MULTI-AGENT SIMULATION USING ADAPTIVE DYNAMIC PROGRAMMING IN THE EXISTENCE OF URBAN CONSOLIDATION CENTERS

Nailah FIRDAUSIYAH¹, Eiichi TANIGUCHI² and Ali Gul QURESHI³

¹Member of JSCE, Ph.D. Student, Dept. of Urban Management, Kyoto University
(Kyotodaigaku-Katsura, Nishikyo-ku, Kyoto 615-8540, Japan)
E-mail: firdausiyah@trans.kuciv.kyoto-u.ac.jp

²Fellow of JSCE, Professor Emeritus at Resilience Research Unit, Kyoto University
(Rohm Plaza, Kyotodaigaku-Katsura, Nishikyo-ku, Kyoto 615-8520 Japan)
E-mail: taniguchi.eiichi.3a@kyoto-u.ac.jp

³Member of JSCE, Associate Professor, Dept. of Urban Management, Kyoto University
(Kyotodaigaku-Katsura, Nishikyo-ku, Kyoto 615-8540, Japan)
E-mail: qureshi.aligul.4c@kyoto-u.ac.jp

This paper aims at developing and exploring a multi-agent system (MAS) with adaptive dynamic programming (ADP) (MAS-ADP) model to support city logistics stakeholders in their decisions for implementing city logistics measures. Adaptive dynamic programming based learning performs better than other learning techniques when agents need to interact in constantly changing environment, such as city logistics. The general framework of MAS-ADP has been described and an ADP model for the freight carrier learning agent has been developed. We investigated the performance of the ADP by embedding it in a simulation for evaluating the existence of Urban Consolidation Centers (UCC), and by comparing it with another simulation based on one of the most popular Reinforcement Learning algorithms namely the Q-learning. It has been found that ADP-based simulation performed better on three evaluation criteria; 1) cost saving, 2) accuracy, and 3) stability and adaptability, as compared to the one using Q-learning. It was observed that ADP-based simulation resulted in 1.5% less on the delivery cost experienced by the agents as compared to the Q-learning based simulation. In highly uncertain environment, ADP-based simulation was found to be 22% more accurate in terms of the difference between the real cost and the expected total delivery cost. Also, it was proved that the ADP is more adaptive to the changing environment and that it is more stable in the action selection by performing 22% less changes in actions from direct delivery system to the joint delivery system with UCC or vice versa as compared to the Q-learning.

Key Words : *adaptive dynamic programming, multi-agent simulation, city logistics, urban consolidation center*

1. INTRODUCTION

Urban freight transport has become an important issue in urban and transportation planning due to high population density in urban areas as well as due to the social, economic, and environmental problems associated with it. Recently, urban freight transport is facing two difficult problems; first is the unpredictable situation of goods delivery due to the parking issues, traffic congestion, and other restrictions in the urban area. It means that the delivery environment in the urban area is uncertain and unpredictable, which effects directly to the operational cost as well as the action selection in presence of optional solutions or policies. In contrast, freight transport's agents such

as freight carriers, are expected to provide higher services at lower costs. The second issue is the involvement of multiple agents in city logistics system, such as freight carriers, shippers, customers, and administrator. All of these key stakeholders in urban freight transport have their own specific objectives and tend to behave in a different manner to any urban freight policy¹⁾. The stakeholders also interact and influence each other in the city logistics environment. It also makes the environment dynamic and unpredictable (shown in **Fig. 1**, which is modified from Tamagawa, et al.²⁾).

Overall, city logistics is defined as the process of fully optimizing the logistics and transport activities with the support of advanced information systems in

urban areas considering the traffic environment, the traffic congestion, the traffic safety, and the energy savings within the framework of a market economy³⁾. The main challenge for the city logistics is, how to provide an efficient and environmentally friendly urban freight transport by considering multi-agent problems within the uncertain environment that has been described in the previous paragraph.

In order to achieve these aims, numerous city logistics initiatives have been proposed and implemented in several cities, including Urban Consolidation Centers (UCC), parking restrictions, load factor controls, etc. Decision support tools (DST) are needed to help public decision makers and practitioners to make decisions and evaluate the city logistics policies before they can be effectively deployed due to their manifold implications on different city logistics stakeholders⁴⁾. These DSTs are mainly based on modeling, optimization, simulation, and evaluation procedures. This research will be focused on modeling and simulation only, and does not attempt to do optimization that may involve multiple objectives, as discussed earlier.

There has been many attempts to develop multi-agent simulations to analyze decision making process of various stakeholders in city logistics, but almost all of them rely on Q-learning^{2), 5), 6)}. However, based on previous research experiences, which will be described in more detail in the next section, it has been found that adaptive dynamic programming based learning performs better agents need to interact in uncertain environment. As explained earlier, stakeholders (agents) of city logistics also have to deal with uncertain environment, therefore this research aims at developing ADP models in the multi-agent simulation to evaluate city logistics policies, such as UCC. The model could be used as DST to achieve better outcomes in the decision process of designing and implementation of city logistics policies as well as to save training cost and time especially when one need to deal with many uncertain variables.

Fig.1 Interactions and actions of four stakeholders in the city logistics.

(1) Multi-agent simulation and Markov decision processes

In simulations based on Multi-Agent Systems (MAS), the behavior of the modeled multi-agent system and dynamic aspects of its environment are implemented with autonomous agents⁷⁾. Based on their individual objectives, agents take decisions, which determine their current and future gains; these decisions also shape the environment. Clearly, at any time t , the decision of an agent depends on its earlier decision process, which is called the Markov Property⁸⁾. Therefore, the decision making in the MAS is called the Markov Decision Process (MDP). In the following we describe the details of MDP which is the base of the MAS, and it will be used in this research to illustrate the interaction and decision process of agents within the city logistics environment.

