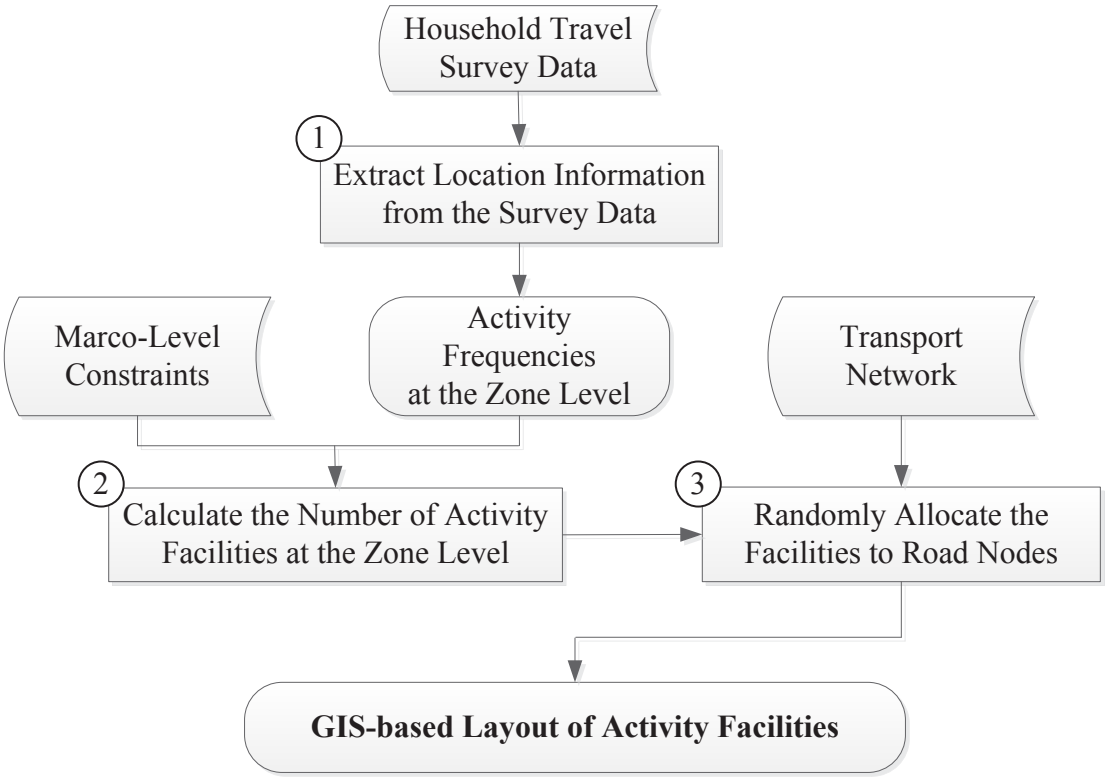


Figure 4: Flowchart of generating GIS-based activity facilities

(2) Synthesis methods for transport facilities

A transport network is composed of links (roads) and nodes (intersections). The transport facilities in this paper include refueling stations and parking lots which can be grouped into link- and node-based transport facilities, respectively. Specifically, refueling stations are generally located next to road links and are used to accommodate the road link-based demand that is traffic flow, and thus they belong to the link-based transport facility (Zhuge & Shao, 2018a); while parking lots are essentially located at specific points, such as workplaces, shopping centers and schools, and the demand for them can be aggregated at points. Thus they belong to the node-based transport facility (Zhuge & Shao, 2018a). In order to accommodate the node- and link-based demand, two traditional approaches, namely p-median and flow-refueling models, have been widely used to locate transport facilities (Chen, Kockelman, & Khan, 2013; Frade, Ribeiro, Gonçalves, & Antunes, 2011; Hodgson & Rosing, 1992; Kim & Kuby, 2012; Kuby & Lim, 2005; Kuby et al., 2009; Li & Huang, 2015; Lim & Kuby, 2010; Upchurch & Kuby, 2010; Wang & Lin, 2009; Zhuge & Shao, 2018a). Therefore, the proposed synthesis methods for node- and link-based facilities will be developed based on these two models, respectively.

(a) Node-based transport facility synthesis method: Activity-based P-media model

The node-based transport facility synthesis method is to generate parking lots based on a so-called activity-based P-media model that is a variant of the traditional P-media model (Hodgson & Rosing, 1992; ReVelle & Swain, 1970; Upchurch & Kuby, 2010). The P-median model can be formulated as a mathematical problem that seeks to locate p facilities with the objective of minimizing the demand-weighted distance from stations to users (Hodgson & Rosing, 1992; ReVelle & Swain, 1970; Upchurch & Kuby, 2010; Zhuge & Shao, 2018a), and the demand was generally calculated based on the traditional four-step method; while the activity-based P-media model in this paper firstly calculates the potential demand for each candidate node in the network based on an activity-based simulation and then selects

those nodes whose demand (D_{Node}) exceed the pre-defined threshold (T_{Node}). The node-based transport facilities will be created based on these selected nodes, and the facility capacity will be determined by Equation (5). Note that the daily plans of each agent, which is the main input of the activity-based simulation, is generated using the method to be introduced in 3.5.

$$C_{Node} = \frac{D_{Node} - T_{Node}}{T_{Node}} \cdot \phi_{Node} \tag{5}$$

Where, D_{Node} denotes the demand at the node; ϕ_{Node} is a scaling factor; Both T_{Node} and ϕ_{Node} need to be set before the facility capacity is calculated. These two parameters can be estimated, given the macro-level constraint on the total number of facilities. In principal, the estimation can be done with various methods, such as Genetic Algorithm-based parameter estimation method. This project uses a simple method that is briefly introduced as follows: either of the parameters can be firstly fixed by being arbitrarily set to an appropriate value, and then an enumeration method can be further used to try different values for the other parameter within a specific range until an optimal value is found to minimize the gap between generated and target total facility capacities (e.g., the total number of parking spaces).

(b) Link-based transport facility synthesis method: Activity-based flow-refueling model

The traditional flow-refueling model can be formulated as a mathematical problem that attempts to intercept as much specific flow as possible (Hodgson, 1990; Hodgson & Rosing, 1992). Similar to the activity-based P-media model, the activity-based flow-refueling model calculates the amount of intercepted flow based on an activity-based simulation, rather than the four-step method used in the traditional model (Zhuge & Shao, 2018a). Then the amount of the flow intercepted will be used to locate facilities in the following way:

Step1: the candidate node with highest amount of flow will be selected as a facility location and a specific capacity that is calculated by Equation (5) will be set.

Step2: those candidate nodes located within a specific radius (R_{min}) to the selected node will be removed from the candidate lists, in order to avoid any competitions between the selected node and the other nodes close to it.

The model will repeat Steps 1 and 2 until the target total number of transport facilities are reached.

(3) Generating disaggregate prices for activity facilities

The prices here particularly refer to the real estate prices of residential facilities, including both rents and selling prices, as SelfSim currently only simulates the residential location choice behavior of household agents. Prices are important factors that can influence many forms of individual behavior in SelfSim, including residential relocation and travel behavior. However, such disaggregate data is difficult to access and even not available in some cases. In response to this, a price synthesis method is developed to generate the prices for each residential facility using both macro-level constraints and the household travel survey data.

The synthesizer generates disaggregate prices for each residential (or "home") facility, considering both spatial features and affordability. Specifically, it has been commonly recognized that the distance to city center is closely correlated with the housing price (Chen & Hao, 2008; McMillen, 2002; van Bergeijk, 2012), and thus it can be considered as a geographical factor in the price generation; Furthermore, the house price to income ratio is an important factor in the housing market and has been widely used to check the affordability (Lau & Li, 2006; Zhuge, Shao, Gao, et al., 2016). The procedure of generating the disaggregate prices here is composed of three steps: First, the household travel survey data and the house price to income ratio will be used to generate initial disaggregate selling prices without considering any spatial features, such as the distances between facilities and city centers; Second, some spatial features

will be used to adjust the initial prices in a specific way; Third, the price to rent ratio, which is another important indicator in the housing market, will be used to calculate the rents for each residential facility with a simple multiplication. The detailed introduction to each step is as follows:

Step 1: **Generating initial selling prices for each residential facility**

For a residential facility *i*, an initially estimated price (IEP_i) will be calculated by multiplying the average income of people living in the facility (AI_i) by the ratio of house price to income ($R_{Price2Income}$), which is formulated as Equation (6). The average income can be extracted from the household travel survey data, and the ratio is macro statistical data and is generally available in statistical yearbooks. In order to meet the macro-level constraint on the average house price of the study area (*AP*), the estimated price *IEP* needs to be scaled up or down by multiplying a scaling factor ($\boldsymbol{\beta}_{PriceScaling}$). The scaling process is formulated as Equation (7), and the scaling factor can be computed with Equation (8).

