Transportation Research Part C 117 (2020) 102706

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

# Development of people mass movement simulation framework based on reinforcement learning

Yanbo Pang[a,*], Takehiro Kashiyama[a], Takahiro Yabe[a,b], Kota Tsubouchi[c], Yoshihide Sekimoto[a]

[a] Institute of Industrial Science, University of Tokyo, Ce507, 4-6-1, Komaba, Meguro-Ku, Tokyo 153-8505, Japan
[b] Lyles School of Civil Engineering, Purdue University, United States
[c] Yahoo Japan Corporation, Japan

ARTICLE INFO

*Keywords:*
Travel demand modeling
Reinforcement learning
Mobility data
Citywide people mass movement simulation

ABSTRACT

Understanding individual and crowd dynamics in urban environments is critical for numerous applications, such as urban planning, traffic forecasting, and location-based services. However, researchers have developed travel demand models to accomplish this task with survey data that are expensive and acquired at low frequencies. In contrast, emerging data collection methods have enabled researchers to leverage machine learning techniques with a tremendous amount of mobility data for analyzing and forecasting people's behaviors. In this study, we developed a reinforcement learning-based approach for modeling and simulation of people mass movement using the global positioning system (GPS) data. Unlike traditional travel demand modeling approaches, our method focuses on the problem of inferring the spatio-temporal preferences of individuals from the observed trajectories, and is based on inverse reinforcement learning (IRL) techniques. We applied the model to the data collected from a smartphone application and attempted to replicate a large amount of the population's daily movement by incorporating with agent-based multi-modal traffic simulation technologies. The simulation results indicate that agents can successfully learn and generate human-like travel activities. Furthermore, the proposed model performance significantly outperforms the existing methods in synthetic urban dynamics.

## 1. Introduction

Recently, the rapid urbanization and the consequent population rise have presented urban planners with various challenges such as traffic congestion, longer commutation, accidents, loss of public space, and disaster management. Travel demand models are used to forecast the response of transportation demand to the variations in the attributes of the transportation system and differences in the characteristics of the people using the transportation system (Bhat and Frank, 1999). Over the last few decades, travel demand modeling approaches have shifted from traditional trip-based approaches (Ruiter and Moshe, 1978; Adler and Ben-Akiva, 1979) to activity-based approaches (Bowman and Moshe, 2001). Several transportation researchers integrated features such as activity participation, time duration choice, the order of the activities, and transport mode choice to travel to the activity location at the individual level. However, developing such models is highly dependent on household travel or time-use surveys (Moshe and Bierlaire,

* Corresponding author at: Institute of Industrial Science, the University of Tokyo, Ce507, 4-6-1, Komaba, Meguro-Ku, Tokyo 153-8505, Japan.
*E-mail addresses:* pybdtc@iis.u-tokyo.ac.jp (Y. Pang), ksym@iis.u-tokyo.ac.jp (T. Kashiyama), tyabe@purdue.edu (T. Yabe), ktsubouc@yahoo-corp.jp (K. Tsubouchi), sekimoto@iis.u-tokyo.ac.jp (Y. Sekimoto).

https://doi.org/10.1016/j.trc.2020.102706
Received 30 April 2019; Received in revised form 8 May 2020; Accepted 17 June 2020
Available online 09 July 2020

1999), which are updated synchronously at infrequent intervals, and incur significant expenses and time. The results obtained from the collected data are limited to typical scenarios. Additionally, the spatial–temporal interactions in activity sequences are often straightforward; the location choice is modeled on the traffic analysis zone (TAZ) level, and time is split into several periods (such as morning peak, evening peak, and others), which cannot provide enough details to decision-makers.

In contrast, with the explosion of the information and communications technology (ICT) and Internet of Things technologies (IoT), emerging data collection methods have enabled researchers to unravel individual mobility patterns and to generate models that can reproduce the time-varying characteristics in human trajectories. For example, high-quality geolocated data such as call detail records (CDRs), global positioning system (GPS), and social media data, have quickly surpassed traditional high-cost data (i.e., census and travel surveys). Currently, such datasets are used as primary data resources and have promoted a series of data-driven approaches for analyzing and modeling human mobility (Gonzalez et al. 2008; Zheng et al. 2008; Song et al. 2010; Cho et al, 2011; Isaacman et al., 2012; Huang et al., 2018), estimating traffic volume (Toole et al. 2015; Polson and Sokolov, 2017; Wu et al. 2018), forecasting crowd congestion (Fan et al. 2014; Zhao et al. 2018), and disaster response (Lu et al., 2012; Song et al. 2016; Yabe et al. 2017). The huge volume of human mobility data also enables researchers to apply machine learning techniques for predicting the next destination (Feng et al. 2018; Altaf et al., 2018), detecting transport mode (Zheng et al. 2014; Witayangkurn et al. 2013), and activity recognition (Shan et al., 2017; Pang, 2018; Arora and Doshi, 2011; Goullias et al., 1990; Yuxi, 2017; National Land Numerical Information, 2020; Smart and Kaelbling, 2002; Zhang et al., 2014). Recently, several studies have exploited the generative power of such deep learning models. Song et al. (2016) built an intelligent mobility prediction and simulation system based on a multi-task long short-term memory (LSTM) model. Yin et al. (2017) analyzed and modeled the daily activity pattern using an input–output Hidden Markov Model (IOHMM) model and CDR data. Ouyang et al. (2018) integrated location embedding with generative adversarial networks to model and generate synthetic human mobility trajectories. However, both of these approaches inevitably require copious amounts of labeled training data to train the models before they begin to give persuasive results.

As a promising branch of machine learning, reinforcement learning (RL) has already been applied to sequential decision-making problems in various fields (Sutton and Barto, 2018). Unlike other machine learning algorithms that need to learn from a training dataset and then apply the trained model to the new dataset, RL is an autonomous, self-teaching system that essentially learns through trial-and-error. The agents employing RL generate different episodes (sequence of actions) and learn from the feedback, irrespective of whether that leads to a good result, and then reinforce the actions that work; otherwise, they decrease the probability of choosing these actions. The model can be straightforwardly applied in solving sequential decision-making problems in control theory. Examples include traffic flow management (Zhu and Ukkusuri, 2014), autonomous driving (Zhu et al., 2018; Ye et al., 2019), and route planning (Ramos et al., 2018; Nazari et al., 2018). Recently, several works have adapted the RL approach to model and synthesize daily activity schedules (Janssens et al., 2007; Yang et al. 2014; Pang et al., 2018). Furthermore, Feygin (2018) introduced the inverse reinforcement learning (IRL) approach for the daily activity schedule planning and recovered the absent reward function from the observed expert demonstrations. However, both of these studies only formulate a daily activity (i.e., home, work, school, and other) as an Markov decision process (MDP), few studies have incorporated both location choice and spatial mobility patterns in directly reconstructing the individual daily trajectories directly.

Given this background, in this study, we aim to develop a novel RL-based travel demand modeling and simulation framework that can model individual daily travel behavior decision-making and replicate the people flow dynamics on a citywide level. The advantages of introducing RL are manifold. First, it provides a framework for an intelligent control of agent behavior instead of manually setting the behavior rules. Second, benefitting from the development of deep learning, RL can resolve the problem of larger choice sets. Third, the reward is a natural metaphor of an individual's motivation; it can also represent individual preferences to generate heterogeneous populations.

The main contributions of this study are as follows.

- We developed an RL-based model for people's daily travel behavior in cities and generate synthetic trajectories to replicate people mass movement.
- We introduced IRL to recover human travel behavior preferences that can capture the spatio-temporal pattern and context features of human mobility from real GPS trajectories.
- We applied our approach to two cities in Japan and verified the generated trajectories using a large GPS dataset collected through smartphones.

## 2. Literature review

### 2.1. Travel demand modeling

Over the last few decades, researchers have made significant progress in modeling and estimating the travel demand. Generally, travel demand models are developed to forecast the response of transportation demand to changes in the attributes of people using the transportation system. Correctly put, these travel demand models are used to predict the travel characteristics and utilization of transport services under alternative socioeconomic scenarios, for alternative transport services as well as for land-use configurations (Adler and Ben-Akiva, 1979). Previously, the travel demand approaches used trips as the basic unit of modeling. In detail, a procedure involving of four separate steps is developed; first, estimating the total inflow and outflow of each zone in the target area (trip generation); second, assigning trips to each zone pair (trip distribution); third, determining transport mode of each trip (mode choice), and finally, assigning trips to the road network (trip assignment) (Kitamura, 1984; Goullias et al., 1990). However, trip-based

travel demand approaches failed to model individual daily schedules with trip-chain structure because all the trips generated from the models are separated, and there is no behavior rule to combine them in a rational order. To fill this gap, an activity-based travel demand approach was developed (Kitamura and Fujii, 1998; Bhat and Frank, 1999; Bowman and Moshe, 2001). Therefore, the travel demand is a result of participating in activities at different places and times. Recently, the most contemporary travel demand models developed and employed by regional transportation agencies are the activity-based approach, which employ self-reported surveys to collect information on individuals' socio-demographics, their daily activity purposes, times, and locations. Travel behaviors are regarded as derivatives of activities at different places such as home, work, shopping, and others. Primarily, the discrete choice models are widely used for modeling activity choice, departure time choice, transport model, and other behavioral factors. However, developing activity-based models requires detailed travel behavior surveys; moreover, the data collection is usually expensive and involves significant delay. Because of these limitations, only a typical day's travel demand can be modeled and estimated, which matches the current requirements from the demand side. Furthermore, most of activity-based models consist of multiple modules such as population synthesis, daily travel pattern, workplace choice, tour generation, trip mode choice, trip time and so on, which increases data requirements, model complexity, and computational burden (Donnelly, 2010).

### 2.2. Human mobility modeling with emerging data

With the development of the technologies such as IoT and ICT, individual travel footprints can be sensed and recorded using numerous services and devices. CDRs, which are the most well-used data source of the population, are collected using mobile phone carriers. A record is generated every time a telephone call, text message, or Internet data exchange that passes through these devices. The nearest base tower number records location of the device carrier. The spatial resolution of such records varies from several hundred meters in the central part of an urban area to a few kilometers in rural areas. Conversely, due to the popularization of smartphones, GPS data are also widely used and collected from location-based services such as navigation, check-ins, recommendation, disaster alerts, and advertising (Zheng et al. 2014). Compared to CDR data, the spatial resolution of GPS is less than 10 m in most devices, which enables researchers to mine more detailed information such as the carrier's travel speed, transportation mode, and points of interest. According to the high population coverage, these data are well-studied and leveraged to exploit knowledge and information such as human mobility patterns (Gonzalez et al. 2008; Song et al. 2010; Huang et al. 2018), link traffic volume (Toole et al. 2015; Polson and Sokolov, 2017; Wu et al. 2018), and dynamic crowd density (Fan et al. 2014). By leveraging these powerful data sources, organizations and companies such *Replica* and *StreetLight Data* collect de-identified mobile location data and replicate travel measures to support planning agencies for decision-making.

