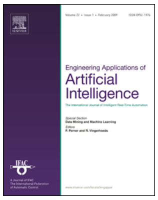
Engineering Applications of Artificial Intelligence 37 (2015) 207–220

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



Designing of an intelligent self-adaptive model for supply chain ordering management system

Ahmad Mortazavi [a], Alireza Arshadi Khamseh [a,*], Parham Azimi [b]

[a] *Industrial Engineering Department, Faculty of Engineering, Kharazmi University, Tehran, Iran*
[b] *Department of Industrial and Mechanical Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran*

ARTICLE INFO

Article history:
Received 14 December 2013
Received in revised form
4 September 2014
Accepted 4 September 2014
Available online 10 October 2014

Keywords:
Supply chain
Ordering policy
Agent based simulation
Reinforcement learning
Markov decision process
Simulation-based optimization

ABSTRACT

One of the challenging issues in supply chain management is the coordination of ordering processes, especially in dynamic situations. In recent years, reinforcement learning algorithms are considered to be efficient techniques for solving such problems. In this paper, an agent-based simulation technique has been integrated with a reinforcement learning algorithm and has been applied to model a four-echelon supply chain that faces non-stationary customer demands. This approach leads to the development of a novel and intelligent simulation-based optimization framework, which includes a detailed simulation modeling of supply chain behavior. Finally statistical methods, including the Var technique, are used for the risk evaluation and sensitivity analysis have been provided to support the decision making process under uncertainty.

1. Introduction

The literature on supply chain management entails decision making processes in several areas such as procurement, production, inventory control, distribution, and so on. Also the decision making process in a supply chain is performed under three perspectives of operational, tactical and strategic decisions, so the nature of a supply chain management system (SCMS) makes it as a complex problem which has attracted the attention of several researchers.

In a SCMS, inventory control is one of the most important problems which cover nearly 50% of the total costs of the supply chain (Lancioni, 2000). One of the most important decisions about inventory control is the supply chain ordering management (SCOM), which is the main focus of this research. In fact, SCOM is an integrated approach to determine the ordering policy of each supply chain actors. In such a problem, the purpose of the decision maker is to minimize inventory costs and satisfying customers demand with a high service level, simultaneously.

A supply chain consists of independent and interacting actors with non-linear behaviors which have the ability of decision making. Also, there are several uncertainties such as uncertainty in demand, lead time etc. which cause a supply chain to be an instance of a complex adaptive system (CAS). Thus, an agent based modeling (ABM) technique can be considered to be a powerful bottom up technique for modeling a supply chain (Macal and Michael, 2009), especially in SCOM problems (Chaharsooghi et al., 2008). The ABM technique has the ability to model the detailed dynamic behaviors of complex systems, and there are several successful reports of using ABM techniques in industrial societies. For instance, the Boeing Company applied ABM for automating of its service supply chain in the case of perishable products Brintrup et al., 2009). In another project Honeywell laboratories employed ABM in its dynamic distribution SCM (Wagner et al., 2003).

For a real supply chain faced with complex and stochastic demand patterns, which leads to a time-varying state that evolves over time; so the mathematical and ordinary optimization techniques are unable to solve such problems. An artificial intelligence method may be helpful here to model a learning mechanism and providing the self-adapting agents. In this research, the main effort is dedicated to embedding a reinforcement learning (RL) algorithm (Sutton and Barto, 1998) in an agent based simulation model to develop an intelligent learning model in order to solve a SCOM as a Markov decision process (MDP).

The paper is organized as follows. Section 2 provides the related literature review. Section 3 describes the problem and the modeling process, while Section 4 is dedicated to implementation of the reinforcement learning algorithm. Section 5 is about performance analysis of the problem. Finally, conclusions and suggestions for future studies are discussed in Section 6.

