Multiagent-Based Simulation of Temporal-Spatial Characteristics of Activity-Travel Patterns Using Interactive Reinforcement Learning

Citation: Yang, Min, Yingxiang Yang, Wei Wang, Haoyang Ding, and Jian Chen. "Multiagent-Based Simulation of Temporal-Spatial Characteristics of Activity-Travel Patterns Using Interactive Reinforcement Learning." Mathematical Problems in Engineering 2014 (2014): 1–11.

As Published: http://dx.doi.org/10.1155/2014/951367

Publisher: Hindawi Publishing Corporation

Persistent URL: http://hdl.handle.net/1721.1/96101

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Research Article

Multiagent-Based Simulation of Temporal-Spatial Characteristics of Activity-Travel Patterns Using Interactive Reinforcement Learning

Min Yang,[1] Yingxiang Yang,[2] Wei Wang,[1] Haoyang Ding,[1] and Jian Chen[1]

[1] *School of Transportation, Southeast University, Si Pai Lou No. 2, Nanjing 210096, China*
[2] *Department of Civil and Environment Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139-4307, USA*

Correspondence should be addressed to Min Yang; transtar2002@163.com

Received 6 September 2013; Revised 6 November 2013; Accepted 9 December 2013; Published 30 January 2014

Academic Editor: Geert Wets

We propose a multiagent-based reinforcement learning algorithm, in which the interactions between travelers and the environment are considered to simulate temporal-spatial characteristics of activity-travel patterns in a city. Road congestion degree is added to the reinforcement learning algorithm as a medium that passes the influence of one traveler's decision to others. Meanwhile, the agents used in the algorithm are initialized from typical activity patterns extracted from the travel survey diary data of Shangyu city in China. In the simulation, both macroscopic activity-travel characteristics such as traffic flow spatial-temporal distribution and microscopic characteristics such as activity-travel schedules of each agent are obtained. Comparing the simulation results with the survey data, we find that deviation of the peak-hour traffic flow is less than 5%, while the correlation of the simulated versus survey location choice distribution is over 0.9.

1. Introduction

Over the few last decades, activity-based approaches has become the main theme in transportation demand modeling, taking the place of trip-based approaches. Trip-based approach has several drawbacks: trip generation is fixed and independent of the transportation system; travel demand is generated from the need of activity participation; and the space and temporal relationship of all trips and activity patterns is ignored. Such drawbacks brought activity-based approach into transportation demand modeling.

The first activity-based approaches began in the 1970s [1–3]. Those pioneering studies explored choices and constraints in travel demand. Since that time, activity-based modeling has flourished. Various methodologies have been introduced and they can be classified into three categories.

The first category is utility-maximizing model (or econometric model) which suggests that individuals seek to maximize their cumulative utilities when performing activities. Those models link individual or household's sociodemographics, transportation policies, and other environmental factors to their activity and travel patterns. Econometric models ranging from discrete choice models (such as multinomial logit and nested logit mode) to hazard duration models remain to be a powerful approach in activity-travel analysis [4–7].

The second category is computational process model (CPM) which focuses on using context-dependent choice heuristics to model individual's decision process. A computational process model is a set of condition-action rules that specify how a decision is made. One precursor in CPM is the time-space prism method. Hägerstrand [3] introduced the three-dimensional space-time models. In such models limited resources of time and space became constraints on each individual's behavior alternatives [8]. The techniques used in more recent studies include decision trees, neural networks, and Bayesian networks [9–11].

The combination of the above two approaches leads to hybrid models. Hybrid models concentrate on the integration of econometric models and CPM. Decision-tree is combined with parametric modeling [12]; random utility maximization is incorporated into activity scheduling model [13]. New algorithms such as reinforcement learning are also introduced into the field.

Reinforcement learning integrates the concepts of reward (utility) maximization and context-dependent choice heuristics. The applications of reinforcement learning include robotics, game theory, dispatching system, and financial trading [14–17]. Tan used reinforcement learning to formalize an automated process for determining stock cycles by tuning the momentum and the average periods. The total experimental results from the five stocks are able to beat the market by about 50 percentage points [18]. Lahkar and Seymour studied reinforcement learning in a population game. Agents in a population game revise mixed strategies using the cross rule of reinforcement learning [17]. In addition, formulation of economic dispatch as a multistage decision making problem is carried out using reinforcement learning by Jasmin et al. [15]. Applying reinforcement learning in transportation demand modeling has several advantages. First, the imitation of human learning through trial and error interactions with a dynamic environment helps to explain behavioral mechanisms [19]. The RL mechanism is distinguished from other computational cognitive mechanisms by its emphasis on learning by an individual from direct interaction with individual's decision environment in the presence of an explicit goal and feedback and without relying on any exemplary supervision. Secondly, it does not need an expert-system to inform it what selection is right and what is wrong. Thirdly, it could react to unforeseen events and take both long-term learning and short-term dynamics into account. Among the first attempts, Charypar and Nagel built the basic model of activity time plans using q-learning and got quite realistic results [20]. This model was then modified to allocate both time and location choice of activity-travel pattern [21]. Because q-learning generally takes a long time to converge and the curse of dimensionality occurs when the problem gets complex, q-learning was combined with the regression tree method to form a new algorithm called q-tree [22].

The above-mentioned researches show several aspects that need further development.

(i) In most of the reinforcement-learning-based studies, though the format of reward function has been scrutinized, the rewards are based on assumption values and are hard to be acquired from survey data, so that the result is hard to be put into practical use.

(ii) In many of the multiagent systems, "multi" means several components of the system such as road, intersection, and traveler rather than multiple travelers. Interactions of travelers are neglected.

(iii) The result analysis is often limited within individual activity-travel schedule. Macroscopic characteristics such as traffic flow distribution are often ignored.

In this study we propose an interactive reinforcement learning algorithm in which individuals not only receive information from the environment, but also give feedback to the environment. We did this by adding road congestion degree, which is determined by travelers' decisions, to the algorithm. The dynamic environment is a medium that passes the influence of one traveler's decision to others. The self-organization effect shown through this mechanism makes the system reach a dynamic equilibrium. This algorithm not only ensures rationality of each single traveler's behavior, but also obtains aggregated temporal-spatial traffic features such as traffic flow distribution and the distribution of activity locations. We also seek a compromise between the well-established theoretical reward function form and the quality of data we could truly get from practical surveys. The simplified reward function makes the algorithm immediately applicable.

