Simulating Human Mobility with Agent-based Modeling and Particle Filter Following Mobile Spatial Statistics

Mingfei Cai
mfcai@iis.u-tokyo.ac.jp
Department of Civil Engineering, The University of Tokyo
Meguro, Tokyo, Japan

Yanbo Pang
Center for Spatial Information Science, The University of Tokyo
Meguro, Tokyo, Japan
pybdtc@iis.u-tokyo.ac.jp

Takehiro Kashiyama
Institute of Industrial Science, The University of Tokyo
Meguro, Tokyo, Japan
ksym@iis.u-tokyo.ac.jp

Yoshihide Sekimoto
Center for Spatial Information Science, The University of Tokyo
Meguro, Tokyo, Japan
sekimoto@csis.u-tokyo.ac.jp

ABSTRACT

Human mobility datasets collected from various sources are indispensable for analyzing, predicting, and solving emerging urbanization and population issues. However, such datasets are only available to the public after aggregation and anonymous processing. In recent years, agent-based modeling approaches have addressed this problem by reproducing synthetic human mobility data through simulation. However, the development of such agent models typically requires a large amount of personal location histories as training data for parameter learning, leading to cost and privacy concerns. To overcome this disadvantage, we attempted to explore optimal parameters using a particle filter to alleviate the strict requirement of the data. We tested our method in a local city in Japan using aggregated real-time observation data collected from mobile phone service companies. The results show that the proposed model can achieve satisfactory accuracy using low-resolution data and can therefore be easily used by local governments for municipal applications.

CCS CONCEPTS

• **Information systems**; • **Computing methodologies → Modeling and simulation**; • **Networks → Location based services**;

KEYWORDS

human mobility simulation, aggregated data, agent-based model, particle filter

ACM Reference Format:

Mingfei Cai, Yanbo Pang, Takehiro Kashiyama, and Yoshihide Sekimoto. 2021. Simulating Human Mobility with Agent-based Modeling and Particle Filter Following Mobile Spatial Statistics. In *29th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '21), November 2–5, 2021, Beijing, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3474717.3484203

1 INTRODUCTION

In recent years, the knowledge of human mobility has become indispensable for describing urban systems. It can serve as an auxiliary for city planning in various aspects. Simulation is a powerful approach for learning the underlying semantics of human mobility. It can replicate and predict social phenomena using techniques such as agent-based modeling. Real-time human mobility in a large area can be perceived owing to the widespread use of devices. We can utilize multisource data to accurately understand mobility using suitable tools to efficiently mine diversified data.

There are various models for human mobility simulation, such as empirical models [7] and data-driven models [8]. However, most of them require a large amount of individual-based data to ensure accurate performance; these data may be expensive or even unavailable in certain areas. Furthermore, even though trajectory data are anonymous and preprocessed, private information may still be leaked from underlying features, such as locations of overnight stays. In addition, privacy, security, or commercial confidentiality limit the use of mobility data sources outside the academic field. Communities cannot access data. Even for local governments, which may benefit the most from human mobility reconstruction, it is unrealistic to utilize these expensive high-resolution data owing to a limited budget.

Accordingly, none of the aforementioned models are suitable for real-world municipal applications. Generally, the demand for human mobility data from local governments has the following characteristics:

- **Low cost**. The price of training data should not exceed the budget.
- **High resolution**. The resolution of data must at least maintain the individual features of the population within a governing district.
- **Real-time data**. The data must be real time to perform time-series analysis.

SIGSPATIAL '21, November 2–5, 2021, Beijing, China

ACM ISBN 978-1-4503-8664-7/21/11…$15.00
https://doi.org/10.1145/3474717.3484203

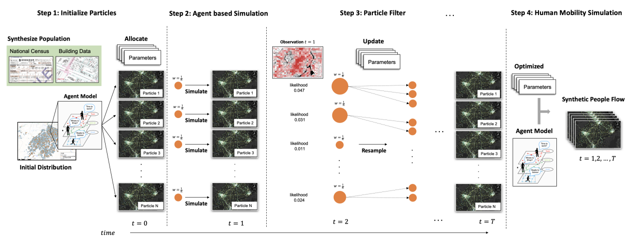

Figure 1: Overview of this study.

Most of the previous research has focused more on the accuracy of simulation results than on cost and actual applications. Our goal was to use low-cost data to reconstruct human mobility for local governments. In this study, rather than using trajectory data for training, we use real-time aggregated mesh data to facilitate human mobility reconstruction. We select Nanto in Toyama Prefecture for analysis. It is a typical local city with an appropriate size. Additionally, a large amount of observation data is not available for this city, unlike Tokyo and Osaka. We synthesize the population from the national census and implement an agent-based model to retrieve the resolution. Thus, we can disaggregate mesh observations and investigate individual behaviors. Moreover, we apply a particle filter (PF) to explore the potential solution space and efficiently search for the best case.