* Correspondence to: Industrial Engineering Department, Faculty of Engineering, Kharazmi University, No. 49, Mofateh Ave., Tehran 15719-14911, Iran.
E-mail addresses: ahd.mortezavi@gmail.com (A. Mortazavi), alireza.arshadikhamseh@gmail.com, arshadi.kh@khu.ac.ir (A. Arshadi Khamseh), p.azimi@qiau.ac.ir (P. Azimi).

http://dx.doi.org/10.1016/j.engappai.2014.09.004
0952-1976/© 2014 Elsevier Ltd. All rights reserved.

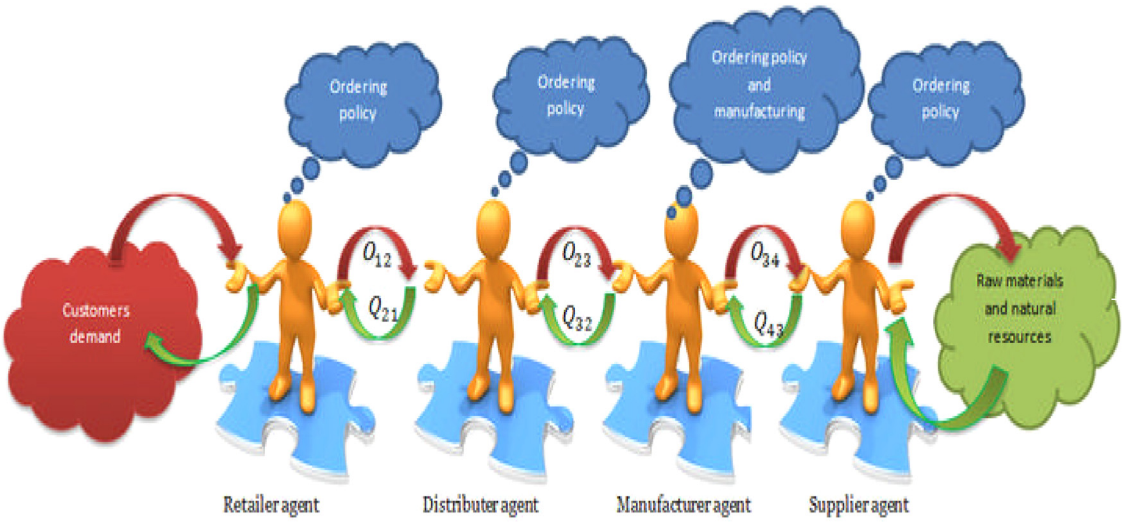


Fig. 1. Structure of agent based supply chain.

2. Literature review

As the main contributions of this paper are twofold, i.e. modeling of supply chain as a multi-agent system and developing of a RL algorithm for solving of SCOM as a Markov decision problem, we have categorized the relevant literature in two subsections including the agent based modeling of a supply chain and then, the application of reinforcement learning methods in the inventory optimization problems.

2.1. Agent based modeling of supply chain

ABM is a well-known architecture for modeling of distributed systems. In this paradigm, components of system are modeled as agents with the ability of decision making and communication. Such agents are autonomous and self-organizing with specific rules to make decisions. Thus, the ABM technique is suitable for modeling of complex systems with non-linear behavior.

To the best of our knowledge, there are two main areas in SCM in which ABM technique is applicable. The first area is scheduling and production systems (Leitão, 2009; Aissani et al., 2012; Trentesaux, 2009; Tehrani Nik Nejad et al., 2010) and the second area, which forms our focus in this paper, is the ordering policy. One of the first research works in agent based modeling and simulation of supply chain is presented by Swaminathan et al. (1998). In this research, ABM is employed for designing of a decision support system with the aim of reengineering of the supply chain. Agents represent the supply chain entities, i.e. the retailer, the distribution center, etc. while several inventory control policies are considered. Strader et al. (1998) applied ABM to simulate of an assembly supply chain which is associated with computer and electronics industries. They investigated the impact of information sharing on orders fulfillment. Later Fox et al. (2001) used ABM to develop two different supply chain architectures with the aim of supply chain coordination and managing of perturbation. In the other research, Pathak et al. (2004) used a simulation framework based on the ABM method for simulation of growth dynamic in an adaptive SCM.