The rest of this paper is organized as follows. Section 2 introduces the algorithm of modified multiagent-based q-learning. Section 3 is devoted to the analysis and calculation of the survey data. Section 4 shows the temporal-spatial simulation results of Shangyu city's traffic system. Section 5 concludes the findings of this paper and discusses future research directions.

2. Multiagent-Based Q-Learning Method

2.1. Reinforcement Learning. Multiagent system focuses on the analysis of several agents' dynamic and complex collective behavior. Because multiagent system has no global control and each agent may get incomplete information, the system must learn repetitively to improve the performance. Reinforcement learning is a major method of this kind. Kaelbling et al. [19] define reinforcement learning as the problem faced by an agent that must learn behavior through trial and error interactions in a dynamic environment. Moreover, the consequences of actions change over time and depend on the current and future state of the environment. Reinforcement learning has the potential to deal with this uncertainty through continuous observations of the environment and through consideration of indirect and delayed effects of actions.

Basic concepts concerning reinforcement learning include the following.

(i) Agent: in this paper, an agent means a traveler.

(ii) State: a vector (activity, start time, duration, location, and congestion degree) denotes an agent's state. The vector is denoted as (a, s, d, l, and vc) for brief.

(iii) Location: the unit of location is traffic zone which is an area that has multifunctions including leisure, shopping, and working.

(iv) Activity: activities include home, work, maintenance, and leisure.

(v) Action: there are 4 actions, staying at current activity or move to one of the other 3 activities. The same as the way activities are represented; actions are denoted as h, w, s, and l for brief.

(vi) Duration and start time: time variables should be discrete in q-learning. The unit of time slot is 15 min, which divides a day into 96 slots. 24 pm is connected with 0 am. Because the number of state should be finite, the longest duration of an activity is limited to 24 hours. Hence, both duration and start time could be represented as a number from 1 to 96.

(vii) Policy: it means how an agent's action may bring it from one state to another.

(viii) Reward function: it is defined as the immediate feedback an action brings.

(ix) Value function: it shows the total feedback an action may bring both immediately and afterward.

(x) Q-value: the Q-value of an action a, given a state s, denotes the expected utility of an agent taking action a in state s.

(xi) Congestion degree: in order to show influence of agents' behavior on the environment, a variable of congestion degree vc is added to the state of environment. Because environment is represented by discrete variables in q-learning, vc should also be a discrete variable. It is defined as $vc = \text{ceil}(5 \cdot (v/c))$. MATLAB function ceil () rounds $r(s_t, a_t)$ to the nearest integer towards infinity. V is the traffic volume of a given OD pair and C is the capacity of the OD pair. Because the capacity is hard to be measured directly, we assume that each OD pair's capacity is 4000 considering the total population and the size of Shangyu city.

Reinforcement learning tasks are generally treated in discrete time steps. A teach time step t, the agent observes the current state s_t stand chooses a possible action at to perform, which leads to its succeeding state $s_{t+1} = \delta(s_t, a_t)$. The environment responds by giving the agent a reward $r(s_t, a_t)$. These rewards can be positive, zero, or negative. It is probable that these preferable rewards come with a delay. In otherwords, some actions and their consequential state transitions may bring low rewards in short-term, while it will lead to state-action pairs later with a much higher reward.

For this reason, the task of the agent is to learn a policy π according to the state S and the action A to receive the maximal accumulative rewards. Given a random policy π from a random state s_t, the accumulative reward can be formulated as follows:

$$V^{\pi}(s_t) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots = \sum_{i=0}^{\infty} \gamma^i r_{t+i}, \qquad (1)$$

where r_{t+i} represents the scalar reward received i steps in the future and γ is the discounting factor. The agent only receives the immediate reward if γ is set to zero.

2.2. Q-Learning Algorithm. The agent needs to learn the optimal policy $\pi^*(s)$ that maximizes the accumulative reward. Unfortunately, it is required that the knowledge of immediate reward function r and state transition function δ are known in advance. In reality, however, it is usually impossible for the agent to predict in advance the exact outcome of applying a random action to a random state. In other words, the domain knowledge is probably not perfect. q-learning is then devised to select optimal actions even when the agent has no knowledge about the reward and state functions.

We define $\widehat{Q}$ as the estimation of true Q-value. The q-learning algorithm maintains a large table with entries to each state-action pair. When it starts, the value of $\widehat{Q}(s, a)$ is initially filled with random numbers. The agent repeatedly observes its current state s, chooses a possible action a to perform, and determines its immediate reward $r(s, a)$ and resulting new state $\delta(s, a)$. The $\widehat{Q}(s, a)$ value is then updated according to the following rule:

$$\widehat{Q}(s, a) \longleftarrow r(s, a) + \gamma \max_{a'} \widehat{Q}(s', a'). \qquad (2)$$

That is to say, the $\widehat{Q}$-value of the current state-action pair is refined based on its immediate reward and the $\widehat{Q}$-value of its next state. The agent can reach a globally optimal solution by repeatedly selecting the action that maximizes the local values of Q for the current state.

This is only a brief introduction of q-learning and detailed introduction could be found in reference [20]. The process can be described as follows:

(1) initialize the Q-values,

(2) select a random starting state s which has at least one possible action to select from,

(3) select one of the possible actions. This action leads to the next state,

(4) update the Q-value of the state-action pair according to the update rule above,

(5) go back to Step 3 if the new state has at least one possible action, if not, go to Step 2.

2.3. Reward Function. Previous researchers in this domain constructed their reward functions based on activity start time, duration, length of travel, and so on [20, 22]. This method is adopted by us and our reward function contains the following parts.

2.3.1. Reward Based on Attraction Degree of Zones. In this paper a location is a zone that has multiple land use functions. In reality, people sometimes prefer to travel for a long time downtown to go shopping because the land use characteristics make downtown more attractive. To quantify this, the reward based on attraction degree of zones is added to the reward function. It is only for maintenance and leisure activities because home and work have fixed locations. We assume the more maintenance activities are conducted in a zone, the higher attraction degree this zone has. This also applies for leisure activities.