$$IEP_i = AI_i \cdot R_{Price2Income} \tag{6}$$

$$FP_i = \beta_{PriceScaling} \cdot IEP_i \tag{7}$$

$$\beta_{PriceScaling} = \frac{AP}{\sum_{i=1}^{I} IEP_i \Big/ I} \tag{8}$$

Step 2: Adjusting the initial selling prices with the consideration of facility locations

The initial selling prices generated in Step 1 may be biased due to the missing spatial constraint. For instance, some people with high income may prefer to live in rural districts where the real estate price is generally lower. Therefore, it may be problematic if only affordability (or household income) is used for the price generation. In order to deal with the bias, the distances between houses and city centers will be used to adjust the initial selling prices. The adjustment is described as follows:

- **Step 2-1**: for each residential facility, the distances to the city centers (one city may have one or more centers or sub-centers) will be firstly calculated and then the weighted sum will be calculated as the final distance to the centers.
- **Step 2-2**: Clustering the final distances of all facilities into a specific number of groups (N) using a k-mean clustering algorithm (Hartigan & Wong, 1979).
- **Step 2-3**: Accordingly, generating N sections of housing prices with equal interval using the maximum and minimum prices in the initial price database.
- **Step 2-4**: Matching the clustered groups and the divided sections based on the group and section indices (e.g., linking group 3 to section 3) and then assigning the distance-based prices to each facility.
- **Step 2-5**: Averaging the distance- and income-based prices, resulting in an average price for each facility.
- **Step 2-6:** Scaling the average prices properly in the same way introduced in Step 1 using Equations (7) and (8), in order to meet the macro-level constraint on the target average house price of the study area, resulting in the final prices.

Step 3: Generating disaggregate rents for each facility based on the selling prices

The disaggregate rents can be generated with a much simpler method based on the generated disaggregate selling prices. Specifically, the method firstly calculates the initial rents for each facility by multiplying the selling price by the price to rent ratio of the study area, which is another important indicator in the housing market, and then scaling the initial rents in the same way the selling prices are scaled, in order to match the target and generated average rents of the study area, using Equations (7) and (8) as well.

3.5 Linking individuals to the physical environment: Generating initial daily plans

Individuals in a synthetic population can be linked to the GIS-based physical environment through the generation of initial daily plans. A daily plan contains the information on the activities that the agent plans to perform throughout a whole day, as well as the travel from one activity facility to the next. The generated initial daily plans could be further optimized through the execution of the daily plans. Specifically, the execution of daily plans can result in the traffic states, which can be in return used to adapt the plans to the dynamic traffic flow. For example, agents may want to reschedule their activities (e.g., change activity locations) in order to avoid traffic congestion, given the traffic states. MATSim (Multi-Agent Transport Simulation) is a typical platform for such optimization (Horni et al., 2016). The procedure of generating initial daily plans is composed of two steps below: Step 1 is to determine the home facility for each household with the constraint of facility capacity; Step 2 is to generate initial daily plans for each member in the household, given the residential location. Each household in the synthetic population can be processed in turn as follows:

Step 1: Determine the residential facility for each household

Residential location could heavily influence many decisions of agents, for instance, on the choices of other activity facilities (e.g., workplaces) and transport modes. In the synthetic population, each household will be allocated with a residential facility according to the following rule: Since the synthetic population is generated by randomly selecting and duplicating the households in the household travel survey data according to the household weights, the original residential location of the selected household (generally available at the zone level) can be a good base for the synthetic household to search for a new residential location, as these two households are similar in attributes and therefore the synthetic household could be allocated to the same location or to another location close to the original one. However, due to the constraint of residential facility capacity, some households may not be able to find residential locations at or near to the original location. Alternatively, a home facility that is closest to the original location can be allocated.

Step 2: Generate initial daily plans for each individual

The initial plan generator uses the original plan skeleton extracted from the household travel survey data as a base and then adds the missing information to the plan, including activity locations, resulting in an initial daily plan. To distinguish between the daily plan and plan skeleton, their definitions are given as follows:

(1) Daily Plan: Each person in a synthetic population, apart from those (e.g., babies) who are not able to perform activities independently, has a daily plan that is composed of activity and travel information. The activity information includes activity type, activity duration, arrival time and departure time. The travel information is a set of links which connect one activity location to the next with the attached information on the transport modes used.

(2) Plan Skeleton: As shown by Figure 5, a plan skeleton is derived from the daily plan but lacks information on activity location and travel information.

The daily plan generator searches for an activity location for each plan skeleton with both time and capacity constraints. Specifically, the travel time between two adjacent activity locations in the original plan will be used as a time constraint, which means the new travel time between two new adjacent activity locations needs to be as close to the original one as possible. The capacity constraint means that the number of activities performed at a facility cannot exceed its capacity. It is worth noting that initial daily plans do not contain detailed information on travel (e.g., routes), but the information can be added, for example, in the MATSim simulation (Horni et al., 2016) when the plans are executed.

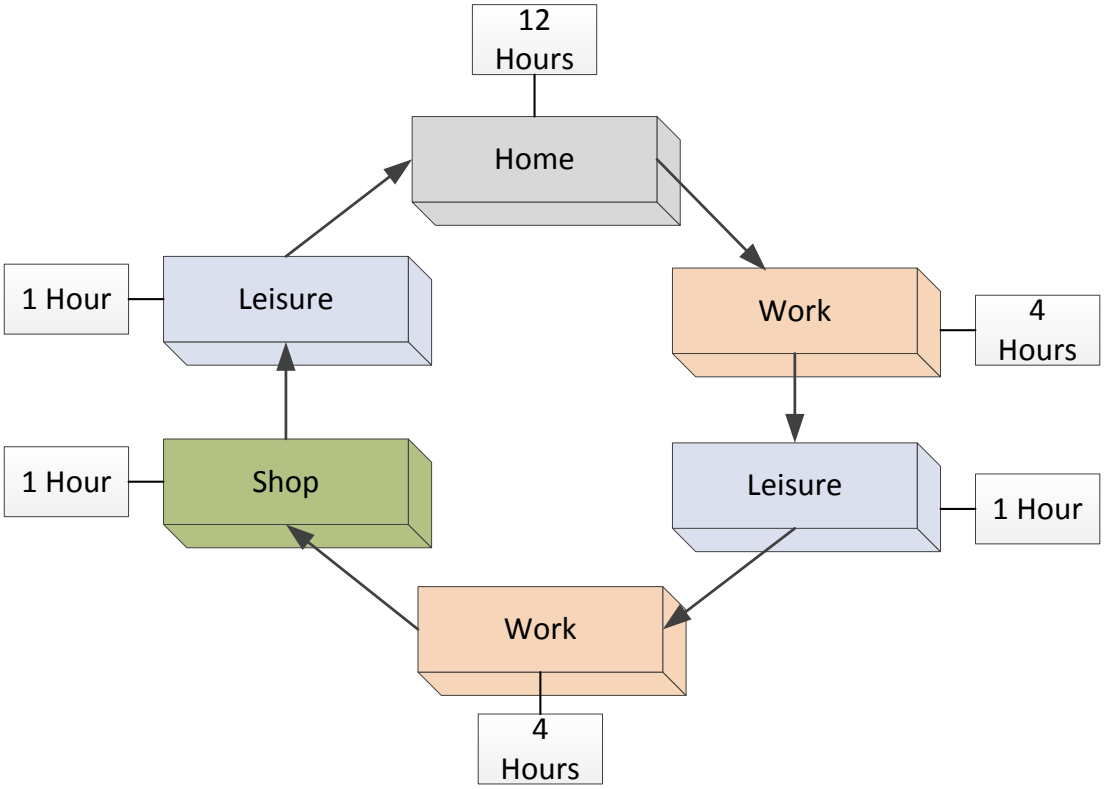
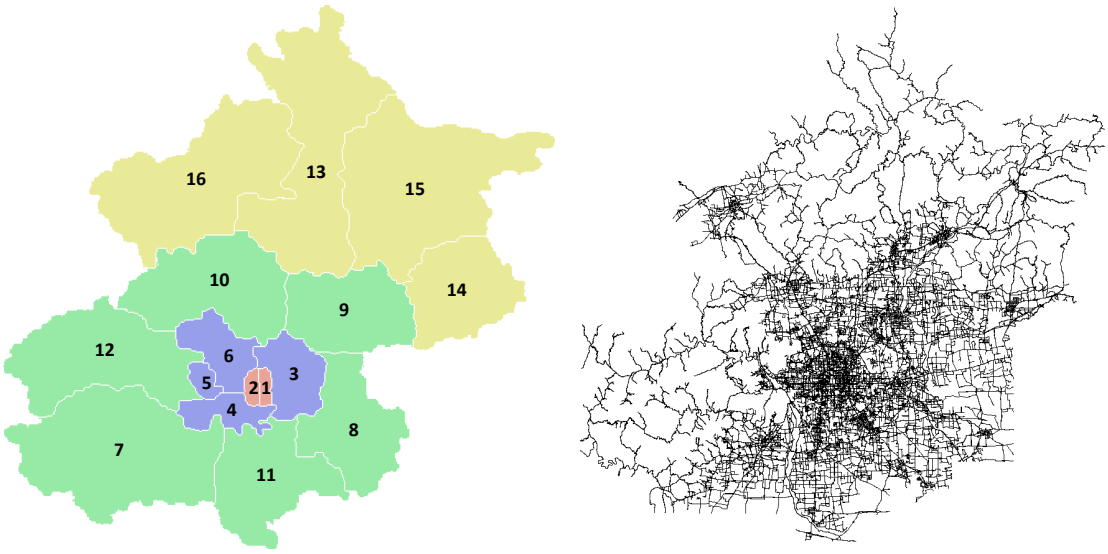