The Markov Decision Process consists of state (S), action (A), transition (T), and reward (R). Here, S is a set of all states within the system; A is a set of all actions which can possibly be taken by an agent, and T is a function describing the transition from one state to the next state. At any given time t the system can be in a state $s_t \in S$, it takes an action $a_t \in A$ determined by the policy π , i.e., $a_t = \pi(s_t)$, for the transition to the next state $s' = s_{t+1}$, which is denoted by $s_{t+1} = T(s_t, a_t)$, and at the same time receives a reward signal $r_{t+1} = r(s_t, a_t, s_{t+1}) \in R$. The function T and R can be described as equations (1) and (2), respectively.

$$T(s'|s, a) = T\{s_{t+1} = s' | s_t = s, a_t = a\} \quad (1)$$

$$R(s'|s, a) = E\{r_{t+1} | s_t = s, a_t = a, s' = s_{t+1}\} \quad (2)$$

The agent's policy π is a mapping of the actions that an agent will take in any given state. Thus, the objective of the agent is to determine a policy, which shall result in the optimum utility. The policy can be deterministic $\pi: S \rightarrow A$ or stochastic $S, A \rightarrow [0, 1]$. Therefore it is possible to define the value of a state s of a policy π as $V^\pi(s)$, which is known as the state-value function for policy π . The state value function (V) is agent's long-term utility for any given state. Therefore, utility not only refers to an immediate reward that can be received after doing an action but also to the sum of future rewards that can be expected either when following the agent's policy or when choosing the action that impacts the highest immediate reward. There are the two ways to calculate the value function; value iteration as equation (3) or policy iteration as equation (4).

$$V(s) \leftarrow R(s) + \gamma \max_{a \in A(s)} \sum_{s'} T(s'|s, a) V(s') \quad (3)$$

$$V(s) \leftarrow R(s) + \gamma \sum_{s'} T(s'|s, \pi(s)) V(s') \quad (4)$$

(2) Reinforcement learning (RL) and ADP

The MDP is usually represented by reinforced learning (RL) in MAS⁷⁾. The RL relates with the iterative learning that what to do and how to map situations to actions by learning a value (reward) received from the interaction with the environment. Temporal Difference (Q-learning and Sarsa), Monte Carlo Method, and ADP are common RL algorithms. ADP is used in this study to learn or evaluate an agent's interaction with, and perception of its uncertain environment that optimize its behavior (action), learning from the feedback received (reward). We investigate the performance of the ADP as a learning method in the city logistics environment (Urban Consolidation Centers (UCC)) by embedding it in a multi-agent simulation and by comparing it with a popular RL algorithms commonly used in the evaluation of city logistics schemes, namely the Q-learning.

2. LITERATURE REVIEW

(1) Evaluation models for city logistics schemes using MAS

According to Taniguchi et al.⁹⁾ decision support tools are needed to help public decision makers and practitioners to deal with City Logistics nuisances (mainly traffic congestion, and parking issues). The combination between MAS model with reinforced learning (RL) algorithms has been used for evaluating the behavior of stakeholders, who are affected by the implementation of a city logistics policy. In MAS environment, there are a number of agents that come together and interact, cooperate, coordinate, and negotiate with each other to reach their intended objectives. In past, the MAS with Q-learning algorithms have been used to evaluate the dynamic usage of UCC¹⁰⁾, load factor control and road pricing⁵⁾, e-Commerce¹¹⁾, truck ban and discounting motorway tolls²⁾, time windows restrictions¹²⁾, and Joint Delivery System (JDS) with parking restrictions⁶⁾. These researches constructed a model that can be used to evaluate implications of city logistics measures by considering the behavior of the key stakeholders, namely the carriers, shippers, administrator, and residents relating to urban freight transport.

MAS with Monte Carlo Method was used by Taniguchi, et al.¹³⁾ to model the effects of city logistics schemes with simulation model based on the dynamic vehicle routing and scheduling problem. The results indicated that implementing a truck ban in the environmentally damaged areas and dis-

counting motorway tolls in the urban motorway network will have a large environmental effect, and it will lead to an acceptable environment for all stakeholders.

It can be observed that most of the MAS research in city logistics use Q-learning to represent the decision making of the agents, they considered. A comparative study conducted by Fagan and Meier¹⁴⁾, proved that ADP performs particularly well in the criteria of adaptability and stability in multi-agent environment compared to other RL algorithms (i.e., Q-learning, Sarsa, and Temporal difference). Similar to their area of application (intelligence traffic systems), the city logistics environment also presents a very dynamic and uncertain environment, therefore, it can be expected that the ADP can improve the quality of the multi-agent simulations in city logistics as compared to the ones, which use Q-learning. Therefore, this research aims at the adaptation and evaluation of the ADP as the learning algorithm in city logistics.

(2) ADP for evaluating city logistics measures

ADP is a learning model in RL part that can be used in the simulation field and optimal control field. As an optimal control tool, Zhang, et al.,¹⁵⁾ has described that ADP scheme is suitable for applications to the systems with strong coupling, strong nonlinearity, and high complexity. It has also been concluded that the ADP is capable to deal with uncertainty^{16), 17)}, stochastic environment^{18), 9)}, and non-linearity¹⁵⁾. ADP has been widely implemented at the confluence of control problem²⁰⁾, intelligence traffic systems^{14), 21)}, and robotics²²⁾.