Consider

$$\text{attract}_{i,j} = \frac{n_{i,j} - n_{i,\text{avg}}}{n_{i,\max} - n_{i,\text{avg}}}, \qquad (3)$$

where $n_{i,j}$ is the number of leisure activities or maintenance activities conducted in zone j, i is the activity type, n_{avg} is the average leisure or maintenance activities conducted among all zones, and $n_{\max}$ is the maximum leisure or maintenance activities conducted among all zones. The reward is $r_{\text{attract}(i,j)} = 50 * \text{attract}_{i,j}$.

2.3.2. Reward Based on Activity Duration. When an agent conducts an activity and the duration is within a reasonable range, it should get a fairly large accumulative reward. When the duration is less than the expected value, the marginal benefit is positive, while if the duration is more than the expected value, the marginal benefit is negative.

Consider

$$r_{\text{duration}(i)} = \begin{cases} 100 + \dfrac{d_{\min(i)} - d}{d_{\min(i)}}, & \left(d < d_{\min(i)}\right), \\ 50, & \left(d_{\min(i)} \leq d < d_{\text{avg}(i)}\right), \\ -50, & \left(d_{\text{avg}(i)} < d \leq d_{\max(i)}\right), \\ -100, & \left(d > d_{\max(i)}\right), \end{cases} \tag{4}$$

where $d_{\min(i)}$, $d_{\text{avg}(i)}$, and $d_{\max(i)}$ represent the reasonable minimum, maximum, and average duration of activity i. They are, respectively, the 5%, 50%, and 95% percentile duration of activity i in the survey data.

2.3.3. Reward Based on Activity Start Time. Each activity's start time distribution is calculated using the survey data. To make the distribution curve more smooth in order to diminish the effect of randomness, we use polynomial functions (use C to denote) to fit the curve. Then function C is normalized.

Consider

$$r_{\text{start time}(i)} = C_i\left(s\right), \tag{5}$$

where i represents the type of activity, while s is the start time of the activity. The range of s is $(1, 96)$.

2.3.4. Reward Based on Travel Time. Some scholars define travel-time-based reward as $r_{\text{travel}} = -c * (bt)^{\alpha}$ [23]. This form is adopted by us, but it needs some modifications because the influence of congestion degree is taken into account. t is no longer a fixed value decided by the length between zones, but it relates to the congestion degree of the OD pair. We use the widely accepted impedance function in China [21]:

$$U = \begin{cases} U_0\left(1 - \dfrac{0.6V}{C}\right), & \dfrac{V}{C} < 0.9, \\ \dfrac{U_0}{7.4V/C}, & \dfrac{V}{C} \geq 0.9, \end{cases} \tag{6}$$

where U is the actual speed, while U_0 is free flow speed. t_0 is the free flow travel time. Actual travel time t could be defined as $t = t_0 * U_0/U$.

2.4. Flow-Chart of Calculation. When q-learning is applied in this paper, the process described below could be shown in Figure 1. The whole process is separated into 3 steps.

> Step 1 is to utilize travel diary survey data to extract typical activity patterns and form different kinds of agents according to their activity patterns. Also utilizing the survey data, the reward function for different kinds of agents is calculated.
>
> Step 2 is to estimate the value function (in this algorithm: Q-values) through trial and error until the Q-value matrix converges.
>
> Step 3 is to add agents on the network and then use the Q-values to decide the activity-travel schedule of each agent. In the end, temporal-spatial characteristics of the simulation result and each agent's activity-travel schedule are calculated and recorded.

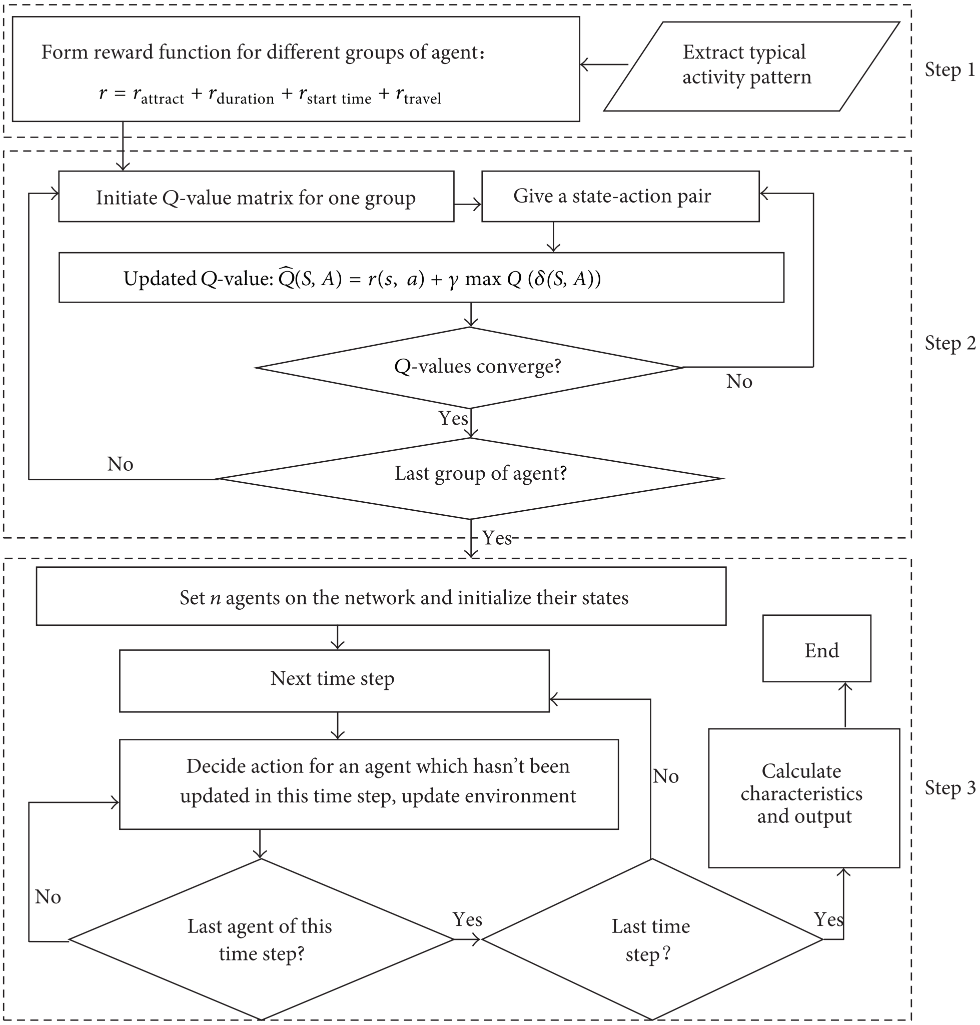
Taking congestion degree into account could enable interactions among agents and let them cooperate and compete in the environment. In the simulation of a network, a number of agents are set on the network and their states are initialized. Then at each time step, agents decide their actions by choosing an action that brings maximum Q-value one by one according to the congestion degree and other aspects of the environment. Their actions would in turn influence the congestion degree therefore would influence other agents' actions. In this way, all agents' activity-travel schedules could be decided.