Figure 5: An example of plan skeleton

4 Case study: Beijing, China

4.1 Scenario description

The capital of China, Beijing was used as a case study, and the virtual city creator was used to create a virtual Beijing in 2010 with a population scaling factor of 0.04, meaning that one synthetic individual or facility in the virtual city represents twenty-five individuals or facilities in reality. In 2010, Beijing was composed of 16 administrative districts with a population of 19.61million. Its land area was 16,410.54 square kilometers (Dong, Shao, Ma, Zhuge, & Li, 2012; Zhang et al., 2018). The 16 administrative regions were Dongcheng, Xicheng, Chaoyang, Fengtai, Shijingshan, Haidian, Fangshan, Tongzhou, Shunyi, Chanpin, Daxing, Mentougou, Huairou, Pinggu, Miyun and Yanqing, which are numbered 1-16 in Figure 6-(a). The input data for the creator mainly includes the 2010 Beijing household travel survey data, a road network (see Figure 6-(b)) and some macro-level constraints extracted from the Beijing Statistical Yearbook 2011 (which recorded the information in 2010) and 2010 Beijing's Sixth National Population Census Data Bulletin.



(a) District IDs (b) Road Network in 2010

Figure 6: Maps of 16 administrative districts and the road network of Beijing in 2010

4.2 Generating a synthetic population: GA-based population synthesis

The GA-based population synthesizer was applied to create a virtual population using both household travel survey data and macro-level constraints. Table 1 shows the variables of interest used for the population synthesis, as well as the synthetic results. The variables are from both person and household levels. Specifically, variables of age and gender are person attributes, and the number of households and car ownership are household attributes. According to the small relative errors in Table 1, it can be concluded that the population synthesizer is capable of fitting both person- and household-level target frequencies very well.

Table 1: Variables of interest and their constraints and synthetic results

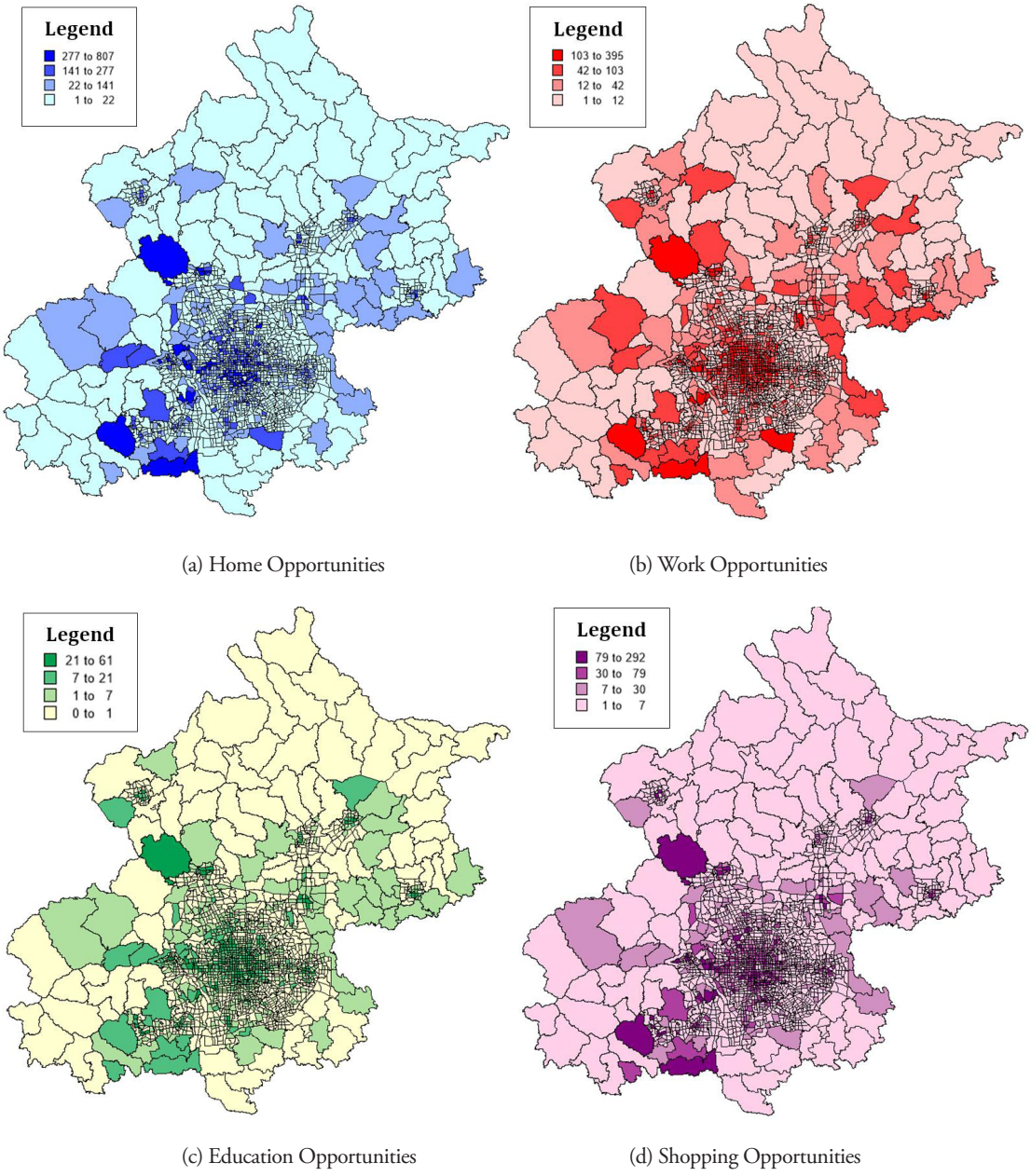
Variables	Target Frequency	Synthetic Frequency	Relative Error
Age: 0-19	2,758,000	2,765,765	0.28%
Age: 20-29	5,011,000	5,044,130	0.66%
Age: 30-39	3,546,000	3,569,809	0.67%
Age: 40-49	3,285,000	3,308,339	0.71%
Age: 50-64	3,303,000	3,318,602	0.47%
Age: >=65	1,709,000	1,712,413	0.20%
Male	10,126,000	10,198,819	0.72%
Female	9,486,000	9,520,239	0.36%
Number of Households	6,680,552	6,632,295	-0.72%
Number of Cars	3,743,814	3,728,613	-0.41%

4.3 Creating the physical environment: Generating activity facilities and transport infrastructures

This case study only considers five types of activity facilities, namely "home", "work", "education", "leisure" and "shopping" facilities. Since the data sources for these five facility types are only available at the district level, the facility synthesis method introduced in Section 3.4 was applied to generate these disaggregate facilities.

(1) Initial number of activity opportunities in each traffic zone

The initial number of activity opportunities by activity type was extracted from the survey data by counting the number of activities performed at each zone. The spatial distributions of the opportunities by activity type are shown by Figure 7. It can be found that most of the activities took place at the central districts and the central areas of the outer districts.