However, none of these previous researches has used the ADP in the multi-agent simulation field, particularly in the area of city logistics which represents a highly uncertain environment. Hardin²³⁾, concluded that the learning and adaptation make the system more robust to imperfect knowledge of the environment. The stability characteristics of a model are also important in decision making due to the urgency of proposing efficient action selection. Therefore, in this study, we will develop and explore the ADP in a multi-agent simulation in a city logistics environment. As mentioned earlier in this section, this study will compare the ADP with one of the famous learning model, namely Q-learning, in terms of cost saving, accuracy, stability and adaptability.

3. MODELS

(1) Problems

The core idea of this research is to evaluate city logistics measures using multi-agent models with

ADP. The key stakeholders in the city logistics systems, who are influenced by policy measures ought to take actions so as to optimize their objectives based on their perception and assessment from the data, and also from the interaction with the environment. As a consequence, agents will receive a reward associated with their actions, from the environment in form of delivery costs, profits, subsidies, NOx levels, etc.

As the preliminary research, this research developed and tested ADP in a simple case of an agent-based system by simulating freight carrier as the learning agent under the implementation of UCC and urban parking management as illustrated in the general research framework in **Fig. 2**. To make the environment more realistic, variations in key data such as demand, number and location of customers, and parking cost have been considered.

The architecture of MAS-ADP inside an agent is illustrated in **Fig. 3**, which has been modified from Fagan and Meier¹⁴⁾.

As stated in the research problem, freight carrier is considered as the learning agent in this research. Here, freight carrier is typically a company that specializes in the last mile delivery of goods from depot to the customers.

Fig. 2 General research framework.

Fig. 3 ADP architecture as the preliminary research framework (modified from Fagan and Meier¹⁴⁾).

As shown in **Fig. 4**, the freight carrier will choose an action whether to opt for a joint delivery systems (JDS) with UCC or deliver goods directly to their customer (DD) based on the minimal immediate rewards (delivery cost) associated with these two actions. UCC operator proposes the UCC fees per parcel to be paid by the freight carrier based on the number of the parcel that are given to UCC for further delivery. In order to support the UCC operation, parking permission and subsidies will be given by the administrator to the UCC operator's truck to enter the city center. In this preliminary research, the freight carrier will learn and take an action (learning agent), while other agents such as customer, UCC operator, and administrator are not the learning agent so they are assumed incapable of performing any action to change the environment.

(2) Framework of the model

Figure 5 shows a framework of the model developed in this research. The model consists of two sub-models, which are; 1) learning model for stakeholders, and 2) the model for vehicle routing problem with soft-time window (VRPSTW). VRPSTW model calculates the delivery cost for each freight carrier.

Fig. 4 Interactions and behavior of four agents.

Fig. 5 Framework of the model.

Two multi-agent simulations based on the ADP and Q-learning will be developed and compared in this study. These two learning models have the same function, i.e. to evaluate the behavior (action) of stakeholders by updating as well as learning the received (reward) value from the interaction with the environment (calculated by VRPSTW), and to take an action of using JDS with UCC or the direct delivery (DD).

(3) Developments of freight carrier's MAS-ADP models

The ADP method interacts with the environment to learn the reward and transition value. It follows the model of the rewards (equation (2)) and model of the transition (equation (1)) from the previous experiences, and then use current transfer probability $t(s_t, a_t)$ and current reward function $r(s_t, a_t)$ from the environment to update the model as equations (5) and (6).

$$T(s_t, a_t) \leftarrow T(s_t, a_t) + \alpha(t(s_t, a_t) - T(s_t, a_t)) \quad (5)$$

$$R(s_t, a_t) \leftarrow R(s_t, a_t) + \alpha(r(s_t, a_t) - R(s_t, a_t)) \quad (6)$$

Here, α is the learning rate of an agent, which is set to be a number between $0 < \alpha < 1$. The learning rate of 1 means that the agent will consider the most current information while 0 means agent does not learn.

Freight carrier's utility value is the total delivery costs that depends on the action selection, i.e., utility value is the sum of delivery cost and parking cost if freight carrier chooses to deliver goods directly to their customer. Otherwise, the utility value is the total fee paid by freight carrier to UCC operator as the reward for choosing joint delivery with UCC. Freight carrier's objective is to minimize the total cost of delivering goods to customers, which is defined later in the paper.

The MAS-ADP algorithm for updating the utility value function for freight carriers is formulated as in equation (7), which is modified from equation (4):

$$V_f(s_{f,t}) \leftarrow R_f(s_{f,t}, \pi_{f,t}) + \gamma_f \sum_{s_{f,t+1}} T(s_{f,t+1} | s_{f,t}, \pi_{f,t}) V_f(s_{f,t+1}) \quad (7)$$

where $V_f(s_{f,t})$ is the expected delivery cost obtained by freight carrier f in the state $s_{f,t}$. $R_f(s_{f,t}, \pi_{f,t})$ is the expected reward when policy $\pi_{f,t}$ is taken from state $s_{f,t}$. The parameter γ_f is the discount rate for freight carrier f , which is set to be a number between $0 < \gamma < 1$. The discount rate set at 1 means that the agent will consider the long term reward, while 0 means that the agent concerns only on the current rewards. $V_f(s_{f,t+1})$ is the expected delivery cost received by freight carrier f in the next state $s_{f,t+1}$.

