Evaluation of Urban Consolidation Centers for Sustainable City Logistics using Adaptive Dynamic Programming based Multi-Agent Simulation

Nailah FIRDAUSIYAH[1], Eiichi TANIGUCHI[2] and Ali Gul QURESHI[3]

[1]Member 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
[2]Fellow 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
[3]Member 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 evaluating the impacts of Urban Consolidation Centers (UCC) for sustainable city logistics using Adaptive Dynamic Programming (ADP) based multi-agent simulation. ADP based learning performs better in the accuracy, stability, and adaptability of the outcomes than other learning techniques when agents need to interact in constantly changing environment, such as city logistics. The ADP models for the freight carrier and UCC operator as the learning agents have been developed. Economic efficiency and environment friendliness criteria were used to evaluate the sustainability of UCC. The results proved that the implementation of UCC as a sustainable city logistics scheme is efficient in reducing 8% of the total delivery cost for freight carrier, and reducing 36% of the total emissions released to the environment. It is also showed that the use of learning agents is essential to demonstrate the successful implementation of the UCC, as it is only in the learning-based simulation, UCC operator could get a profit. Our simulation analysis also confirmed that as compared to widely used reinforced learning algorithms (Q-learning), ADP brings in the increased accuracy, stability and adaptability to the evaluations' results of UCC.

Key Words : *sustainability, city logistics, urban consolidation centre, adaptive dynamic programming, multi-agent system*

1. INTRODUCTION

Sustainable city logistics 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. 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 economy[1)]. The harmonization of economic efficiency and environmental friendliness in city logistics is essential for ensuring sustainable development in urban areas[2)], which faces two difficult problems. First is the efficiency of goods delivery within the uncertain environment (due to the parking issues, traffic congestion, and other restrictions in the urban area) that directly effects the operational cost as well as the action selection in presence of optional solutions or policies. 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 policy[3)]. These stakeholders also interact and influence each other in the city logistics environment, which makes the environment unpredictable. Therefore, the main challenge for the city logistics is to provide a sustainable urban freight transportation while considering multi-agent problems within the uncertain environment.

In order to achieve theses aims, numerous city logistics initiatives have been proposed and implemented in several cities, including the Urban Consolidation Centers (UCC)[4)]. It is important to evaluate

the city logistics policies before they can be effectively deployed due to their manifold implications on different city logistics stakeholders[5]. For that purpose, decision support tools (DST) are needed to help public decision makers and practitioners to make decisions, acceptable to all parties. These DSTs are mainly based on modeling, optimization, simulation, and evaluation procedures.

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[6,7,8]. However, based on previous research experiences, which will be described in more detail in the next section, it has been found that ADP based learning performs better in the accuracy of the outcomes when agents need to interact in uncertain environment, such as city logistics. Therefore, in order to have an accurate evaluation of the UCC, an ADP-based multi-agent simulation has also been developed, which can be used as a DST to achieve better outcomes in the decision process of designing and implementation of sustainable city logistics policies.

2. LITERATURE REVIEW

(1) Evaluation models for evaluating city logistics measures

Multi-agent systems (MAS) based on the reinforced learning (RL) algorithms have been used for evaluating the behavior of stakeholders, who are affected by the implementation of a city logistics policy. In MAS environment, multiple agents come together and interact, cooperate, coordinate, and negotiate with each other to reach their intended objectives. Various other city logistics policies have been evaluated using MAS with Q-learning such as load factor control and road pricing[6], e-Commerce[7], truck ban and motorway toll discounts[8]. These researches used Q-learning to model evolving behavior of the key stakeholders, namely the carriers, shippers, administrator, and residents relating to urban freight transport. The MAS with Q-learning algorithms have also been used to evaluate the dynamic usage of UCC[9]. The results indicated that the main policy measures that contribute to the successful functioning of the UCC are road pricing, operational subsidies and the application of time windows. Another study evaluated the Joint Delivery System (JDS) with parking restrictions using MAS with Q-learning[10] and the results showed that JDS with UCC and car parking management have the potential for improving environmental issues related with the urban freight. MAS with Monte Carlo Method was used by Taniguchi et. al.[11] 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 discounting motorway tolls in the urban motorway network will have a large environmental impact, resulting in 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. A comparative study conducted by Fagan and Meier[12], proved that ADP performs particularly well on the criteria of accuracy, adaptability and stability in multi-agent environment compared to other RL algorithms (i.e., Q-learning and Sarsa). 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.

(2) ADP for evaluating city logistics measures

Hardin[13], concluded that the learning and adaptation make the system more robust to imperfect knowledge of the environment. 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.,[14] 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[15]. ADP has been widely implemented at the confluence of control problem[16], intelligence traffic systems[12], and robotics[17]. 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. Therefore, in this study, we will develop and use the ADP based multi-agent simulation to evaluate the UCC for sustainable city logistics as illustrated in the general research framework (**Fig.1**).