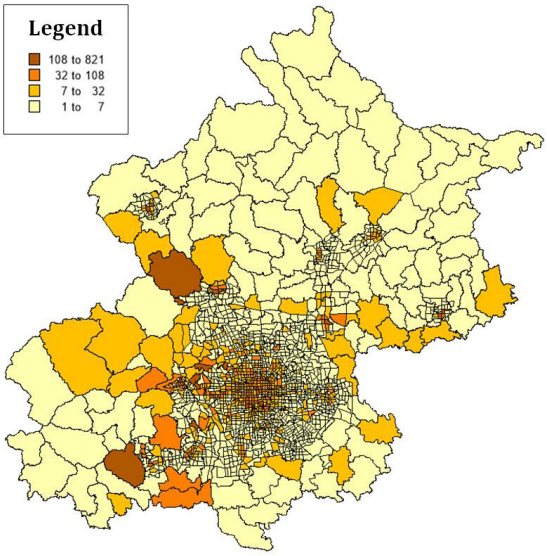


(a) Home Opportunities

(b) Work Opportunities

(c) Education Opportunities

(d) Shopping Opportunities

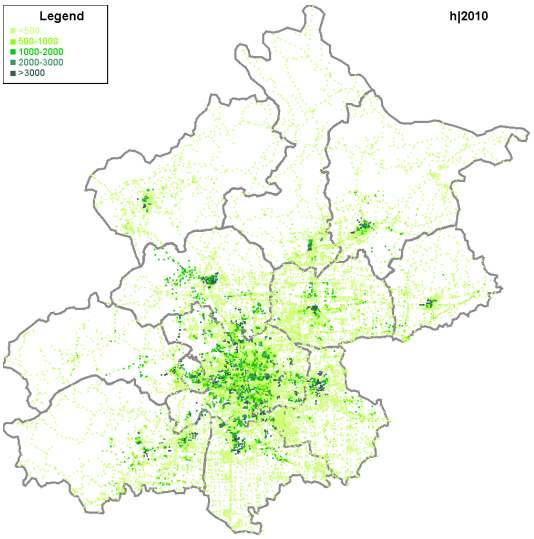


(e) Leisure Opportunities

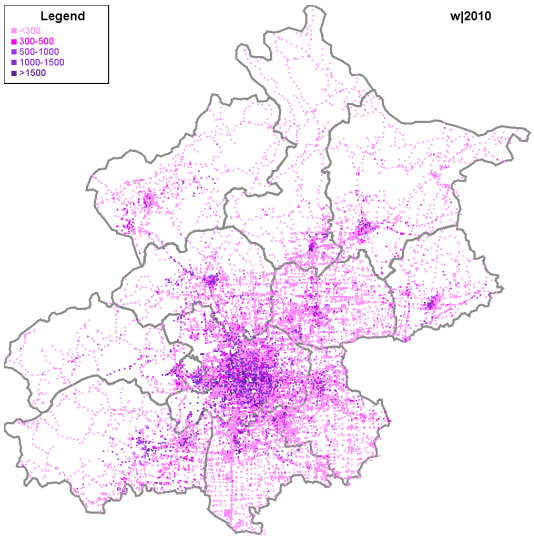
Figure 7: Spatial distributions of activity opportunities

(2) Synthetic activity facilities with capacity information attached

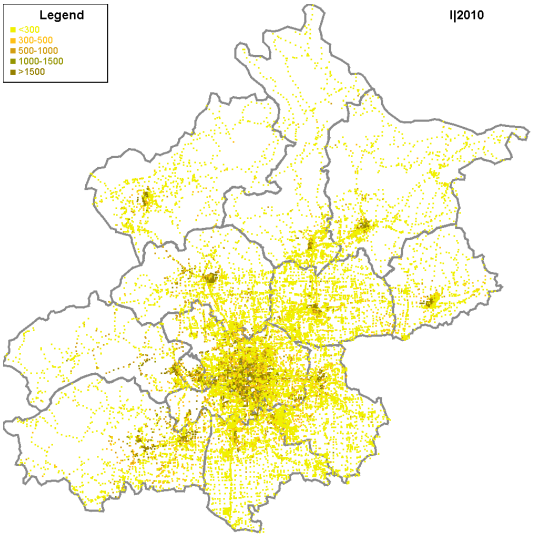
Given the initial number of activity opportunities in each zone and the total number of opportunities for each facility type, the synthetic activity facilities can be generated using the synthesis method introduced in Section 3.4. Figure 8 shows the spatial distribution of synthetic facilities by type. Figure 9 shows the distributions of activity facility capacities. Note that those facilities that currently have no capacities were also mapped in Figure 8 (as capacities may be allocated in the subsequent simulations), but were not counted in Figure 9. Table 2 in Appendix illustrates how to estimate the total number of activity opportunities for the activity types of home, work, leisure and shop and also lists the data sources used for the estimation. In terms of the facilities for education, they were generated using the data on the locations of educational institutions available on the website of the Beijing Municipal Education Commission, in addition to the macro-level constraints and initial number of activity opportunities. It can be found from the maps that the central districts and the central areas of the outer districts tend to have more activity opportunities, which is consistent with the information (or the initial number of activity opportunities) extracted from the household travel survey data.