Moreover, these emerging data sources also enable researchers to apply machine learning techniques to human mobility modeling and analysis. The objectives of these studies revealed multiple dimensions of human mobility such as next place prediction (Song et al., 2004; Prasad and Agrawal, 2010; Altaf et al., 2018), detecting transport mode (Zheng et al. 2014; Witayangkurn et al. 2013) and activity recognition (Shan et al., 2017; Yin et al., 2017). Recently, various generative models have been proposed for generating synthetic mobility trajectories. Song et al. (2016) built an intelligent mobility prediction and simulation system based on a multi-task LSTM model. Yin et al. (2017) and Lin et al. (2017) analyzed and modeled the daily activity patterns with input–output HMM model and the CDR data. Ouyang et al. (2018) integrated location embedding with generative adversarial networks to model and generate synthetic human mobility trajectories. However, both of these approaches inevitably require copious amounts of labeled training data to train the models before they begin to give persuasive results.

### 2.3. Reinforcement learning

RL is based on an agent interacting with the environment, learning an optimal policy through trial and error for sequential decision-making problems in various fields such as robot control (Smart and Kaelbling, 2002; Heess et al. 2017; Schulman et al. 2017), video games (Mnih et al., 2015), and system optimization (Hester and Stone, 2009). The RL theory provides interpretable, psychological, and neuron-scientific perspectives on human behavior, regarding how humans plan their actions in a given environment (Sutton and Barto, 2018). The framework of RL provides a mathematical formalization of intelligent decision making that is powerful and broadly applicable for agent control and can be straightforwardly applied in solving sequential decision-making problems in control theory. Examples include traffic flow management (Zhu and Ukkusuri, 2014), autonomous driving (Zhu et al. 2018; Ye et al. 2019), and route planning (Ramos et al., 2018; Nazari et al., 2018). Recently, several works have adapted the RL approach to model and synthesize daily activity schedules (Janssens et al., 2007; Yang et al. 2014; Xiong, 2014). However, for an extended period, their applications are limited to domains in which agents behave in low-dimensional state spaces with a well-defined reward function (Bertsekas, 2012).

Another promising method is IRL. The objective of IRL is to infer the underlying reward structure guiding an agent's behavior based on observations as well as a model of the environment (Finn et al., 2016; Arora and Doshi, 2011). This may be done either to learn the reward structure for modeling purposes or to provide a method to enable the agent to imitate a demonstrator's specific behavior. Several prior works in this domain rely on the parameterization of the reward function based on the hand-crafted features. Previous studies such as (Ng and Stuart, 2000; Ratliff et al. 2006; Ziebart et al., 2008) expressed the reward function as a weighted linear combination of hand-selected features. Recently, few studies have attempted to fill the gap between traditional discrete choice models (Ermon et al. 2015), human preferences (Pang, 2018; Christiano et al., 2019) and RL. Notably, researchers in the transportation domain have extracted activity sequences from the CDRs data and integrated an activity-based approach with IRL to infer the structural model of travel behaviors (Feygin, 2018). However, the formulation of the model only considers limited activity types

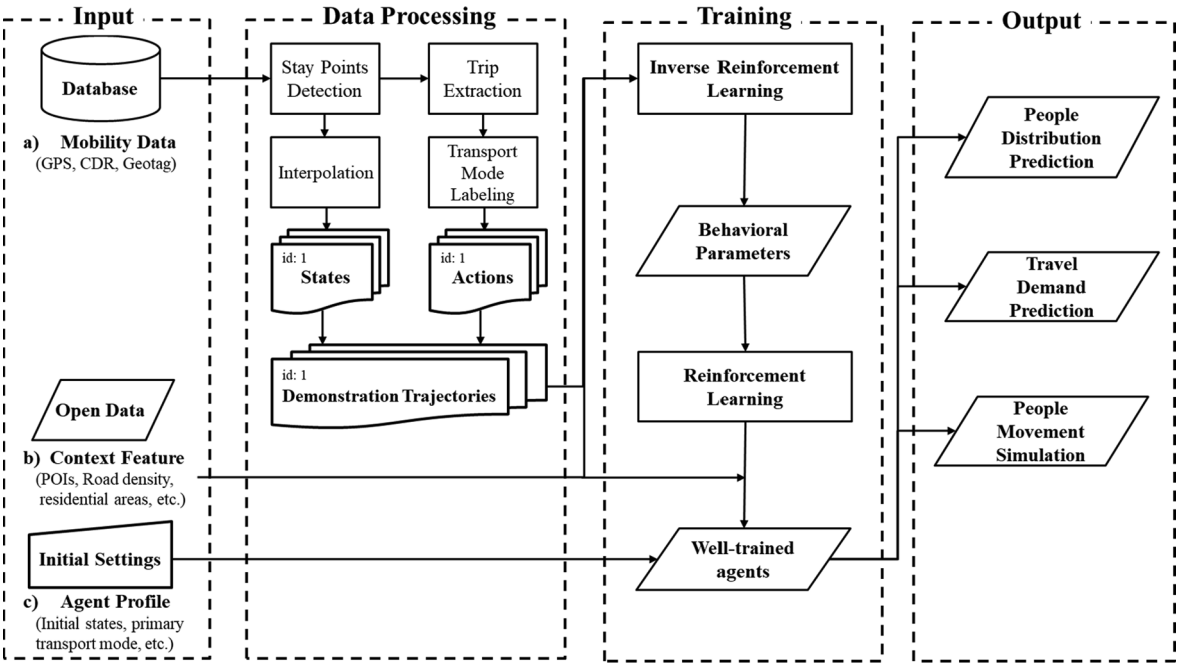


**Fig. 1.** Diagram of the modeling framework.

(home, work, travel by car, and travel by bus) as states, and transitions between such states are considered as actions. Features such as location choice and spatial mobility patterns are not leveraged and incorporated with models to directly reconstruct the individual daily trajectories directly.

In this study, we expand the use of the RL framework to model and simulate people's daily travel behaviors. Unlike most of existing travel demand modeling approaches, our method focuses on the problem of inferring agents' spatio-temporal preferences from observed trajectories based on IRL techniques and attempts to replicate a large amount of the population's daily movement through incorporating of with agent-based multi-modal traffic simulation technologies.

## 3. Methodology

In this study, our aim was to model and simulate the daily human travel behavior based on an RL framework. Although RL enables the agents to learn in an interactive environment through trial-and-error while using feedback from their own experience, there is insufficient knowledge for designing a perfect reward function whose optimization would generate a human-like behavior; therefore, a straightforward RL may not lead to a realistic simulation result. One solution is to mimic the behavior of human beings and characterize the set of reward functions. Hence, human behavioral observations were utilized as training data. As shown in Fig. 1, the end-to-end framework comprises of three parts: developed data processing, agent modeling with parameter training, and agent-based travel micro-simulation.

### *3.1. Data processing*

#### *3.1.1. Mobility data*

The RL agents interact with their environment in discrete time steps and output a series of state-action pairs as trajectories. In this study, the training data should satisfy the following conditions:

- The training data should be represented as a sequence of time-stamped points, each of which specifies a traveler's location.
- The training data should have sufficient temporal and spatial granularity to provide travelers with movement-related decision-making at each time step.

In this study, the GPS data collected from the Yahoo Japan Corporation were used. All data were provided by users who agreed to upload their locations for research purposes through a disaster alert application. The temporal granularity of this dataset is sparser compared to that of many other GPS datasets (such as GeoLife), which are logged in a dense representation, such as every 1–5 s; however, it is similar to that of various CDR datasets used in recent mobility studies.

We first extracted the demonstration data collected from mobile phones and processed the data that enabled the RL agent to learn travel behavior decision making from the observations. In the first step, it was necessary to extract travelers' stay points and infer

their location at each time step. We leveraged the method proposed by (Zheng et al. 2014), which was used to detect stay points and extract travel sequences to cope with the data sparsity issue. We used 1000 m and 30 min as a discretized threshold. Further, the trips' transport modes were classified as "walk", "vehicle" and "train" based on previous work (Witayangkurn et al. 2013). Therefore, the stay points and trips were regarded as states and actions (these are detailed in the following section). The demonstration trajectory of an individual $i$ can be represented as:

$$D^i = \{(s_0^i, a_0^i), (s_1^i, a_1^i), \cdots, (s_T^i, a_T^i)\}. \tag{1}$$

This method overcomes the data sparsity issue. Consequently, it has the potential to be applied with various datasets that are open for research purposes.

#### 3.1.2. Context feature data

To generate feature descriptors $f(\bullet)$, we designed seven environmental factors, namely, the number of offices, the number of employers, the number of schools, the number of evacuation facilities, the number of amusement facilities, the length of roads, railway stations, and the residential density. We leveraged the data in the National Land Numerical Information Database provided by the

Ministry of Infrastructure, Land, and Transport and Tourism of Japan (National Land Numerical Information, 2020). Furthermore, we aggregated these data into each mesh grid ($1km \times 1km$); therefore, each grid is represented by a feature vector.

### 3.2. Model architecture

The MDP is commonly used to model the sequential decision process of a rational agent. In this study, it is used to describe the travel decision-making of a person. In a typical RL problem, all elements of the MDP are assumed to be known, and the task is to estimate an optimal policy$\pi(a|s)$, which by observing rewards maps a state s to an action a. Formally, an MDP is defined as:

$$M = \{\mathscr{S}, \mathscr{A}, T(\bullet, \bullet), R_\theta(\bullet, \bullet)\}. \#$$

**States:** We divide an area into $n$ disjoint cells using a mesh grid size of $r$ as state space $\mathscr{S} = \{s_1, s_2, \cdots, s_n\}$, and $s_t \in \mathscr{S}$ denotes the state at time $t$. Each state can be mapped into a d-dimensional vector $\phi_s$ with a feature descriptor $f(\bullet)$ based on the observed context features.

**Actions:** $\mathscr{A}$ is a set of actions representing all possible travel behaviors. An action $a \in \mathscr{A}$ is denoted as $a = [destination, mode]$ where $a$ indicates the trip destination and transport mode. All locations can be potential destinations for agents. We found that the structure of the road and railway network constrains the accessibility for some areas (i.e., some rural areas are not reachable by railway), making it possible to filter out such useless actions to reduce the action space and computational complexity. The subset action of s is denoted as $\mathscr{A}_s \in \mathscr{A}$.