In later research works, there are various models of supply chains supported by ABM technology. In some of these research works, financial decisions are investigated by the ABM approach. For instance, Sun et al. (2012) used agent based simulation for mitigating of a bankruptcy propagation through the supply chain by considering operational parameters in financial decision making, While Li et al. (2010) used agent based simulation for analyzing the dominant players' behavior of supply chains. The latter research showed that the profit stability is feasible in spite of decreasing the sales price.

Table 1
Mean of customers demand in 12 week period.

Week	1	2	3	4	5	6	7	8	9	10	11	12
Mean of daily demand	1	2	4	5	6	7	7	6	5	4	2	1

In recent years, many researchers have reported works in the area of inventory control. Wang et al. (2008) used agent based simulation to analyze the impact of radio frequency identification (RFID) technology on the improvement of an inventory system of LCD supply chain in Taiwan. Wang et al. (2011) developed a multi-agent model of supply chain to investigate (RFID) impact on inventory system based on total inventory cost, inventory turnover and the bullwhip effect.

Considering of highly dynamic customer demands, supply chain coordination, especially, ordering management, is important for optimum performance of supply chain. On the other hand, simulation can be considered to be an efficient tool for decision making and risk evaluation of supply chains under uncertainty (Galasso and Thierry, 2009). In this area, some researchers used simulation optimization techniques, including agent based simulation and meta-heuristic algorithms. For instance, Pan et al. (2009) used simulation optimization based on agent based simulation and genetic algorithm for obtaining the optimal reordering strategy in apparel supply chains which experience dynamic customer demands. In another work, Sinha et al. (2011) applied agent based simulation, assisted by the co-evolutionary particle swarm optimization (PSO) algorithm to coordinate a petroleum supply chains, while Brintrup (2010) employed NSGAII with a bi-objective function to optimize a supply chain using the ABS. In this research, maximization of supply chain revenue and minimization of lead time are considered to be objective functions while manufacturing policy and supplier selection are decision variables.

Although meta-heuristic algorithms are well known techniques for simulation optimization, there are several potentially suitable techniques in this area. Among other methods, RL is one of the most efficient techniques to solve dynamic optimization problems. Furthermore, RL is based on multi-agent modeling paradigms and hence it can be potentially advantageous for optimization of agent based simulation models. But using RL for agent based simulation optimization, is relatively rare in related literature, which is hence a major contribution of this paper.