Freight carrier will update the expected reward R and expected transfer probability T using the equation (8) and equation (9) below,

$$T(s_{f,t}, a_{f,t}) \leftarrow T(s_{f,t}, a_{f,t}) + \alpha_f(t_f(s_{f,t}, a_{f,t}) - T(s_{f,t}, a_{f,t})) \quad (8)$$

$$R(s_{f,t}, a_{f,t}) \leftarrow R(s_{f,t}, a_{f,t}) + \alpha_f(r_f(s_{f,t}, a_{f,t}) - R(s_{f,t}, a_{f,t})) \quad (9)$$

Here, $r_f(s_{f,t}, a_{f,t})$ is the immediate reward, and $t_f(s_{f,t}, a_{f,t})$ is the immediate transfer obtained by freight carrier based on the two possible actions $a_{f,t}$ viz., direct delivery (equation (10)), and JDS (equation (11)).

$$r_f(s_{f,t}, a_{f,t}) = O_{f,t} + p_{f,t,k}, \quad (10)$$

$$r_f(s_{f,t}, a_{f,t}) = UCCfee_{f,t} \quad (11)$$

where $O_{f,t}$ is the operational delivery cost when freight carrier f decides to deliver goods directly to its customer as the action $a_{f,t}$ on time t , and $p_{f,t,k}$ is equal to $\sum_{i \in C} p_i$ which is the total additional parking cost for a freight carrier f using vehicle k at time t that serves a customer i . Operational delivery cost $O_{f,t}$ is the total cost to do goods delivery by a freight carrier, calculated using the Vehicle Routing Problem with Soft Time Windows (VRPSTW) defined later in the paper.

The Freight carriers are considered to use VRPSTW to optimize their delivery costs and they will utilize the VRPSTW for decision making in the reward model. The second possibility of immediate reward that freight carrier can possibly receive is the $UCCfee_{f,t}$ which is the consequence of choosing a joint delivery system with UCC. $UCCfee_{f,t}$ is the multiplication of UCC fee offered by UCC with the total number of demand (parcels) that freight carrier gives to UCC.

The definition of state in this research is taken as the weekdays in a week (i.e. from Monday to Friday, total five states).

(4) Development of freight carrier's MAS Q-learning models

In order to evaluate our ADP model, we firstly compare the simulation results of ADP with Q-learning model. Similar to the ADP, Q-learning is also one of the Reinforcement Learning (RL) algorithms, that learns an action value function Q instead of a state-value function (V) and thus does not require a model of the environment, therefore, it is known as a model-free algorithm.

Q-learning was developed by Watkins and Dayan²⁴; the simplest form of one-step Q-learning is shown in equation (12), which is equivalent to the equation (4) in ADP.

$$Q(s, a) \leftarrow Q(s, a) + \alpha [r_{s,a} + \gamma \max_{a'} Q(s', a')] \quad (12)$$

Q-learning uses algorithms that adjust estimated values $Q(s, a)$ of a state with immediate reward $r_{s,a}$ and the maximum estimated value of the next state $\max_{a'} Q(s', a')$. The Q-learning algorithm for updating the action-value function for freight carriers is formulated as in equation (13),

$$Q(s_{f,t}, a_{f,t}) \leftarrow (1 - \alpha_f) Q_f(s_{f,t}, a_{f,t}) + \alpha_f \left[r_{f,t} + \gamma_f \max_{a_{f,t+1} \in A_f} Q_f(s_{f,t+1}, a_{f,t+1}) \right] \quad (13)$$

Where $Q(s_{f,t}, a_{f,t})$ is the expected delivery cost obtained by freight carrier f in the state $s_{f,t}$ when freight carrier chooses an action $a_{f,t}$ in state $s_{f,t}$. The immediate reward is given by $r_{f,t}$ when action $a_{f,t}$ is taken from state $s_{f,t}$. The parameter γ_f is the discount rate for freight carrier f . The Maximum utility $Q(s_{f,t+1}, a_{f,t+1})$ means the minimum expected delivery cost received by freight carrier f in the next state $s_{f,t+1}$ for all possible actions $a_{f,t+1} \in A_f$.

As mentioned earlier, in this research, there are two possible actions for freight carrier f regarding goods delivery to its customers at time t ; Joint Delivery System (JDS) with UCC, and direct delivery system. Therefore, the immediate delivery cost $r_{f,t}$ that a freight carrier will receive will be expressed as equation (7) and equation (8) (similar to the ADP).

(5) Vehicle routing and scheduling problem with soft time windows (VRPSTW)

As stated in the previous section, the utility value (delivery cost) of freight carrier will be calculated using the VRPSTW. The VRPSTW is relevant to this problem because of the nature of city logistics problems that follow a course of visiting a number of the customer, meeting the collection or delivery time by a group of trucks that are operated by a company.

A directed graph $G = (V, A)$ represents the VRPSTW. The vertices set V consists of the depot vertex 0 and set of customers $C = \{1, 2, \dots, n\}$. The arc set A includes all feasible arcs $(i, j), i, j \in V$. Variable cost c_{ij} and time t_{ij} are linked with each arc $(i, j) \in A$. A set of vehicles (symbolized with K) with capacity (q) are located at the depot to service customer's demands. Demand d_i with $d_0 = 0$ is related with vertices V . A time window $[a_i, b_i]$ representing the earliest and the latest service time, while b_i' is the extension of latest service time (Fig. 6). c_l is the unit late arrival penalty cost and c_e is the unit early arrival penalty cost. However, based on a routing decision, the modified arc costs c'_{ijk} depends on the service time s_{jk} at customer $j \in C$ by a vehicle $k \in K$.