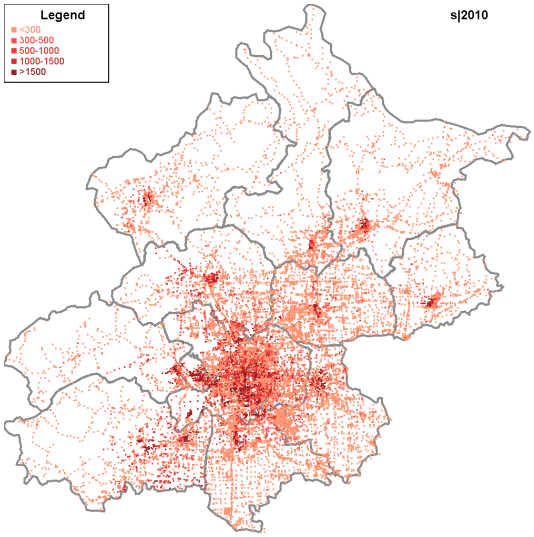
(a) Home Facilities



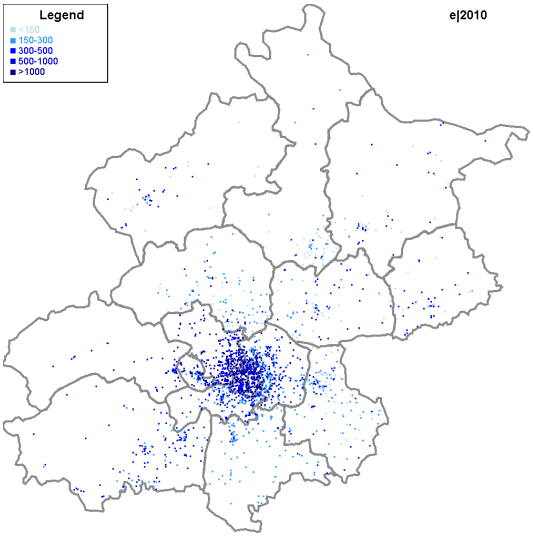
(b) Work Facilities



(c) Leisure Facilities

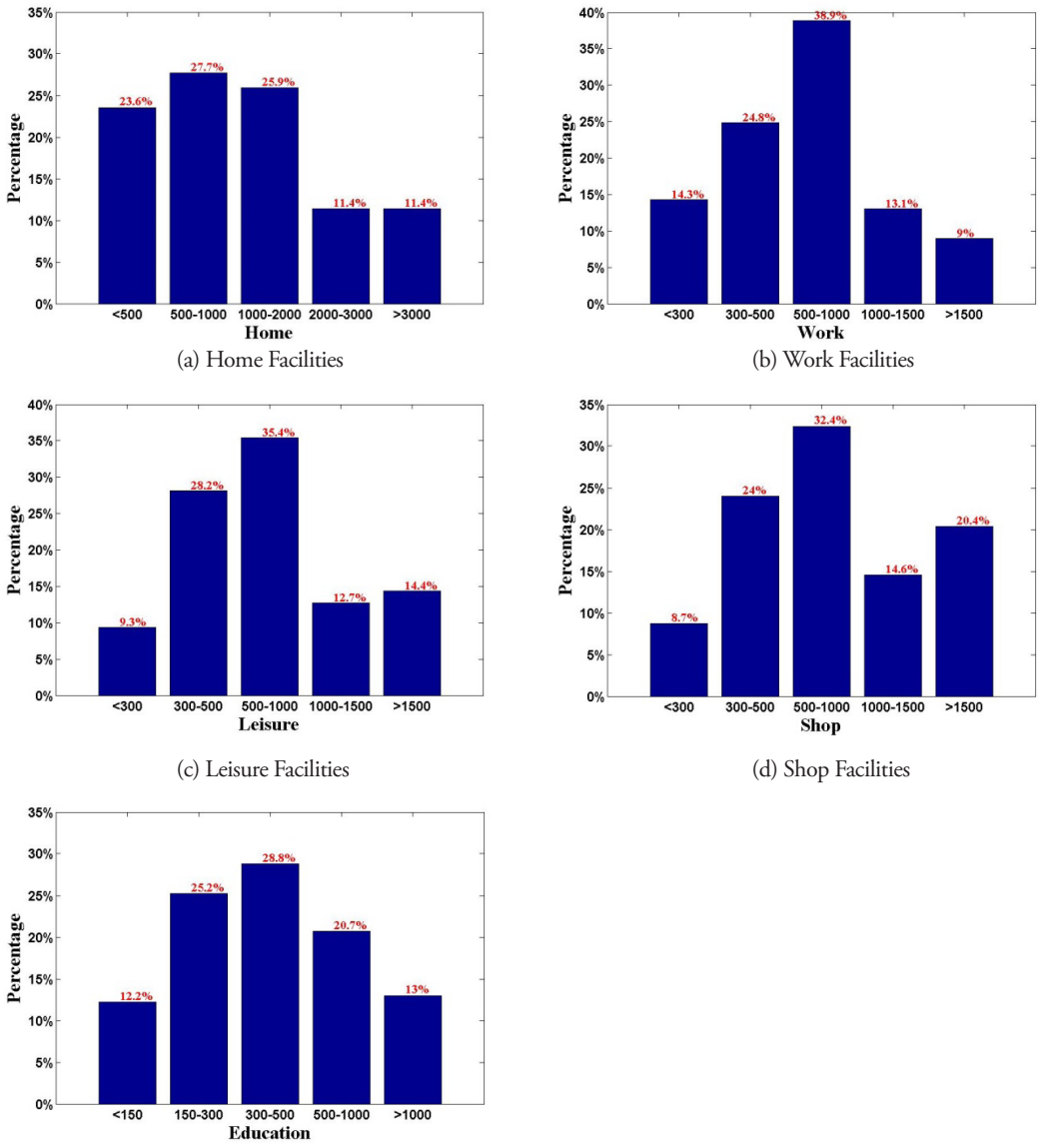


(d) Shop Facilities



(e) Education Facilities

Figure 8: Spatial distributions of synthetic activity facilities by type in 2010



(a) Home Facilities

(b) Work Facilities

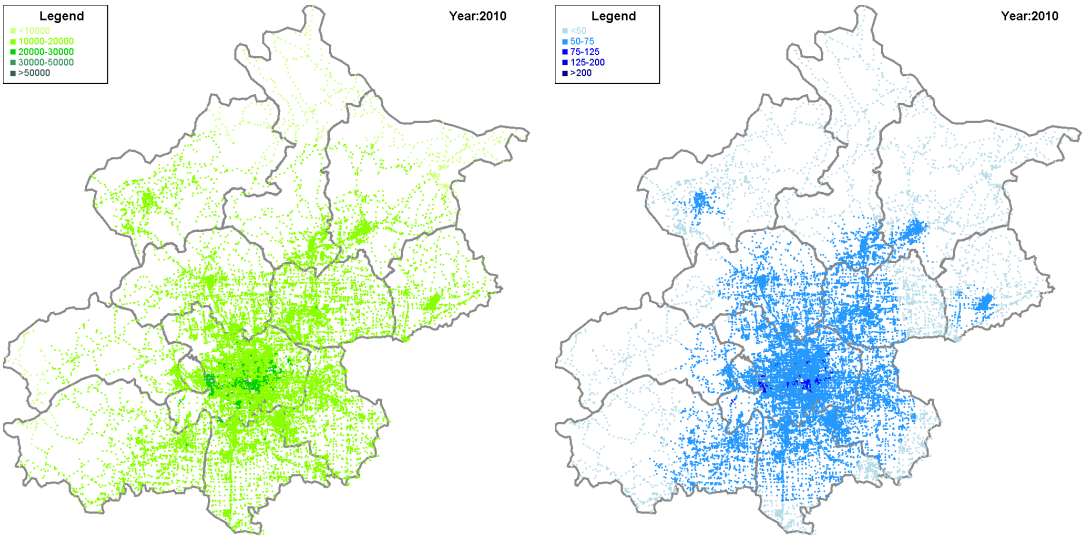
(c) Leisure Facilities

(d) Shop Facilities

(e) Education Facilities

Figure 9: Distributions of synthetic activity facility capacities in 2010

In addition to the capacity, the real estate prices, including selling price and rent, are another important attributes to residential facilities, as the prices could influence the residential locations, for example. The synthesis method for facility prices described in Section 3.4 was applied to generate the 2010 disaggregate selling prices and rents for each home facility, which are shown in Figure 10-(a) and -(b), respectively. It can be seen from the maps that the residential facilities with higher prices (either selling prices or rents) tended to be located at the central districts where more activity opportunities (e.g., shopping activities) were available, as shown by Figure 8.

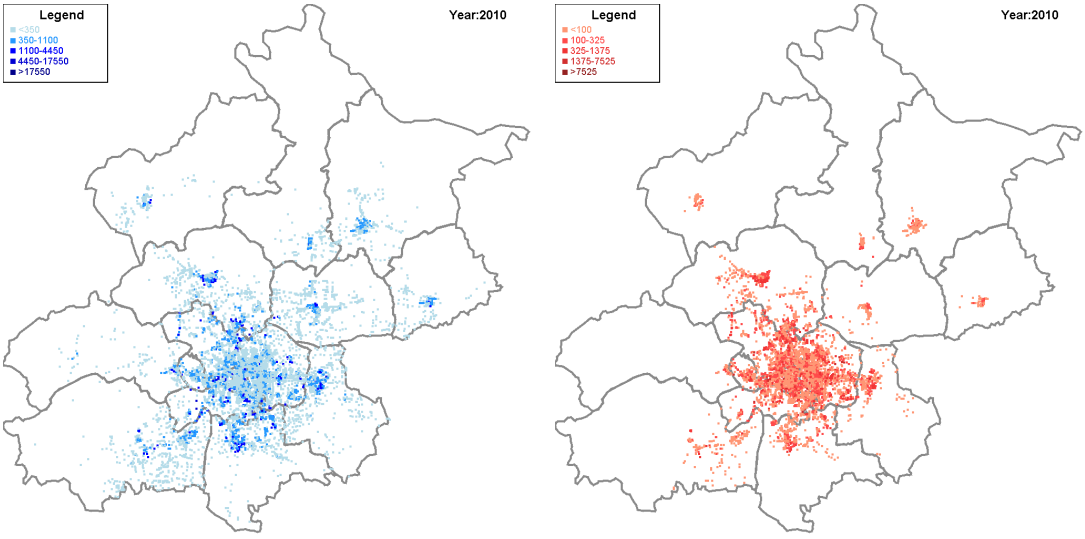


(a) Selling Price (Yuan/Square Meter) (b) Rent (Yuan/Square Meter)

Figure 10: Maps of synthetic real estate prices in 2010

(3) Synthetic transport facilities with capacity information attached

The node-based transport facility synthesizer introduced in Section 3.4 was employed to generate parking facilities, given the total number of parking spaces. For the private parking lots, it was assumed that all vehicle owners have dedicated parking spaces located at their home facilities. The synthetic private and public parking spaces are shown in Figure 11-(a) and -(b), respectively. From the maps, it can be found that those parking lots located at the central districts tend to have higher capacities (or more parking spaces).



(a) Private Parking Spaces (b) Public Parking Spaces

Figure 11: Map of synthetic parking spaces by capacity in 2010

The refueling stations in 2010 were synthesized using the activity-based flow-refueling model introduced in Section 3.4 with a constraint on the total number of refueling posts at stations, which was estimated as 4296 based on the report of "Planning for the Development of Refueling Stations from 2009 to 2015"(BMCCM, 2009). Figure 12 shows the map of the synthetic refueling stations. It can be found that the synthetic stations have tended to be located at the central areas of districts. The reasons are discussed as follows: on one hand, it should be noted that the capacity of one synthetic refueling station is 1 refueling post that represents 25 refueling posts in reality, as the population scaling rate is set to 0.04, which means all the facilities, including both transport and activity facilities, need to be scaled down accordingly in the simulation; on the other hand, although there were some refueling stations located in rural areas of each district in the real world, the number was relatively small. As a result, those stations in rural areas need to be aggregated to the stations located at the central areas of districts where the traffic flow tends to be much heavier (meaning that the refueling demand tends to be higher).