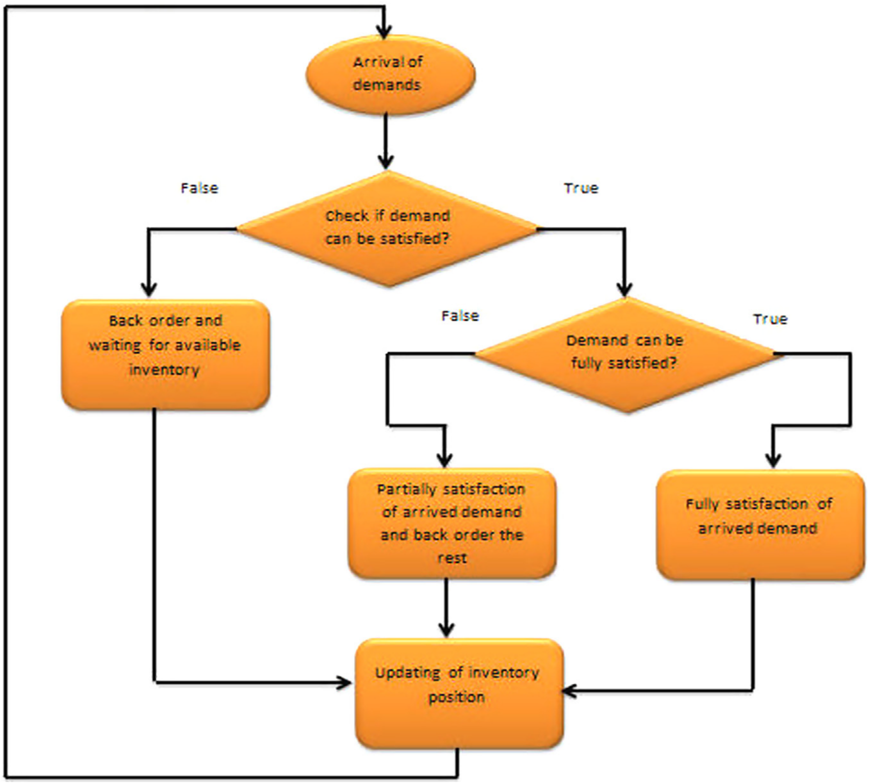

Arrival of demands
Check if demand can be satisfied?
False
True
Back order and waiting for available inventory
Demand can be fully satisfied?
False
True
Partially satisfaction of arrived demand and back order the rest
Fully satisfaction of arrived demand
Updating of inventory position

Fig. 2. Logic of retailer behavior.

2.2. Application of RL algorithm in inventory optimization

Reinforcement learning is an area of machine learning which is inspired by behaviorist psychology and is also based on stochastic dynamic programming. In an RL algorithm setting, software agents interact with the environment and based on the rewards or punishments received, learn near-optimum or optimum policies for selecting actions. RL is an applied solution methodology for a vast spectrum of problems such as those encountered in robotics (Duan et al., 2007), production scheduling (Wang and Usher, 2005), economics (Krause et al, 2006), etc.

Bertsekas and Tsitsiklis (1996) and Sutton and Barto (1998) explained reinforcement learning methodology, thoroughly. RL provides an integrated framework for a supply chain management (Pontrandolfo et al., 2002). In this area, Giannoccaro and Pontrandolfo (2002) modeled a SCOM as a MDP problem for a three-echelon supply chain. They considered holding, shortage, ordering and pipeline costs and also assumed that demand follows a stationary exponential distribution. Chaharsooghi et al. (2008) also applied a RL algorithm for solving a SCOM problem as a MDP on the so-called beer game. In their beer game model, the cost structure consists of holding and shortage costs and a deterministic sequence of numbers was considered as a demand pattern. They also compared the performance of RL with a heuristic algorithm based on GA. Kwak et al. (2009) applied a RL algorithm for a vendor managed inventory (VMI) system of two-echelon supply chain for minimizing the total inventory costs.

Unlike the former works, Jiang and Sheng (2009) considered learning an optimum inventory policy with the aim of service level satisfaction instead of cost minimization; so they proposed a case based reinforcement learning (CRL) algorithm for a two-echelon multi-retailer and multi-customer supply chain. Later, Kim et al. (2010) employed RL for attaining a satisfactory service level in n-echelon supply chain.

The contribution of our paper is threefold. First of all, an ABM model of a four-echelon supply chain is developed in detail. In the developed model, a complex and non-stationary stochastic pattern is used to model customer demand in realistic way. The second contribution is in developing and embedding a RL algorithm for learning of ordering policy. Finally risk analysis is performed based on a value at risk (VAR) technique (Hull, 2012), which can help perform an in-depth analysis of the supply chain's performance.

3. Problem description and modeling

A linear supply chain consisting of four echelons is considered in this paper. In each echelon, there is one actor which modeled as an agent. Actor agents are, the retailer, the distributer, the manufacturer and the supplier; which they are shown in Fig. 1 and their indexes are $i=1$, 2 ,3 and 4 ,respectively. For our supply chain model, there are three groups of variables which are described in what follows:

$S_i(t)$: state variable of agent i which is inventory position in time step of t, $i=1, 2, 3, 4$.

$O_{ij}(t)$: action variable of agent i which is order quantity of i-th agent to the upstream agent j, in time step of t, $i=1, 2, 3, 4$; $j=i+1$.

$m(t)$: amount of manufactured goods in time step of t in manufacturer agent.

$Q_{ij}(t)$: amount of quantity which agent i ships to downstream agent j in time step of t, $i=1, 2, 3, 4$, $j=i-1$.