3. Data Analysis and Process

3.1. Data Survey. We utilized the travel diary survey data from Shangyu city conducted in 2006. The survey includes individual/household sociodemographics and travel records. Travel records include trip starting and ending times, origin and destination, mode used, and trip purpose. Trip purpose is divided into nine categories, including work, school, official business, shopping, socializing-recreation, serving passengers, personal business, returning home, and returning to work. Among these purposes, work, school, and official business are named commute activity or simply work. Shopping, serving passengers, and personal business are called maintenance activities. Socializing-recreation is called leisure activity. Maintenance and leisure activities generally are named none-working activities. Hence, the 9 categories of activities could be divided into 4 types: work, maintenance activity, leisure activity, and staying at home.

Shangyu city has a population of 204,900. 4,101 residents from 1,564 households are surveyed. After deleting the incorrect statistics, data from 3,368 people are used, representing 82.1% of the people surveyed. 486 students account for 14.4% of the valid data. Because students' activity-travel schedules are rather fixed and the main focus of this paper is on working and none-working groups, the students' data are not considered. Thus, the data obtained from the remaining 2,883 people, accounting for 85.6% of all the valid data, are used for the analysis.

3.2. Typical Activity Patterns. The first step of processing valid survey data is to extract typical activity patterns. A tour is defined as the travel from home to one or more activity locations and back to home again [4]. An activity pattern here is defined as all tours an individual conducted in a single day. In the valid data, 10 of the patterns are shared by more than 20 samples. We call these 10 activity patterns typical activity patterns and the description of them can be seen in Table 1. They take up 2397 of the 2882 valid samples. Agents could be



Form reward function for different groups of agent：
$r = r_{\text{attract}} + r_{\text{duration}} + r_{\text{start time}} + r_{\text{travel}}$
Extract typical activity pattern
Step 1
Initiate Q-value matrix for one group
Give a state-action pair
Updated Q-value: $\widehat{Q}(S, A) = r(s, a) + \gamma \max Q (\delta(S, A))$
Q-values converge?
No
Yes
Step 2
No
Last group of agent?
Yes
Set n agents on the network and initialize their states
End
Next time step
Decide action for an agent which hasn't been updated in this time step, update environment
No
Calculate characteristics and output
Step 3
No
Last agent of this time step?
Yes
Last time step？
Yes

Figure 1: Three Steps of multiagent-based q-learning simulation.

classified according to their activity patterns. We take these 10 typical patterns to form 10 types of agents. Patterns which include working activity are called commuting patterns, and others are called none-working patterns. The characteristics of these 10 patterns are described as in Table 1 (the 4 activities are written as h, w, s, and l for brief).

3.3. Reward Function Calculation. The reward function has been constructed in Section 2.3. The paragraphs below show the values of parameters used in the reward function, calculated from the survey data. Furthermore, ten different types of agents have their own parameters, respectively, though the functional forms are the same.

3.3.1. Attraction Degree of Zone. The attraction degrees of zones are listed in Figure 2, next to it is the traffic zone division of Shangyu city. Because these degrees are decided by land use characteristics of different zones, to different groups of people the attraction degrees are the same.

It is quite clear that zones 2, 8, and 13 are the center of leisure activity, while zones 3 and 5 are the center of maintenance activity. This result corresponds to the land use characters of Shangyu because these zones are in the center of Shangyu.

3.3.2. Reward Based on Duration. To make the results more realistic, we calculate the rewards based on duration of 10 typical activities patterns according to the definition in Section 2.3. The unit of these parameters is 15 min. The relatively small value of standard deviation shows that people who belong to the same group share much similarity in behavior, at least in the duration of activity.

Where are the statistics?

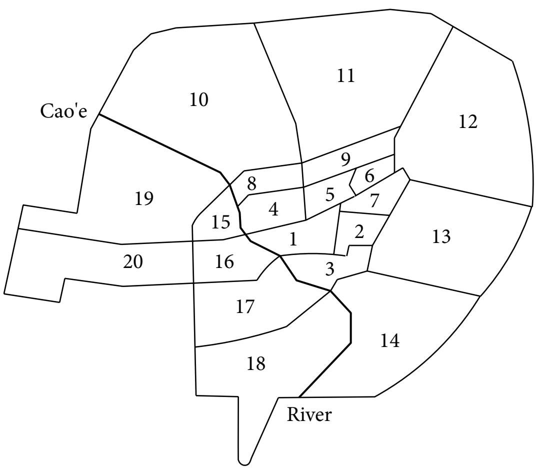
3.3.3. Reward Based on Activity Start Time. The process of calculating this reward has been stated in Section 2.3. Use polyfit function in MATLAB to fit every activity's start time distribution of each group into smooth curves.

In Figure 3 min is not 0.

The start time-duration-reward graphs of the four activities are shown in Figure 3.

Table 1: Description of typical activity patterns.