**Transition function:** The transition function is the system dynamics. This function is a probability distribution over the next possible successor states, given the current state and action. In this study, because the definition of action has determined a specific destination, which represents the next state, we set the transition probability as 1.

**Reward function:** The reward function is indispensable for computing an agent's policy $\pi(a|s)$. An appropriate representation of the reward leads the agent to generate a desirable behavior. However, a "desirable behavior" is complicated to define and varies from person to person. Therefore, in this study, a multi-attribute reward function $R(s, a; \phi)$ is proposed, which defines the immediate reward of action $a$ at state $s$. The feature vector $\phi$ is generated by feature descriptors $f(\bullet)$, which incorporate the information from the current state $s$, in addition to the destination, and travel cost from action $a$ considering the transport mode.

### 3.3. Training procedure

In the framework presented, there are two parts to the requisite procedures for agent training. The first is reward function parameters, which represent travelers' travel behavioral preferences using IRL from demonstration trajectories. Further, with the recovered reward function, the agent learns policies from the environmental inputs by using RL. As the name implies, inverse reinforcement is under the notion of RL; therefore, the algorithm and details about RL are first provided, and then, and later IRL is introduced.

#### 3.3.1. RL using proximal policy optimization

RL provides a mathematical formalization of intelligent decision making that is powerful and broadly applicable for optimal control (Sutton and Barto, 2018). The fundamental idea of RL is to learn a policy $\pi$ that represents the probability of choosing action $a$ in state $s$ in order to optimize the future reward feedback from the environment. The expected future reward at time $t$ can be represented as:

$$R_t = E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} \cdots +] = E\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k}\right]. \tag{2}$$

where $\gamma \in (0, 1)$ is the discount factor. The natural way of selecting an optimal action is to find the action that brings back highest future reward. This action-value can be represented as:

$$Q^{\pi}(s, a) = E_{\pi}[R_t|(s_t = s, a_t = a)] = E_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k}\right]. \#$$ (3)

The RL approach aims to derive an optimal policy that leads to an expected feedback higher than any other policies over all states corresponding to an optimal action-value $Q^*(s, a)$. In our case, the number of states and actions are very large and using classical tabular algorithms such as Q-learning and value iteration may suffer from the consequences such as the curse of dimensionality. Recently, many approaches in large-scale state and action spaces, leveraged the power of neural networks for the approximation of the action value function. The most promising methods are proximal policy optimization (PPO) (Schulman et al. 2017), deterministic policy gradient (DDPG) (Lillicrap et al. 2015), and soft actor-critic (SAC) (Haarnoja et al. 2018). These follow the actor-critic architecture: In the actor module, the policy chooses actions for the agent and the critic module evaluates how good these actions are. In this study, PPO was chosen to solve RL for several reasons. First, it outperforms most of state-of-the-art algorithms in 3D locomotion, and other forms of robot control as the default choice of in OpenAI projects. Second, it computes an update at each step which minimizes the cost function while ensuring that the deviation from the previous policy is relatively small. This characteristic is crucial for training human-like agents aiming to model the daily travel behavior. In addition, PPO's hyperparameters are robust for a large variety of tasks, and they boast higher performance and low computational complexity.

As one of policy gradient methods, PPO alternates between sampling data through interaction with the environment and optimizing a "surrogate" objective function using stochastic gradient ascent. Policy gradient methods start from estimating the policy gradient, and then apply a stochastic gradient ascent algorithm for the estimation. The gradient is typically formulated as:

$$\hat{g} = \hat{E}\left[\nabla_{\theta} log \pi_{\theta}(a_t|s_t)\widehat{A_t}\right]$$ (4)

where $\pi_{\theta}$ is a policy under parameter $\theta$, and $\widehat{A_t}$ is an estimator of the advantage function at time $t$. The advantage function represents how good an action is compared to the other actions; therefore, good actions are reinforced through positive rewards, and bad actions are punished using negative rewards.

The fundamental difference between PPO and classical actor-critic methods is that PPO updates its actor neural network based on the Advantage value (TD error) estimated by the critic neural network as follows:

$$\widehat{A_t} = -V(s_t) + r_t + \gamma r_{t+1} + \cdots + \gamma^{T-t+1} + \gamma^{T-t} V(s_T) \#(5)$$ (5)

where t specifies the time index in$[0, T]$. The actor parameters are updated by the clipped surrogate objective as

$$L^{CLIP}(\theta) = \widehat{\mathbb{E}_t}[\min(r_t(\theta)\widehat{A_t}, clip(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\widehat{A_t})]$$ (6)

where $r_{\theta} = \frac{\pi_{\theta}(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)}$, and $r(\theta_{old}) = 1$. $L^{clip}$ The update rate is decided based on the ratio of the old parameter and to the new parameter in an interval$[1 - \epsilon, 1 + \epsilon]$. Such an objective assures the avoidance of updating too much noise and a considerably slower update speed.

A PPO algorithm that uses fixed-length trajectory segments is shown below. In each iteration, each of the N (parallel) actors collects T time steps of data. Later, we constructed the surrogate loss on these N timesteps of data and optimized it with minibatch stochastic gradient descent for K epochs.

### 3.3.2. Reward function approximation

To imitate human behavior, agents should receive a higher reward when they choose an appropriate action that a human would choose in the same situation. Recovering the hidden reward $r$ from a set of demonstrations $\mathbb{D}$ can help to understand the human action preferences and enable the agents to reproduce higher levels of human-like actions in the simulation. Various approaches using structured maximum margin prediction (Ratliff et al. 2006), feature matching (Ng and Stuart, 2000), and maximum entropy IRL (Ziebart et al., 2008) have been widely used to recover the cost function.

The maximum entropy IRL approach is used as a foundation, and the model employed to model the daily travel behavior decision making is further extended. The principle benefits of the maximum entropy paradigm include the ability to handle expert suboptimality as well as stochasticity by operating on the distribution over possible trajectories. In this formulation, the probability of user preference for any given path is generated assuming proportionality of the probability to the exponential of the reward along the trajectory, i.e.,

$$P(\zeta|R) \propto \exp\{\sum^{T} R(s_t, a_t)\}$$ (7)

The distribution takes this form because the given exponential distribution maximizes the entropy subject to a fixed mean value. In this study, we used the following linear parameterized representation was used: $R(s, a) = \phi \top \boldsymbol{f}(s, a)$

Further, $f$ can be applied to each state-action pair in the demonstrations. Every $\zeta$ in the demonstration $D$ is a sequence of $T$-step state-action pairs. The feature space $\phi$: $\mathbb{R}^N$ can then be applied to the trajectory as:

$$R(\zeta|\theta) = \sum_{\{(s,a)\in\zeta\}} \phi^{\top} f(s, a) = \phi^{\top} \boldsymbol{f}_{\zeta}$$ (8)

In this study, the reward structure is restricted by stipulating that the states with similar features should have similar rewards. To this end, the data provided by National Land Numerical Information were used as features for the approximation of the reward function. The National Land Numerical Information represents numerical data prepared from information related to national lands to

support the promotion and formulation of land planning such as the Comprehensive National Development Plan, National Land Use Planning, and National Spatial Strategy, as described in Section 3.1.2.

The task of solving the IRL problem to derive an optimal reward weight vector $\phi$ from the demonstrated trajectories can be learned by maximizing the joint posterior distribution by observing the expert demonstrations $\mathbb{D}$ as follows:

$$\phi^* = arg\max_{\phi} L(\phi) = \sum_D logP(\zeta|\phi) \tag{9}$$

This function is convex for deterministic MDP, and the optima can be obtained using gradient-based optimization methods (Dertsekas, 2012). The gradience is the difference expected feature counts between demonstrations and the agents' trajectories, which can be represented as:

$$\nabla L(\phi) = \boldsymbol{f} - \sum_{\zeta} P(\zeta|\phi)\boldsymbol{f}_{\zeta} = \boldsymbol{f} - \sum_{s_i,a_i} D_{s_i,a_i}\boldsymbol{f}_{\boldsymbol{s_i,a_i}} \tag{10}$$

where the $D_{s_i,a_i}$ is the state-action pair visitation frequency. Further the details are described by Ziebert et al. (2008).

It was found that fewer state-of-the-art IRL studies have considered time factors or time series because the standard benchmarks and previous tasks, such as urban navigation and activity forecasting, are not sensitive to time change (Ziebart et al. 2008; Song et al., 2013; Feygin 2018). However, travel behavior is highly correlated with time. For instance, business areas are more attractive for commuters during daytime than during the nighttime. Similarly, working schedules for most people result in peak transportation in the mornings and evenings during weekdays. Despite such typical correlations between time and travel behavior, some specific issues will cause an unrealistic travel behavior. However, the linear function makes it difficult to model the influence of time variation because all features are highly correlated to it. In order to tackle this problem, we introduced a temporal dimension into a reward function, and assigned a set of parameters to the reward function for each discretized time step. Finally, the expected action visitation frequencies were computed by enumerating all paths and probabilistically counting the number of paths and times in each path in which the particular state is visited. The outline of the IRL is described in Algorithm 1 shown in Fig. 2.

## 4. Experimental results

### 4.1. Study areas

The mobility of each individual is unique in the geographical space. Therefore, to examine model performance in different urban layouts, the model was instantiated in two different areas in Japan, as shown in Table 1. Tokyo comprises Japan's largest domestic and international hub for rail, ground, and air transportation. The transport network in the Tokyo area includes public and private rail and highway networks, airports for international, domestic, and general aviation, buses, motorcycle delivery services, pedestrians, bicycling; and commercial shipping. Commuter rail ridership is very dense, at 6 million people per line mile annually, with the

**Algorithm 1:** Maximum Entropy Inverse Reinforcement Learning

**Input :** The demonstration state action frequencies $\mu_{s_i,a_i}$ , feature parameter vector $f$, state space $S$, action space $A$, and discount factor $\gamma$

**Output:** Optimal reward function parameter $\phi^*$

Initialize $\phi^0$ with random number;

**Iterative model refinement;**

**for** $n = 1: N$ **do**

  **for** $t = 1 : T$ **do**

    $r^t = f(s,a)^T \times \phi^t$;

    **Solve MDP with current reward;**

    $\pi^t = \texttt{approximatevalueiteration}(r_t, S, A, \gamma)$ ;

    $D_{s,a} = ExpectedEdgeFrequencyCalculation[38](f, S, A, \pi)$;

    **Determine Maximum Entropy loss and gradients;**

    $\nabla\mathcal{L}_{\backslash}{}^t = (\mu^a_{demo} - D_{s,a}) \times f(s,a)$;

    **Update parameter with gradient;**

    $\theta^{t+1}$ = update weights with $\theta^t, \nabla\mathcal{L}_{\backslash}{}^t$

**Fig 2.** Outline of the IRL.

**Table 1**
Details of the study area.

| | Zones | Observed Population | Training Data | Destination Set | Transport Mode Set | Period |
|---|---|---|---|---|---|---|
| **Tokyo** | 23 special wards | 100,000 | 1000 | 1003 | Walk, Vehicle, Train | 6:00–23:30 |
| **Hiroshima** | 4 wards | 5200 | 1000 | 213 | Walk, Vehicle, Train | 6:00–23:30 |

highest utilization among the automotive urban areas. To verify that this approach applies to different types of urban situations, another case study was set up in the Hiroshima eastern area, whose total population and urban density are less than those of the Tokyo region. Transport in the east of Hiroshima is also different from that in Tokyo because only one train line connects this area with central Hiroshima.