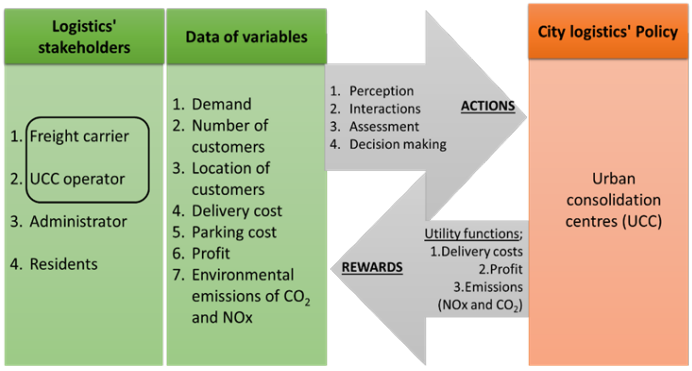

Fig. 1 General research framework

3. FRAMEWORK OF THE LEARNING PROCESS

Fig. 2 shows the learning process framework of the models developed in this research. The framework consists of two sub-models, which are; 1) learning model for stakeholders using ADP and Q-learning, and 2) the model for vehicle routing problem with soft-time window (VRPSTW). VRPSTW model calculates the delivery cost for each freight carrier and UCC operator. The two learning models (ADP and Q-learning) have the same function, i.e. to evaluate the behavior (action) of stakeholders by updating as well as learning based on the received (reward) value from the interaction with the environment (calculated by VRPSTW), and to take an action.

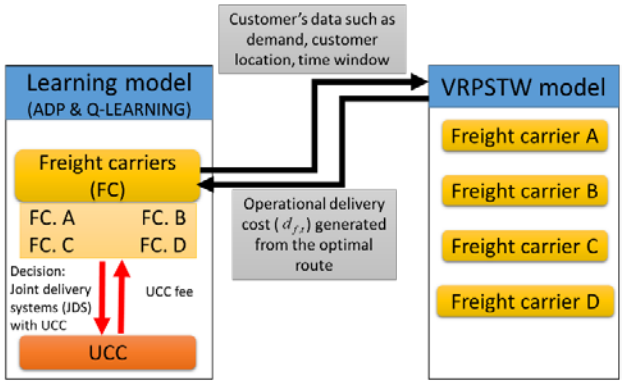

Fig. 2 Framework of the learning process

(1) MAS-ADP models for the learning agents

Recognizing agent's objective is important in modeling its learning behavior. Considering, freight carrier as a company that specializes in the last mile delivery of goods from depot to the customers, its objective is set to minimize the total cost of delivering goods to customers as equation (1), which is calculated by the VRPSTW. For more details on the VRPSTW formulation and solution algorithms, readers are referred to Qureshi, et al.[18].

$$Min \sum_{k \in K} \sum_{(i,j) \in A} c_{ij}' x_{ijk} \tag{1}$$

The UCC operator has been considered as a private or public company that consolidates and delivers the goods from the UCC to customers. Therefore, the UCC operator's objective is to maximize the profit, defined by equations (2) to (4).

$$Max\ E\lfloor B_u(UCCfee) \rfloor = UCCfee * E\left[D_f(UCCfee)\right] - E[d_u] \tag{2}$$

Subject to

$$0 \le UCCfee \le UCCfee_{max} \tag{3}$$

$$0 \le E[D(UCCfee)] \le D_{f,max} \tag{4}$$

Where $E\lfloor B(UCCfee) \rfloor$ is the expected profit of UCC operator based on the proposed price (*UCCfee*) by the UCC operator; $UCCfee_{max}$ is the maximum price that the UCC operator can propose, $E\left[D_f(UCCfee)\right]$ is the expected demand received from freight carrier *f*, $D_{f,max}$ is the maximum receivable demand from freight carrier *f*, and $E[d_u]$ is the expected delivery cost for the UCC operator *u* to deliver goods to the customer (calculated by VRPSTW model).

The MAS-ADP algorithm for updating the utility value function for freight carrier and UCC operator is formulated as in equation (5):

$$V_{la}(s_{la,t}) \leftarrow R_{la}(s_{la,t}, a_{la,t}) + \gamma_{la} \sum_{s_{t+1}} T\left(s_{la,t+1} \middle| s_{la,t}, a_{la,t}\right) V_{la}(s_{la,t+1}) \tag{5}$$

where $V_{la}(s_{f,t})$ is the expected delivery cost obtained by the learning agents *la* (freight carrier and UCC operator) when the agent chooses an action $a_{la,t}$ in the state $s_{la,t}$. $R_{la}(s_{la,t}, a_{la,t})$ is the expected reward when action $a_{la,t}$ is taken from state $s_{la,t}$. The parameter γ_{la} is the discount rate for the learning agent *la*, which is set to be a number between $0 < \gamma < 1$. A discount rate of 1 means that the agent will consider the long term reward, while 0 means that it only considers the current rewards. $V_{la}(s_{la,t+1})$ is the expected delivery cost received by the learning agents *la* in the next state $s_{la,t+1}$. The learning agents will update the expected reward $R_{la}(s_{la,t}, a_{la,t})$ and expected transfer probability $T_{la}(s_{la,t+1} | s_{la,t}, a_{la,t})$ using the equation (6) and equation (7) below,