Activity pattern	Number of samples	Ratio	Description
hwh	1069	37.1%	Simple work pattern with only primary tour
hwsh	26	1.0%	Having other stops when getting off work, with only primary tour
hwhsh	44	1.5%	With a secondary tour, primary tour is simple work pattern
hwhwh	568	19.7%	Work tour with home-based subtour
hwswh	26	1.0%	With a subtour during work
hsh	357	12.4%	Simple maintenance tour
hlh	111	3.9%	Simple leisure tour
hshlh	32	1.1%	With both maintenance and leisure tours, the prior one is a maintenance tour
hshsh	105	3.6%	With two maintenance tours
hlhsh	59	2.0%	With both maintenance and leisure tours, the prior one is a leisure tour



Attraction degree	Zone number	1	2	3	4	5
	Leisure attraction degree	−0.15	0.58	−0.19	0.29	−0.17
	Maintenance attraction degree	−0.09	0.05	0.43	−0.14	0.82
	Zone number	6	7	8	9	10
	Leisure attraction degree	−0.21	−0.26	0.70	0.35	−0.11
	Maintenance attraction degree	−0.17	−0.15	−0.13	−0.10	0.09
	Zone number	11	12	13	14	15
	Leisure attraction degree	−0.11	−0.23	0.64	−0.23	−0.15
	Maintenance attraction degree	0.15	−0.12	−0.08	−0.17	−0.05
	Zone number	16	17	18	19	20
	Leisure attraction degree	−0.05	−0.09	−0.19	−0.28	−0.13
	Maintenance attraction degree	−0.15	−0.07	−0.09	0.13	−0.13

Figure 2: Attraction degrees of zones and traffic zone division of Shangyu city.

4. Simulate Temporal-Spatial Features of Multipleagents

4.1. Assumptions and Preparations for the Simulation. To simulate traffic conditions in Shangyu, the first step is to expand the number of agents from the size of the sample to the proportion of population these types of agent take up in Shangyu. By calculation, the 2397 samples in the survey should be expanded to a population of 145684 people. Apart from the already existed data of 2397 people, we need to establish 143287 people's attribute data. Each people's attributes include activity pattern and home and work locations (if this person works). To make the distribution of each attribute in the newly established data the same as the survey data, the procedure of establishing one person's attributes could be as follows.

(1) Randomly generate a natural number from 1 to 2397. The activity pattern of this people will equal to that of the number i people in the survey data.

(2) Likewise, the attribute of home and work locations can be decided by randomly choosing one from the 2397 survey samples.

Initialize each agent's state and simulate 1000 time steps. We take the last 96 time step to analyze. Both each individual agent's activity-travel schedule and spatial-temporal characteristics are analyzed.

4.2. Simulation Results of Activity-Travel Schedules. Each agent's activity-travel schedule in one day is recorded. We randomly choose one agent from each pattern and show his/her activity-travel schedule in a day (from 0 am to 24 pm). The

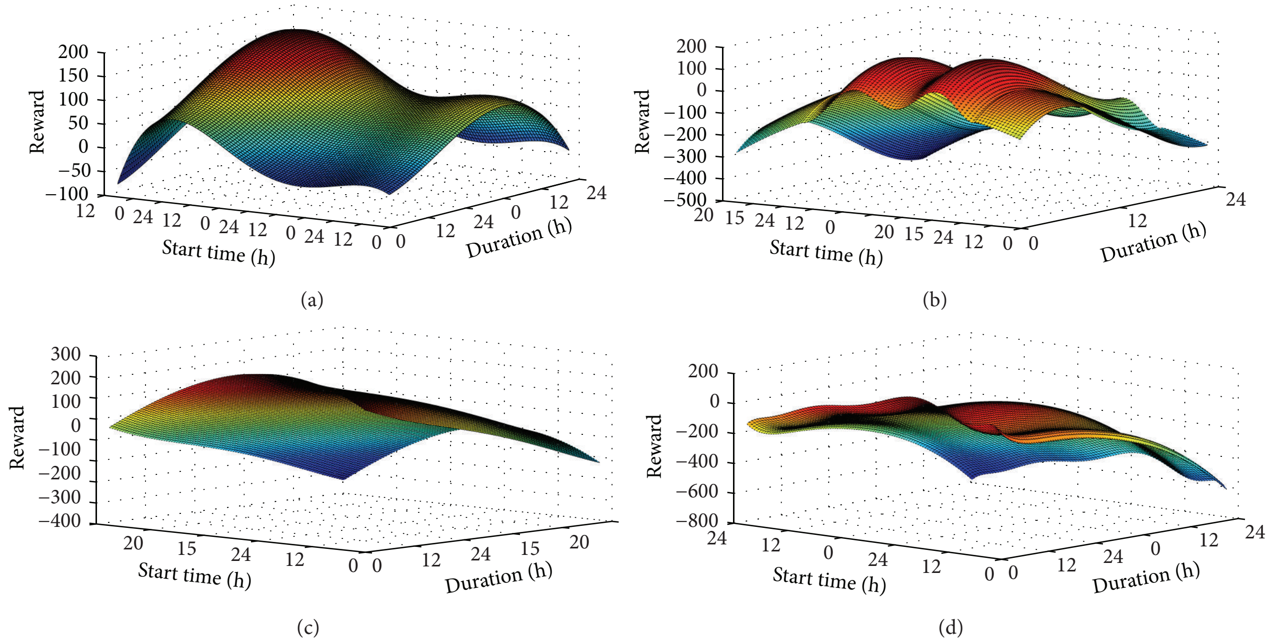


FIGURE 3: (a) Home reward function. (b) Working reward function. (c) Leisure reward function. (d) Shopping reward function.

TABLE 2: Activity-travel schedules.

Agent type	Activity-travel schedule
hwh	h (00:00–08:00, 05) w (08:30–16:45, 08) h (17:30–24:00, 05)
hwsh	h (00:00–07:00, 12) w (07:30–17:00, 11) s (17:15–18:15, 11) h (18:45–24:00, 12)
hwhsh	h (00:00–07:45, 01) w (08:15–15:30, 02) h (15:45–19:00, 01) s (19:15–19:45, 03) h (20:00–24:00, 01)
hwhwh	h (00:00–06:30, 11) w (07:30–11:15, 07) h (12:00–13:00, 11) w (13:45–17:30, 07) h (18:30–24:00, 11)
hwswh	h (00:00–07:15, 04) w (07:45–11:45, 06) s (12:15–12:45, 05) w (13:15–16:30, 06) h (17:15–24:00, 04)
hlh	h (00:00–05:45, 03) l (06:00–07:00, 02) h (07:30–24:00, 03)
hlhsh	h (00:00–04:45, 02) l (05:15–06:15, 08) h (06:45–07:00, 02) s (07:30–08:00, 03) h (08:30–24:00, 02)
hsh	h (00:00–06:30, 15) s (07:00–07:30, 05) h (08:00–24:00, 15)
hshlh	h (00:00–06:00, 04) s (06:30–06:45, 08) h (07:15–07:30, 04) l (08:00–08:30, 08) h (09:00–24:00, 04)
hshsh	h (00:00–05:15, 17) s (06:30–07:00, 03) h (08:00–16:30, 17) s (17:15–17:45, 03) h (18:30–24:00, 17)

result is shown in Table 2. The first part in each parenthesis is activity time and the second part is activity location. The table shows that no abnormal sequence, such as staying at one activity for too long or conducting activities in improper time, occurs in these 10 examples. One flaw is that to avoid the morning peak of commute agents, the none-working agents' trips are generally a little bit earlier than the peak shown by the survey.