I_{il}^{h}: amount of holding inventory for agent i in l-th day of week, $i=1, 2, 4$; $l=1, 2, 3,\ldots,7$.

I_{il}^{hr}: amount of holding raw inventory for agent i in l-th day of week, $i=3$; $l=1, 2, 3,\ldots,7$.

I_{il}^{hf}: amount of holding finished inventory for agent i in l-th day of week, $i=3$; $l=1, 2, 3,\ldots,7$.

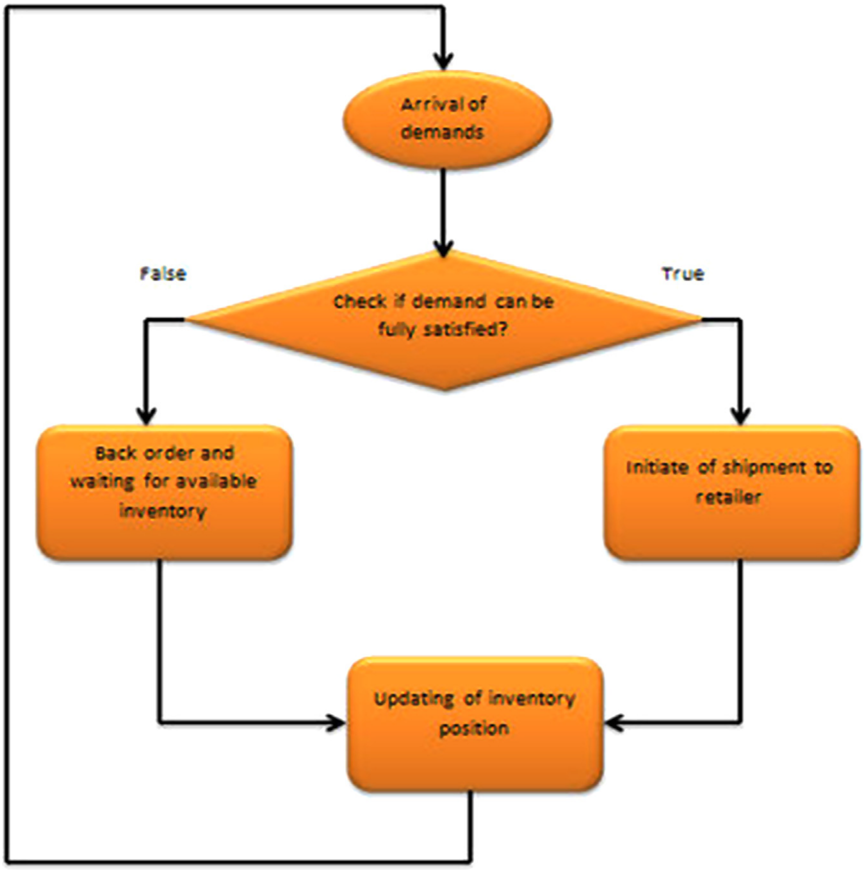

Arrival of demands
Check if demand can be fully satisfied?
False
True
Back order and waiting for available inventory
Initiate of shipment to retailer
Updating of inventory position

Fig. 3. Logic of distributor behavior.

I_{il}^{b}: amount of backlog inventory for agent i in l-th day of week, $i=1, 2, 3, 4$; $l=1, 2, 3,\ldots,7$.

Demand of customers follows a non-stationary distribution. Arriving of customers in the retailer is a Poisson process with a mean of 2 customers in each day – in a way that amount each customer demands follows a Poisson distribution with a variable mean. The mean of customer demand changes trough 12 weeks data are provided in Table 1.