From the Yahoo GPS dataset, the users who provided sufficient data points were first extracted. The rate of time slots (30 min per unit) for which a user was observed (as GPS logs have reported) out of the total number of slots in a day was set. This is because the data collection mechanism, where the user's location is detected when the phone stops staying at the current location and starts moving (iPhone) or automatically updates the location in every 30 min information when the service function is active (Android). In addition, there is a significant signal loss in the period between 0:00 to 5:00 because of the phone being switched off or people staying inside high buildings for a long time. Consequently, the simulation period was set between 6:00 and 23:30 for optimal model representation and performance.

The observed population was approximately 100,000 in Tokyo and 5200 in the Hiroshima eastern area from a typical observed weekday. In addition, 1000 trajectories were randomly selected as the training set, and the rest of the remaining trajectories were chosen as the test set. To evaluate the accuracy, the synthesis data followed the same population distribution (initial location distribution) and population size as the test set and both were later compared using some metrics, which are explained in the following section.

### 4.2. Reward function estimation results

In this section, we provide the estimation results of the reward function based on the IRL. The first problem we need to address is how many trajectories are required for the IRL algorithm to derive stable reward function parameters. We set 10 experiments with varying data for training amounts from 100 to 1000. As shown in Fig. 3, we visualized the reward function parameters of 5 context features (the number of offices, employees, the number of schools, the night population, and the number of evacuation facilities within a 1 km$^2$ area) over time with different amounts of training data amounts. Clearly, when the training datasets are less than 500, the parameter curves show erratic fluctuation, which is not in accordance with real people's behavior preferences. Taking the experimental result for n = 100 as an example, we can see the night population density feature curve fluctuates dramatically within a

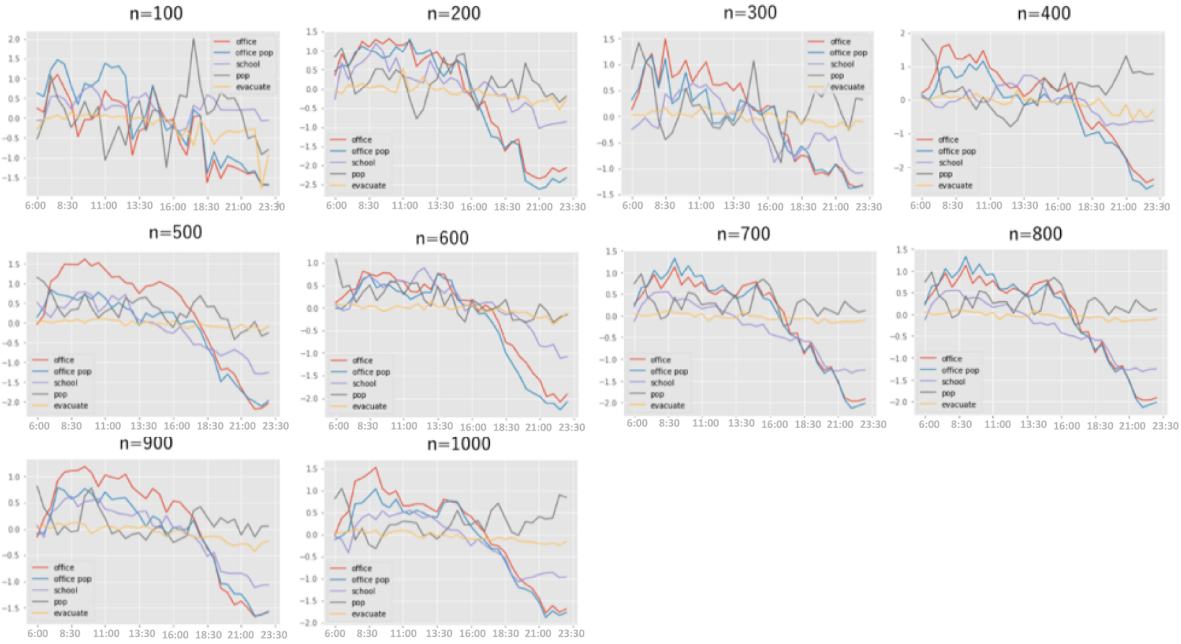


**Fig.** 3 Reward function parameter estimation results for different number of samples. The X-axis represents the time of the day (from 6:00 to 23:59, discretized for 30 min); Y-axis represents the weight of each feature as the number of offices, employees, schools, the nighttime population, and the number of evacuation facilities within a 1 km$^2$ area.

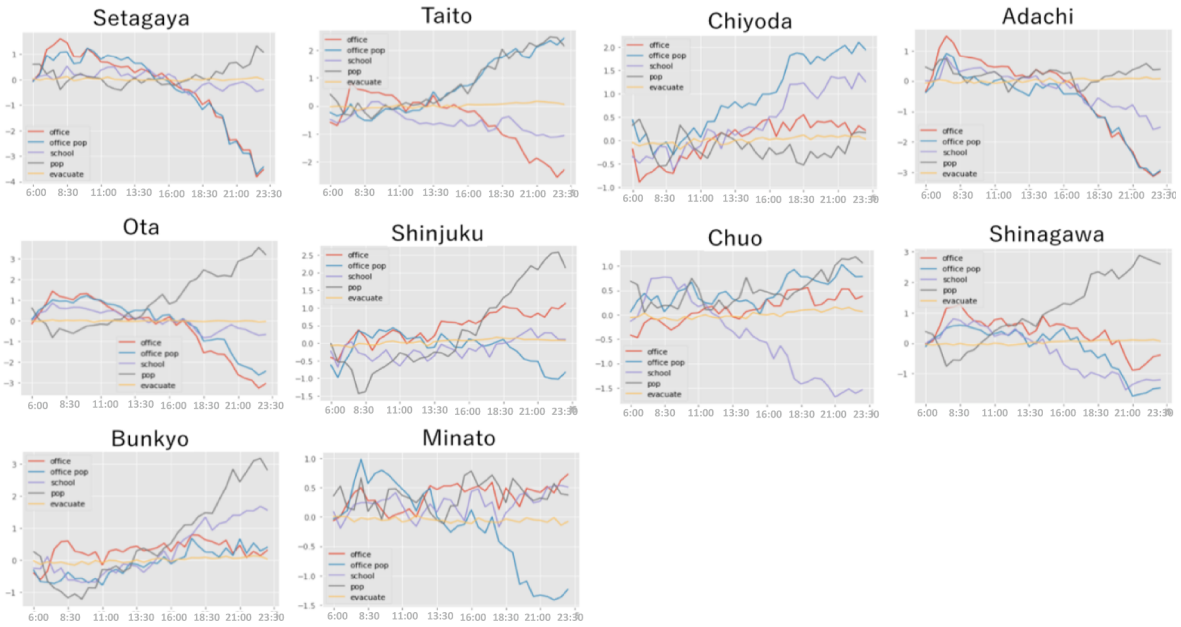


**Fig.** 4 Reward function parameter estimation results for samples derived from different areas. The X-axis represents the time of the day (from 6:00 to 23:59, discretized over 30 min time slots), Y-axis represents the weight of each feature as the number of offices, employees, schools, the nighttime population, and the number of evacuation facilities within a 1 $km^2$ area.

short period, which means the preference of choosing a place with a lot of several houses, changes at each time step. As the amount of training data amount increases, the parameter curves become smooth and reasonable. As shown in Fig. 3 for n = 1000, the weight of the office count feature increases in the morning peak period and starts decreasing in the afternoon. On the contrary, the influence of the night population shows an opposite trend. This result reflects an ordinary commuter's daily travel pattern (which is also easily observed from the dataset). Therefore, in the following part of this thesis, we choose 1000 samples to infer the reward function.

Furthermore, we tested the differences in the reward function parameters between different areas. Using the same dataset, we classified the trajectories based on their initial state (location at 0:00, was regarded as surveyor's home) into 23 zones of Tokyo. Moreover, for each zone, we randomly chose 1000 trajectories to train the reward model. As shown in Fig. 4, we can easily see that the reward function curve shows a different pattern between different zones. "Setagaya" is located in the west of Tokyo and has the highest residential population density, where the curve of office count (residential population density) feature's weight increases (decreases) in the morning and starts decreasing (increasing) after the morning peak. The results reflect a typical commuter's daily movement pattern, where people from this area can receive a higher reward by choosing commercial regions (which have more office facilities and less residential population) in the morning and show opposite patterns in the evening. Another trend can be seen in Fig. 4 "Shinjuku", which is a major commercial and administrative center, houses the northern half of the busiest railway station in Tokyo. The reward function curves show the opposite pattern compared to the result of "Setagaya". In other areas such as "Bunkyo" and "Chiyoda", the weight of the "school" feature becomes more influential than those of different regions.

### 4.3. Baselines and matrices

We further implemented four non-trivial baseline models for comparison. We choose the *first-order Markov chain model* as the first baseline, which defines the current state as the current location, and the next step choice probability is only dependent on the current location through space-complexity, $O(n^2)$ ,where n is the total number of the total locations. The second baseline is *the time-dependent Markov chain model*, which assumes that the transition probability is dependent on a discrete time step. We also considered the *discrete choice model*, which includes the choice of time of day, destination, and transport mode of trip. Each choice is influenced by the expected maximum utility with identical features, as described above. Lastly, we implemented the recurrent neural network Zhang et al. (2014) as an actor for temporal prediction, with identical features as described above. In the training process of the RNN model, features of each location were fed forwarded into a hidden layer, together with the previously accumulated hidden state. For the first-order Markov chain model, the time-dependent markov model, and the RNN model, we discretize the target area with 1 km grid and split time into 30 min as a step.