Overall, considering the behavior of all agents, we can maintain accuracy at the macroscopic and microscopic levels. The results indicate that the proposed model can simulate human trajectories with high accuracy while considering mesh observations and open-source census data. Figure 1 presents an overview of this study. The contributions of this study are as follows:

- We develop an agent-based model with multilevel decision-making rules, which can reconstruct citywide human trajectories using low-cost data and is suitable for local governments.
- We design a PF that is customized for human mobility to describe the non-linear relationship between aggregated mesh data and discrete trajectories, thereby solving a one-to-many dilemma.
- We test the results based on diversified metrics for human mobility to comprehensively analyze our output and prove its validity.

The remaining sections are organized as follows. Section 2 introduces the methodology. Section 3 describes the experimental details. Section 4 presents the results. Section 5 summarizes the paper.

2 METHODOLOGY

2.1 Agent-based Model

Agent-based modeling is a powerful simulation technique for disaggregating data and analyzing travel behaviors at the individual level [2]. As the simulated trajectories generated from agent models do not refer to an individual's actual behavior, the results may be open to the public without privacy concerns. Most individual daily travel patterns follow simple rules [10], which are of seventeen unique types [9]. In this study, we further simplify these travel patterns into three types depending on the Nationwide Person Trip Survey in Japan [1] and assign these travel patterns to the synthetic population based on demographic attributes, as follows:

- **Workers** follow the home-work-home activity chain. This type comprises 70% of the population aged 18–65 years. The location choice procedure assigns a working place to agents.
- **Students** follow the home-school-home activity chain. They are aged 6–18 years. Each agent is assigned the school that is the closest to a home.
- **Others** follow the home-random place-home activity chain. This type includes homemakers, freelancers, and other non-commuters.

We use the discrete choice model [6] to determine the destination for each agent. In particular, the utility of individual n associated with alternative i in choice set C_n is given by

$$U_{in} = V_{in} + \varepsilon_{in} \tag{1}$$

where ε_{in} is a random term that captures the unobserved characteristics of individuals and alternatives as uncertainty, and V_{in} is the deterministic part of the utility value.

An agent evaluates the utility of alternatives in the choice set, which captures the preference. The attributes consist of the night population, number of companies and places of business, number of employees, and travel distance. Therefore, the probability of alternative i being chosen by individual n is derived using the multinomial logit model [4], as follows:

$$Pr(i \mid C_n) = \frac{e^{\mu V_{in}}}{\sum_{j \in C_n} e^{\mu V_{jn}}} \tag{2}$$

Similarly, there is no direct information about the departure/return time choices of individuals. Instead, we capture the departure time distribution from the Nationwide Person Trip Survey in Japan and sample random variables from the distribution to determine the behavior.

Based on the Nationwide Person Trip Survey in Japan until 2005, we assign the transportation mode of "**Vehicles**" to the trips with distances of over 1 km and "**Walk**" to the remaining trips. We use the shortest-path algorithm to calculate the route [5] considering the transportation mode and derive the movement speed from the road network conditions.

2.2 Particle Filter

A PF is a filtering method that can approximate a probability density function (PDF) with an intractable integral for the posterior using Monte Carlo approaches [3]. It is suitable for nonlinear and non-Gaussian dynamic systems such as cities. In this study, we use the PF to analyze the results of the agent-based model. However, the setting of the human mobility problem is complex. Rather than monitoring one object, the proposed method must consider all agents within one city, thereby significantly increasing the dimensionality. Thus, we must design a customized PF for the human mobility problem.

In our setting, each particle represents one possible simulated world, including the sequence of state vector $\boldsymbol{x}_C$, that is, the trajectories and the other settings. The state refers to the position set of agents.

In the sampling phase, the PF generates # particles and lets the agent-based model determine the next position of all individuals. Then, in the update phase, the PF uses real-time observation $\boldsymbol{o}_C$ to correct the prior probability distribution from the agent-based modeling and obtain the posterior probability distribution. Every particle is weighted to measure the likelihood of matching the actual population distribution. The result is the expectation of state $\boldsymbol{x}$.