Having activity-travel schedules of all agents, we could move our analysis further to macroscopic temporal-spatial characteristics of the traffic.

4.3. Temporal Characteristics of the Simulation Result. The traffic flow distributions of this paper's algorithm and the traditional algorithm which has not taken interactions between agents are compared in Figure 4. Both methods show apparent morning and evening peak. But in the traditional method environment is static, which means one agent's action will not affect other agents' choices; it is natural that agents of the same attributes all do the same activity at the same time and zone. Therefore, the traditional method's distribution of flow is ladder-like, which means peak hour flow is very large.

By comparison, because the congestion degree is taken into account, in this paper's method, some agents avoid traveling in the rush hour because it will lead to lower rewards. As a result, the peak hour flow is much lower. Even agents of the same attributes would have different activity-travel schedules because the environment is dynamic. Thus, the behavior of the whole population is not isolated but has interactions.

Traffic flow distributions of the 2397 samples' survey result and their corresponding agents' simulation result are shown in Figure 5. The simulation result matches the survey result well. Their peak hour flow deviation is less than 5%.

To show the features of different patterns' traffic flow distribution, we could mark different traffic patterns' flows with different colors as is shown in Figures 6(a) and 6(b).

Table 3: Comparison of two methods' PHR values.

OD pair	Morning peak		Evening peak	
	Traditional method	New method	Traditional method	New method
(3, 11)	32.25%	17.93%	27.79%	11.42%
(5, 16)	24.43%	11.34%	31.64%	15.03%
(4, 17)	32.73%	21.52%	46.19%	20.49%
(5, 11)	25.50%	8.65%	28.61%	15.01%
(9, 10)	22.18%	16.79%	18.77%	10.12%
(9, 5)	46.23%	18.11%	17.98%	12.38%

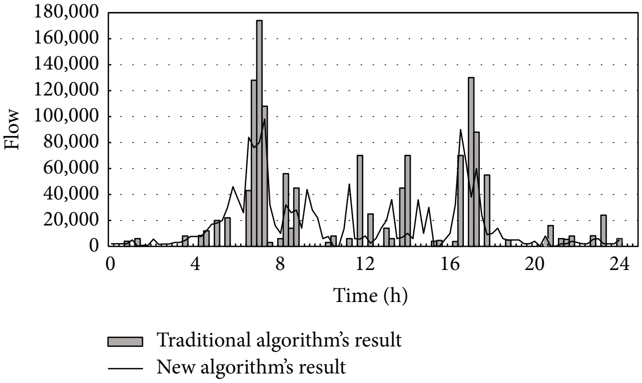


Figure 4: Comparison of traffic flow distribution between new method and survey data (flipped with Figure 5).

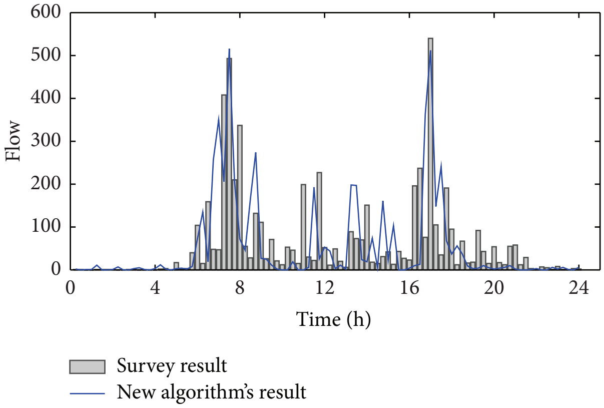


Figure 5: Comparison of traffic flow distribution between traditional method and new one.

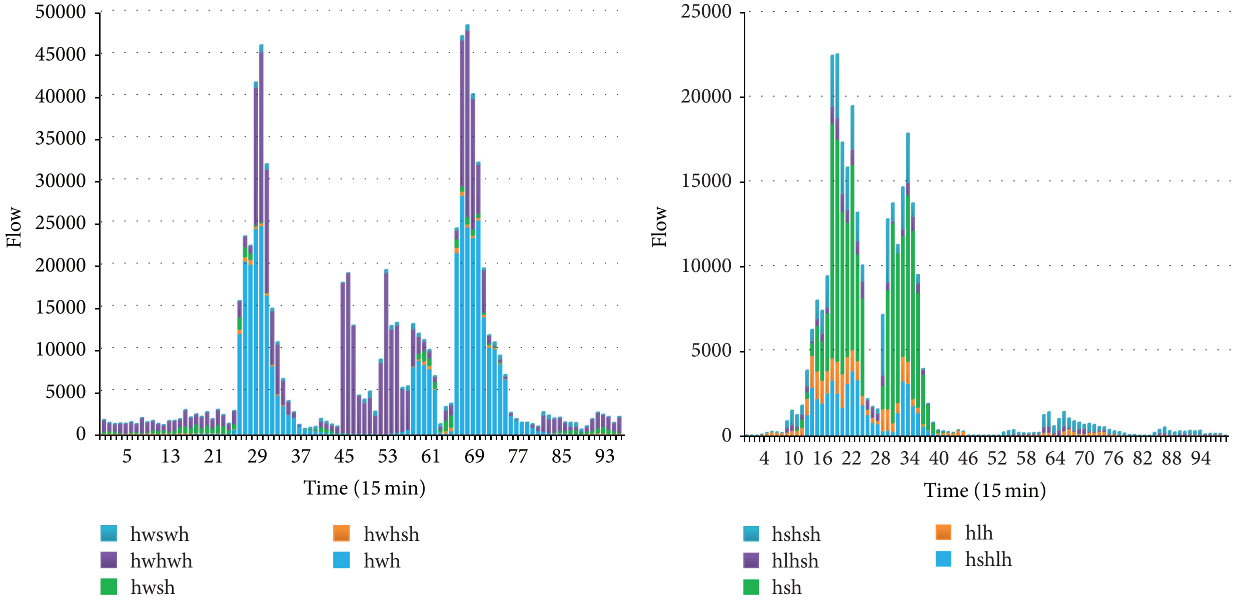
Figure 6(a) shows the flow distribution of the 5 commute patterns. It shows clear morning and evening peaks, at about 7 am and 6 pm, respectively. Compared with Figure 3 we could find out that these commuting patterns, especially pattern hwh and hwhwh, account for a large percentage of morning and evening peaks' flow. The peak at noon is caused by pattern hwhwh agents who go home at noon. On the whole, pattern hwh and hwhwh are the determinants of commuting patterns' flow distribution, and other commuting patterns have too few people to influence the trend.

