Computers & Industrial Engineering 160 (2021) 107596

Contents lists available at ScienceDirect

Computers & Industrial Engineering

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

Reinforcement learning enabled dynamic bidding strategy for instant delivery trading

Chaojie Guo [a], Russell G. Thompson [a], Greg Foliente [a], Xiaoshuai Peng [b,*]

[a] Department of Infrastructure Engineering, The University of Melbourne, Parkville, Victoria 3010, Australia
[b] Department of Operation and Logistics Management, School of Business Administration, Northeastern University, Shenyang 110167, China

ARTICLE INFO

Keywords:
Instant delivery
Sequential auctions
Reinforcement Learning
Bidding strategies

ABSTRACT

Due to the great potential to enable collaboration and improve consolidation, auctions have been identified as a possible effective option to improve the efficiency of instant delivery. Instant delivery markets are complex and dynamic systems influenced by highly random demand. Conventional bidding strategies require perfect market information and cannot be adjusted effectively according to the evolution of requests. To address this problem, this paper proposes an auction-based trading platform to enable freight transportation procurement and develops a Reinforcement Learning (RL) enabled dynamic bidding strategy to optimize carrier's behavior in sequential auctions. In the RL enabled dynamic bidding strategy, three RL algorithms, including Q-learning, Deep Q Network and experience replay based Q-learning are used to improve carrier's bidding ability. The simulation results demonstrate that compared with the conventional bidding strategy, the RL enabled dynamic bidding strategies with any of the three RL algorithms can help carrier secure more auctions and gain more profit in a competitive marketplace. In addition, the advantages of the RL enabled dynamic bidding strategies are more obvious and the performance is more stable in more uncertain market environments.

1. Introduction

The boom of app-based services is reshaping freight transportation in urban areas. Instant delivery is a fast-growing segment in marketplaces which emphasizes the limited timeframe of freight delivery requests (Dablanc et al., 2017). More and more retailers offer online services and promise delivery in hours to facilitate people's life within urban areas (Lafkihi et al., 2019). For the operation of instant delivery, direct pickup and drop off is the main solution to meet the requirement of fast response. However, such solutions result in additional operational costs due to the failure to achieve economies of scale (Zhou and Lin, 2019). Currently, the challenge for logistics companies who carry out instant delivery is to improve the utilization of their facilities without violating service levels. Consolidating networks with multiple pickups before drop off's is considered a good strategy for saving operating costs (Hong et al., 2019). In addition, collaborative freight transportation systems, which are able to reduce operational costs significantly for the entire system (Thompson and Hassall, 2012), provide a promising alternative to optimize instant delivery markets in urban areas. The emerging concept of the Physical Internet is encouraging more consolidative and collaborative hyperconnected schemes for city logistics to improve transportation efficiency (Crainic and Montreuil, 2016).

Due to time sensitivity as well as the demand growth of instant delivery, freight transportation is fluctuating much greater than ever before in urban areas. In a dynamic transportation marketplace, auctions have been suggested as an effective option to enable efficient trading (Handoko et al., 2014), achieve collaboration (Guo et al., 2021; Karaenke et al., 2019) and improve consolidation and optimization opportunities for freight transportation (Figliozzi et al., 2005). Auction mechanisms for freight transportation trading have been well studied in the literature. An extensive survey can be found in Xu et al. (2018). However, previous work is more focused on designing auction rules for risk-neutral participants or investigating optimal bidding strategies for carriers in one-time auction contexts. In the uncertain environment of instant delivery, demand from the marketplace arrives stochastically over time. Therefore, auctions are performed sequentially over time. In addition, carriers face two complicated problems in auction-based collaboration. The first is to minimize operational costs, while the second is to maximize profit (Figliozzi et al., 2005). Cost minimization relates to optimizing fleets routing and scheduling problems which have

* Corresponding author.
E-mail addresses: chaojieg1@student.unimelb.edu.au (C. Guo), rgthom@unimelb.edu.au (R.G. Thompson), greg.foliente@unimelb.edu.au (G. Foliente), xspeng912@163.com (X. Peng).

https://doi.org/10.1016/j.cie.2021.107596
Received 4 November 2020; Received in revised form 19 July 2021; Accepted 29 July 2021
Available online 2 August 2021

been widely studied in the literature. However, very limited studies have investigated flexible and dynamic bidding strategies to enable carriers to optimize their offered bids based on interaction with the environment, therefore, to maximize their profits in sequential auctions. Moreover, in highly dynamic instant delivery marketplaces, carriers need to effectively complete the transportation task bidding according to the information that is collected via real-time interaction with the environment. RL can enable actions to be modified based on interactions with the environment (Mendel and McLaren, 1994). Therefore, it is able to help carriers learn and adapt to the changing instant delivery marketplace for formulating a bidding strategy which can maximize their profits. Although, RL has shown great potential in learning the optimal bidding strategies in auction markets (Cheng et al., 2019; Zhang et al., 2020), there is a lack of studies that have investigated the performance of RL in auctions for instance delivery trading. How RL can dynamically improve carriers' bidding strategies in on-line auctions when competing with other carriers is still not well understood.

Based on the issues identified above, this paper considers an auction-based instant delivery trading platform (AITP) to support instant delivery planning in the stochastic and dynamic competitive marketplace. In the AITP, shippers first offer requests. Then depending on the incremental service costs which are obtained by minimizing their operational costs, carriers compete for these requests to maximize their profits through sequential on-line auctions over a rolling horizon. In order to optimize carrier's behavior in sequential auctions, three RL based dynamic bidding strategy are designed. The contributions of this paper can be summarized below. First, a new dynamic bidding based framework is proposed to optimize the instant delivery, which is capable of tackling two complicated problems, i.e., minimizing operational costs and maximizing profit involved in the sequential auctions. Second, different RL algorithms, including Q-learning, Deep Q Network (DQN) and an experience replay based Q-learning, are evaluated for designing dynamic bidding strategies to improve carrier's bidding ability. To the best of our knowledge, we are the first to introduce these three RL algorithms into instance delivery trading auctions and explore the performance of different RL algorithms in dynamically improve carriers' bidding strategy in a competitive market. Last but not least, the viability of the approach is demonstrated through extensive simulation experiments, and the value of RL based dynamic bidding strategy is established, i.e., the three RL algorithms achieve substantial improvement regarding the learning carriers' winning times and profits gained compared with the conventional bidding strategy.

The rest of this article is organized as follows. Section 2 reviews the relevant literature. We describe the framework of AITP in Section 3. In Section 4, three types of RL algorithms are introduced. Simulation results and discussion are presented in Section 5. Our conclusions and future research directions are given in Section 6.

2. Literature review

2.1. Auction-enabled freight transportation request trading

Driven by the rapid growth of ecommerce marketplaces, instant delivery is developing rapidly. Researchers have studied the instant delivery problem for different scenarios, such as order picking and delivery planning (Zhang et al., 2019), crowdsourcing-based system design (Du et al., 2019) and vehicle-UAV operation schemes (Gu et al., 2020). Recently, to further improve the efficiency of instant delivery, auctions have been suggested as a promising alternative. Auction-enabled freight transportation request trading mechanisms have been thoroughly studied to optimize vehicle utilization and reduce excessive costs per delivery (Xu et al., 2017). Auction-enabled trading frameworks for transportation can be classified into two types: (1) request trading between shippers and carriers, and (2) request trading among carriers to reallocate requests. This article is focused on auction-enabled trading between shippers and carriers. Figliozzi et al. (2003a) proposed an auction-based framework for procurement of truckload transportation services in a dynamic context. Based on this framework, extended problems have been discussed including optimal assignment strategies using sequential auctions (Figliozzi et al., 2004) and optimal bidding strategies by evaluating opportunity costs (Figliozzi et al., 2006). Mes et al. (2007) claim that a well-designed auction-enabled management approach for scheduling transportation requests performs better than traditional optimisation methods. A dynamic threshold policy for shippers in transportation procurement auction has been investigated by Mes et al. (2009). The policy aimed at reducing shipper's cost for transportation procurement was shown to achieve a 20–30% cost saving per request. Then, Mes et al. (2013) extended the work of Mes et al. (2009) to the scenario where both shippers and bidding carriers adopt profit maximizing strategies in auctions. The simulation showed that transportation costs can be reduced by 10–20% for the entire system. Triki et al. (2014) investigated an auction-enabled full truckload transportation request trading problem focusing on combinatorial auctions, where bid generation and pricing problems were integrated into a probabilistic optimization model and heuristic approaches were developed to solve the model.

From the above discussion, we find that most of the auction-enabled freight transportation request trading research has been based on optimization which requires more accurate and complete market information. However, due to the randomly fluctuation of instant delivery demand, shippers' requests present a dynamic unstable evolution, the chance to obtain accurate and complete information about shippers' requests (e.g., freight type, volume, pickup and drop off locations as well as time windows) converges to zero. To address this problem, this paper introduces the RL method and uses adaptive nature of RL to help carriers adapt to the evolution of requests and dynamically improve their bidding strategy.