The inventory system in the entire supply chain is based on a periodic reviewing policy. In this problem, each actor of supply chain reviews inventory position $S_i(t)$ (state variable) once a week, to determines his/her state and places an order (take action) with a variable quantity $O_{ij}(t)$ (action variable) to the upstream actor and also is responsible for the replenishment order of the downstream actor $Q_{ij}(t)$ with inventory in hand which may change his/her state for next time step. Modeling of each agent is done by any logic software and is described in details in the following:

3.1. Modeling of retailer

In the retailer module, there is a daily stream of customer demand. Each customer's demand should be satisfied immediately or it is treated as back order and incurs a penalty to the retailer. In this stage of the supply chain, partial demand satisfaction is permitted. In fact, with consideration of the inventory levels, the retailer satisfies entire demand; otherwise it satisfies a possible fraction of the demand and the customer waits until replenishment of the retailer inventory.

The retailer capacity is about 200 units of products and four cost are considered in the retailer agent. The holding cost C_h is about 1 for each unit of products in each day while the back order cost C_b is 5 for each unit of back order per day. The order setup cost C_o is 30 for each order and the transportation cost C_t is 3 for each unit of products. The lead time for this agent is zero, because there is no transportation for delivering customer demand. So, in the case where inventory is available, it is shipped to customers directly from the retailer. Fig. 2 shows the logic of retailer behavior.

According to Fig. 2 in the case when there is inventory in hand, the customer demand reduces inventory position and changes the state variable of the retailer agent. This agent also takes action by placing an order to the distributer agent.

3.2. Modeling of distributer

In the distributer agent, the order of the retailer arrives once every seven days (one week). Weekly orders of retailers affect both the retailer and the distributer state variables. Each order increases the retailer inventory position while decreasing distributor inventory position. This agent fulfills the order immediately; otherwise it treats them as back orders completely until the next replenishment of inventory. This agent has a capacity of 600 units of products and the lead time for delivering of orders to the retailer is a stochastic variable which has the uniform distribution: UNIF (0.25, 0.5) day. Like the retailer agent, there are four types of costs for the distributor agent: C_h is 1.25, C_b is 6, C_o is 40 and C_t is 4 units. The distributor agent also takes actions by placing orders weekly to the manufacturer agent. Fig. 3 shows the logic of distributor behavior.

3.3. Modeling of manufacturer

The manufacturer agent's behavior is more complex in comparison to that of other agents. This agent receives weekly orders of the distributor and has to fulfill received orders or pay a penalty for