$$T_{la}(s_{la,t}, a_{la,t}) \leftarrow T_{la}(s_{la,t}, a_{la,t}) + \alpha_{la}(t_{la}(s_{la,t}, a_{la,t}) - T_{la}(s_{la,t}, a_{la,t})) \tag{6}$$

$$R_{la}(s_{la,t}, a_{la,t}) \leftarrow R_{la}(s_{la,t}, a_{la,t}) + \alpha_{la}(r_{la}(s_{la,t}, a_{la,t}) - R_{la}(s_{la,t}, a_{la,t})) \tag{7}$$

Here, α is the learning rate of a learning 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 at all. $r_{la}(s_{la,t}, a_{la,t})$ is the immediate reward, and $t_{la}(s_{la,t}, a_{la,t})$ is the immediate transfer obtained by the learning agents *la* based on the possible actions $a_{la,t}$.

The first learning agent, freight carrier (i.e. $la = f$)

can choose two possible actions, viz., direct delivery (DD) or JDS, with corresponding immediate rewards given by equation (8) and equation (9), respectively.

$$r_f(s_{f,t}, a_{f,t}) = O_{f,t} + p_{f,t,k} \quad (8)$$

$$r_f(s_{f,t}, a_{f,t}) = UCCfee_{f,t} \quad (9)$$

where $O_{f,t}$ is the operational delivery cost (calculated by VRPSTW model) when freight carrier *f* decides to deliver goods directly to its customer (i.e. $a_{f,t}$ = DD) on time *t*, and $p_{f,t,k}$ is equal to $\sum_{i=C} p_i$ which is the total additional parking cost for a freight carrier f to be paid to serve customers $i \in C$. The set of customers is represented by *C*. The second possibility of immediate reward (equation (9)) that freight carrier can possibly receive is the consequence of choosing the joint delivery system with UCC (i.e. $a_{f,t}$ = JDS). It is obtained by multiplying the UCC fee offered by UCC operator with the total number of demand (parcels) that freight carrier gives to the UCC operator.

The second learning agent, the UCC operator (i.e. *la* = *u*) has three options of actions, which are price up, flat price, and price down. The immediate reward of the UCC operator will be converted to percentage profit and compared with the desired percentage profit margin which is assumed as 9% in this research. The UCC operator will make decision ($a_{u,t}$) based on the rules explained by equation 10:

$$a_{u,t} = \begin{cases} Price\ Down, if\ r_u(s_{u,t}, a_{u,t}) < 0\% \\ Price\ Up, if\ r_u(s_{u,t}, a_{u,t}) > 9\% \\ Flat\ Price, if\ 0\% \le r_u(s_{u,t}, a_{u,t}) \le 9\% \end{cases} \quad (10)$$

Depending on the selected action, the UCC operator will update the immediate profits as one of the equations (11) to (13).

$$r_u(s_{u,t}, a_{u,t}) = (UCCfeeUp_t * D_t) - d_u \quad (11)$$

$$r_u(s_{u,t}, a_{u,t}) = (UCCfeeFlat_t * D_t) - d_u \quad (12)$$

$$r_u(s_{u,t}, a_{u,t}) = (UCCfeeDown_t * D_t) - d_u \quad (13)$$

where $UCCfeeUp_t$, $UCCfeeFlat_t$ and $UCCfeeDown_t$ are the UCC prices offered by the UCC operator to freight carrier in state $s_{u,t}$, D_t is the demand received by the UCC operator in state $s_{u,t}$, d_u is the operational delivery cost for the UCC operator (calculated by VRPSTW model).

(2) MAS Q-learning models for the learning agents

As mentioned earlier, Q-learning-based MAS has been extensively used in the evaluation of the city logistics policies (including the UCC) and as one of the objectives of this research is to come up with a more accurate evaluation tool (i.e. the ADP-based MAS), we give a brief description of the Q-learning[19] before comparing the results of the two MAS models. The action-value function for each learning agent *la* is updated using equation (14) in Q-learning, which is equivalent to equation (5) in ADP.

$$Q(s_{la,t}, a_{la,t}) \leftarrow (1-\alpha_{la}) Q_{la}(s_{la,t}, a_{la,t}) + \alpha_{la}\left[r_{la,t} + \gamma \min_{a_{la,t+1} \in A_{la}} Q_{la}(s_{la,t+1}, a_{la,t+1}) \right] \quad (14)$$

where $Q(s_{la,t}, a_{la,t})$ is the expected delivery cost obtained by the learning agent *la* when it chooses an action $a_{la,t}$ in state $s_{la,t}$. The immediate reward is given by $r_{la,t}$ when action $a_{la,t}$ is taken in state $s_{la,t}$; $\min_{a_{la,t+1} \in A_{la}} Q_{la}(s_{la,t+1}, a_{la,t+1})$ is the minimum expected delivery cost received by the learning agent *la* in the next state $s_{la,t+1}$ for all possible actions. As mentioned earlier, in this research, there are two possible actions for freight carrier *f* (i.e. JDS with UCC, and DD). Therefore, the immediate delivery cost $r_{la,t}$ that a freight carrier will receive will be expressed as equation (8) and equation (9) (similar to the ADP). Similarly, the $r_{la,t}$ value for the UCC operator (*la* = *u*) will also be calculated using equations (11) to (13) (same as ADP). It can be emphasized here that the main difference between the ADP and the Q-learning is the update function equation (5) vs. equation (14).