2.2. Reinforcement learning based strategies for freight transportation

As a method of artificial intelligence, RL refers to an actor/agent who interacts with an uncertain environment and learns to optimize its behavior by evaluation from experience (Sutton and Barto, 2018). Due to unpredictable demand, operational costs and multiple stakeholders, the environment of freight transportation is fluctuating. In this context, Firdausiyah et al. (2019) developed a multi-agent based model to simulate city logistics operations, and used RL methods to optimize the decision-making processes of freight carriers and the operator of an urban consolidation center. Similar RL based decision support strategies for city logistics involving interactions between freight carriers and an urban consolidation center were investigated by Firdausiyah et al. (2017) and Firdausiyah et al. (2020). In contrast to these studies, this paper focuses on RL based dynamic bidding strategy of freight carriers involved in the competitive instant delivery trading marketplace. Another application of RL in freight transportation was the road pricing problem (Tamagawa et al., 2010; Teo et al., 2012; Teo et al., 2014). To the best of our knowledge, there has been limited studies that have investigated the application of RL in auction-based freight transportation trading problems. Wang et al. (2018) proposed a centralized truckload request matching framework where the brokerage offers a price for a request, then both the shipper and carrier have to make decision on whether accept the quote. The brokerage was a learning agent who learns from the responses of shippers and carriers for proposals. In this paper, we consider an AITP in which shippers offer requests and carriers compete for these requests by submitting bids independently. In this system, a carrier is a learning agent who can learn from their behavior to formulate the optimal bidding strategy. Figliozzi et al. (2003b) and Figliozzi et al. (2005) considered the procurement of truckload services through sequential auctions in a decentralized context. Although the competitive freight carriers can learn from their behavior to adjust bidding strategies and to improve profit in the two studies, they only considered one RL algorithm. In this paper, except for

Q-learning and experience replay based Q-learning algorithms, we use DQN to represent the decision making of agents, which can handle more complex and changeable environments and is more suitable for the stochastic and dynamic instant delivery marketplace. In addition, Figliozzi et al. (2003b) and Figliozzi et al. (2005) assumed that the RL based dynamic biding strategy was only applied to one carrier and other competitive carriers behave in a conventional way. In this paper, we extend their work by allowing multiple carriers to learn simultaneously from their behavior to improve their optimal bidding strategies.

3. Problem statement

The primary focus of this study is to investigate the dynamic biding strategies for competitive carriers in the AITP. To do so, we consider an instant delivery marketplace within a metropolitan area where shippers (e.g., grocery, pharmacy, restaurants and individuals) procure transportation services from competitive carriers to delivery their freight. All deliveries must be finished within a limited period on the same day. In order to enable real-time trading for instant delivery services an AITP which operates dynamically based mainly on price is considered. In the AITP, multiple shippers publish their requests with corresponding attributes including type of freight (e.g., food, document and medicine), pickup and drop off locations and time windows. Time windows are described by the earliest pickup time and latest drop off time (since the timeframe of instant delivery is very limited, we ignore the latest pickup time and earliest drop off time). Each published request triggers one auction and multiple carriers participate in the auction to compete for the request (bidding) within a decentralized management system. The AITP determines the winning carrier. We assume that auctions are performed one at a time to simply the model (Figliozzi et al., 2005).

In the AITP, requests arrive sequentially and dynamically over time. Each published request at epoch t is denoted as u_t. Let $e, e\in \mathrm{E}$ denote the set of competitive carriers. Each carrier e operates a set of fleets f_e for fulfilling real time requests. When a new request u_t is observed from the AITP, each carrier submits a bid with a monetary amount of $b^e_{u_t}$ (every carrier must bid). The AITP determines who is the winner to perform the shipment. The shipper pays the winning carrier a service fee of S_{u_t}. Let $c^e_{u_t}$ denote the incremental cost for serving the new request of carrier e. The profit gained by the winning carrier for fulfilling request u_t is $S_{u_t} - c^e_{u_t}$. The bid price is usually determined based on the incremental cost, here, we assume each carrier determines their bid price following a two-stage pricing mechanism (Berger and Bierwirth, 2010) which is formulated as follows:

$$b^e_{u_t} = p_1 + p_2 * c^e_{u_t} \tag{1}$$

$$c^e_{u_t} = \beta * l^e_{u_t} \tag{2}$$

where p_1 is base rate, p_2 is tour length dependent rate and p_1 is profit factor. β is the operational cost per kilometer which consists of the costs of freight vehicles and employees. $l^e_{u_t}$ is the incremental tour length for serving request u_t by carrier e. Estimation of $l^e_{u_t}$ involves a branch of the Vehicle Routing Problem which is the dynamic pick up and delivery problem with time windows (DPDTW). For the DPDTW problems, the cheapest insertion procedure from Mitrović-Minić and Laporte (2004) is adopted to estimate $l^e_{u_t}$ and (re)planning routes when carriers achieve new requests from the AITP in real-time. We assume the service capacity is able to meet all requests, namely, there are enough freight vehicles for carriers.

Due to the highly dynamic nature of the instant delivery marketplace, we assume that carriers can adjust their profit factor p_1 at any time to flexibly determine their bidding prices in sequential auctions. To improve carriers' bidding ability, a RL based dynamic biding strategy is applied to one carrier e^{learn} (Figliozzi et al., 2005). When a new request u_t is published in the AITP at epoch t, the learning carrier observes their current state S_t, which is described by a two-tuple $\left(W_t, c^{e^{learn}}_{u_t}\right)$, where W_t denotes the winning rate in auctions (winning times/total biding times). The observed information S_t is the input of the learning model, the output is a selected optimal profit factor p_1 from all candidate values. Then the bidding price $b^{e^{learn}}_{u_t}$ is determined based on the two-stage pricing mechanism. The framework of considering one learning carrier in the AITP is illustrated in Fig. 1. The general procedure of the method is shown the Fig. 2.

4. Design of the RL based dynamic bidding strategy

4.1. Q-learning

Q-learning is a model free algorithm that learns based on experiencing the consequences of actions (Watkins and Dayan, 1992). A Q-table should be designed properly with a predefined state set S and action set A. The matrix of a Q-table describes the expected utility for every possible finite action a under a certain state s. In our context, the states are associated with factors involving a winning rate W_t and a serving cost $c^{e^{learn}}_{u_t}$. Both are fluctuating with uncertainty, ranging over the state set at each epoch t. We classify both W_t and $c^{e^{learn}}_{u_t}$ into three levels for designing the Q-table, therefore, there are nine grouped states in total. The action is to determine the value of p_1. We define three candidate values for p_1, a minimum value, a medium value and a maximum value. The action set is noted as $A = \{minimump_1, mediump_1, maximump_1\}$. For each auction at epoch t, the learning carrier e^{learn} has to select a profit factor p_1 from the action set. The Q-table is used to determine the optimal decision based on the action value function $Q(s_t, a_t)$. A reward r_t will be received immediately along with the outcome of the auction. Namely, if the learning carrier secures the request, a reward r_t will be the profit gained from the acquired request. Otherwise, the reward r_t will be zero if the learning carrier fails to gain the request in the auction. The Q-table is updated at each epoch t according to the following equation:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right] \tag{3}$$

where

α = learning rate of the learning carrier, $(0 < \alpha < 1)$
γ = discount rate for the learning carrier, $(0 < \gamma < 1)$
$\max_a Q(s_{t+1}, a)$ = maximum Q-value of the next state s_{t+1} for all candidate actions a

The learning rate of 1 implies that the learning carrier fully considers the current information while 0 implies the learning carrier does not learn from experience at all. The discount rate of 1 implies that the learning carrier will consider the most long term reward, while 0 implies the carrier only considers the current reward (Firdausiyah et al., 2019). Algorithm 1 outlines details of Q-learning as applied in our context.