Figure 6(b) shows none-working agents' traffic flow distribution, which is totally different from commuting agents': there is no such dominant pattern. On the contrary, all none-working patterns contribute to the formation of figure's shape. Two peaks of the flow are all in the morning, at about 5 am and 9 am, respectively. The survey result shows that 42.9% of the none-working groups are retired people in Shangyu. In China the elderly usually like to go out to do some exercises early in the morning and food markets usually open very early; this explains why both the survey result and the simulation result show that none-working people's travel peak is in the morning. Agents of pattern hlh tend to go out early at 5 am, while the flow of pattern hsh almost distributes evenly from 5 to 10. In China, most people tend to stay at home in the evening, especially the none-working people so there is not much traffic in the evening as in Figure 6(b).

Table 3 shows two methods' comparison of peak hour ratio (PHR). Peak hour ratio is defined as the ratio of peak hour flow and the traffic flow of a whole day. In China, the measured PHR is often between 10% and 15%. Because there are too many OD pairs, the table listed the results of 6 OD pairs which have the largest traffic flow as representative. For all OD pairs, the original method's average PHR is 30.5% and the result of the new method is 16.2%. It is clear that the latter is closer to reality.

4.4. Spatial Characteristics of the Simulation Result. In the traditional method, because congestion degree is not taken into account and attraction degree's effect is quite distinct, all agents conduct their maintenance and leisure activities at the zone that has maximum attraction degree: all the 20094 leisure activities are conducted in zone 8, while all the 70433 maintenance activities are conducted in zone 5. We need to mention that because Shangyu is a small city and the distances between zones are not very long; the influence of distances between OD pairs is subtle. After taking into account congestion degree, the choice of location is much more dispersed. Agents would choose to conduct their activities in other zones which have lower attraction degrees when center zones are crowded. Finally, 5845 leisure activities are conducted in zone 8, which accounts for 29.0% of all leisure activities. 22392 maintenance activities are conducted in zone 5, accounting for 31.7% of all maintenance activities.

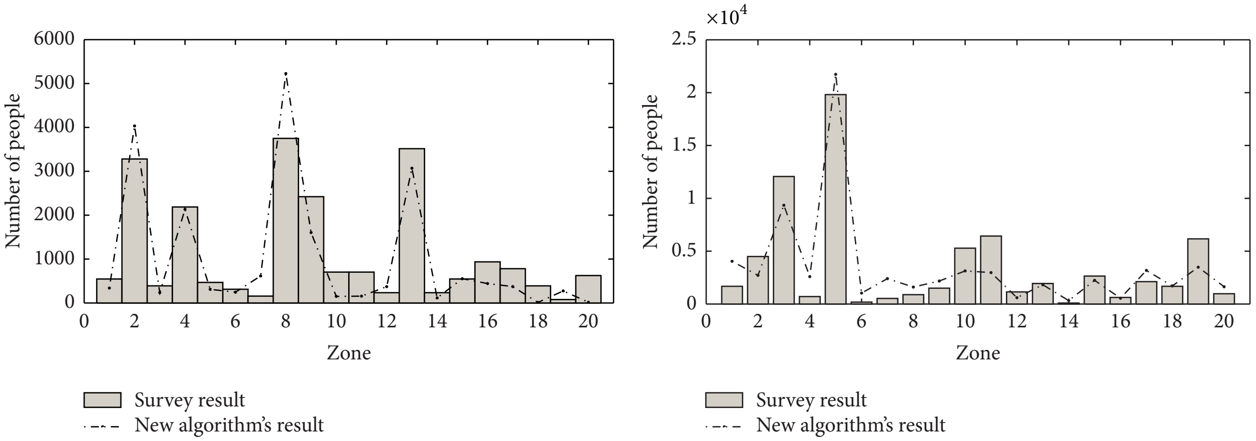
The choice of activity zones is shown in Figure 7. The survey data's activity location distribution is calculated and then it is extended the same proportion that the samples are



(a)

(b)

Figure 6: (a) Commute agents' traffic flow distribution. (b) None-working agents' traffic flow distribution.



(a)

(b)

Figure 7: (a) Choice leisure activity location. (b) Choice of maintenance activity location.

extended to show how the 145684 people's choice of location would be like according to the survey data.

It is compared with the simulation result and the figure shows that the simulation result is quite close to the extended survey result. The correlation coefficient between the survey data's leisure activity location distribution and the simulation result is 0.921. And the correlation coefficient between survey data's and the simulation result's maintenance location distribution is 0.902.

5. Conclusions and Future Directions

In this paper we use a modified multiagent-based reinforcement learning algorithm to simulate the traffic condition of Shangyu city. Both the spatial-temporal features of the entire population and the activity-travel schedule of single individuals are analyzed. The main findings are listed as follows.

(i) This paper's method takes the congestion degree between OD pairs into account, which enables agents' actions to influence the environment. Thus, agents' actions have interactions with each other. Because of this interaction, both the spatial-temporal features of the entire population and the activity-travel schedule of single agent are close to actual situations.

(ii) Because in this paper agents are no longer separated individuals but an integrity that interacts with each other, the spatial-temporal features of the whole population, such as traffic flow distribution, PHR factor, and location choice distribution, could be calculated, which is rarely seen in previous research in this field.

(iii) Survey data are utilized throughout the whole process, including the setting of traffic zones, extraction of typical activity patterns, formation of agents, and reward functions. The utilization of the survey data makes the simulation result closer to the actual situation in Shangyu; therefore, the simulation result has practical meanings and could be further utilized in transportation planning and management. For example, it could be used in TDM policy effect analysis.