Fig. 6 Penalty cost function for the VRPSTW.

The c'_{ijk} is calculated by equation (14). The maximum limit of b_i' is formulated as equation (15).

$$c'_{ijk} = \begin{cases} c_{ij}, & \text{if } a_j \leq s'_{jk} \leq b_j \\ c_{ij} + c_l(s'_{jk} - b_j), & \text{if } b_j < s'_{jk} \leq b'_j \\ c_{ij} + c_e(a_j - s'_{jk}), & \text{if } a'_j \leq s'_{jk} < a_j \end{cases} \quad (14)$$

$$b'_i = \min \left[b_0 - t_{i0}, b_i + \frac{(c_{0i}, c_{i0})}{c_l} \right] \quad (15)$$

The VRPSTW can be formulated as following;

$$\min \sum_{k \in K} \sum_{(i,j) \in A} c'_{ijk} x_{ijk} \quad (16)$$

Subject to;

$$\sum_{k \in K} \sum_{j \in V} x_{ijk} = 1 \quad \forall i \in C \quad (17)$$

$$\sum_{i \in C} d_i \sum_{j \in V} x_{ijk} \leq q \quad \forall k \in K \quad (18)$$

$$\sum_{j \in V} x_{0jk} = 1 \quad \forall k \in K \quad (19)$$

$$\sum_{i \in V} x_{ihk} - \sum_{j \in V} x_{hjk} = 0 \quad \forall h \in C, \forall k \in K \quad (20)$$

$$\sum_{i \in V} x_{i0k} = 1 \quad \forall k \in K \quad (21)$$

$$a'_i \leq s'_{ik} \leq b'_i \quad \forall i \in V, \quad \forall k \in K \quad (22)$$

$$a_i \leq s_{ik} \leq b_i \quad \forall i \in V, \quad \forall k \in K \quad (23)$$

$$s_{ik} + t_{ij} - s_{jk} \leq (1 - x_{ijk}) M_{ijk} \quad \forall (i, j) \in A, \quad \forall k \in K \quad (24)$$

$$x_{ijk} \in \{0, 1\} \quad \forall (i, j) \in A, \quad \forall k \in K \quad (25)$$

Decision variable in the above formulation is x_{ijk} which represents whether the arc (i, j) is used ($x_{ijk} = 1$) or not ($x_{ijk} = 0$). The other decision variable is the arrival time s'_{jk} . The objective function minimizes the delivery cost (equation 16) and it is subjected to some constraints (equation 17 to 25) which ensure that all routes must start and end at the central depot, respecting the time windows and vehicle capacity. For more details on the VRPSTW formulation and solution algorithms, readers are referred to Qureshi, et al.¹²⁾.

4. CASE STUDY

We applied MAS-ADP model in a simulation for evaluating UCC. A square topology-based, hypothetical network is used for testing the simulations based on ADP and Q-learning model within MAS, as illustrated in **Fig. 7**. Four carriers (A, B, C, D), and 20 customers are involved in this network. UCC is a given city logistics policy by the government added to this hypothetical city. As the preliminary research, our simulation is started with the simple case, which uses freight carrier as the only learning agent.

In our simulation, the MAS-ADP models are iterated for 24 episodes (24 weeks), each include 5 weekdays from Monday to Friday as states. The agent will make decision as an action on every state (day) by considering the number of customer, location of the customer, parking cost, and demand received from their customer every day.

This research uses some assumptions as listed in **Table 1**. In order to test the capability of the two learning models, ADP and QL, we gradually increase the uncertainty level of the environment from low to high as listed in the following description;

1. Case 0 (base case)

The uncertainty level of the environment in case 0 is set to be low. It is assumed that the demand and the parking cost are changing in every day, but remains the same in data pattern for every episode. **Table 2** shows the demand data, which is different in every day, but the same pattern is iterated for every episode, i.e., the customers' demand for freight carrier A in case 0 is the same value on every Monday for all episodes (week 1 to 24). Similarly, the data of parking cost is shown in **Table 3**. In case 0, the number of customers remains 20 and their location is fixed as shown in **Fig. 7**. The base case helps to understand how ADP and Q-learning behave in the environment with high repetition of data.

Fig. 7 Test road network.

2. Case 1

In case 1, the uncertainty level of the environment is increased to moderate level by assuming that the demand and the parking cost for each freight carrier are fluctuating in every state as well as in every episode. An example of demand data for freight carrier A in case 1 is shown in **Table 4**; it can be observed that the customers' demand on Mondays of different weeks (episodes) are different. An example data of parking cost for freight carrier A is shown in **Table 5**. Number and location of the customers remains the same as in case 0.

3. Case 2

The city logistics environment is full of highly uncertain variables that might affect either the operational cost or the action selection. The uncertainty level of the environment in case 2 is high to represent the real conditions of the city logistics environment. All considered variables such as demand, parking cost, and number of customer are set to be different on every state and every episode. Therefore, case 2 also uses the demand data (**Table 4**) and parking cost data (**Table 5**), similar to case 1. As the number of customer is also set to be different in every state and every episode in case 2 (as shown in **Table 6**), **Fig. 8** shows some examples of the corresponding change in the customer's location as well.

Table 1 Simulation assumptions.