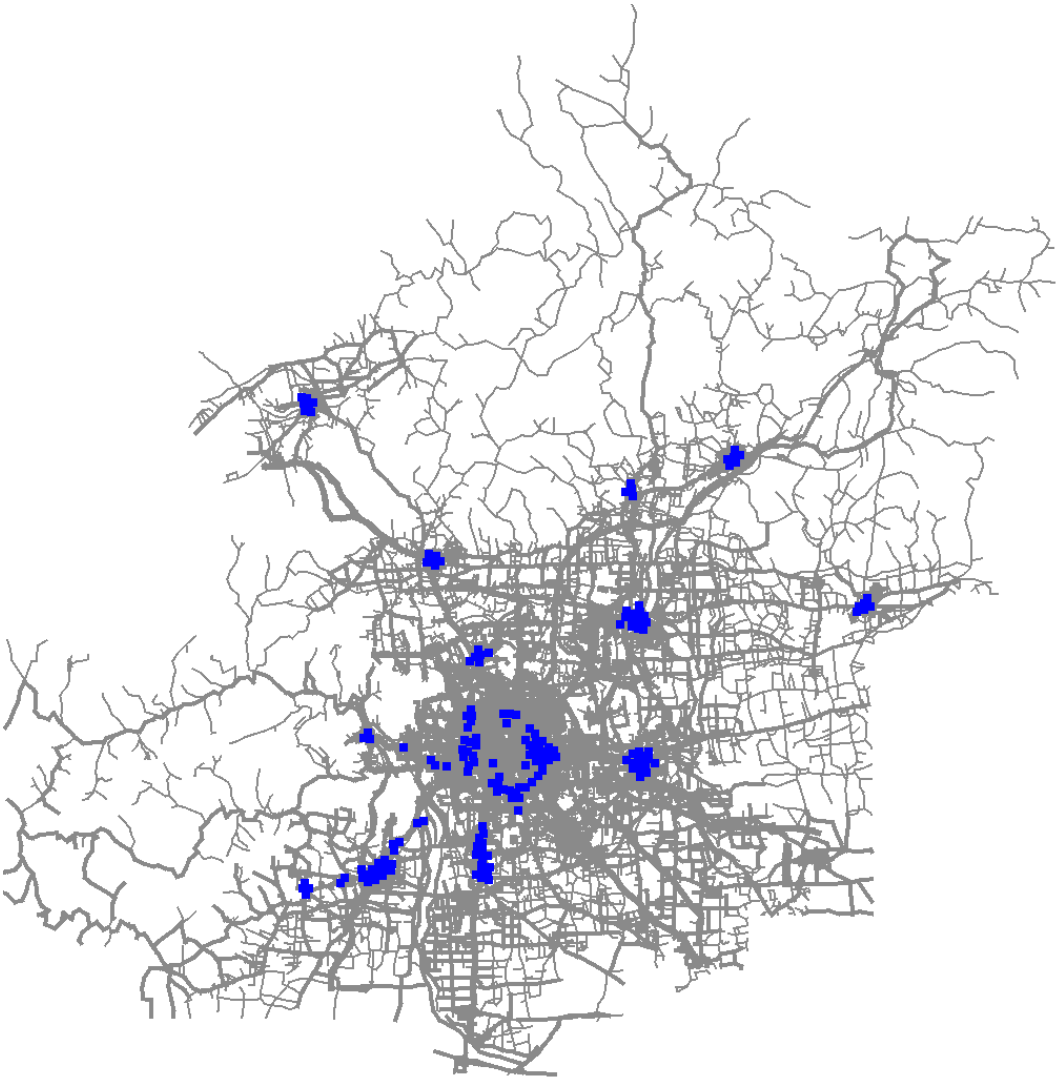
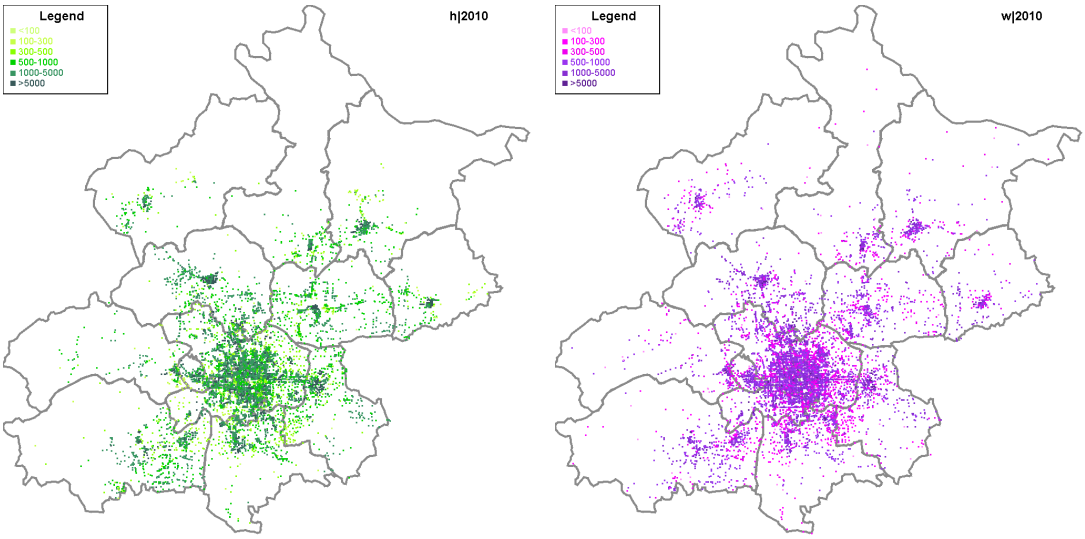


Figure 12: Map of synthetic refueling stations in 2010

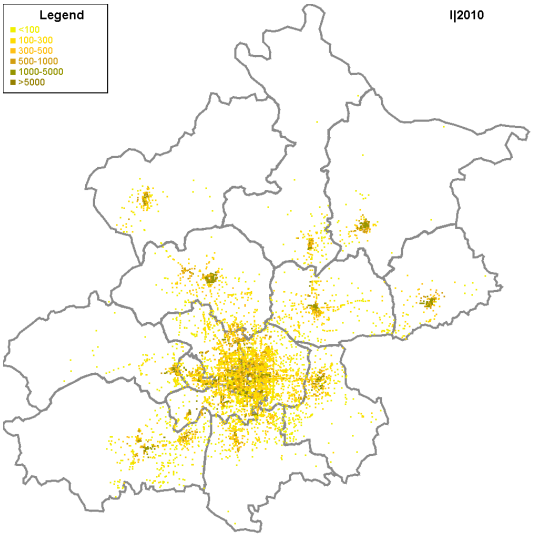
4.4 Linking individuals to the physical environment: Generating initial daily Plans

As introduced in Section 3.5, the generation of initial daily plans for each agent in the synthetic population can be viewed as a process of linking individuals to the physical environment that is composed of both activity and transport facilities. In this case study, only 4% of the whole population, which is around 788,000 agents, is generated, because it is computationally expensive to process the whole population. The number of generated initial daily plans for the synthetic population is around 598, 000 (it should be noted that some agents do not have daily plans, especially children aged below 8), which comprises around 1,676,000 trips and 2,274,000 activities. Among the trips, about 19% of them are car-based. The number of trips per person is around 2.8. The activity locations in the resulting initial daily plans are further mapped and shown by Figure 13. Figure 14 shows the distribution of the numbers of activities performed.