(iv) Data used in this paper come from the survey of a typical small city in east China. Both the survey data and the simulation results have distinct Chinese characteristics. For example, maintenance and leisure activity are conducted mostly in the morning and people tend to stay at home in the evening; commuting groups have few leisure and maintenance activities during weekdays. These features provide materials for future research of Chinese traffic.

The above mentioned analysis of the simulation result shows that this paper's simulation method could better reflect actual traffic conditions. Both the macroscopic spatial-temporal features and the microscopic activity-travel schedule render this method valid. The veracity of the simulation result and the utilization of survey data enable this method to better service practical transportation planning and management.

Because of the limitations of the survey data and the algorithm, several aspects of the research can be improved in the future.

(i) Route choice in the current model is simplified. The travelers "jump" directly from the origin to the destination, while the influence on the intermediate regions is neglected.

(ii) In this paper, the reward function contains four different parts; they are, respectively, based on attraction degree of zones, activity start time, duration, and travel time. When accumulated, the weights of them are considered to be equal. However, in reality, these factors have different effects on people when they make the decision on their trips. So one future direction is to calculate these weights according to the survey data, making the simulation results more accurate.

(iii) Road impedance varies greatly according to the type of traffic mode, since different modes have different occupation rates of roads and their speed are also different. As a result, it is better to take traffic mode of each agent into consideration when calculating congestion degree.

(iv) Reaction to uncertain events is a special characteristic of reinforcement learning. In this paper we are focusing on the most probable or the "average" state of the system. But it is also interesting to explore how the agents would react to radical changes of the environment and how do they interact with each other under this circumstance.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research is supported by the National Basic Research 973 program (2012CB725400) and Natural Science Foundation of China (51378120 and 51338003). Also Fundamental Research Funds for the Central Universities and Foundation for Young Key Teachers of Southeast University are appreciated.

References

[1] M. Fried, J. Havens, and M. Thall, "Travel behavior—a synthesized theory," Final Report, NCHRP, Transportation Research Board, 1977.

[2] F. S. Chapin, *Human Activity Patterns in the City*, John Wiley & Sons, New York, NY, USA, 1974.

[3] T. Hägerstrand, "What about people in regional Science?" *Papers of the Regional Science Association*, vol. 24, no. 5, pp. 6–21, 1970.

[4] J. L. Bowman and M. E. Ben-Akiva, "Activity-based disaggregate travel demand model system with activity schedules," *Transportation Research A*, vol. 35, no. 1, pp. 1–28, 2001.

[5] F. S. Koppelman and C.-H. Wen, "The paired combinatorial logit model: properties, estimation and application," *Transportation Research B*, vol. 34, no. 2, pp. 75–89, 2000.

[6] M. E. Ben-Akiva and S. R. Lerman, *Discrete Choice Analysis: Theory and Application to Travel Demand*, The MIT Press, London, UK, 1985.

[7] C. R. Bhat, "A hazard-based duration model of shopping activity with nonparametric baseline specification and nonparametric control for unobserved heterogeneity," *Transportation Research B*, vol. 30, no. 3, pp. 189–207, 1996.

[8] P. M. Jones, *Understanding Travel Behavior*, University of Oxford, Transport Studies Unit, Oxford, UK, 1983.

[9] M. G. Karlaftis and E. Vlahogianni, "Statistical methods versus neural networks in transportation research: differences, similarities and some insights," *Transportation Research C*, vol. 19, no. 3, pp. 387–399, 2011.

[10] T. A. Arentze and H. J. P. Timmermans, "A learning-based transportation oriented simulation system," *Transportation Research B*, vol. 38, no. 7, pp. 613–633, 2004.

[11] T. Arentze, F. Hofman, and H. Timmermans, "Reinduction of Albatross decision rules with pooled activity-travel diary data and an extended set of land use and cost-related condition states," *Transportation Research Record*, vol. 1831, pp. 230–239, 2003.

[12] T. Arentze and H. Timmermans, "Parametric action decision trees: incorporating continuous attribute variables into rule-based models of discrete choice," *Transportation Research B*, vol. 41, no. 7, pp. 772–783, 2007.

[13] K. M. Nurul Habib, "A random utility maximization (RUM) based dynamic activity scheduling model: application in weekend activity scheduling," *Transportation*, vol. 38, no. 1, pp. 123–151, 2011.

[14] B. Fernandez-Gauna, J. M. Lopez-Guede, and M. Graña, "Transfer learning with partially constrained models: application to reinforcement learning of linked multicomponent robot system control," *Robotics and Autonomous Systems*, vol. 61, no. 7, pp. 694–703, 2013.

[15] E. A. Jasmin, T. P. Imthias Ahamed, and V. P. Jagathy Raj, "Reinforcement learning approaches to economic dispatch problem," *International Journal of Electrical Power & Energy Systems*, vol. 33, no. 4, pp. 836–845, 2011.

[16] A. Agung and F. L. Gaol, "Game artificial intelligence based using reinforcement learning," *Procedia Engineering*, vol. 50, pp. 555–565, 2012.

[17] R. Lahkar and R. M. Seymour, "Reinforcement learning in population games," *Games and Economic Behavior*, vol. 80, pp. 10–38, 2013.

[18] Z. Tan, C. Quek, and P. Y. K. Cheng, "Stock trading with cycles: a financial application of ANFIS and reinforcement learning," *Expert Systems with Applications*, vol. 38, no. 5, pp. 4741–4755, 2011.

[19] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: a survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996.

[20] D. Charypar and K. Nagel, "Q-learning for flexible learning of daily activity plans," *Transportation Research Record*, vol. 1935, pp. 163–169, 2005.

[21] W. Wang, *Transportation Engineering*, Southeast University Press, Nanjing, China, 2000.

[22] M. Vanhulsel, D. Janssens, and G. Wets, "Calibrating a new reinforcement learning mechanism for modeling dynamic activity-travel behavior and key events," in *Proceedings of the 86th Annual Meeting of the Transportation Research Board*, Transportation Research Board, Washington, DC, USA, 2007.

[23] D. Janssens, Y. Lan, G. Wets, and G. Chen, "Allocating time and location information to activity-travel patterns through reinforcement learning," *Knowledge-Based Systems*, vol. 20, no. 5, pp. 466–477, 2007.