Item	Value
Working time	8 AM to 8 PM
Time window	60 minutes per customer
Capacity of the truck	200 parcels/ truck
Waiting charge (C_e) for early arrival	1 Yen/ minute
Penalty charge (C_l) for late arrival	5 Yen/ minute

Table 2 Total demand of customers (parcels) for freight carrier A in case 0.

		States (days)				
		Mon	Tue	Wed	Thu	Fri
Episodes (weeks)	1	768	539	650	482	602
	2	768	539	650	482	602
	⋮					
	24	768	539	650	482	602

Table 3 Total parking cost of customers (Yen) for freight carrier A in case 0.

		States (days)				
		Mon	Tue	Wed	Thu	Fri
Episodes (weeks)	1	390	420	420	430	440
	2	390	420	420	430	440
	⋮					
	24	390	420	420	430	440

Table 4 Total demand of customers (parcels) for freight carrier A in case 1 and case 2.

		States (days)				
		Mon	Tue	Wed	Thu	Fri
Episodes (weeks)	1	768	539	650	482	602
	2	649	486	492	459	460
	⋮					
	24	269	380	380	334	427

Table 5 Parking cost (Yen) for freight carrier A in case 1 and case 2.

		States (days)				
		Mon	Tue	Wed	Thu	Fri
Episodes (weeks)	1	390	420	420	430	440
	2	340	300	360	270	350
	⋮					
	24	270	260	250	220	340

Table 6 Number of customers for freight carrier A in case 2.

		States (days)				
		Mon	Tue	Wed	Thu	Fri
Episodes (weeks)	1	18	16	20	20	14
	2	17	17	17	13	17
	⋮					
	24	11	13	12	12	17

Fig. 8 Customer's location for freight carrier A in case 2.

5. RESULTS AND DISCUSSIONS

All simulations are done in MATLAB with the different settings of learning rate and discount factors for ADP and Q-learning models based on the results of a sensitivity analysis that has been done prior to the case study. The learning rate and the discount factor for ADP have been used as 0.2 and 0.6, respectively; whereas, the learning rate and the discount factor for Q-learning have been set as 0.2 and 0.8, respectively.

Fig. 9 Experienced delivery costs for all freight carriers (FC) in case 0.

In this study, we performed two separate simulations using ADP and Q-learning and compared the outcomes of both simulations using the following criteria; 1) cost saving, 2) accuracy and 3) stability and adaptability. The differences between these two simulations arise from the different action selection depending on the decision of action or behavior suggested by the learning model.

Equations (7) and (13) give the “expected cost” for each possible action in each state by the ADP and Q-learning algorithms. Based on this expected cost an agent chooses an action. “Cost experienced” by the freight carrier depends on this choice and may be given by equations (10) or (11). To calculate the cost saving, we compared the difference in the experienced cost for the agent under the ADP-based simulation and Q-learning-based simulation.

The second comparative criterion, “accuracy”, refers to the closeness of the gap between the expected costs obtained in the ADP and Q-learning based simulations to the corresponding costs experienced by the freight carrier. Smaller gap means more accurate method. Finally, adaptation enables an agent to make the right and the stable decisions by learning the new information from the environment.

To calculate the third comparative criterion, “stability”, we compared the number of changes in the selection of action suggested by ADP-based simulation and Q-learning based simulation. Fewer changes in action (policy) selection by an agent means better stability.

(1) Results in case 0 (base case)

The base case represents a low level of uncertainty of the environment, and used a high rate of data repetition for all variables.

Fig. 9 shows that both ADP and Q-learning result in the same experienced delivery costs in the simulation of case 0, but the percentage of the gap between the experienced costs and expected cost for ADP-based simulation (24.9%) is lower than the Q-learning (44.2%), on average. This means ADP can improve the simulation of city logistics by about 19 points as compared with Q-learning in case 0 as shown in **Fig. 10** and **Fig. 11**.

Fig. 10 The accuracy gaps between experienced costs and expected costs by ADP in case 0.

Fig. 11 The accuracy gaps between experienced costs and expected costs by QL in case 0.

Fig. 12 Experienced delivery costs (Yen) for all freight carriers (FC) in case 1.

Fig. 13 The accuracy gaps between experienced costs and expected costs by ADP in case 1.

Fig. 14 The accuracy gaps between experienced costs and expected costs by QL in case 1.

Fig. 15 Number of changes in the selection of action made by freight carrier (FC) in case 1.

Fig. 16 Variation of changes in action selection for freight carrier A performed by ADP and QL in case 1.

The gap between expected delivery costs with the experienced delivery costs for freight carrier A is only 17%, which is the lowest accuracy gap among agents modeled by ADP (**Fig. 10**). In the high repetition data and strong coupling on every episode, both ADP and Q-learning have the same reaction in the action selection resulting in the same experienced cost due to the stability of the environment. Therefore, in the simulation, the actions selection of both ADP and Q-learning in choosing action JDS with UCC and direct delivery (DD) was the same.

(2) Results in case 1

While demand and parking cost are changed randomly, a constant number of customers and fix locations of the customers are used in case 1. Therefore, the uncertainty level for case 1 is moderate. **Fig. 12** shows the experienced delivery costs by a freight carrier for ADP-based simulation is 0.99% lower than Q-learning based simulation, on the average. It means that using ADP as the freight carriers' behaviour learning model is better than using the Q-learning, as the former choice can save almost 1%, on average, of the total delivery costs; thereby improving the quality of the city logistics simulation.