Algorithm 1. Q-learning for bidding strategy in the AITP

Initialize Q-table
For epochs $t = 1, 2, \cdots, T$ **do**:
  Identify current winning rate W_t and serving cost $c^{e^{learn}}_{u_t}$
  Get current state s_t
  If $t > 1$:
    Update Q-table:
    $Q(s_{t-1}, a_{t-1}) \leftarrow Q(s_{t-1}, a_{t-1}) + \alpha\left[r_{t-1} + \gamma \max_a Q(s_t, a) - Q(s_{t-1}, a_{t-1})\right]$
  End
  Select action a with $a^* = arg\max_a Q(s_t, a)$ and ϵ-greedy policy
  Get bid price based on the two-stage pricing mechanism and submit bid
  Receive an immediate reward r_t
End

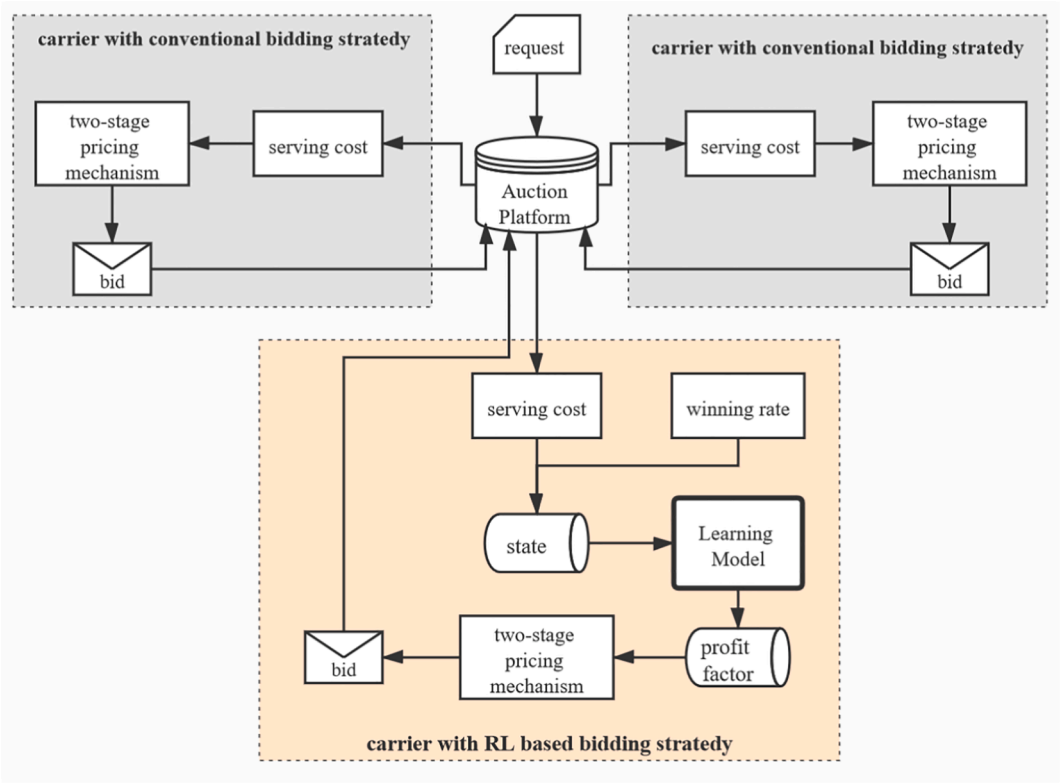


Fig. 1. Framework for considering one learning carrier in the AITP.

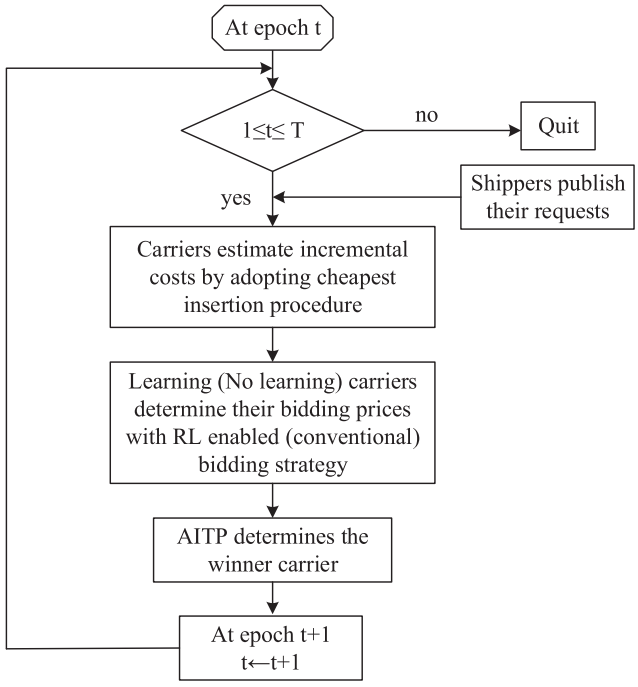


Fig. 2. General procedure of the AITP.

4.2. Deep Q network

In the highly uncertain market of instant delivery logistics, the state in each epoch fluctuates significantly. It is not easy to group the infinite type of states into a finite number of states properly for a classical Q-table. In addition, a Q-table with a high number of dimensions works inefficiently. The Deep Q Network (DQN) has been proposed to deal with problems associated with an infinite number of states (Mnih et al., 2015). The Q value for each possible action is predicted based on the observed state by a neural network which is parametrized by $Q(s,a;\theta)$ instead of a predefined Q-table. The parameter vector θ contains the weights and biases of the multiple layers of the neural network. Here, the directly observed parameters which include the winning rate W_t and the serving cost $c_{u_t}^{elearn}$ represent the state of epoch t for the DQN to predict the Q-value. Therefore, the state of each epoch varies over an infinite set. Rather than updating values of the Q-table directly at each epoch, the DQN consists of an evaluate Q-network and a target Q-network. The evaluate Q-network estimates the action value while the target Q-network trains and adjusts the evaluate Q-network in the target direction. Fig. 3 illustrates the architecture of DQN for this model. Algorithm 2 shows the implementation of DQN for carrier's bidding strategy in the AITP.

Algorithm 2. DQN for bidding strategy in the AITP

Initialize evaluate Q-network
For epochs $t = 1, 2, \cdots, T$ **do**:
  If $t \bmod \partial = 0$:
    target Q-network = evaluate Q-network
  End
  Observe current winning rate W_t and serving cost $c_{u_t}^{elearn}$ as input $state_t$
  Store $[s_{t-1}, a_{t-1}, r_{t-1}, s_t]$ in memory pool$M$
  Select action a with $a^* = arg\max_a Q^{eval}\left(s_t, a; \theta_t^{eval}\right)$ and ϵ-greedy policy
  Get biding price based on the two-stage pricing mechanism and submit bid
  Receive an immediate rewardr_t
  If $t \bmod \omega = 0$:
    Train evaluate Q-network
    Sample a random minibatch of experience from$M$
    Calculate the target Q-value
    $Q^{target} = r + \gamma \max_a Q^{target}\left(s', a'; \theta_t^{target}\right)$
    Update evaluate Q- *network* with gradient descent (AdamOptimizer is used here)
    $L_t(\theta_t) = E\left[\left(Q^{target} - Q^{eval}\right)^2\right]$
  End
End

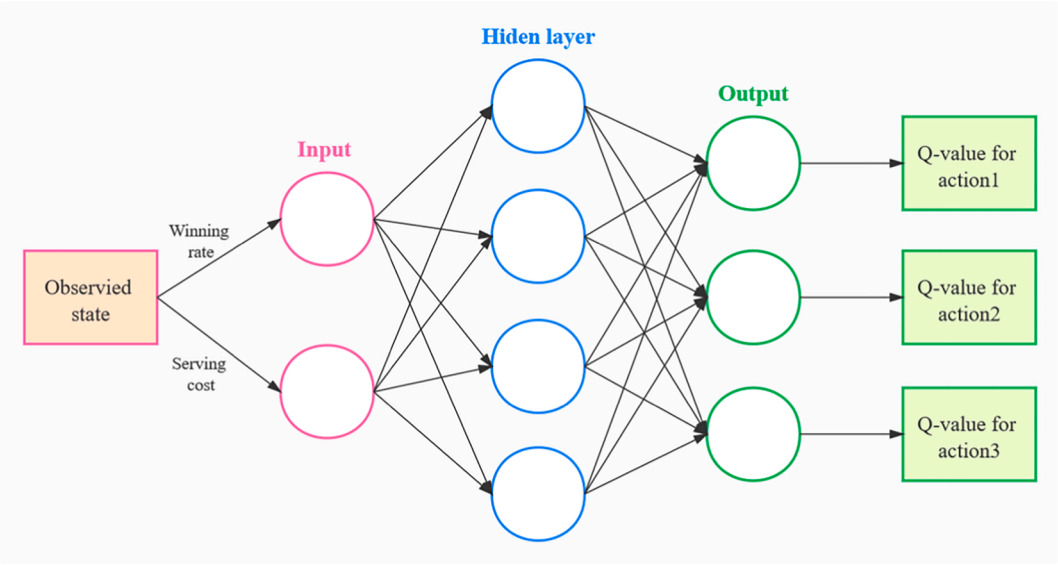


Fig. 3. Architecture of the DQN for bidding strategy in the AITP.

4.3. Experience replay based Q-learning

Traditional Q-learning is known to be unstable which may lead to overfitting the learning model (Mnih et al., 2015). Based on this consideration, the same authors used an Experience Replay (ER) method to break the correlation between samples and improve learning efficiency.

In this section, we introduce an ER-Q-learning method that adapts Q-learning with an ER optimizer. Inspired by Deep Q Network (DQN), we define an evaluate a Q-table and a target Q-table. The evaluate Q-table is updated at each epoch while the target Q-table is held fixed and is only reset by copying the evaluate Q-table directly every ∂ epochs. The ER optimizer is used to train the evaluate Q-table every δ epochs according to function (4).