To evaluate the performance of the proposed model, negative log-loss (NLL) was used as the probabilistic comparison metric. The NLL:

$$NLL(s) = E_{\pi(a|s)}[-log \prod_{t} \pi(a_t|s_t)]$$

**Table 2**
Performance evaluation on individual trajectory generation.

| | NLL | | Distance Error | | Accuracy | |
|---|---|---|---|---|---|---|
| | Tokyo | Hiroshima | Tokyo | Hiroshima | Tokyo | Hiroshima |
| First-order MC model | 12.43 | 3.20 | 4.54 | 3.28 | 0.35 | 0.37 |
| Time-dependent MC model | 10.45 | 2.91 | 3.67 | 2.50 | 0.39 | 0.40 |
| Discrete Choice model | 13.51 | 4.11 | 5.09 | 3.96 | 0.23 | 0.31 |
| Recurrent neural network | – | – | 5.20 | 4.41 | 0.12 | 0.14 |
| **Proposed Method** | 8.72 | 3.56 | 3.08 | 2.08 | 0.43 | 0.52 |

The NLL is the expectation of the log-likelihood of a trajectory $s$ under a policy $\pi(a|s)$. In the example, this metric measures the probability of drawing the demonstrated trajectory from the learned distribution over all possible trajectories. The distance between two trajectories was also calculated as a physical measure of the distance error. Given two trajectories A and B with the same number of points, $Dist(traj_A, traj_B) = avg\left(dist\left(p_i^t, p_j^t\right)\right)$, where $dist(p_i, p_j)$ is the Euclidean distance between points a and point b, and $n$ represents the uniform discrete time slot. Finally, the Jaccard similarity coefficient was used as a measure of *accuracy* (Real and Vargas, 1996).

### 4.4. Comparison results

Table 2 shows how the proposed method outperformed the other baseline models. Although both the baseline methods and the proposed method need observed trajectories for training, the results showed that the proposed method was capable of handling the situation in which input does not appear in the training data. In contrast, the RNN actor achieves quite high performance on the training dataset. However, when the simulated environment is larger than the training dataset, the actor cannot give the correct answer when the input is not from any of the classes in the training data.

Scores in Hiroshima were better than those in the Tokyo area. This does not mean that training in Hiroshima was more successful but that the area in Hiroshima is much smaller than that in Tokyo, agents face fewer choices when choosing actions, and synthetic trajectories are much closer to real trajectories.

#### 4.4.1. Hourly population distribution in the simulation results

The main objective of this study was to enable the RL agent to generate synthetic trajectories without compromising the observation data (i.e., raw GPS data or products derived from them). However, it is a tough task to infer whether a synthetic trajectory is "accurate" or "realistic" because it is not known which specific trajectory in the test data should be compared. Based on the synthetic trajectories, one can easily calculate the population distribution over time to examine whether the agents are located in correct locations compared to the real data. Population distribution is also an essential source for travel demand estimation and human mobility management. In Figs. 5 and 6, the population distribution that was tested for two study areas from 6:00 a.m. to 11:00 p.m. between the test dataset ($x$-axis) and synthetic dataset ($y$-axis) is depicted. Moreover, for both of the two areas, the simulation results replicate the population quite closely. At each time step, agents in the Tokyo area need to pick out a proper destination from more than 1000 possible choices, and sometimes there may exist several locations that share very similar features corresponding to the same reward, which may confuse the agents. However, the overall correlation coefficients were still higher than 0.8 over all times in Tokyo. The estimation results for the Hiroshima area are provided in Fig. 6. Compared to the Tokyo area, the layout of the Hiroshima area is more concrete, and agents only face approximately 200 location choices so that the agents are more natural to choose a proper destination. Overall, the proposed method shows promising results that successfully estimate the population distribution in citywide level areas.

The population distribution with the root mean square error (RMSE) and root mean square percentage error (RMSPE) are shown in Fig. 7. In the Tokyo area, for the period in the morning commuting hours, the result shows relatively high RMSE and low RMSPE. Over time, the RMSE decreases to a lower level, whereas RMSPE increases dramatically due to a widely dispersed population at nighttime.

#### 4.4.2. Transport usage in the simulation result

Another concerning output of the simulation result is the transport system usage situation. An accurate prediction of public transit and road network usage plays a vital role in the planning and management of the transportation system. The number of passengers on each mesh grid with different transport modes was calculated to examine whether agents choose correct actions over a time period. Figs. 8 and 9 show the actual vehicle users and train users in the Tokyo area compared with the test dataset. The results show a strong positive correlation with the test dataset during commuting hours. Owing to the significantly higher total number of users, one can clearly see that the railway plays a primary role in urban transport in the Tokyo area. However, the correlation decreased drastically at the end of the day (from 9:00 p.m.) because of the lower average passenger amount compared to the ground truth. Notably, the movements were overestimated in this period because some agents struggled with finding the way home. This result also corresponds to the dispersed pattern of population distribution in the same time period. From Fig. 10 it can be observed that the estimated number

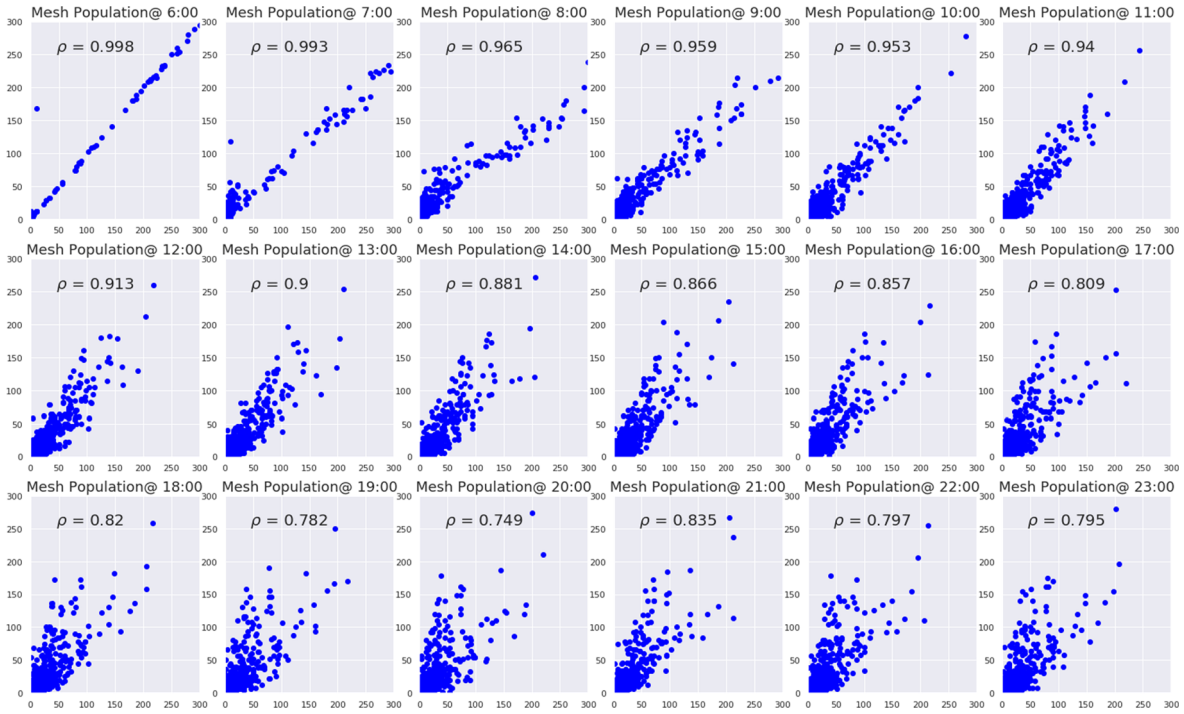


**Fig.** 5 Scatterplots of population distribution between population in synthetic dataset (y-axis) and population in test dataset (x-axis) from 6:00 a.m. to 11:00 p.m. in Tokyo. Each plot represents a 1 $km^2$ area in the city.

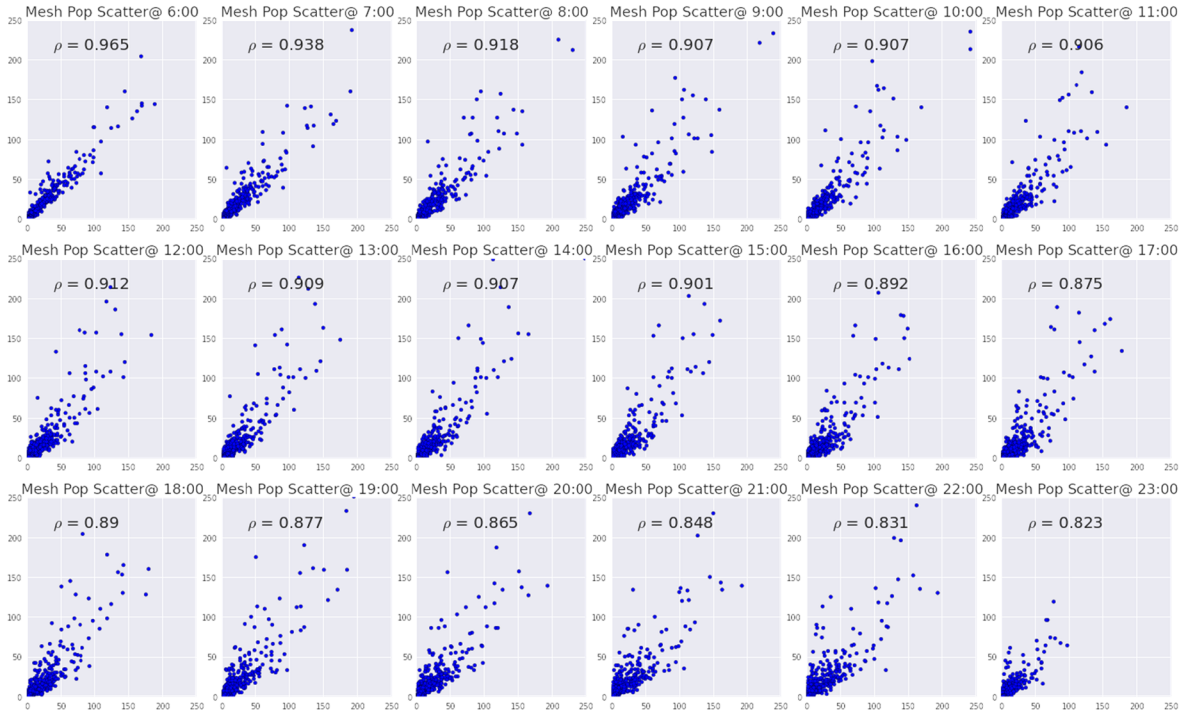


**Fig.** 6 Scatterplots of population distribution between population in synthetic dataset (y-axis) and population in test dataset (x-axis) from 6:00 a.m. to 11:00 p.m. in Hiroshima. Each plot represents a 1 $km^2$ area in the city.

of travelers is overestimated by both transportation modes, implying that agents may obtain more reward by visiting more places than staying at their previous location. One reason for this problem is that the MDP-based RL agents cannot take their past behaviors into consideration for judging the future action plans; consequently, the agents may underestimate the overall cost on a whole day