The maximum cost savings were as much as 1.5%. On the average, percentage of the gap between expected costs with the experienced costs in the ADP-based simulation is also lower (30.7%) than the Q-learning based simulation (54.4%) as shown in **Fig. 13** and **Fig. 14**. The gap between expected delivery costs with the experienced delivery costs for freight carrier D was only 29%, the lowest accuracy gap among agents modelled by ADP (**Fig. 13**). The lowest accuracy gap generated by QL was 53% for freight carrier B and D (**Fig. 14**).

In the simulation of case 1, as shown in **Fig. 15**, the number of changes in actions (from direct delivery to JDS with UCC or vice versa) in ADP-based simulation is less than the Q-learning based simulation, which means that ADP is more adaptive to the changing environment by providing stable action selection behaviour for the modelled agent. On average, the number of changes in action selection

performed by ADP is 16% lower than the Q-learning based simulation.

In addition, **Fig. 16** shows the variation in action selection for freight carrier A in ADP and Q-learning based simulations. In this figure, number 0 shows a decision of direct delivery, whereas, 1 represents JDS with UCC. We can see from **Fig. 16** that both ADP and Q-learning guided the agent to different decisions on action selection. Q-learning based simulation is less stable in the pattern of action selection as it decisions vary a lot from choosing direct delivery to JDS with UCC or vice versa as compared to the ADP-based simulation. In case 1, ADP-based simulation resulted in a change of action 41 times out of 120 days, while Q-learning based simulation required change of action 53 times out of 120 days.

(3) Results in case 2

Case 2 represents close to the realistic city logistics environment, which uses the high uncertainty level of simulation data (no data repetition for any considered variable). This case was aimed to investigate the behaviour of ADP and Q-learning when agents need to interact in constantly changing environment.

Fig. 17 Experienced delivery cost (yen) for all freight carrier (FC) in case 2.

Fig. 18 The accuracy gaps between experienced costs and expected costs by ADP in case 2.

Fig. 19 The accuracy gaps between experienced costs and expected costs by QL in case 2.

Fig. 17 shows the experienced delivery costs for a freight carrier for ADP-based simulation is 1.5% lower than the Q-learning based simulation, in the average for all freight carriers. On the average, percentage of the gap between expected costs with the experienced costs by ADP-based simulation is lower (51.7%) than Q-learning (74.5%) as shown in **Fig. 18** and **Fig. 19**, which means ADP can improve the quality of simulation by 22.8 points in case 2. Both ADP and Q-learning based simulations have dropped in the accuracy criterion as compared with case 0 and case 1, as the uncertainty level is increased in case 2.

In the simulation of case 2, as shown in **Fig. 20**, the number of changed actions (from direct delivery to JDS or vice versa) in ADP is less than Q-learning, which means ADP is more adaptive to the changing environment by providing stable action selection. In case 2, ADP is 22% more stable in the actions selection compared with Q-learning in average for all freight carrier. The variation of changes in action selection for freight carrier A in ADP and Q-learning based simulations in case 2 are given in **Fig. 21**. In this figure also, number 0 shows a decision of direct delivery, whereas, 1 represents JDS with UCC. **Fig. 21** shows that both ADP and Q-learning based simulations result in the different decision on action selection. In case 2, out of 120 days, ADP suggest a change in action for 37 times as compared to 50 times suggested by Q-learning. It means that Q-learning based simulation is less stable in the pattern of action selection.

The evaluation results can be summarized as **Table. 7**.

Fig. 20 Number of changes in the selection of action made by freight carrier (FC) in case 2.

Fig. 21 Variation of changes in action selection for freight carrier A performed by ADP and QL in case 2.

Table 7 Evaluation results of ADP and Q-learning.

Case	Evaluation criteria								
	Total saving of delivery cost (Yen)			Accuracy gap to the real delivery cost (%)			Adaptability and Stability in action selection (number of changes in the selection of action)		
	ADP	QL	Cost difference	ADP	QL	Difference	ADP	QL	Difference
0	457836	457836	0	24.9%	44.2%	19.3%	0	0	0
1	482726	487552	0.9%	30.7%	54.4%	23.7%	41	53	12
2	360237	365804	1.5%	51.7%	74.5%	22.8%	37	50	13

6. CONCLUSIONS AND FUTURE WORK

This paper presented the general framework of the ADP models in evaluating the policy of UCC. It was found that ADP-based simulation performed better in all evaluation criteria (cost saving, accuracy, and stability and adaptability) as compared to Q-learning based simulation for the three cases. As stated in the introduction, a model should accommodate the agent's objective. It was observed that simulation using ADP resulted in 1.5% less experienced cost as compared to the simulation done using Q-learning. The freight carrier as a business company always try to minimize the cost so to maximize the profit. A 1.5% reduction in the cost is meaningful in the business strategy because it may increase the profit significantly. In highly uncertain environment, ADP-based simulation can reduce the accuracy gap by 22 points in terms of the difference between experienced cost and the expected total delivery cost compared with Q-learning based simulation. This can immensely improve the simulation quality for city logistics policy evaluations. Also, it was proved that the ADP is more adaptive to the changing environment and that it is more stable in the action selection by performing 22% less changes than Q-learning based simulation and still getting lower experienced costs. The adaptability and stability are important in decision making due to the need of choosing a stable and an efficient action.

The study of MAS-ADP model is still in its early stages, especially in city logistics area. As the general research framework illustrates, other variable of the environment setting (such as changing travel time) will be evaluated in the future. Similarly, ADP models will be developed for other stakeholders (UCC operator, customer, administrator, and residents) and interactions within the multi-agents environment will be done in the future. Simultaneous evaluation of multiple policies within MAS using ADP is also an interesting area to be investigated as

the future work.