Let's define a memory pool M with a certain size to store experience at each epoch t which consists of state s, action a, reward r, and the new state s' of the learning carrier. Each observed experience $[s, a, r, s']$ is added into the memory pool M. When performing the ER optimizer, a minibatch of stored experiences $[s, a, r, s'] \sim M$ are selected randomly from the memory pool as training data. The new observed experience will replace the earliest one when the memory pool is full. Algorithm 3 outlines the ER-Q-learning as applied to the bidding strategy in AITP.

$$L_t = E_{[s,a,r,s'] \sim M}\left[\left(r + \gamma \max_a Q^{target}(s', a') - Q^{eval}(s, a)\right)^2\right] \quad (4)$$

Algorithm 3. ER-Q-learning for bidding strategy in the AITP

Initialize evaluate Q-table
For epochs $t = 1, 2, \cdots, T$ **do**:
  If $t \bmod \partial = 0$:
    target Q-table = evaluate Q-table
  End
  Achieve current winning rate W_t and serving cost $c_{u_t}^{e_{learn}}$
  Get current state s_t
  Store $[s_{t-1}, a_{t-1}, r_{t-1}, s_t]$ in memory pool M
  If $t > 1$:
    Update evaluate Q-table:
    $Q^{eval}(s_{t-1}, a_{t-1}) \leftarrow Q^{eval}(s_{t-1}, a_{t-1}) + \alpha\left[r_{t-1} + \gamma \max_a Q^{eval}(s_t, a) - Q^{eval}(s_{t-1}, a_{t-1})\right]$
  End

(continued on next column)

(continued)

  Select action a with $a^* = arg\max_a Q^{eval}(s_t, a)$ and ϵ-greedy policy
  Get biding price based on the two-stage pricing mechanism and submit bid
  Receive an immediate reward r_t
  If $t \bmod \delta = 0$:
    Executive ER optimizer
    Sample a random minibatch of experience from M
    Calculate the target Q-value
    $Q^{target} = r + \gamma \max_a Q^{target}(s', a')$
    Update evaluate Q-table with gradient descent (AdamOptimizer is used here)
    $L_t = E\left[\left(Q^{target} - Q^{eval}\right)^2\right]$
  End
End

5. Computational study

5.1. Simulation setup

Here we investigate instant delivery within an urban area where multiple carriers compete for instant delivery requests through the AITP. Competitor's actions are affected by different auction mechanisms. In First Price Auctions, the carrier who offers the lowest bid wins the request and the shipper pays the winning carrier for the service with an amount of the lowest bid, where $S_{u_t} = b_{u_t}^{min}$. The all candidate profit factor p_1 namely the possible actions should be equal or larger than zero, the reason is that a negative p_1 is possible to generate negative profit. For the Second Price Auction, the carrier with a lowest price wins the auction, however, shipper pays the amount equal to the second lowest price, where $S_{u_t} = b_{u_t}^{second\_min}$. A negative value of p_1 can still achieve a positive profit in a Second Price Auction since the payment is dependent on competitors' bids (Figliozzi et al., 2005). Both the First Price Auction and Second Price Auction are examined in our simulations. In addition, the instant delivery marketplace is simulated over a 20×20 km^2 and a 50×50 km^2 area, respectively. The four simulation environments examined are outlined in Table 1 for understanding the performance of the learning carrier with multiple learning algorithms in multiple scenarios.

Retailers and individuals are both called customers who are randomly distributed within the study area. A customer can be either a shipper or a receiver. It is assumed that there are one and two hundred customer nodes in the 20×20 km^2 area and the 50×50 km^2 area, respectively. Requests arrive dynamically and are assumed to follow a

Table 1
Simulation environments for the AITP.

	Auction mechanism		Study area(km^2)	
	First Price	Second Price	20×20	50×50
1	*		*	
2	*			*
3		*	*	
4		*		*

Poisson process. Pickup and delivery locations of requests are randomly generated from customer nodes. We consider three carriers for all simulation environments and carrier 1 is assigned to be the learning carrier while other carriers use a stable biding strategy with a fixed value of p_1. Each carrier has a depot, all vehicles start and end at their own depot. Fig. 4 illustrates an operating example. In order to let each carrier have similar operational competitiveness, we located their depots in the center of the study area, as Fig. 4 shows. Considering the urgency of instant delivery, we assume each request should be completed in one hour from its required earliest pickup time. The publishing time of each request is assumed to be its earliest pickup time. In addition, linear travel distances are assumed for simplicity. For each environment, the simulation operates for 100 days with each day including 200 requests. At the beginning of each day, the state of each carrier is initialized so no request has been contracted, and no vehicle is operating before the sequential auctions. All requests must be finished in the same day. The total of twenty thousand requests are generated randomly before the simulation. Four biding strategies for the learning carrier (carrier 1) has been examined during the simulation, including, (i) no learning (stable strategy which is the same with other carriers'), (ii) Q-learning enabled biding strategy, (iii) ER-Q-learning enabled biding strategy, and (iv) DQN enabled biding strategy. As shown in Fig. 3, in the neural network, the hidden layer has 4 units and use the ReLU activation function. The last layer is a linear output layer with 3 units, one for each action. The other parameters used in this neural network and simulations are summarized in Table 2.

Table 2
Simulation assumptions and parameters.

Parameters	Values
Speed of vehicle:	45 km/h for $20 \times 20 km^2$ area, 60 km/h for $50 \times 50 km^2$ area
Operational cost per kilometre (β):	1
Tour length dependent rate (p_2):	1
Base rate (p_1) for conventional bidding strategy:	3 (for both First Price Auction and Second Price Auction)
Action set (candidate p_1) for the learning carrier:	[0, 3, 6] for First Price Auction, [−3, 0, 3] for Second Price Auction
Learning rate (α):	0.2 (for all simulations)
Discount rate (γ):	0.8 (for all simulations)
Target Q-table/network replaced step (∂):	100 (for both ER-Q-learning and DQN)
Size of memory pool (M):	500 (for both ER-Q-learning and DQN)
Size of minibatch:	30
ER optimizer called step (δ):	20
DQN training step (ω):	2

5.2. Simulation results

The simulation seeks to investigate how different RL algorithms effect the performance of the learning carrier in the AITP. The evaluation criteria are focused on the number of winning times in sequential auctions as well as the profit gained by offering instant delivery service for the learning carrier (carrier 1) from the AITP. In addition, the standard deviation (*Std*) ($\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}$) of carrier 1's profit per day is presented to show the stability of their bidding performance during the 100 days simulated. A larger value of *Std* indicates that each day's profit fluctuates more fiercely around the average value.

For the simulation environment of the First Price Auction, Table 3 summarizes the performance of the learning carrier (carrier 1) over the simulated 100 operational days with different bidding strategies (no learning, Q-learning, ER-Q-learning and DQN). The simulation was carried out on both the 20×20 km^2 area and the 50×50 km^2 area. It

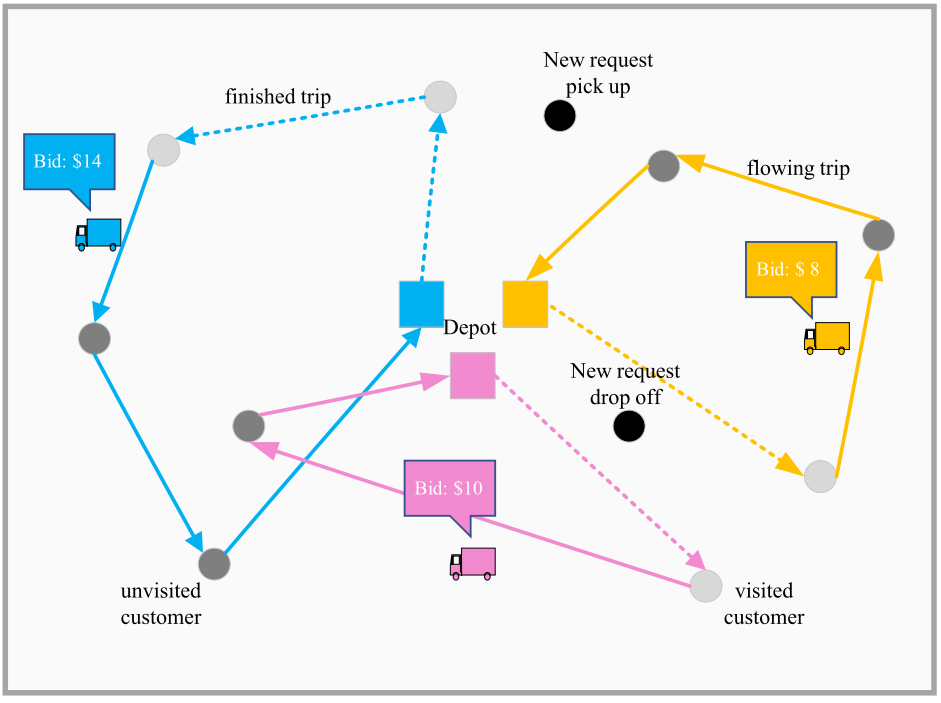
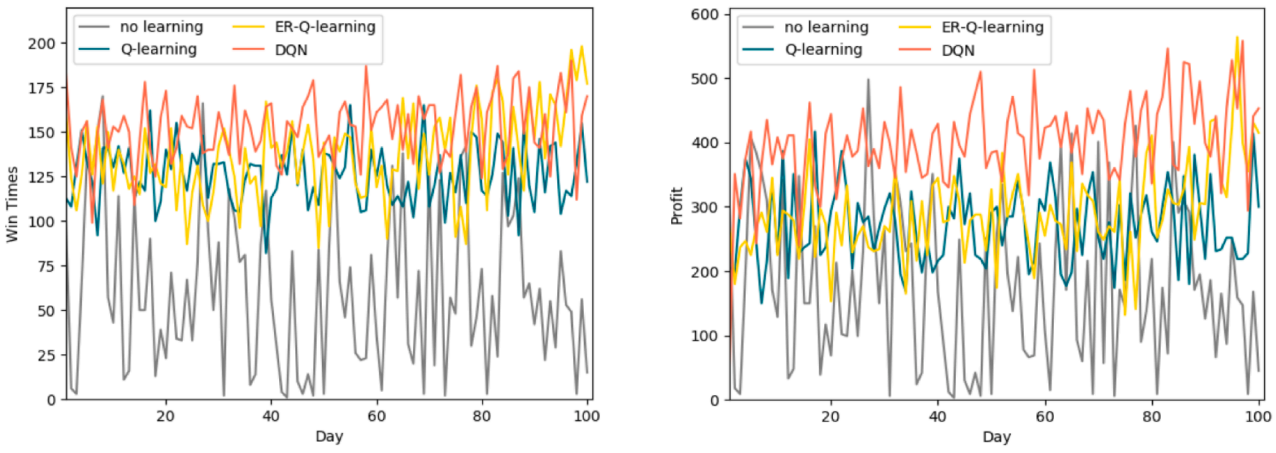