(3) Environmental emissions model

Environmental emissions are considered as the negative effect of city logistics. In order to evaluate the benefit of using UCC for the environment, we calculated the carbon di-oxide (CO_2), oxides of nitrogen (NOx) and suspended particulate matter (SPM) produced by the trucks. These three environmental emissions are estimated using equation (15) to (17)[20].

$$CO_2 = l_{ij}(278.448 + 0.048059 v_{ij}^2 - 5.1227 v_{ij} + \frac{2347.1}{v_{ij}} \quad (15)$$

$$NO_x = l_{ij}(1.06116 + 0.000213 v_{ij}^2 - 0.0246 v_{ij} + \frac{16.258}{v_{ij}} \quad (16)$$

$$SPM = l_{ij}(0.03442 + 0.000039391 v_{ij}^2 -$$

$$0.0036777v_{ij} + \frac{1.2754}{v_{ij}} \quad (17)$$

where,
CO_2: expected carbon oxide emissions in grams
NO_x: expected nitrogen oxide emissions in grams
SPM: expected suspended particulate matter in grams
l_{ij}: length of road link between nodes *i* and *j* in kilometres
v_{ij}: speed of vehicle travelling on road link between nodes *i* and *j* in kilometres per hour

4. CASE STUDY

A square topology-based, hypothetical network (**Fig. 3**) is used for evaluation of the UCC based on the simulations using ADP and Q-learning models within MAS. Four carriers (A, B, C, D), one UCC, and 20 customers are involved in this network. In our simulation, the MAS models are iterated for 24 episodes (24 weeks), each include 5 weekdays from Monday to Friday as states. The agents will make decisions as actions in every state (day) by considering the fluctuating parking cost, UCC fees, and demand received from their customer every day.

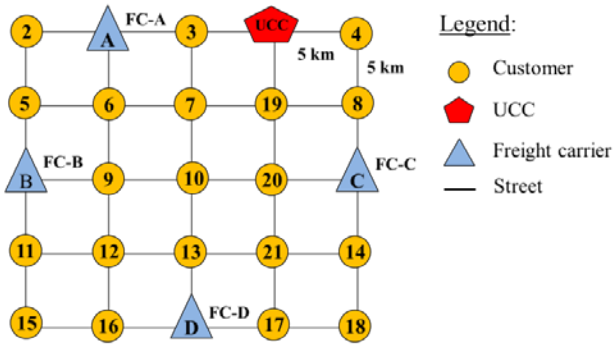

Fig. 3 Test road network

This research uses some assumptions as listed in **Table 1**;

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	I Yen/ minute
Penalty charge (c_l) for late arrival	5 Yen/ minute

5. RESULTS AND DISCUSSIONS

All simulations are done using MATLAB. 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. These values are based on the results of a sensitivity analysis that has been done prior to the case study. In this study, we performed two separate simulations using ADP and Q-learning to evaluate the impacts of UCC for sustainable city logistics using the following criteria; 1) economics' efficiency for each learning agent, i.e., cost saving for the freight carrier, and profitability for the UCC operator; 2) environmentally friendliness. This study also evaluated the accuracy, stability and adaptability of the outcomes of both simulations on evaluating the UCC within multi-agent and uncertain environment. The differences between these two simulations arise from the different action selection depending suggested by the learning model.

(1) Accuracy of the learning models

Accuracy of the outcomes obtained in the ADP and Q-learning is important for the learning models to evaluate a sustainable city logistics scheme (such as UCC in this study). The "accuracy", refers to the closeness of the gap between the expected value obtained in the ADP and Q-learning based simulations to the corresponding value experienced by the learning agent. For example, for the freight carrier (as learning agent), equations (5) and (14) give the "expected cost" for each possible action (JDS or DD) in each state by the ADP and Q-learning algorithms, respectively. Based on this expected cost the agent chooses an action. "Cost experienced" by the freight carrier depends on this choice and is given by equation (8) or (9). Smaller gap between these two costs means more accurate method.

Fig. 4 shows that the percentage gap between expected cost and the experienced cost in the ADP-based simulation is lower (39.6%) than the Q-learning based simulation (47.7%) for freight carrier A; similar pattern were obtained for freight carriers B, C, and D. Similarly, in case of the UCC operator, the percentage gap between the expected profit (equations (5) and (14)) and the experienced profits (equations (11) to (13)) in the ADP-based simulation is also lower (46.4%) than the Q-learning based simulation (51.9%) as shown in **Fig. 5**. It proves that the ADP-based learning can improve the accuracy of the simulation, thereby improving the quality of simulation.