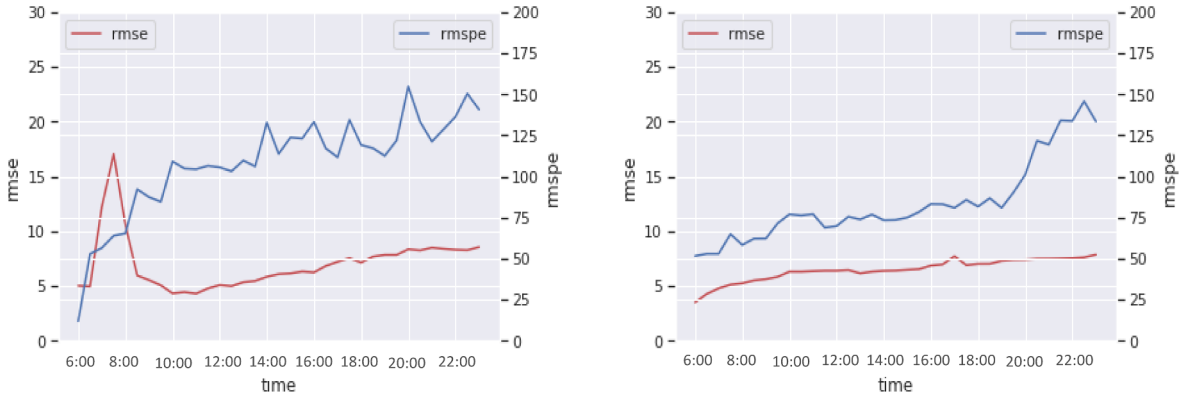


**Fig. 7** Comparison of accuracy of mesh population between the synthetic and test datasets using the RMSE and RMSPE values, and the correlation coefficient in Tokyo (left) and Hiroshima (right).

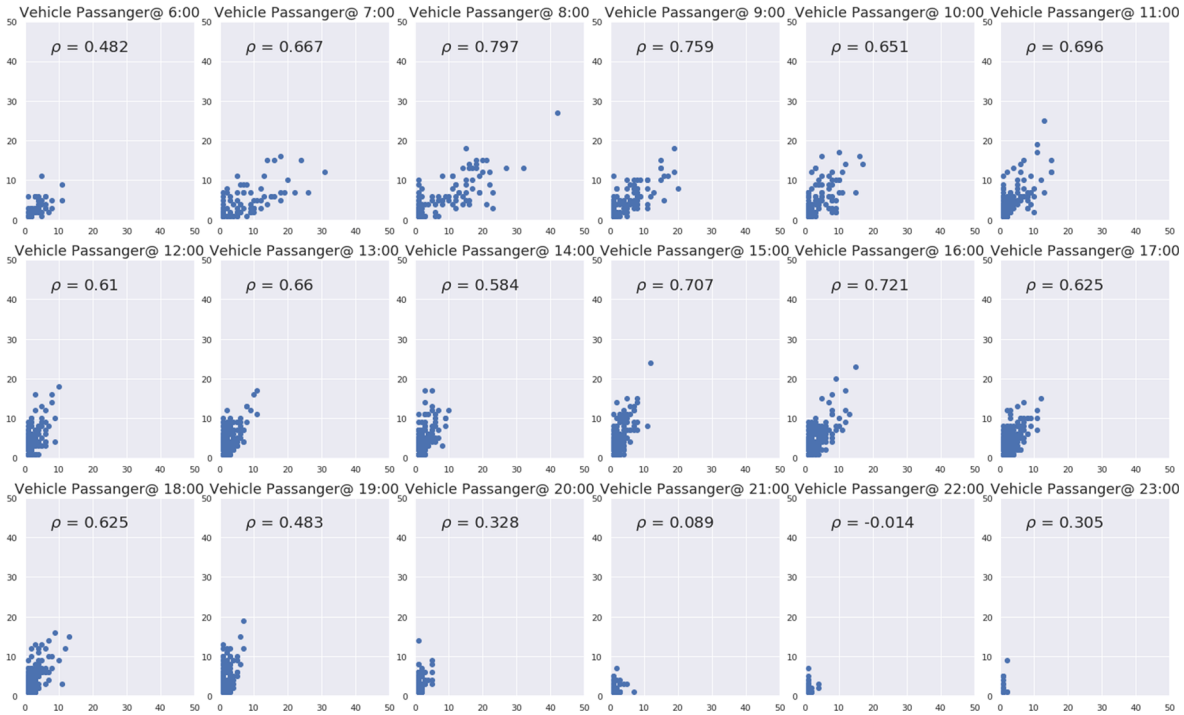


**Fig 8.** Scatterplots of vehicle users between synthetic dataset (*y*-axis) and test dataset (*x*-axis) from 6:00 a.m. to 11:00 p.m. in Tokyo.

when deciding next step behavior. Further improvement by revising the reward function is needed to determine whether there is a better way to formulate human mobility patterns into the reward function structure.

In contrast, the transport usage in the Hiroshima areas shows an entirely different pattern. There is only one railway line in the study area and railway users are barely observed in the training/test dataset. The results obtained for vehicle users and those for total passengers (those travelling by train, vehicle, and by foot) results are shown in Figs. 11 and 12, respectively.

## 5. Discussion

As aforementioned, our approach to modeling and simulating people's daily travel is an effective way to replicate people mass movement. Specifically, the experimental results indicate that, our approach is able to 1) capture the human travel behavior preferences associated with the spatio-temporal patterns and the context features from anonymous GPS data, 2) intelligently plan sequential locations and transport mode choice over time at the scale of a metropolitan area, 3) reproduce people mass densities and transport usage. Our modeling approach produces synthetic trajectories. Using the generated synthetic dataset, we can directly visualize, analyze and predict human mobility.

Benefitting from the emerging algorithms of RL, the agents could make decisions from a complex environment with a multimodal

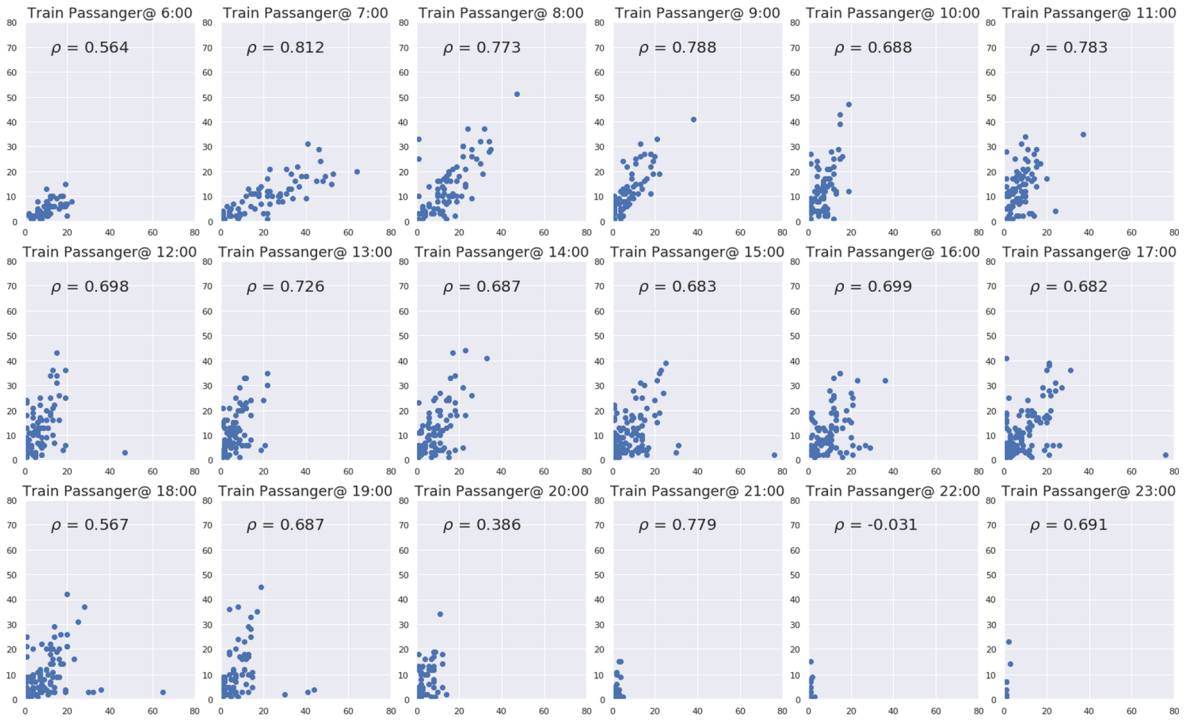


**Fig 9.** Scatterplots of railway users between synthetic dataset (*y*-axis) and test dataset (*x*-axis) from 6:00 a.m. to 11:00 p.m. in Tokyo.

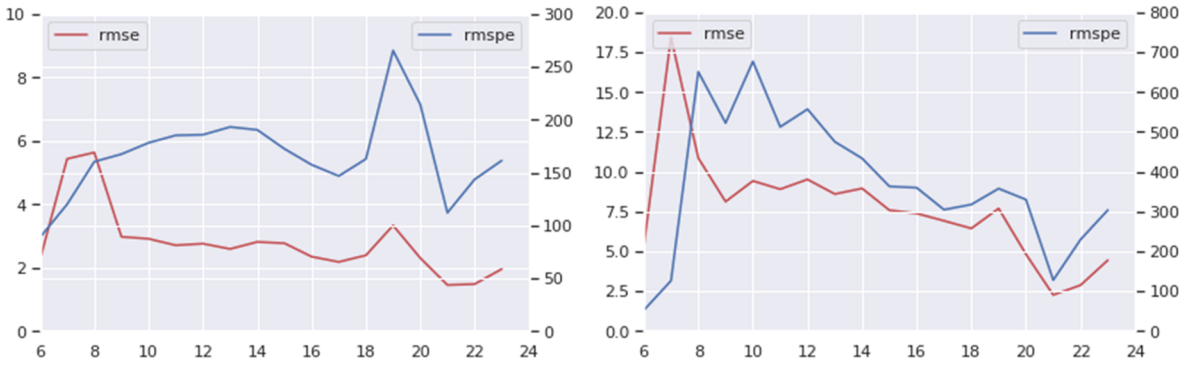


**Fig.** 10 Comparison of accuracy of travelers between the synthetic and test datasets using the RMSE and RMSPE values, and the correlation coefficients for travelers using vehicle (left) and train (right).

transportation system. The pipeline of learning behavioral rules was developed as a problem of recovering the behavior preferences from the observed trajectories based on RL and generating a human-like behavior policy for the agent-based modeling and the simulation. The framework outputs synthetic trajectories in the same format as that of the training data. It is also easy to derive a series of products such as dynamic population distribution, travel demand estimation and anomaly detection. A synthetic dataset was reproduced using the agent-based modeling framework and a dataset capable of replicating the urban dynamics was evaluated. Moreover, the input data for the method were not limited to the GPS data of smartphones. Any data that capture the timestamps of the beginning and end of commuting activities could be applied.