To prevent the degeneracy problem, which implies that only a small number of particles contribute to the result, the PF uses multinomial resampling to preferentially select particles with a higher probability to replace those with a lower probability. In addition, the PF retains sufficient particles with a lower probability to ensure that it can detect strongly nonlinear behaviors. Thus, the PF can select the results that are closer to the truth from agent-based modeling. We perform the calculation at different time steps to recursively evaluate the weight using Equation 3.

$$F_C^8 \propto F_{C-1}^8 \frac{?(\boldsymbol{o}_C|\boldsymbol{x}_C^8)?(\boldsymbol{x}_C^8|\boldsymbol{x}_{C-1}^8)}{@(\boldsymbol{x}_C^8|\boldsymbol{x}_{C-1}^8, \boldsymbol{o}_C)} = F_{C-1}^8 ?(\boldsymbol{o}_C|\boldsymbol{x}_C^8) \tag{3}$$

In this study, we adopt Equation 4 to calculate the likelihood $?(\boldsymbol{o}_C|\boldsymbol{x}_C^8)$, between the predicted and observed population distribution. An advantage of this method is that a mesh with a larger population in the ground truth will matter more.

$$?(\boldsymbol{o}_C|\boldsymbol{x}_C^8) = \frac{\boldsymbol{o}_C^2}{||\boldsymbol{o}_C - \boldsymbol{x}_C^8||^2} \tag{4}$$

3 EXPERIMENT

The target area of the experiment is Nanto, Toyama Prefecture, which is located near to the center of Japan. The area of Nanto is $668.64\ :<^2$, and it is a typical local city with a simple urban structure and no nearby metropolitan region; therefore it is appropriate for the analysis.

NTT Docomo Inc. provides aggregated mobile phone data. It preprocesses raw call detail record data and converts it to aggregated data by enumerating the number of people within the scope to conceal private information. The aggregated data have the format of timestamp, mesh ID, and population. The data focus on a 500 m × 500 m mesh, and the definition of the mesh is provided by the Ministry of Land, Infrastructure, Transport, and Tourism of Japan. We use the data for October 15, 2019, which include users of the NTT Docomo telecommunication network. Open-source data are also used. The model uses the Japan National Census to initialize agents and the Nationwide Person Trip Survey in Japan to determine the behaviors in the agent-based model. These two sets of data are obtained from the Statistics Bureau of Japan [11]. Moreover, Zerin Co., Ltd. [12] provides infrastructure data for the initialization of agents and al data for the determination of features.

We generate 53601 agents for the simulation. The number of facilities in Nanto and Toyama Prefecture is 34053 and 351571, respectively. The simulation starts at 5 AM and ends at 9 PM. The time interval for agents is 6 min. We use multi-observation to analyze the consequences in detail. The experiment is performed on the Amazon Web Services EC2 m5.8xlarge instance with 32 CPU cores and 128 GB memory.

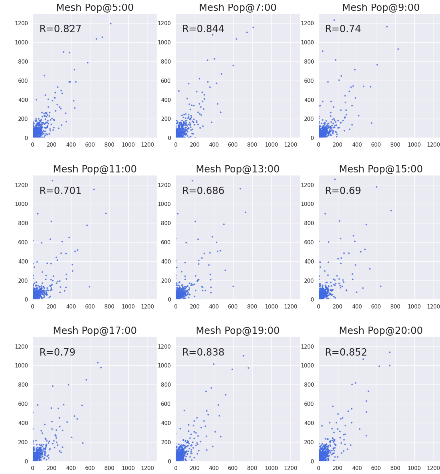

Figure 2: The Pearson Correlation Coefficient between the observation and simulation. Each dot represents a 500m × 500m grid in the target area, while x-axis is the observation and y-axis is the simulation. R is the value of the COR, for which the closer to one, the better result. The points close to the y-axis mean the overestimate error and those close to the x-axis mean the underestimate error. From the COR, the accuracy degenerates in the middle of day.

4 RESULTS

First, we analyze the correlation coefficient (COR) between the predicted and observed data. The scatter plot for different hours is shown in Figure 2. The COR is more than 0.8 early in the morning and late at night. This is because most agents stay at home; hence, movement patterns are not related to the population distribution. The COR is approximately 0.7 in the middle of the day; this is acceptable for use of coarse-grained training data. The COR decreases significantly at 9 AM; this is when most commuters are on the trips to their destinations. The distribution of points shows that the number of points close to the y-axis increases during the rush hour, which leads to a low COR. This indicates that the proposed model overestimates agents in several meshes. As the scatter plot does not elucidate the spatial distribution, we compare the results on the map.

Figure 3 illustrates the comparison of the predicted and observed population distribution. Note that there are four reddish clusters within the city, which are the downtown areas with numerous facilities. The size of the clusters in the ground-truth data is smaller than the predicted size. This suggests that the proposed model can find the center of the district but is insensitive to its border. In addition, the proposed model simulates a smaller population in the core urban area. The difference between the predicted and observed population is remarkable from the rush hour until the afternoon,