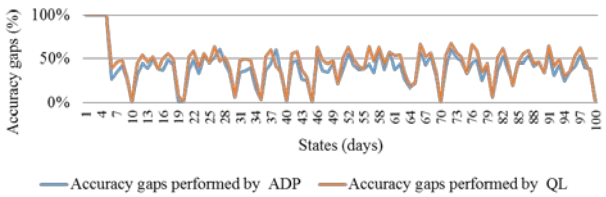

Fig. 4 Accuracy gaps (%) in case of freight carrier A

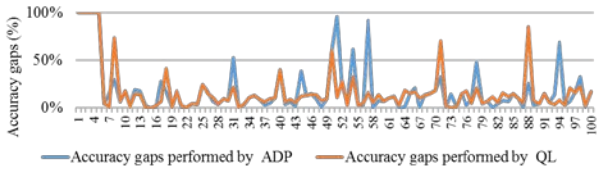

Fig. 5 Accuracy gaps (%) in case of UCC operator

(2) Stability and adaptability of the learning models

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.

In the simulation, as shown in **Fig. 6**, the number of changed actions (from direct delivery to JDS or vice versa) in ADP is less than Q-learning especially in case freight carrier A and B, which means ADP is more adaptive to the changing environment by providing stable action selection. In case freight carrier A and B, ADP is 7.5% more stable in the actions selection compared with Q-learning in average, while in case freight carrier C and D, both ADP and Q-learning have the same reaction in the number of changed actions. 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. In addition, **Fig.7** shows the variation in action selection, i.e., 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.7** that both ADP and Q-learning guided the freight carrier A 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 freight carrier A, ADP-based simulation resulted in a change of action 41 times out of 120 days, while Q-learning based simulation required change of action 45 times out of 120 days.

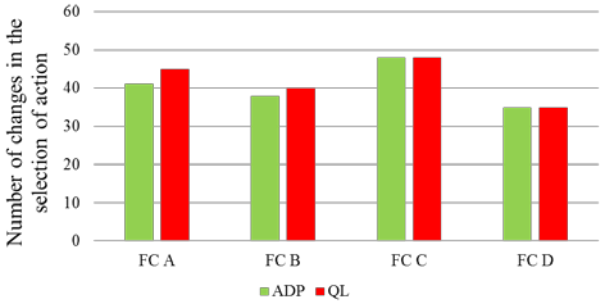

Fig. 6 number of changes in action selection for freight carriers

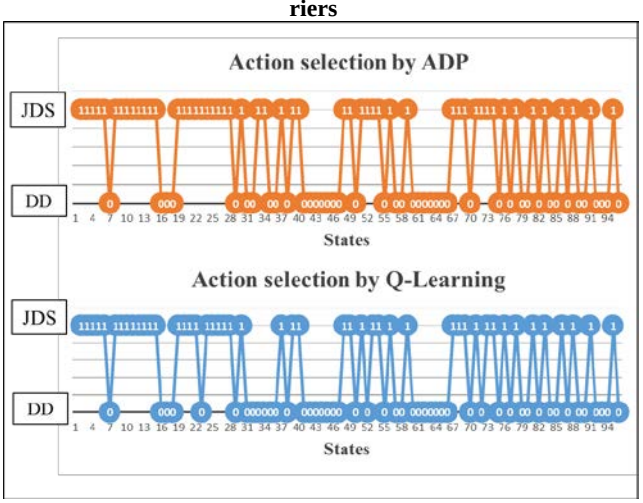

Fig. 7 variation in the selection of action for freight A

In the simulation result, as shown in **Fig.8**, the number of changes in actions (from increasing the UCC fee (price up) to decreasing the UCC fee (price down) and using flat price or vice versa) in ADP-based simulation and the Q-learning based simulation is almost same. In case of UCC operator, ADP-based simulation resulted in a change of action 37 times out of 120 days, while Q-learning based simulation required less change of action 36 times out of 120 days. Both ADP and Q-learning guided the UCC operator to different decisions on action of UCC fee selection. The UCC fee suggested by the ADP is always 5% lower in average than UCC fee suggested by Q-learning (**Fig. 12**). It makes the possibility of choosing JDS with UCC by freight carrier is higher that direct delivery. Using the decision of ADP, the UCC operator will get more profits as well as reduce the emissions released to the environment.