Fig. 4. Operating example for auction based freight transportation procurement.

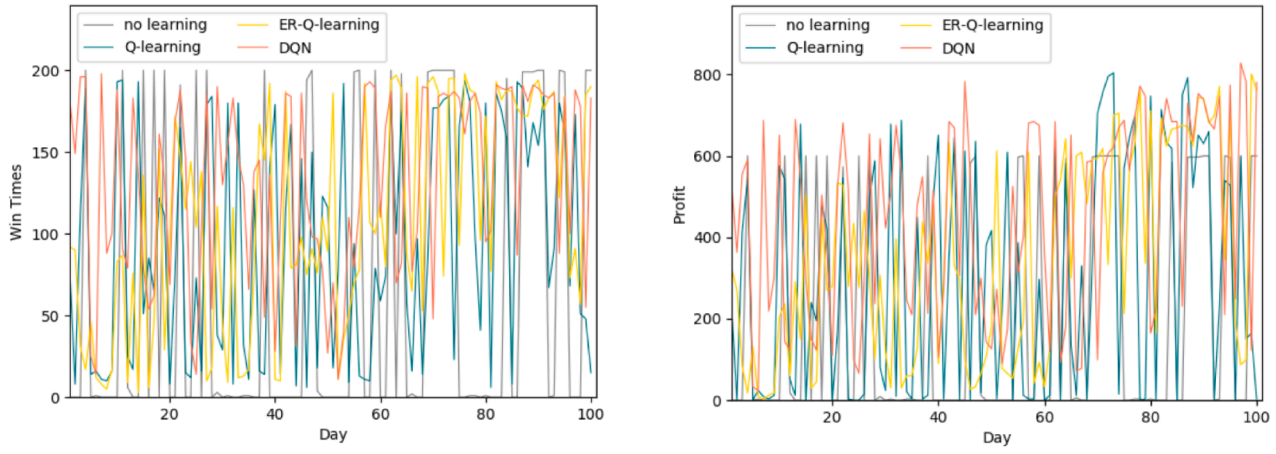
Table 3
Performance of the learning carrier (carrier 1) with multiple bidding strategies in First Price Auction.

	Time (s)	Win times	Profits			*Std* profit of carrier 1	Actions of carrier 1
		Carrier1	**Carrier1**	Carrier2	Carrier3		[times of selecting action1, times of selecting action2, times of selecting action3]
		Study area: 20×20 km^2					
No learning	88.31	**6062**	**17,636**	20,315	16,049	124.7	Fixed setting: 3
Q-learning	95.56	**12,603**	**27,003**	10,242	11,949	61.8	[3644, 14485, 1871]
ER-Q-learning	6245.23	**13,681**	**29,097**	9288	9649	69.7	[4653, 13993, 1354]
DQN	142.89	**15,263**	**40,050**	6954	7257	70.5	[1940, 17158, 902]
		Study area: 50×50 km^2					
No learning	134.22	**6395**	**19,185**	24,558	16,257	288.7	Fixed setting: 3
Q-learning	145.34	**9440**	**29,934**	19,074	14,313	278.2	[4542, 7703, 7755]
ER-Q-learning	8003.04	**10,869**	**34,737**	13,338	15,745	252.6	[4832, 7732, 7436]
DQN	213.21	**13,239**	**44,556**	12,416	11,865	241.9	[4380, 7746, 7774]

can be found that a significant improvement of carrier 1′s performance is achieved when it uses the RL enabled dynamic bidding strategy to make decisions compared with using the conventional bidding strategy. Regarding the number of times carrier 1 has won from the total twenty thousand auctions simulated, carrier 1 secures 30.3% (6062 times) and 32% (6395 times) of the auctions over the study area of 20×20 km^2 and 50×50 km^2 respectively, when it adopts the same conventional bidding strategy as the other carriers. In contrast, RL enabled dynamic bidding strategy enable carrier 1 to secure 63% of the auctions with Q-learning, 68.4% of the auctions with ER-Q-learning, and 76.3% of the auctions with DQN for the 20×20 km^2 study area. The *Std* of carrier 1′s profit per day decreased by half which implies a more stable performance for carrier 1. However, the improved performance by RL enabled dynamic bidding strategy is not that substantial when simulation is



a. Study area: 20×20 km^2



b. Study area: 50×50 km^2

Fig. 5. Performance of the learning carrier each day with multiple bidding strategies in the First Price Auction.

conducted over a larger area but still achieved a favorable outcome, securing 47.2% with Q-learning, 54.3% with ER-Q-learning, and 58.7% with DQN auctions for the 50×50 km^2 study area. Along with the growth of winning times, the profit of carrier 1 increased significantly as well.

Fig. 5.a illustrates the performance of the learning carrier per day with different bidding strategies in the First Price Auction. For the 20×20 km^2 study area, it is clear that when there is no learning method adopted, carrier 1 wins less than half of total requests per day and gains less than $300 of profit in most days. The performance curves are upward when RL techniques are applied. The upward curves present a significant improvement of bidding power for carrier 1. In addition, DQN shows better performance than Q-leaning and ER-Q-learning in this environment. Two reasons account for this: First, DQN is able to use data in a more efficient way for optimizing the Q-network (Mnih et al., 2015). In the DQN algorithm, the entire parameters of Q-network are trained and are updated frequently using historical data sets. While for Q-learning and ER-Q-learning, only a certain Q-value is updated which is associated with the state and selected action of each epoch. Secondly, DQN is superior to Q-learning for reinforcing models with infinite states (Mnih et al., 2015). In the context of the AITP, we group the infinite states into nine groups in total for Q-learning and ER-Q-learning. While DQN predicts the Q-value using a neural network based on the observed states directly. The performance curves for the 50×50 km^2 study area look quite messy and are illustrated in Fig. 5.b. The large value of *Std* in Table 3 also demonstrates the great fluctuating performance. This is because requests from the marketplace over a larger area are more uncertain than that over a smaller area due to the larger fluctuation of randomly distributed pickup and drop off locations. Therefore, learning approaches perform less efficiently in a more unpredictable environment but still outperform the no learning situation.

For the simulation environment of the Second Price Auction, the performance of the learning carrier (carrier 1) with different bidding strategies and different simulation environments are summarized in Table 4. The results show that the improvement of carrier 1′s performance is substantial with the support of RL enabled dynamic bidding strategy compared with the conventional bidding strategy for both study areas. With the conventional bidding strategy, carrier 1 secures 27% (5517 times) of the auctions in the simulation over study area of 20×20 km^2. In contrast, RL techniques enable carrier 1 to secure roughly 98% of the auctions with Q-learning, ER-Q-learning, and DQN. Also, a dramatic decrease of *Std* has been achieved by the RL techniques. It can be noticed that carrier 1 selects the action $p_1 = -3$ more frequently than other possible actions and action $p_1 = 3$ is selected the least number of times. The substantial advantage of RL algorithms was achieved for the simulation over study area of 50×50 km^2, where nearly 100% auctions were secured by carrier 1. The selection among different possible actions are distributed more evenly than that with the 20×20 km^2 study area.