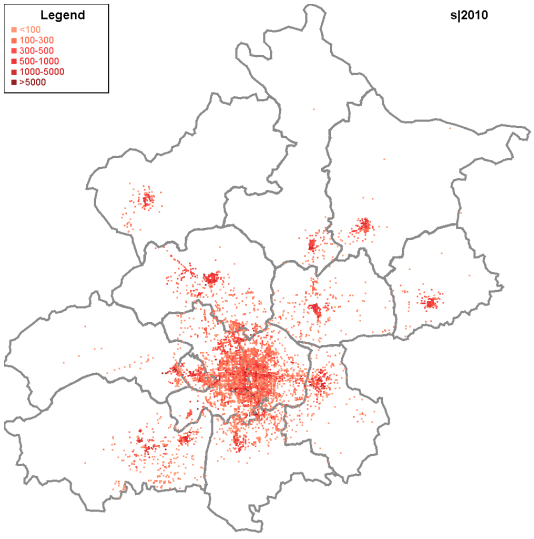


(a) Home Activities

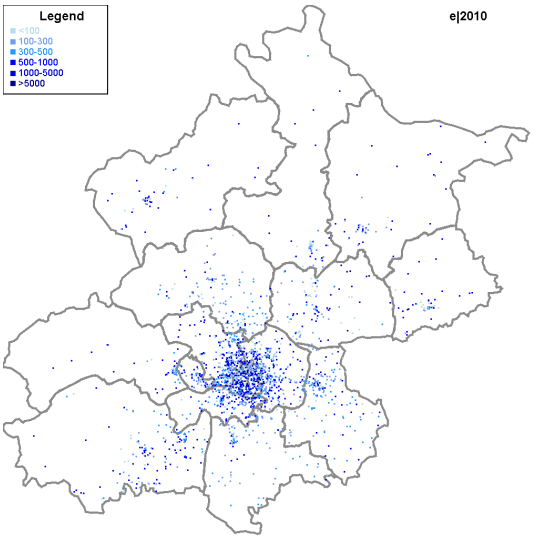
(b) Work Activities



(c) Leisure Activities

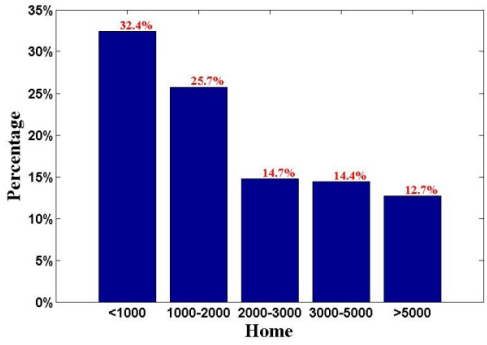


(d) Shopping Activities

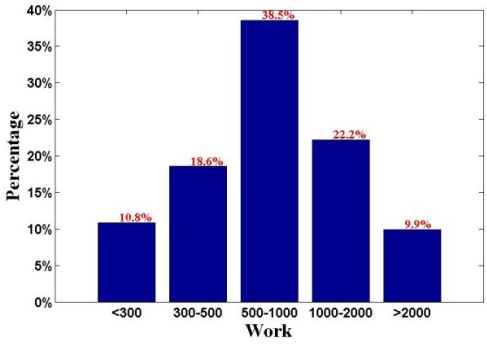


(e) Education Activities

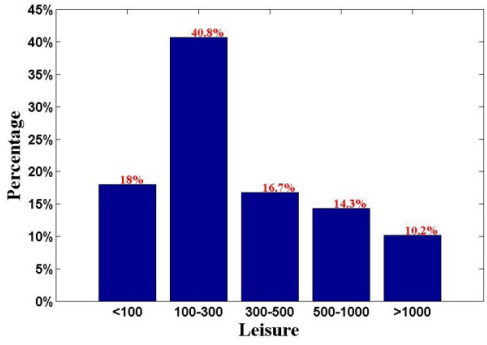
Figure 13: Spatial distributions of activity locations by activity type in 2010



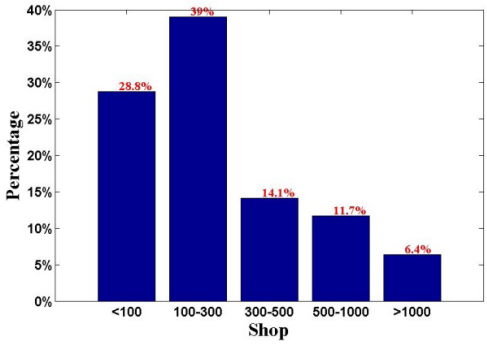
(a) Home Activities



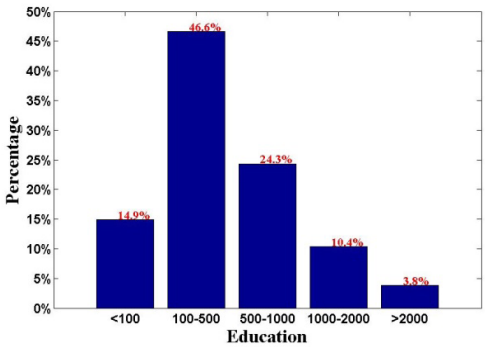
(b) Work Activities



(c) Leisure Activities



(d) Shopping Activities



(e) Education Activities

Figure 14: Distributions of the number of activities by activity type in 2010

5 Conclusions

This paper proposes a virtual city creator that is a set of disaggregate data synthesis methods, including a population synthesizer, an activity facility synthesizer, a transport facility synthesizer and a daily plan generator. These synthesis methods are run in a specific sequence to create a virtual city containing individuals, households, transport and activity facilities, as well as their attributes and linkages. The creator has been applied to Beijing, China to generate an agent- and GIS-based virtual Beijing in 2010 mainly using the household travel survey data. Specifically, the GA-based population synthesizer in the creator was firstly used to generate a synthetic population comprising individuals and households, with the target frequencies well matched at both individual- and household- levels. Then the activity facility synthesizer was used to generate the GIS-based layouts of five typical activity facilities (namely "home", "work", "education", "leisure" and "shopping" facilities) with the capacity information attached. Similarly, the transport facility synthesizer was used to generate the GIS-based layouts of parking lots and refueling stations also with the capacity information attached. Finally, the daily plan generator was used to generate initial daily plans for each individual in the population by linking them to different activity facilities. The above individuals, households, activity facilities, transport facilities, and their attributes and linkages make up a virtual Beijing in 2010.

In theory, the virtual city creator can be applied to any cases where the general input data is available, including household travel survey data, a road network and some relevant macro-level constraints. Among them, the household travel survey data is the most important, and the extent to which the resulting virtual city can represent the real world heavily depends on the data quality. Therefore, the quality of the household travel survey data needs to be checked before the data is used for creation. In addition, population scaling, which uses a small fraction of agents and facilities for a simulation and then scales up the simulation results accordingly, has been a commonly used approach to speeding up agent-based large-scale simulations that contain a large number of agents. For instance, the case study in this paper created a virtual Beijing with a population scaling factor of 0.04. However, it remains unknown how such a population scaling method may influence the resulting virtual city and further the subsequent simulations. It is therefore suggested to choose an appropriate scaling factor when the virtual city creator is applied.

The future work on the virtual city creator will be focused on the following aspects: (1) Model Validation. Although some macro-level constrains (e.g., average real estate price) have been used to ensure that the resulting virtual city can well represent the real world at the macro level, the extent to which the virtual city can represent at the mesoscopic or microscopic levels still remains to be further explored.

For example, the distributions of activity facility capacities (Figure 9) and real estate prices (Figure 10) can be further compared to the real distributions, so as to better understand the model performance. However, such real data is generally difficult to access. (2) Supplementary Big Data. Big data, which is recently popular, generally contains rich disaggregate information and thus can potentially serve as a good dataset for the virtual city creation. Therefore, it should be possible and useful to further improve the proposed creator by extra using social media data for the synthesis of activity facilities, for example. In general, social media data is collected in an unorganized way and thus is very likely to be unrepresentative. However, it could be a good supplement to the traditional household travel survey data which generally contains a limited number of samples, but is collected in a well-organized way. (3) Adding Social Networks. The creator is currently able to link individuals in the population to the physical environment at the micro level. It can be further extended by incorporating an agent-based social network generator that is capable of linking individuals with social ties (or friendships), resulting in spatial social networks for the whole population.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant No. 51678044). We would also thank Dr. Mike Bithell for discussing with us about modelling.

References

Abraham, J. E., Stefan, K. J., & Hunt, J. (2012). *Population synthesis using combinatorial optimization at multiple levels.* Paper presented at the Transportation Research Board 91st Annual Meeting, Washington, DC.

Bar-Gera, H., Konduri, K. C., Sana, B., Ye, X., & Pendyala, R. M. (2009). *Estimating survey weights with multiple constraints using entropy optimization methods*. Paper presented at the Transportation Research Board 88th Annual Meeting, Washington, DC.

Barthelemy, J., & Toint, P. L. (2013). Synthetic population generation without a sample. *Transportation Science, 47*(2), 266–279.

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

BMCCM. (2009). *Planning for the development of refueling stations from 2009 to 2015.* Beijing: Beijing Municipal Commission of City Management (BMCCM).

Chen, J., & Hao, Q. (2008). The impacts of distance to CBD on housing prices in Shanghai: A hedonic analysis. *Journal of Chinese Economic and Business Studies*, *6*(3), 291–302.

Chen, T. D., Kockelman, K. M., & Khan, M. (2013). *The electric vehicle charging station location problem: A parking-based assignment method for Seattle*. Paper presented at the Transportation Research Board 92nd Annual Meeting, Washington, DC.

Chu, Z., Cheng, L., & Chen, H. (2012). *A review of activity-based travel demand modeling*. Paper presented at the The Twelfth COTA International Conference of Transportation Professionals, Beijing, China.