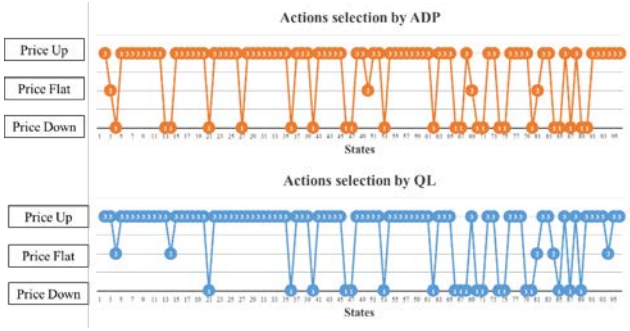

FIGURE 8 variation in the selection of action for UCC operator

(3) Sustainability criteria: economics' efficiency

a) Freight carrier

The meaning of efficiency for freight carrier is delivery of goods at lower cost to the customers. To calculate the efficiency for freight carrier, we compared the difference in the experienced cost for the freight carrier in the existence of UCC and without UCC. The existence of UCC provides more alternatives of goods' delivery for freight carrier, which are direct delivery, and JDS with UCC. The ADP and Q-learning as the learning models will suggest the actions based on the reward received from the environment.

Both ADP and Q-learning based simulation resulted in the lower experienced delivery costs for freight carriers in the case of UCC than the without UCC case (**Fig. 9**). The delivery cost with UCC resulted from ADP based simulation is 8.4% lower on average as compared to the experienced delivery cost without UCC. Corresponding figures for the Q-learning based simulation was 6.7%. **Fig. 10** illustrates more details with the cumulative experienced delivery cost for a freight along the simulation. It means that implementing the UCC as a sustainable city logistics policy is efficient to minimize the delivery costs for a freight carrier. Moreover, using ADP as the freight carriers' behavior learning model is better than using the Q-learning, as the former choice can further save almost 1.7%, on average, of the total delivery costs; thereby improving the quality of the city logistics simulation i.e. ADP-based simulation provides more favorable comparison with the no-policy (without UCC) case.

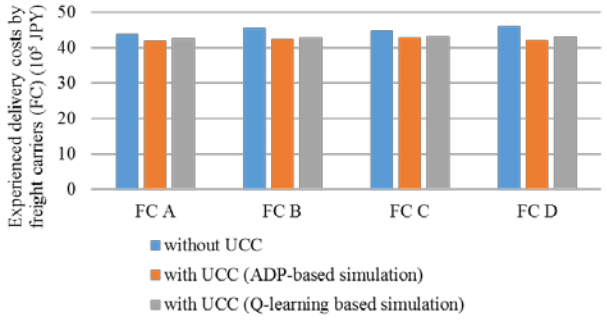

Fig. 9 Experienced delivery costs by freight carriers (FC) with UCC and without UCC

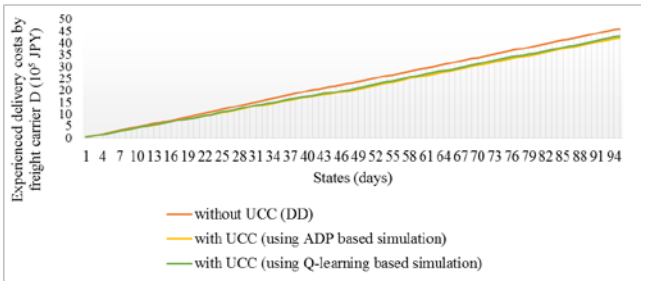

Fig. 10 Simulation results of cumulative experienced delivery cost for the freight carrier D

(b) UCC operator

The economic efficiency for UCC operator means higher profits at lower UCC fee offered to freight carriers to foster more demand. To calculate the profitability, we compared the difference in the experienced profits for UCC operator under the learning environment with freight carrier and without learning. In the learning environment, we assumed the UCC operator will update the UCC fee every day by learning the reward received (profits) based on the business provided by the freight carriers. Without learning (α=0), the UCC operator will not consider the current information from the environment. Therefore, the UCC operator is assumed to offer fixed UCC fee (150 JPY/parcel) to the freight carrier every day. As mentioned earlier, equations (5) and (14) give the "expected profit" for each possible action in each state by the ADP and Q-learning algorithms, respectively. Based on this expected profit, UCC operator chooses an action. "Profit experienced" by the UCC operator depends on this choice and is given by equations (11) to (13).

Fig. 11 clearly shows that the UCC would fail dramatically if a fixed UCC fee policy is followed without learning from the environment, as the cumulative experienced profit received by the UCC operator without learning is way lower than the cumulative experienced profits resulted from learning using either of ADP or Q-learning. A dip in the cumulative experienced profit curve shows a negative profit obtained in that episode. The negative profit means the UCC failed to cover the downstream delivery cost based on the business (demand) received from the freight carriers. If the UCC operator is modelled as a learning agent, it learns from these negative reward values to adjust the UCC fee (may be to attract more demand) and becomes profitable again. It shows the importance of the MAS-based simulations in the evaluation of the city logistics policies where one stakeholder's the action/behavior can seriously impact the other. It is important for the UCC operator to learn from the behavior of the freight carriers (refuse to

join UCC due to high fees) to become profitable.