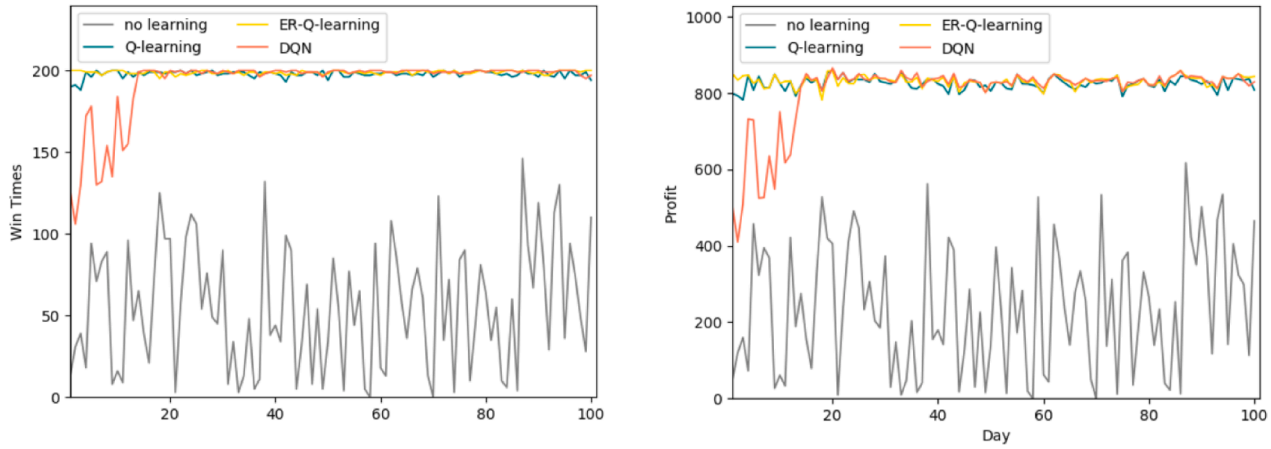
The performance of the learning carrier per day with different bidding strategies in the Second Price Auction is illustrated in Fig. 6. For the 20×20 km^2 study area, carrier 1 wins less than 50% of auctions executed per day and gains less than $500 profit in most days when it does not apply RL enabled dynamic bidding strategy to improve its bidding strategy. In contrast, the number of times each day of carrier 1 wins stays around 200 and the profit per day is around $850 when it uses RL techniques. For the 50×50 km^2 study area, the performance curves fluctuate dramatically with the no learning situation due to the greater uncertainty in the marketplace. However, all the three proposed RL algorithms are able to maintain a stable and high level of performance. It can be noted that for the 20×20 km^2 study area, DQN performs less effective than Q-learning and ER-Q-learning in the first couple of days. The extremely stable performance of DQN is achieved after sufficient training steps. However, this phenomenon does not happen in the 50×50 km^2 study area where DQN performs as good as (or even more stable) than both Q-learning and ER-Q-learning in the first few days. The reason is that the gap of servicing costs for each request which is also one of the factors to determine the state at each epoch is very small and hard to identify the difference over the 20×20 km^2 study area, therefore, more learning steps are needed to train and optimize the learning model. In contrast, the pickup and drop off locations are distributed with longer distances over the 50×50 km^2 study area, therefore, the state gap is easy to determine. Thus, the learning model can reach optimality faster.

It can be seen that the RL enabled bidding strategy performs better in Second Price Auction than that in the First Price Auction. This is because in the Second Price Auction the learning carrier finds its easier to learn that a lower p_1 can be more likely to secure the auction. While the profit gained is more than that based on the bid price since payment is based on the second lower price which is definitely higher than it's own bid price. The actions of the learning carriers for both auction mechanisms can be found in Table 3 and 4. These results demonstrate that the learning carrier selects *minimum* p_1 more frequently in the Second Price Auction.

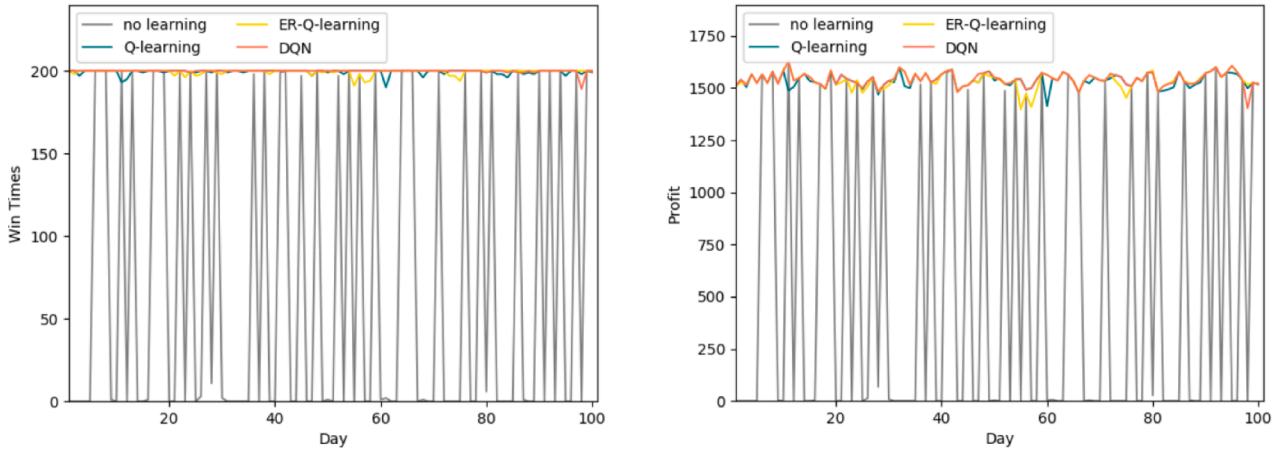
From the column "Time" of Tables 3 and 4, we observe that, although the RL enabled bidding strategies can significantly improve carrier's performance, the computational times of the cases with RL enabled bidding strategies are longer compared with those of the cases with conventional bidding strategy. Comparing the computational time of different RL algorithms, we see that the Q-learning algorithm needs shorter computational time than ER-Q-learning and DQN algorithms. However, as shown in the previous analysis, except for simulation environment of the Second Price Auction over study area of $20 \times$

Table 4
Performance of the learning carrier (carrier 1) with multiple bidding strategies in the Second Price Auction.

	Time (*s*)	**Win times**	**Profits**			***Std***profit of carrier 1	**Actions of carrier 1**
		Carrier1	**Carrier1**	Carrier2	Carrier3		[times of selecting action1, times of selecting action2, times of selecting action3]
		Study area: 20×20 km^2					
No learning	74.24	**5517**	**23,540**	24,136	27,507	166.9	Fixed setting: 3
Q-learning	89.23	**19,776**	**82,592**	373	365	15.9	[8520, 7317, 4163]
ER-Q-learning	6538.59	**19,879**	**83,179**	184	207	14.1	[8267, 8008, 3725]
DQN	156.34	**19,265**	**81,458**	965	1445	86.6	[7413, 6556, 6031]
		Study area: 50×50 km^2					
No learning	143.33	**7217**	**55,559**	60,888	47,104	738.2	Fixed setting: 3
Q-learning	149.74	**19,937**	**153,549**	123	115	32.1	[6780, 7564, 5656]
ER-Q-learning	8021.12	**19,940**	**153,632**	131	141	37.8	[6761, 7932, 5307]
DQN	213.21	**19,983**	**154,211**	133	96	32.6	[7665, 6384, 5951]



a. Study area: $20 \times 20\ km^2$



b. Study area: $50 \times 50\ km^2$

Fig. 6. Performance of the learning carrier each day with multiple bidding strategies in the Second Price Auction.

20 km^2, compared with the other two algorithms, the Q-learning algorithm has the lowest performance in improving the winning time and profit of the learning carrier. For example, in the simulation environment of the First Price Auction over study area of $20 \times 20\ km^2$, compared with the Q-learning algorithm, the ER-Q-learning algorithm can improve the profit of learning carrier 1 by 7.8%, while increasing the winning outcomes by 8.6% (1078 times). However, the computational time of ER-Q-learning algorithm is more than 65 times that of the Q-learning algorithm. By adopting the DQN algorithm, the learning carrier 1 can obtain almost 1.5 times more profit than adopting the Q-learning algorithm through wining more times (2660 more wins). However, the computational time of DQN algorithm only increases by 47.33 s compared with that of Q-learning algorithm. This means that the DQN algorithm can achieve a better balance between performance and computational time.

5.3. Extended experiments

To further investigate how RL enabled dynamic bidding strategy effects the performance of the learning carriers in the AITP, we extended the simulation experiments to cases of different scales. Considering the superiority of the DQN algorithm in balancing the performance and computational time, we use the DQN algorithm in the extended simulation experiments. Moreover, we allow multiple carriers to learn simultaneously in the extended simulation experiments to investigate the performance of RL enabled dynamic bidding strategy in the marketplace where multiple learning carriers compete. The eight examined simulation environments are outlined in Table 5. In Table 5, TNC and NLC represent the total number of carriers and the number of learning carriers involved in the simulation environments. For cases with three carriers, the simulation operates for 100 days with each day including 200 requests. For cases with ten carriers, simulation environments with 100 days and each day including 400 requests are considered. The other settings of simulation environments are same as Section 5.1.