Dong, C., Shao, C., Ma, Z., Zhuge, C., & Li, Y. (2012). Temporal-spacial characteristic of urban expressway under jam flow condition. *Jiaotong Yunshu Gongcheng Xuebao, 12*(3), 73–79.

Ettema, D. (2011). A multi-agent model of urban processes: Modelling relocation processes and price setting in housing markets. *Computers, Environment and Urban Systems, 35*(1), 1–11.

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

Filatova, T., Parker, D., & Van der Veen, A. (2009). Agent-based urban land markets: agent's pricing behavior, land prices and urban land use change. *Journal of Artificial Societies and Social Simulation*, *12*(1), 3.

Filatova, T., Parker, D. C., & van der Veen, A. (2007). *Agent-based land markets: heterogeneous agents, land prices and urban land use change*. Paper presented at the Proceedings of the 4th Conference of the European Social Simulation Association (ESSA'07), Toulouse, France.

Frade, I., Ribeiro, A., Gonçalves, G., & Antunes, A. (2011). Optimal location of charging stations for electric vehicles in a neighborhood in Lisbon, Portugal. *Transportation Research Record: Journal of the Transportation Research Board, 2252,* 91–98.

Gao, J., Zhao, P., Zhuge, C., & Zhang, H. (2012). Research on public transit network hierarchy based on residential transit trip distance. *Discrete Dynamics in Nature and Society*, 2012. doi: 10.1155/2012/390128

Gargiulo, F., Ternes, S., Huet, S., & Deffuant, G. (2010). An iterative approach for generating statistically realistic populations of households. *PloS One*, *5*(1), e8828.

Ge, Y., Meng, R., Cao, Z., Qiu, X., & Huang, K. (2014). Virtual city: An individual-based digital environment for human mobility and interactive behavior. *Simulation, 90*(8), 917–935.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 28*(1), 100–108.

Hodgson, M. J. (1990). A flow—capturing location— allocation model. *Geographical Analysis, 22*(3), 270–279.

Hodgson, M. J., & Rosing, K. E. (1992). A network location-allocation model trading off flow capturing and p-median objectives. *Annals of Operations Research, 40(*1), 247–260.

Horni, A., Nagel, K., & Axhausen, K. W. (2016). *The multi-agent transport simulation MATSim*. London: Ubiquity.

Huang, Q., Parker, D. C., Filatova, T., & Sun, S. (2014). A review of urban residential choice models using agent-based modeling. *Environment and Planning B Planning and Design, 41*(4), 661–689.

Kim, J.-G., & Kuby, M. (2012). The deviation-flow refueling location model for optimizing a network of refueling stations. *International Journal of Hydrogen Energy*, *37*(6), 5406–5420.

Kuby, M., & Lim, S. (2005). The flow-refueling location problem for alternative-fuel vehicles. *Socio-Economic Planning Sciences, 39*(2), 125–145.

Kuby, M., Lines, L., Schultz, R., Xie, Z., Kim, J., & Lim, S. (2009). Optimization of hydrogen stations in Florida using the flow-refueling location model. *International Journal of Hydrogen Energy, 34*(15), 6045–6064.

Lau, K. M., & Li, S.-M. (2006). Commercial housing affordability in Beijing, 1992–2002. *Habitat International, 30*(3), 614–627.

Li, S., & Huang, Y. (2015). *Optimal design of electric vehicle charging infrastructure network–A case study of South Carolina*. Paper presented at the Transportation Research Board 94th Annual Meeting, Washington, DC.

Li, X., Zhuge, C., Zhang, X., Gao, J., & Zhang, H. (2014). Multiobjective optimization model of residential spatial distribution. *Mathematical Problems in Engineering*, 2014, 1-9. doi: http://dx.doi.org/10.1155/2014/167495.

Lim, S., & Kuby, M. (2010). Heuristic algorithms for siting alternative-fuel stations using the flow-refueling location model. *European Journal of Operational Research, 204*(1), 51–61.

Liu, X., & Long, Y. (2016). Automated identification and characterization of parcels with OpenStreetMap and points of interest. *Environment and Planning B: Planning and Design, 43*(2), 341–360.

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

Macal, C. M., & North, M. J. (2010). Tutorial on agent-based modelling and simulation. *Journal of Simulation, 4*(3), 151–162.

Magliocca, N., Safirova, E., McConnell, V., & Walls, M. (2011). An economic agent-based model of coupled housing and land markets (CHALMS). *Computers, Environment and Urban Systems,* 35(3), 183–191.

McMillen, D. P. (2002). The center restored: Chicago's residential price gradient reemerges. E*conomic Perspectives — Federal Reserve Bank of Chicago, 26*(2), 2–11.

Müller, K., & Axhausen, K. W. (2010). *Population synthesis for microsimulation: State of the art.* Zürich: ETH Zürich, Institut für Verkehrsplanung, Transporttechnik, Strassen-und Eisenbahnbau (IVT).

Müller, K., & Axhausen, K. W. (2011). *Hierarchical IPF: Generating a synthetic population for Switzerland*. Zürich: Eidgenössische Technische Hochschule Zürich, IVT.

Parker, D. C., & Filatova, T. (2008). A conceptual design for a bilateral agent-based land market with heterogeneous economic agents. *Computers, Environment and Urban Systems, 32*(6), 454–463.

Pinjari, A. R., & Bhat, C. R. (2011). Activity-based travel demand analysis. *A Handbook of Transport Economics, 10,* 213–248.

ReVelle, C. S., & Swain, R. W. (1970). Central facilities location. *Geographical Analysis, 2*(1), 30–42.

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

Twomey, P., & Cadman, R. (2002). Agent-based modelling of customer behavior in the telecoms and media markets. *Info*, *4*(1), 56–63.

Upchurch, C., & Kuby, M. (2010). Comparing the p-median and flow-refueling models for locating alternative-fuel stations. *Journal of Transport Geography, 18*(6), 750–758.

van Bergeijk, P. A. (2012). *The relation between land price and distance to CBD in Bekasi.* Master's thesis. Erasmus University, Rotterdam.

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

Wang, Y.-W., & Lin, C.-C. (2009). Locating road-vehicle refueling stations. *Transportation Research Part E: Logistics and Transportation Review, 45*(5), 821–829.

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

Zhang, H., Shi, B., Yu, X., Mou, Z., Li, M., Wang, L., & Zhuge, C. (2018). Transfer stability of urban subway network with passenger flow: Evidence in Beijing. *International Journal of Modern Physics B, 32*(14), 1850174.

Zhuge, C., Li, X., Ku, C.-A., Gao, J., & Zhang, H. (2016). A heuristic-based population synthesis method for micro-simulation in transportation. *KSCE Journal of Civil Engineering, 21*(6), 2373–2383. doi:10.1007/s12205-016-0704-1

Zhuge, C., & Shao, C. (2018a). Agent-based modelling of locating public transport facilities for conventional and electric vehicles. *Networks and Spatial Economics*. doi:10.1007/s11067-018-9412-3

Zhuge, C., & Shao, C. (2018b). Agent-based modelling of purchasing, renting and investing behavior in dynamic housing markets. *Journal of Computational Science, 27,* 130–146.

Zhuge, C., Shao, C., Gao, J., Dong, C., & Zhang, H. (2016). Agent-based joint model of residential location choice and real estate price for land use and transport model. *Computers, Environment and Urban Systems*, 57, 93–105.

Zhuge, C., Shao, C., Gao, J., Meng, M., & Ji, X. (2014). Evolution prospect and structure of urban transportation and land use system. *Journal of Transportation Systems Engineering and Information Technology, 14*(2), 19–24.

Zhuge, C., Shao, C., Gao, J., Meng, M., & Xu, W. (2014). An initial implementation of multiagent simulation of travel behavior for a medium-sized city in China. *Mathematical Problems in Engineering*, 2014, 1–11. doi: http://dx.doi.org/10.1155/2014/980623.

Zhuge, C., Shao, C., Li, X., Gao, J., & Zhang, H. (2016). *A genetic algorithm based population synthesizer*. Paper presented at the 14th World Conference on Transport Research, Shanghai, China.

Zhuge, C., Shao, C., Wang, S., & Hu, Y. (2017). Sensitivity analysis of integrated activity-based model: Using MATSim as an example. *Transportation Letters*, 2017, 1–11. doi:10.1080/19427867.2017.1286772.

Zilske, M., Neumann, A., & Nagel, K. (2011). OpenStreetMap for traffic simulation. Retrieved from https://depositonce.tu-berlin.de/bitstream/11303/4976/2/zilske_neumann_nagel.pdf.