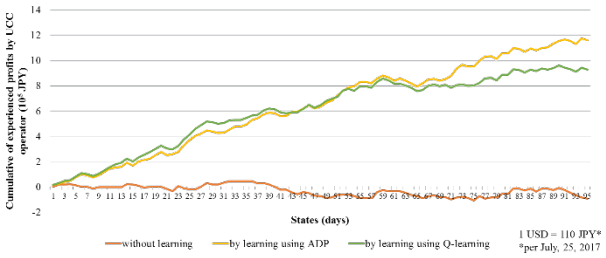

Fig. 11 experienced profits without learning by UCC operator

Moreover, **Fig. 11** shows that using ADP as the UCC operator's behavior learning model is better than the Q-learning. The two simulations suggest different actions of managing the UCC fee level to the UCC operator, which results in the difference of the experienced profits. The UCC fee suggested by the ADP is always 5% lower, on average, than the UCC fee suggested by Q-learning (**Fig. 12**). It increases the possibility of choosing JDS with UCC by freight carrier than the direct delivery under the ADP-based learning. The impact is also evident in the profits, which were 3.7% for ADP as compared to 2.1% for Q-learning (**Fig. 13**).

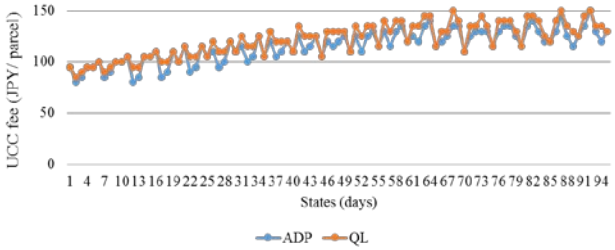

Fig. 12 UCC fee per parcel offered by UCC operator

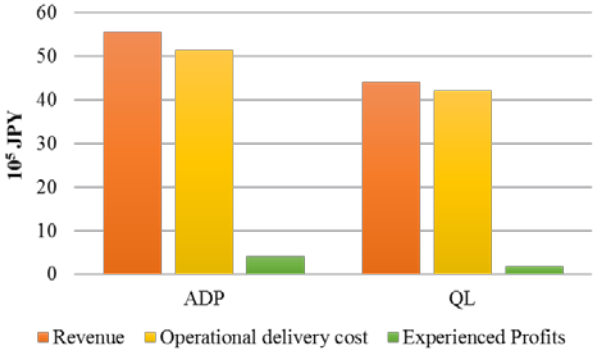

Fig. 13 experienced profits received by UCC operator

(4) Sustainability criteria: environmentally friendliness

We evaluated the impacts of UCC by calculating the total emissions (CO_2, NOx and SPM) from the delivery activities made by freight carriers and the UCC operator with and without UCC using ADP and Q-learning. The existence of UCC will reduce 36% (averaged from ADP and Q-learning results) of total emissions as compared to the condition without UCC. As explained earlier, the differences of the result between these two simulations arise from the different action selection suggested by the learning model. The experienced emission level of CO_2 (**Fig.14**), NO_x (**Fig. 15**), and SMP (**Fig. 16**), obtained under ADP-based simulation is 7.8% lower than the Q-learning based simulation. It means that using ADP as the learning model for both agents is better than using the Q-learning, as the former choice can reduce almost 8%, of the total emissions released to the environment; thereby improving the quality of the city logistics simulation, i.e. ADP-based simulation provides better comparison with the no-policy (without UCC) case.