Table 6 shows the computational times in different simulation environments. Tables 7 and 8 summarize the performance of the learning

Table 5
Simulation environments for the AITP in the extended experiments.

	Carrier		Auction mechanism		Study area(km^2)	
	TNC	NLC	First Price	Second Price	50×50	100×100
1	3	2	*		*	
2	3	2		*	*	
3	10	1	*			*
4	10	1		*		*
5	10	3	*			*
6	10	3		*		*
7	10	5	*			*
8	10	5		*		*

Table 6
The computational times in different simulation environments.

Carrier		Computational time (*s*) Auction mechanism	
Total carriers	Learning carriers	First price auction	Second price auction
3	No learning carrier	134.22	143.33
3	One learning carrier	205.67	213.21
3	Two learning carrier	303.60	306.33
10	No learning carrier	283.90	286.20
10	One learning carrier	408.83	409.21
10	Three learning carrier	541.83	545.33
10	Five learning carrier	591.99	599.37

and no-learning carriers with different auction mechanisms in the cases of different scales. In Tables 7 and 8, the learning carriers are marked in bold. It can be found that, compared with the cases of no learning, the performance of the learning carriers can be significantly improved regardless of the simulation environment (auction mechanism, the total number of carriers and the number of learning carriers). This further demonstrates the advantage of the RL enabled dynamic bidding strategy in improving the carriers' bidding abilities. Similar to the results with small networks, the computational times of the cases with RL enabled dynamic bidding strategy are longer than those of the cases with conventional bidding strategy. Moreover, the computational times increase with the number of learning carriers. However, none of the computational times are more than 600 s. For the operational planning of 100 days, the computational times are reasonable and acceptable.

In Table 8, as we move from the case with no learning carrier to the case with five learning carriers, we can find that, the new carrier who uses the RL enabled dynamic bidding strategy can secure more auctions and gain more profit. Compared with the conventional bidding strategy, in the case with one learning carrier, the profit and win times of new learning carrier 1 increases by 530.1% and 322.9% (10226 times), respectively. In the case with three learning carriers, the profits of new learning carriers 2 and 3 increase by 160.8% and 200.6%, respectively. For the win times, the corresponding values are 100.9% (3427 times) and 111.1% (3537 times). In the case with five learning carriers, the RL enabled dynamic bidding strategy can improve the profits of new learning carriers 4 and 5 by 4.3% and 26.3%, respectively, while increasing their win times by 20.1% (688 times) and 26.2% (1915 times). In addition, Table 8 shows that, the profits and win times of existing learning carriers decrease when other carriers use RL enabled dynamic bidding strategy. However, compared with the conventional bidding strategy, RL enabled dynamic bidding strategy can still help carriers improve their bidding abilities. The conclusion is the same in the case with three carriers.

5.4. Managerial insights

In this section, we provide some managerial insights according to the

Table 7
Performance of the carriers with multiple bidding strategies under different auction mechanisms in cases with three carriers.

No learning carrier					Two learning carriers				
Carrier	First price auction		Second price auction		Carrier	First price auction		Second price auction	
	Profit	Win times	Profit	Win times		Profit	Win times	Profit	Win times
Carrier1	**19,185**	6395	55,559	7217	**Carrier1**	21,234	9510	62,456	9415
Carrier2	24,558	7577	60,888	7869	**Carrier2**	27,500	9531	65,193	8345
Carrier3	16,257	6027	47,104	4914	Carrier3	2877	959	21,536	2240

Table 8
Performance of the carriers with multiple bidding strategies under different auction mechanisms in cases with ten carriers.

No learning carrier					One learning carrier				
Carrier	First price auction		Second price auction		Carrier	First price auction		Second price auction	
	Profit	Win times	Profit	Win times		Profit	Win times	Profit	Win times
Carrier1	6234	3167	13,734	2695	**Carrier1**	39,280	13,393	152,616	20,383
Carrier2	7311	3398	19,498	3569	Carrier2	8026	2342	10,703	1855
Carrier3	6345	3184	15,971	3145	Carrier3	8753	3351	7878	1179
Carrier4	7834	3417	44,468	4327	Carrier4	2388	1796	9249	1444
Carrier5	19,578	7315	55,468	6987	Carrier5	10,421	3907	20,019	3212
Carrier6	6732	3175	15,246	3038	Carrier6	7854	2618	9251	1613
Carrier7	7756	3996	52,156	5789	Carrier7	8338	2796	12,920	2068
Carrier8	8503	4049	25,773	3974	Carrier8	8961	2987	25,430	3855
Carrier9	8973	4821	9989	1853	Carrier9	8715	2905	11,503	1853
Carrier10	8029	3478	47,291	4623	Carrier10	10,715	3905	19,734	2538
Three learning carriers					Five learning carriers				
Carrier	First price auction		Second price auction		Carrier	First price auction		Second price auction	
	Profit	Win times	Profit	Win times		Profit	Win times	Profit	Win times
Carrier1	18,205	6472	58,351	9059	**Carrier1**	12,925	5585	33,164	5406
Carrier2	19,066	6825	119,945	15,192	**Carrier2**	12,495	5405	51,630	9237
Carrier3	19,070	6721	70,375	10,303	**Carrier3**	9352	4633	44,034	7518
Carrier4	6635	2545	926	214	**Carrier4**	8172	4105	39,293	6885
Carrier5	14,916	4972	9444	1782	**Carrier5**	24,720	9230	49,735	8423
Carrier6	5784	1928	1932	448	Carrier6	5580	1860	2021	482
Carrier7	5619	1873	2174	420	Carrier7	7164	2388	1501	327
Carrier8	9708	3236	7973	1355	Carrier8	7995	2665	1860	399
Carrier9	6432	2144	1350	270	Carrier9	4572	1524	2546	530
Carrier10	9852	3284	6652	957	Carrier10	6815	2605	4752	793

experimental results.

As auctions become an effective option to improve the efficiency of instant delivery, how to capture more demand and obtain a stable and high level of profit have led to new challenges for carriers. From the experimental results, compared with the conventional bidding strategy, RL enabled dynamic bidding strategy can help carriers significantly improve profit by increasing winning times. Moreover, the performance of learning carriers is remarkably stable in a volatile marketplace. Therefore, carriers should adopt a RL enabled dynamic bidding strategy to improve their performance and competitiveness. Furthermore, compared with Q-learning and ER-Q-learning algorithms, the DQN algorithm can achieve a better balance between performance and computational time. So, the DQN algorithm enabled dynamic bidding strategy is the best option for carriers.

From analyzing the efficiency of the RL enabled dynamic bidding strategy, we find that the performance of RL enabled bidding strategy is better in the Second Price Auction than that of the First Price Auction. This means that in the Second Price Auction, carriers can gain more profit by adopting the RL enabled dynamic bidding strategy. However, the platform should be careful when a Second Price Auction is executed since learning carriers can secure almost all the auctions in Second Price Auctions. In this case, the platform should take some measures to prevent a monopoly.

With the emergence of the advantage of the RL enabled bidding strategy, more and more carriers will adopt this strategy to improve their performance. The experimental results show that, when multiple carriers adopt a RL enabled bidding strategy simultaneously, the number of win times and profit of existing learning carriers will decrease as new learning carriers join. However, compared with adopting a conventional bidding strategy, carriers can still gain more profit by adopting the RL enabled dynamic bidding strategy. In contrast, compared with the case without considering the RL enabled dynamic bidding strategy, almost all no-learning carriers will suffer a significant decrease in their win times and profits. Therefore, carriers should insist on adopting the RL enabled dynamic bidding strategy to cope with the high uncertainty of the instant delivery trading market. Otherwise, their existing market share will be taken up by those carriers who adopt the RL enabled dynamic bidding strategy.

6. Conclusions

This article investigates a dynamic bidding strategy for freight carriers in an auction based instant delivery trading platform (AITP) in a stochastic and uncertain environment. Three RL algorithms are developed to improve carrier's bidding power in sequential auctions in the AITP: (i) Q-learning, (ii) DQN (iii) ER-Q-learning. Simulations are carried out to evaluate the performance of the RL enabled bidding strategies compared to the conventional bidding strategy. Multiple environment factors are examined for the simulation: (1) auction mechanisms, the online auctions are performed with First Price Auction as well as Second Price Auction environment, (2) size of study areas, simulations are carried out over a $20 \times 20\ km^2$ area and a $50 \times 50\ km^2$ area, (3) number of total carriers and learning carriers, simulations are conducted in the cases with three and ten carriers, and different numbers of carriers are allowed to adopt RL enabled dynamic bidding strategy in each cases. Simulation results indicate that the RL enabled dynamic bidding strategies with any of the three RL algorithms achieve substantial improvement regarding the learning carriers' winning times and profits gained compared with the conventional bidding strategy.