However, our proposed approach has several limitations. First, in this study, the focus was on developing a multi-modal transportation ABM system that requires the training data to provide the transport mode for each travel behavior. However, the sparsity of the source data limited the accuracy of transport mode detection. In particular, the signal loss during taking subway (underground) and low report rate made it difficult to distinguish between railway trips and vehicle trips. This affects the final travel demand estimation accuracy. Second, RL was developed from the MDP, which satisfies the Markov property. However, generally, human being's decisions regarding the next course of step action depend on not only the current state but also on the previous states or actions. For example, if an agent cannot remember its home location, it cannot stop at the correct location, such as at home at the end of the day. Although we can introduce some hand-crafted rules to force agents to return their initial location, or give extra rewards for agents who arrive at home, such kinds of settings decrease the generality of applying the model to different scenarios such as post-

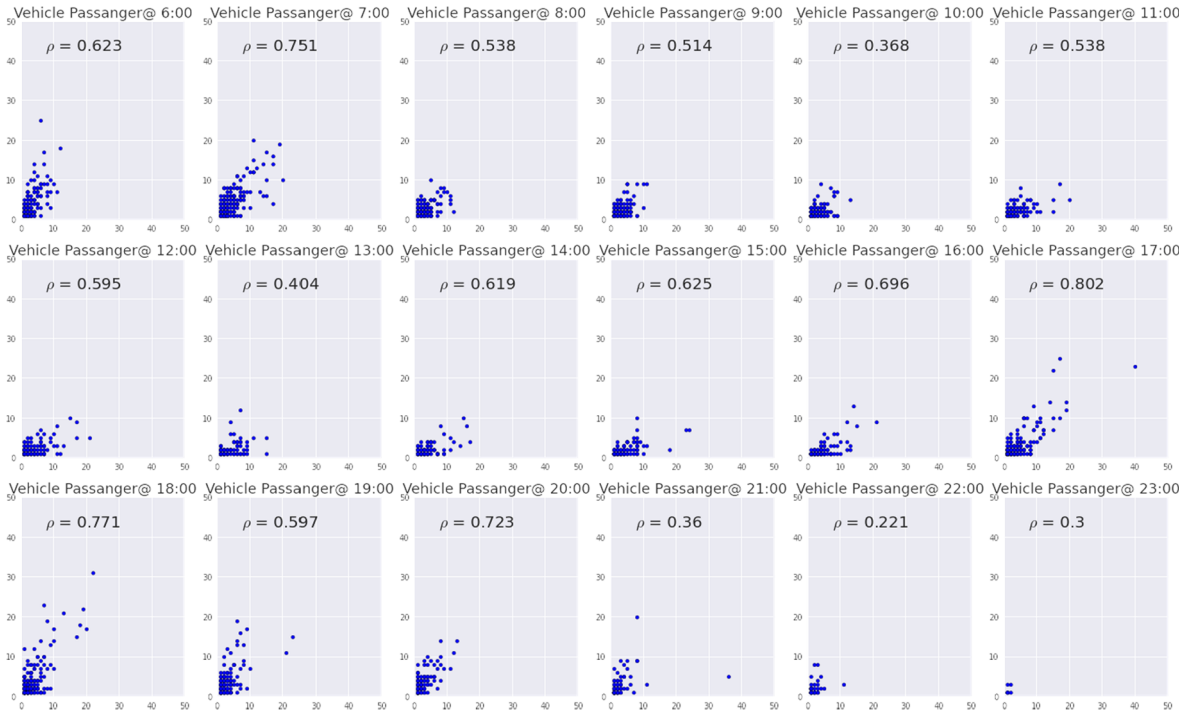


**Fig. 11.** Scatterplots of vehicle users between synthetic dataset (*y*-axis) and test dataset (*x*-axis) from 6:00 a.m. to 11:00 p.m. in Hiroshima.

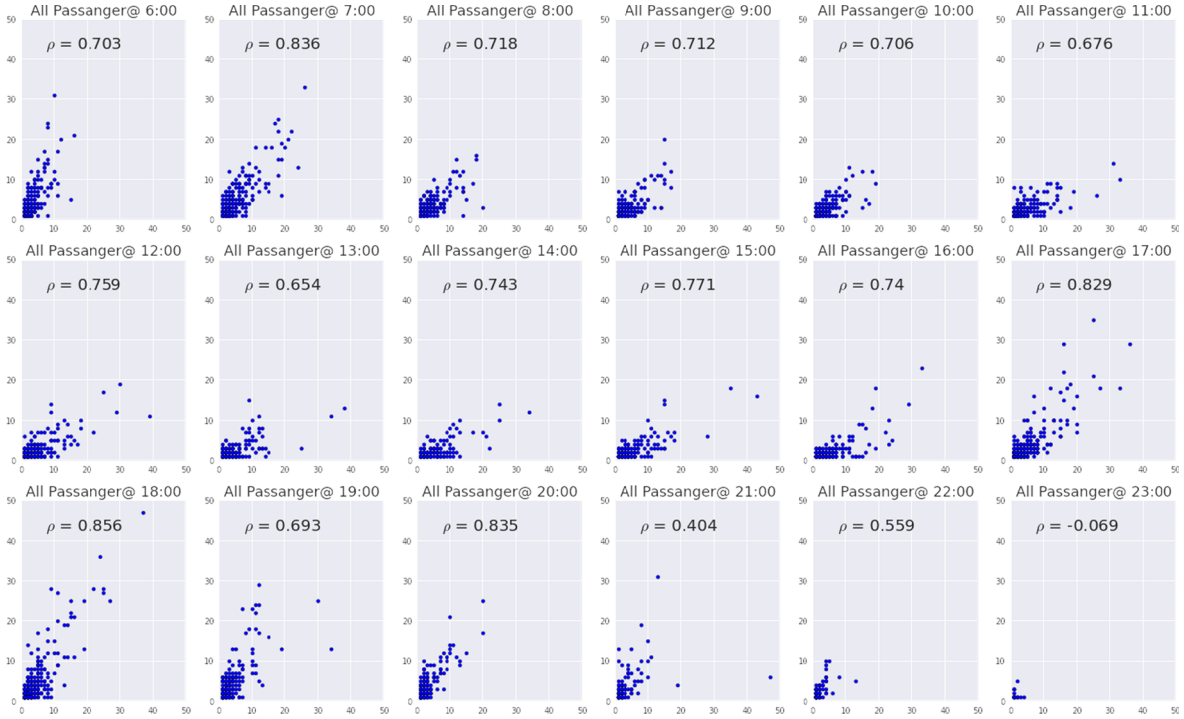


**Fig 12.** Scatterplots of all passengers travelling by all mediums (train, vehicle, and on foot) between synthetic dataset (*y*-axis) and test dataset (*x*-axis) from 6:00 a.m. to 11:00 p.m. in Hiroshima.

disaster simulation where people stay at shelters or other places.

There are also several future research opportunities that this study enables. First, the accuracy of the simulation should be improved. In this study, we tested the proposed method using only demonstration trajectories and environment context without any

other information. Each agent had no extra predefined behavioral rules or characteristics that assisted the travel behavior planning. We believe that by incorporating with population synthesis using census and survey data, and choosing a reward function from corresponding demonstration samples, the accuracy of the simulated results can be improved. For example, an agent labeled as 'commuter' should be assigned by the reward function that is trained using commuters' trajectories.

The second direction is to apply the method to more complex scenarios such as disasters, New Year ceremonies and other rare events. Although in this study we successfully replicated the people movement of people on a normal day, we are motivated to investigate whether this method works for unprecedented scenarios using a pre-trained model. For example, if we want to know how people's movement will be affected by a severe earthquake in some target areas (however, no such heavy disasters have occurred in such areas yet), can we transfer the knowledge or pre-trained model from another area that suffered the same disaster? Recently, several studies have tested the transferability of RL algorithms between different game environments and tasks. We believe that such a trial will provide more practical and promising results for the transportation and disaster prevention management.

Third, in this study, we utilize a linear function to represent the 'reward' of taking travel behaviors and denote the travel behavior preferences at each time of the day using the reward function parameters. We observed that increasing and clustering the demonstration trajectories for training can significantly affects the reward functions. However, there is a lack of evidence to evaluate whether the results revealed real people's preferences. To overcome this issue, introducing probe person survey with both questionnaire survey and mobile location data is necessary.

## 6. Conclusion

In this study, we proposed an RL-based human mobility model for replicating the people mass movement at the scale of a metropolitan area. We introduced IRL to recover human travel behavior preferences that can capture the spatio-temporal pattern and the context features of human mobility using real GPS trajectories, and produce synthetic trajectories that report locations and transport mode choice over time. The generated synthetic trajectory dataset can be utilized to capture and predict within-day people mass movement in various cities where the demonstrated trajectories are available.

We evaluated the proposed approach in two cities with different scales. The experimental results showed that the proposed method significantly outperformed the current simulation methods in generating a more realistic microscopic travel behavior.

## CRediT authorship contribution statement

**Yanbo Pang:** Conceptualization, Methodology, Software, Data curation, Writing - original draft. **Takehiro Kashiyama:** Resources, Visualization. **Takahiro Yabe:** Conceptualization. **Kota Tsubouchi:** Conceptualization. **Yoshihide Sekimoto:** Conceptualization, Formal analysis, Supervision.

## Acknowledgments

This study was supported by the Japan Science and Technology Agency (JST), CREST program, "Creating an innovative earthquake, tsunami damage mitigation big data analysis base by data assimilation and collaboration of large scale, high resolution value simulations" under grant JPMJCR1411. We specially thank Yahoo Japan Corporation for their supporting.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.trc.2020.102706.

## References

Adler, T., Ben-Akiva, M., 1979. A theoretical and empirical model of trip chaining behavior. Transport. Res. Part B: Methodol. 13 (3), 243–257.

Altaf, B., Lu, Y., Xiangliang, Z., 2018. Spatio-temporal attention based recurrent neural network for next location prediction. In: 2018 IEEE International Conference on Big Data (Big Data). IEEE, Seattle, WA, Springer.

Arora, S., Doshi, P., 2011. A survey of inverse reinforcement learning: Challenges, methods and progress. arXiv preprint arXiv:1806.06877.

Bertsekas, D.P., 2012. Approximate Dynamic Programming In: Editor's Initials and Surnames, eds. Dynamic programming and optimal control Vol. II. 4th Edition Athena Scientific, Massachusetts, USA.