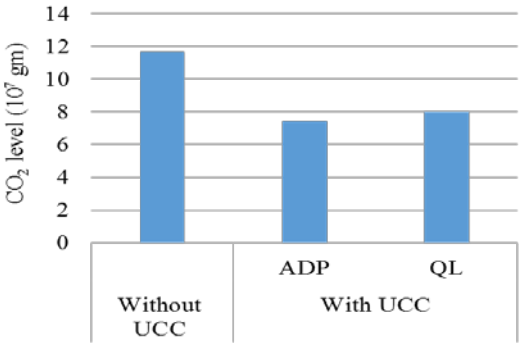

Fig. 14 CO_2 emissions

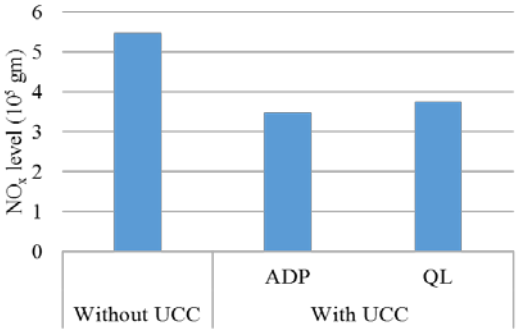

Fig. 15 NO_x emissions

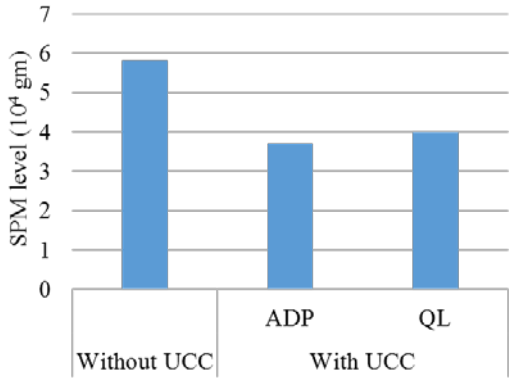

Fig. 16 SPM emissions

5. CONCLUSIONS AND FUTURE WORK

This paper developed the ADP models for evaluating the UCC as a sustainable city logistics policy. Economic efficiency and environmentally friendliness criteria were used to evaluate the sustainability of UCC. The results showed that the implementation of UCC as a sustainable city logistics scheme is efficient by reducing 8% of the total delivery cost for the freight carrier, and reducing 34% of the total emissions released to the environment. It is also showed that the use of learning agents is essential to demonstrate the successful implementation of the UCC, as it is only in the learning-based simulation, UCC operator could get a profit.

In addition, simulations should accommodate the agent's objective. It was observed that simulation using ADP resulted in a further 1.7% less experienced cost as compared to the simulation done using Q-learning. In case of UCC operator, the ADP satisfy its objective by getting higher profits than the Q-learning based simulation. The differences of the results between these two simulations arise from the different action selection suggested by the learning model. Therefore, the accuracy, stability and adaptability of the outcomes is also very important for the learning models, especially within the uncertain environment. It was found that ADP-based simulation improved the accuracy, stability and adaptability of the expected delivery costs for the freight carrier as compared to the Q-learning based simulation.

As shown in the general research framework (Figure 1), other city logistics stakeholders (such as customer, administrator, and residents) will also be considered and modelled using ADP in the future. The model application to the real life applications of UCC (such as in Motomachi, Japan) is planned to be done in the future in order to verify the model application.

REFERENCES

1) Taniguchi, E., and Thompson, R. G. City Logistics I. Institute of Systems Science Research. Kyoto., Japan, 1999.
2) Taniguchi, E., Thompson, R. G., Yamada, T. Logistics systems for sustainable cities: Visions for city logistics. Elsevier Ltd., UK, 2004.
3) Taniguchi, E. & Tamagawa, D. Evaluating city logistics measures considering the behavior of several stakeholders. Journal of the Eastern Asia Society for Transportation Studies, Volume 6, pp. 3062-3076, 2005
4) Browne, M., Allen, J., & Leonardi, J. Evaluating the use of an urban consolidation centre and electric vehicles in central London. IATSS Research 35. Page 1-6, 2011
5) Taniguchi, E., and Thompson, Russell G. City Logistics: Mapping the Future. CRC Press Taylor and Francis Group., UK, 2015.
6) Teo, J. S. E., Taniguchi, E. & 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.
7) 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. In Transportation Research Record: Journal of the Transportation Research Board, No. 2269, Transportation Research Board of the National Academies, Washington, D.C., pp. 127–134, 2012
8) 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.
9) 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, Volume 39, pp. 333-348, 2012.
10) Wangapisit, O., Taniguchi, E., Teo, J. S. E. & Qureshi, A. G. Multi-Agent systems modelling for Evaluating Joint Delivery Systems. Procedia-Social and Behavior Sciences, Volume 125C, pp. 472-483, 2014.
11) Taniguchi, E., Yamada, T., & Okamoto, M. Multi-agent modeling for evaluating dynamic vehicle routing and scheduling systems. Journal of the Eastern Asia Society for Transportation Studies, Volume 7, pp. 933–948, 2007.
12) Fagan, D., and Meier, R. Dynamic Multi-Agent Reinforcement Learning for Control Optimization. Fifth International Conference on Intelligent System, Modelling and Simulation, 2014.
13) Hardin, G. The Tragedy of the Commons. American Association for the Advancement of Science, Vol. 162, No. 3859, pp. 1243-1248, 1968
14) Zhang, Hua-Guang, Xin Zhang, Yan-Hong Luo, and Jun Yang. An overview of research on adaptive dynamic programming. Acta Automatica Sinica, volume 39(4), pp. 303−311, 2013
15) Godfrey, G. & Powell, W. B. An adaptive, dynamic programming algorithm for stochastic resource allocation problems I: Single period travel times. Transportation Science, volume 36(1), pp. 21–39, 2002.
16) Venayagamoorthy G K, Harley R G, Wunsch D C. Dual heuristic programming excitation neuro control for generators in a multimachine power system. IEEE Transactions on Industry Applications, volume 39(2), pp. 382−394, 2003.
17) 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, volume 6, no.2, pp. 1224-1231, 2014.
18) Qureshi, A.G., Taniguchi, E., and Yamada, T. Hybrid insertion heuristics for vehicle routing problem with soft time windows. Journal of the Eastern Asia Society for Transportation Studies, Volume 8, pp. 827–841, 2010.
19) Watkins, C.J.C.H and Dayyan, P. Q.Learning. Machine Learning, 8, pp. 279-292, 1992
20) NILIM. Qualitative appraisal index calculations used for basic unit for computation of CO2, NOx and SPM (in Japanese), 2003

(Received July 31, 2017)