In addition, RL enabled dynamic bidding strategies in Second Price Auctions are more stable and optimal where the learning carrier is able to secure almost all auctions and gain more profit each day. DQN is found to be superior to Q-learning and RE-Q-learning in First Price Auctions. Furthermore, due to the greater uncertainty of demand in the $50 \times 50\ km^2$ study area, performance of the target carrier fluctuates dramatically when it uses the conventional bidding strategy. RL enabled dynamic bidding strategies still improve bidding power significantly especially in the Second Price Auction. Finally, in the scenarios with different numbers of learning carriers, compared with the conventional bidding strategy, the RL enabled dynamic bidding strategy can maintain the performance advantage.

An extension of this work would be to take all stakeholders of the AITP into consideration. Profits of the platform manager and profits of all participating freight carriers as well as cost of shippers could be examined with the application of RL techniques. Also, future work could investigate the effect of different RL strategies in the AITP on the environmental performance of urban freight networks for improving sustainability.

CRediT authorship contribution statement

Chaojie Guo: Conceptualization, Methodology, Software, Visualization, Writing – original draft. **Russell G. Thompson:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Greg Foliente:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Xiaoshuai Peng:** Conceptualization, Methodology, Visualization, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Berger, S., & Bierwirth, C. (2010). Solutions to the request reassignment problem in collaborative carrier networks. *Transportation Research Part E: Logistics and Transportation Review, 46*(5), 627–638.

Cheng, Y., Zou, L., Zhuang, Z., Liu, J., Xu, B., & Zhang, W. (2019). An extensible approach for real-time bidding with model-free reinforcement learning. *Neurocomputing, 360*, 97–106.

Crainic, T. G., & Montreuil, B. (2016). Physical internet enabled hyperconnected city logistics. *Transportation Research Procedia, 12*, 383–398.

Dablanc, L., Morganti, E., Arvidsson, N., Woxenius, J., Browne, M., Saidi, N., 2017. The rise of on-demand 'Instant Deliveries' in European cities. Supply Chain Forum: An International Journal. Taylor & Francis, pp. 203-217.

Du, J., Guo, B., Liu, Y., Wang, L., Han, Q.i., Chen, C., & Yu, Z. (2019). CrowDNet: enabling a crowdsourced object delivery network based on modern Portfolio Theory. *IEEE Internet of Things Journal, 6*(5), 9030–9041.

Figliozzi, M. A., Mahmassani, H., & Jaillet, P. (2005). Auction settings and performance of electronic marketplaces for truckload transportation services. *Transportation Research Record, 1906*, 89–97.

Andres Figliozzi, M., Mahmassani, H. S., & Jaillet, P. (2003a). Framework for study of carrier strategies in auction-based transportation marketplace. *Transportation Research Record, 1854*(1), 162–170.

Figliozzi, M.A., Mahmassani, H.S., Jaillet, P., 2003b. Modeling Carrier Behavior in Sequential Auction Transportation Markets.

Figliozzi, M. A., Mahmassani, H. S., & Jaillet, P. (2004). Competitive performance assessment of dynamic vehicle routing technologies using sequential auctions. *Transportation Research Record, 1882*(1), 10–18.

Figliozzi, M. A., Mahmassani, H. S., & Jaillet, P. (2006). Quantifying opportunity costs in sequential transportation auctions for truckload acquisition. *Transportation Research Record, 1964*(1), 247–252.

Firdausiyah, N., Taniguchi, E., & Qureshi, A. G. (2019). Modeling city logistics using adaptive dynamic programming based multi-agent simulation. *Transportation Research Part E: Logistics and Transportation Review, 125*, 74–96.

Firdausiyah, N., Taniguchi, E., Qureshi, A.G., 2017. Multi-agent Simulation using Adaptive Dynamic Programming in The Existence of Urban Consolidation Centers. 土木学会論文集 D3 (土木計画学) 73(5), I_835-I_846.

Firdausiyah, N., Taniguchi, E., & Qureshi, A. G. (2020). Multi-agent simulation-Adaptive dynamic programming based reinforcement learning for evaluating joint delivery systems in relation to the different locations of urban consolidation centres. *Transportation Research Procedia, 46*, 125–132.

Gu, Q., Fan, T., Pan, F., & Zhang, C. (2020). A vehicle-UAV operation scheme for instant delivery. *Computers & Industrial Engineering, 149*, 106809. https://doi.org/10.1016/j.cie.2020.106809

Guo, C., Thompson, R. G., Foliente, G., & Kong, X. T. R. (2021). An auction-enabled collaborative routing mechanism for omnichannel on-demand logistics through transshipment. *Transportation Research Part E: Logistics and Transportation Review, 146*, 102206. https://doi.org/10.1016/j.tre.2020.102206

Handoko, S. D., Nguyen, D. T., & Lau, H. C., 2014. An auction mechanism for the last-mile deliveries via urban consolidation centre. In 2014 IEEE International Conference on Automation Science and Engineering (CASE). IEEE, pp. 607–612.

Hong, J., Lee, M., Cheong, T., & Lee, H. C. (2019). Routing for an on-demand logistics service. *Transportation Research Part C: Emerging Technologies, 103*, 328–351.

Karaenke, P., Bichler, M., & Minner, S. (2019). Coordination is hard: Electronic auction mechanisms for increased efficiency in transportation logistics. *Management Science, 65*(12), 5884–5900.

Lafkihi, M., Pan, S., & Ballot, E. (2019). Freight transportation service procurement: A literature review and future research opportunities in omnichannel E-commerce. *Transportation Research Part E: Logistics and Transportation Review, 125*, 348–365.

Mendel, J., & McLaren, R. (1994). Reinforcement-learning control and pattern recognition systems. *A Prelude to Neural Networks: Adaptive and Learning Systems*, 287–318.

Mes, M., van der Heijden, M., & Schuur, P. (2009). Dynamic threshold policy for delaying and breaking commitments in transportation auctions. *Transportation Research Part C: Emerging Technologies, 17*(2), 208–223.

Mes, M., van der Heijden, M., & Schuur, P. (2013). Interaction between intelligent agent strategies for real-time transportation planning. *Central European Journal of Operations Research, 21*(2), 337–358.

Mes, M., van der Heijden, M., & van Harten, A. (2007). Comparison of agent-based scheduling to look-ahead heuristics for real-time transportation problems. *European Journal of Operational Research, 181*(1), 59–75.

Mitrović-Minić, S., & Laporte, G. (2004). Waiting strategies for the dynamic pickup and delivery problem with time windows. *Transportation Research Part B: Methodological, 38*(7), 635–655.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … Ostrovski, G. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tamagawa, D., Taniguchi, E., & Yamada, T. (2010). Evaluating city logistics measures using a multi-agent model. *Procedia-Social and Behavioral Sciences, 2*(3), 6002–6012.

Teo, J. S. E., Taniguchi, E., & Qureshi, A. G. (2012). Evaluation of distance-based and cordon-based urban freight road pricing in e-commerce environment with multiagent model. *Transportation Research Record, 2269*(1), 127–134.

Teo, J. S. E., Taniguchi, E., & Qureshi, A. G. (2014). Evaluation of load factor control and urban freight road pricing joint schemes with multi-agent systems learning models. *Procedia-Social and Behavioral Sciences, 125*, 62–74.

Thompson, R. G., & Hassall, K. P. (2012). A Collaborative Urban Distribution Network. *Procedia - Social and Behavioral Sciences, 39*, 230–240.

Triki, C., Oprea, S., Beraldi, P., & Crainic, T. G. (2014). The stochastic bid generation problem in combinatorial transportation auctions. *European Journal of Operational Research, 236*(3), 991–999.

Wang, Y., Nascimento, J.M.D., Powell, W., 2018. Reinforcement learning for dynamic bidding in truckload markets: an application to large-scale fleet management with advance commitments. arXiv preprint arXiv:1802.08976.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine Learning, 8*(3–4), 279–292.

Xu, S. X., Huang, G. Q., & Cheng, M. (2017). Truthful, budget-balanced bundle double auctions for carrier collaboration. *Transportation Science, 51*(4), 1365–1386.

Xu, S. X., Shao, S., Qu, T., Chen, J., & Huang, G. Q. (2018). Auction-based city logistics synchronization. *IISE Transactions, 50*(9), 837–851.

Zhang, J., Liu, F., Tang, J., & Li, Y. (2019). The online integrated order picking and delivery considering Pickers' learning effects for an O2O community supermarket. *Transportation Research Part E: Logistics and Transportation Review, 123*, 180–199.

Zhang, Y., Zhang, Z., Yang, Q., An, D., Li, D., & Li, C.e. (2020). EV charging bidding by multi-DQN reinforcement learning in electricity auction market. *Neurocomputing, 397*, 404–414.

Zhou, W., & Lin, J. (2019). An on-demand same-day delivery service using direct peer-to-peer transshipment strategies. *Networks and Spatial Economics, 19*(2), 409–443.