Bhat, C.R., Frank, S.K., 1999. Activity-based modeling of travel demand. In: Handbook of transportation Science, International Series in Operations Research & Management Science, vol 23. Springer, Boston, MA, pp. 35–61.

Bowman, J.L., Moshe, E.B., 2001. Activity-based disaggregate travel demand model system with activity schedules. Transport. Res. Part A: Policy Pract. 35 (1), 1–28.

Cho, E., Myers, S.A., Leskovec, J., 2011. Friendship and mobility: user movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, San Deigo, CA, USA, http://doi.org/10.1145/2020408.2020579.

Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D., 2017. Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, June 2019, Daejeon, South Korea, Springer, pp. 4299–4307.

Donnelly, R., 2010. Advanced practices in travel forecasting. Transportation Research Board. Nth ed. Name of Publication.

Ermon, S., Xue, Y., et al., 2015. Learning large-scale dynamic discrete choice models of spatio-temporal preferences with application to migratory pastoralism in East Africa. In: AAAI'15: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Austin, Texas, USA, 25-30 Janurary. pp. 644–650.

Fan, Z., Xuan, S., Ryosuke, S., 2014. CitySpectrum: a non-negative tensor factorization approach. In: UbiComp '14: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM. Seattle, USA, 13-17 September. pp. 213–223.

Feng, J., Li, Y., Zhang, C., Sun, F., Meng, F., Guo, A., Jin, D., 2018. Deepmove: Predicting human mobility with attentional recurrent networks. In: WWW '18: Proceedings of the 2018 World Wide Web Conference. Lyon, France, 23-27 April. pp. 1459–1468.

Feygin, S. 2018. Inferring Structural Models of Travel Behavior: An Inverse Reinforcement Learning Approach (Doctoral dissertation, UC Berkeley).
Finn, C., Sergey, L., Pieter, A., 2016. Guided cost learning: Deep inverse optimal control via policy optimization. International Conference on Machine Learning. In: ICML'16: Proceedings of the 33rd International Conference on International Conference on Machine Learning. New York City, USA 19-24 June, pp. 49–58.
Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L., 2008. Understanding individual human mobility patterns. Nature 453 (7196), 779.
Goullias, K., Pendayala, R., Kitamura, R., 1990. Practical method for the estimation of trip generation and trip chaining.
Haarnoja, T., Zhou, A., Abbeel, P., Levine, S., 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. arXiv preprint arXiv:1801.01290.
Heess, N. et al., 2017. Emergence of locomotion behaviours in rich environments. Viewed 02, May 2020, arXiv preprint arXiv:1707.02286.
Hester, T., Stone, P., 2009. Generalized model learning for reinforcement learning in factored domains. In: AAMAS '09: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems -Volume 2. Budapest, Hungary, May 10-15. pp. 717–724.
Huang, Zhiren, Ling, Ximan, Wang, Pu., Zhang, Fan, Mao, Yingping, Lin, Tao, Wang, Fei-Yue, 2018. Modeling real-time human mobility based on mobile phone and transportation data fusion. Transport. Res. Part C: Emerg. Technol. 96, 251–269.
Isaacman, S., et al., 2012. Human mobility modeling at metropolitan scales. In: MobiSys '12: Proceedings of the 10th international conference on Mobile systems, applications, and services. ACM. Low Wood Bay, Lake District, United Kingdom, June 26-28. pp. 239–252.
Janssens, D., et al., 2007. Allocating time and location information to activity–travel patterns through reinforcement learning. Knowl.-Based Syst. 20 (5), 466–477.
Kitamura, R., 1984. Incorporating trip chaining into analysis of destination choice. Transport. Res. Part B: Methodol. 18 (1), 67–81.
Kitamura, R., Fujii, S., 1998. Two computational process models of activity-travel behavior. Theoret. Found. Travel Choice Model. 251–279.
Lin, Z., et al., 2017. Deep generative models of urban mobility. IEEE Trans. Intell. Transp. Syst.
Yuxi, L., 2017. Deep reinforcement learning: An overview. arXiv preprint arXiv: 1701.07274. Viewed: 02, May 2020.
Lu, X., Bengtsson, L., Holme, P., 2012. Predictability of population displacement after the 2010 Haiti earthquake. Proc. Natl. Acad. Sci. 109 (29), 11576–11581.
Mnih, V., et al., 2015. Human-level control through deep reinforcement learning. Nature 518 (7540), 529.
Moshe, B., Bierlaire, M., 1999. Discrete choice methods and their applications to short term travel decisions. Handbook of Transportation Science. Springer, Boston, MA, pp. 5–33.
National Land Numerical Information. http://nlftp.mlit.go.jp/ksj-e/index.html. (Accessed: 15 March 2020).
Nazari, M., et al., 2018. Reinforcement learning for solving the vehicle routing problem. Adv. Neural Inform. Process. Syst. 2018, 9839–9849.
Ng, A.Y., Stuart, J.R., 2000. Algorithms for inverse reinforcement learning. Icml 1.
Ouyang, K., Shokri, R., Rosenblum, D.S., Yang, W., 2018. A Non-Parametric Generative Model for Human Trajectories. IJCAI. 2018, 3812–3817.
Pang, Y., et al., 2018. Replicating urban dynamics by generating human-like agents from smartphone GPS data. Information Systems. ACM.
Polson, N.G., Sokolov, V.O., 2017. Deep learning for short-term traffic flow prediction. Transport. Res. Part C: Emerg. Technol. 79, 1–17.
Prasad, P.S., Agrawal, P., 2010. Movement prediction in wireless networks using mobility traces. In: CCNC'10: Proceedings of the 7th IEEE conference on Consumer communications and networking conference. IEEE, Las Vegas, Nevada USA, January 9-12, pp. 714–718.
Ramos, G. de O., Bazzan, A.L.C., Da Silva, B.C., 2018. Analysing the impact of travel information for minimising the regret of route choice. Transport. Res. Part C: Emerg. Technol., 88, 257–271.
Ratliff, N.D., Bagnell, J.A., Zinkevich, M.A., 2006. Maximum margin planning. In: ICML '06: Proceedings of the 23rd international conference on Machine learning. ACM. Pittsburgh, Pennsylvania, USA, June 25-29. pp. 729–736.
Real, R., Vargas, J.M., 1996. The probabilistic basis of Jaccard's index of similarity. Syst. Biol. 45 (3), 380–385.
Ruiter, E.R., Moshe, B., 1978. Disaggregate travel demand models for the San Francisco Bay Area. system structure, component models, and application procedures. Transp. Res. Rec. 673.
Schulman, J. et al., 2017. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.
Shan, J., Ferreira, J., Gonzalez, M.C., 2017. Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. IEEE Trans. Big Data 3 (2), 208–219.
Smart, W.D., Kaelbling, L.P., 2002, Effective reinforcement learning for mobile robots. In: Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292). IEEE. pp. 3404–3410.
Song, C., Qu, Z., Blumm, N., Barabási, A.L., 2010. Limits of predictability in human mobility. Science 327, 5968, 1018–1021.
Song, L. et al., 2004. Evaluating location predictors with extensive Wi-Fi mobility data. IEEE Infocom. Vol. 2. Institute of Electrical Engineers Inc (IEEE).
Song, X., Zhang, Q., Sekimoto, Y., Horanont, T., Ueyama, S., Shibasaki, R., (2013, August). Modeling and probabilistic reasoning of population evacuation during large-scale disaster. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 1231–1239.
Song, X., Hiroshi, K., Ryosuke, S., 2016. DeepTransport: Prediction and simulation of human mobility and transportation mode at a citywide level. IJCAI 16.
Sutton, R.S., Barto, A.G., 2018. Reinforcement learning: An introduction. MIT press.
Toole, J.L., Colak, S., Sturt, B., Alexander, L.P., Evsukoff, A., González, M.C., 2015. The path most traveled: Travel demand estimation using big data resources. Transport. Res. Part C: Emerg. Technol. 58, 162–177.
Witayangkurn, A. et al., 2013. Trip reconstruction and transportation mode extraction on low data rate GPS data from mobile phone. In: Proceedings of the international conference on computers in urban planning and urban management (CUPUM 2013).
Wu, X., et al., 2018. 2018 Hierarchical travel demand estimation using multiple data sources: A forward and backward propagation algorithmic framework on a layered computational graph. Transport. Res. Part C: Emerg. Technol. 96, 321–346.
Xiong, Y., 2014. Modelling individual and household activity: travel scheduling behaviours in stochastic transportation networks. Doctoral dissertation. The Hong Kong Polytechnic University.
Yabe, T., Kota, T., Yoshihide, S., 2017. CityFlowFragility: measuring the fragility of people flow in cities to disasters using GPS data collected from smartphones. Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol. 1 (3), 117.
Yang, M., et al., 2014. Multiagent-based simulation of temporal-spatial characteristics of activity-travel patterns using interactive reinforcement learning. Mathem. Probl. Eng. 2014.
Ye, Y., Zhang, X., Sun, J., 2019. Automated vehicle's behavior decision making using deep reinforcement learning and high-fidelity simulation environment. Transport. Res. Part C: Emerg. Technol. 107, 155–170.
Yin, M., Sheehan, M., Feygin, S., Paiement, J.F., Pozdnoukhov, A., 2017. A generative model of urban activities from cellular data. IEEE Trans. Intell. Transp. Syst. 19 (6), 1682–1696.
Zheng, Y., et al., 2014. Urban computing: concepts, methodologies, and applications. ACM Trans. Intell. Syst. Technol. 5 (3), 38.
Ziebart, B.D., et al., 2008. Maximum entropy inverse reinforcement learning. Aaai 8.
Zhang, Y., Dai, H., Xu, C., Feng, J., Wang, T., Bian, J., Liu, T.Y., 2014, June. Sequential click prediction for sponsored search with recurrent neural networks. In: Twenty-Eighth AAAI Conference on Artificial Intelligence.
Zhao, Z., Koutsopoulos, H.N., Zhao, J., 2018. Individual mobility prediction using transit smart card data. Transport. Res. Part C: Emerg. Technol. 89, 19–34.
Zheng, Y., Li, Q., Chen, Y., Xie, X., Ma, W.Y., 2008. Understanding mobility based on GPS data. In: Proceedings of the 10th international conference on Ubiquitous computing. 2008. pp. 312–321.
Zhu, F., Ukkusuri, S.V., 2014. Accounting for dynamic speed limit control in a stochastic traffic environment: A reinforcement learning approach. Transport. Res. Part C: Emerg. Technol. 41, 30–47.
Zhu, M., Wang, X., Wang, Y., 2018. Human-like autonomous car-following model with deep reinforcement learning. Transport. Res. Part C: Emerg. Technol. 2018 (97), 348–368.