REFERENCES

- 1) Taniguchi, E. and Tamagawa, D. : Evaluating city logistics measures considering the behavior of several stakeholders, *Journal of the Eastern Asia Society for Transportation Studies*, Vol. 6, pp. 3062–3076, 2005.
- 2) Tamagawa, D., Taniguchi, E. and Yamada, T. : Evaluating city logistics measures using a multi-agent model, *Procedia–Social and Behavioral Sciences*, Vol. 2, pp. 6002–6012, 2010.
- 3) Taniguchi, E. and Thompson, R. G. : *City Logistics I*, Institute of Systems Science Research, Kyoto, Japan, 1999.
- 4) Taniguchi, E. and Thompson, R. G. : *City logistics - Mapping the Future*, CRC Press Taylor and Francis Group, UK, 2015.
- 5) Teo, J. S. E., Taniguchi, E. and Qureshi, A. G. : Evaluation of load factor control and urban freight road pricing joint schemes with multi-agent systems learning models, *Procedia–Social and Behavioral Sciences*, Vol. 125, pp. 62–74, 2014.
- 6) Wangapisit, O., Taniguchi, E., Teo, J. S. E. and Qureshi, A. G. : Multi-agent systems modelling for evaluating joint delivery systems, *Procedia–Social and Behavior Sciences*, Vol. 125, pp. 472–483, 2014.
- 7) Wojtusiak, J., Warden, T., and Herzog, O. : Machine learning in agent-based stochastic simulation, *Computers and Mathematics with Applications*, <http://dx.doi.org/10.1016/j.camwa.2012.01.079>, 2012.
- 8) Sutton, R. S. and Barto, A. G. : *Reinforcement Learning*, Second edition, pp. 4, The MIT Press Cambridge, England, 2012.
- 9) Taniguchi, E., Thompson, R. G. and Yamada, T. : Incorporating risks in City Logistics, *Procedia–Social and Behavioral Sciences*, Vol. 2, No. 3, pp. 5899–5910, 2010.
- 10) van Duin, J. H. R., van Kolck, A., Anand, N. and Tavasszy, L. A. : Towards an agent-based modeling approach for the evaluation of dynamic usage of urban distribution centers, *Procedia–Social and Behavioral Sciences*, Vol. 39, pp. 333–348, 2012.
- 11) Teo, J. S. E., Taniguchi, E. and Qureshi, A. G. : Evaluation of distance-based and cordon-based urban freight road pricing in e-commerce environment with multiagent model, *Transportation Research Record*, No. 2269, pp. 127–134, 2012.
- 12) Qureshi, A. G., Taniguchi, E. and Yamada, T. : Exact solution for the vehicle routing problem with semi-soft time windows and its application, *Procedia–Social and Behavioral Sciences*, Vol. 2, pp. 5931–5943, 2010.
- 13) Taniguchi, E., Yamada, T. and Okamoto, M. : Multi-agent modeling for evaluating dynamic vehicle routing and scheduling systems, *Journal of the Eastern Asia Society for Transportation Studies*, Vol. 7, pp. 933–948, 2007.
- 14) Fagan, D. and Meier, R. : Dynamic multi-agent reinforcement learning for control optimization, *Fifth International Conference on Intelligent System, Modelling and Simulation*, pp. 99–104, 2014.
- 15) Zhang, H. G., Xin, Z., Yan, H. L. and Jun, Y. : An overview of research on adaptive dynamic programming, *Acta Automatica Sinica*, Vol. 39(4), pp. 303–311, 2013.
- 16) Huang, Z. and Balakrishnan, S. N. : Robust adaptive critic based neuro-controllers for missiles with model uncertainties, *Navigation Control Conference*, Canada, 2001.
- 17) Godfrey, G. and Powell, W. B. : An adaptive, dynamic programming algorithm for stochastic resource allocation problems I, *Transportation Science*, Vol. 36(1), pp. 21–39, 2002.

- 18) Novoa, C. and Storer, R. : An approximate dynamic programming approach for the vehicle routing problem with stochastic demands, *European Journal of Operational Research*, Vol. 196(2), pp. 509–515, 2009.
- 19) Bertsekas, D. P. and Shreve, S. E. : *Stochastic Optimal Control*, Athena Scientific, Belmont, 1996.
- 20) Liu, W. X., Venayagamoorthy, G. K. and Wunsch, D, C. II. : A heuristic-dynamic-programming-based power system stabilizer for a turbogenerator in a single machine power system, *IEEE Transactions on Industry Applications*, Vol. 41(5), pp. 1377–1385, 2005.
- 21) Venayagamoorthy, G. K., Harley, R. G. and Wunsch, D. C. : Dual heuristic programming excitation neuro control for generators in a multimachine power system, *IEEE Transactions on Industry Applications*, Vol. 39(2), pp. 382–394, 2003.
- 22) Karthikeyan, R., SheelaRani, B. and Renganathan, K. : An instant path planning algorithm for indoor mobile a robot using adaptive dynamic programming and reinforcement learning, *International Journal of Engineering and Technology*, Vol. 6, No. 2, pp. 1224–1231, 2014.
- 23) Hardin, G. : The tragedy of the commons, *American Association for the Advancement of Science*, Vol. 162, No. 3859, pp. 1243–1248, 1968.
- 24) Watkins, C. J. C. H. and Dayyan, P. : Q–Learning, *Machine Learning*, Vol. 8, pp. 279–292, 1992.

(Received February 24, 2017)