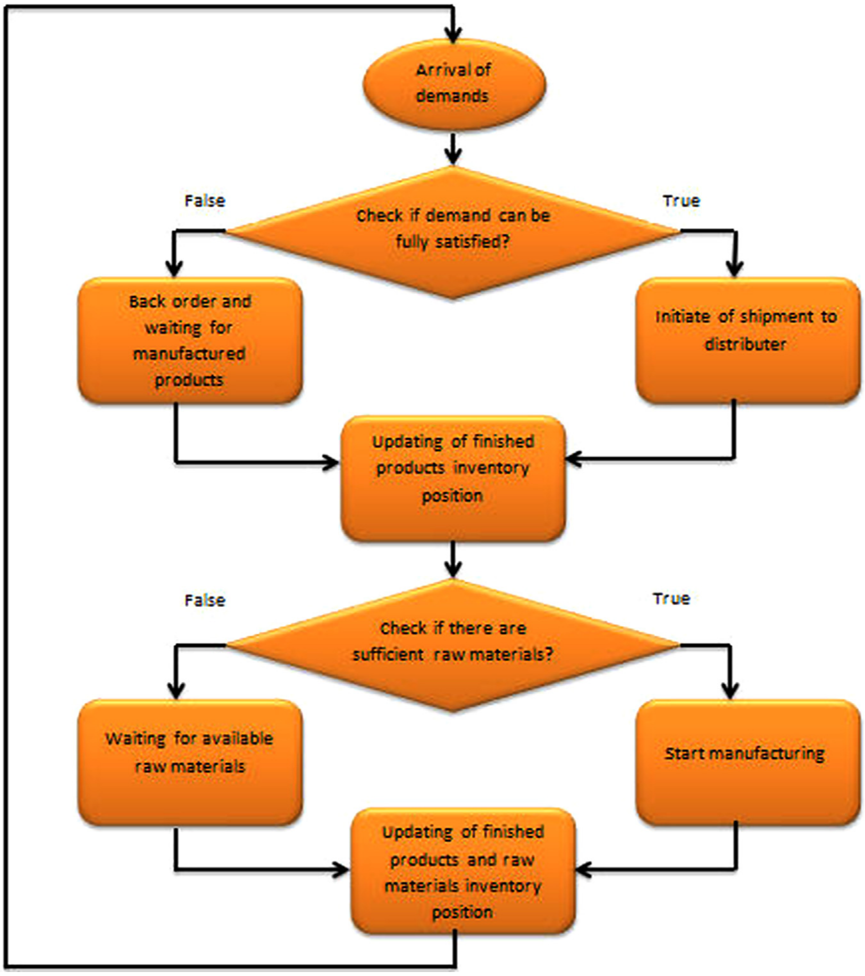

Arrival of demands
Check if demand can be fully satisfied?
False
True
Back order and waiting for manufactured products
Initiate of shipment to distributer
Updating of finished products inventory position
Check if there are sufficient raw materials?
False
True
Waiting for available raw materials
Start manufacturing
Updating of finished products and raw materials inventory position

Fig. 4. Logic of manufacturer behavior.

backordering. This agent has a capacity of 400 units of products. Unlike other agents, the manufacturer holds two different types of inventory. The first type is raw material and the second is of finished goods type, but for ordering, only raw material inventory is reviewed. Another difference between the manufacturer and the other agent is manufacturing operations which involve transferring raw material to finished goods, which are then available for fulfillment of distributer orders. So there are two types of lead times: the first lead time is for manufacturing operations and the second is dedicated to transportation. The manufacturing lead time is 1 h plus 0.01 h for manufacturing of each product and the transportation lead time is a random variable with a uniform distribution: UNIF (0.5, 1) day. But like the other agents, the manufacturer takes actions by placing orders to the upper level agent. On the other hand, fulfillment of received orders decreases its inventory position as state variable while placing of an order to the supplier increases inventory position. For this agent, the holding cost encompasses the holding cost for raw materials, C_{hr}, which is 0.5, and the holding cost for finished goods C_{hf} which is 0.75. The back order penalty C_b is 8 for each back order product in each day. The manufacturing cost C_m is 5 and the transportation cost C_t is 4. Finally, the manufacturing setup cost C_s is 50 and the setup cost for orders C_0 is 40. Fig. 4 explains the logic underlying manufacturer behavior.

3.4. Modeling of supplier

This agent is the representative of suppliers which provide raw material and parts. It has a capacity of 200 units and receives orders weekly which decrease inventory position as state variable. This agent satisfies orders with inventory in hand; otherwise it incurs a back order penalty. The supplier action is procuring and a specific procurement time is considered for providing the manufacturer agent with each raw part. The procurement time is a random number with a uniform distribution: UNIF (0.5, 1) min for each part. The supplier cost includes the holding, back order and transportation costs. C_h is 0.75, C_b is 4 and C_t is 5. Fig. 5 explains the supplier behavior logic.

4. Implementation of reinforcement learning algorithm

In this framework, the decision maker agent experiences several states. In this context, the selected action of the agent and the randomness of the environment together drive the decision maker to the new state. One of the well-known methods for solving MDPs, especially in large scale, is RL Alpaydin (2010). As shown in Fig. 6, the agent interacts with environment, which follows Markov property, via the RL algorithm. In the state s_t, the agent chooses action a_t; then the environment changes its state into state s_{t+1} and also provides the agent with reward r_{t+1}.

In this situation, the Markov property must exist and is formulized as follows (Gosavi, 2003):

$$Pr\{s_{t+1}=s', r_{t+1}=r|s_t, a_t, r_t, s_{t-1}, a_{t-1}, r_{t-1}, \dots, s_0, a_0, r_0\}$$
$$= Pr\{s_{t+1}=s', r_{t+1}=r|s_t, a_t\} \quad (1)$$

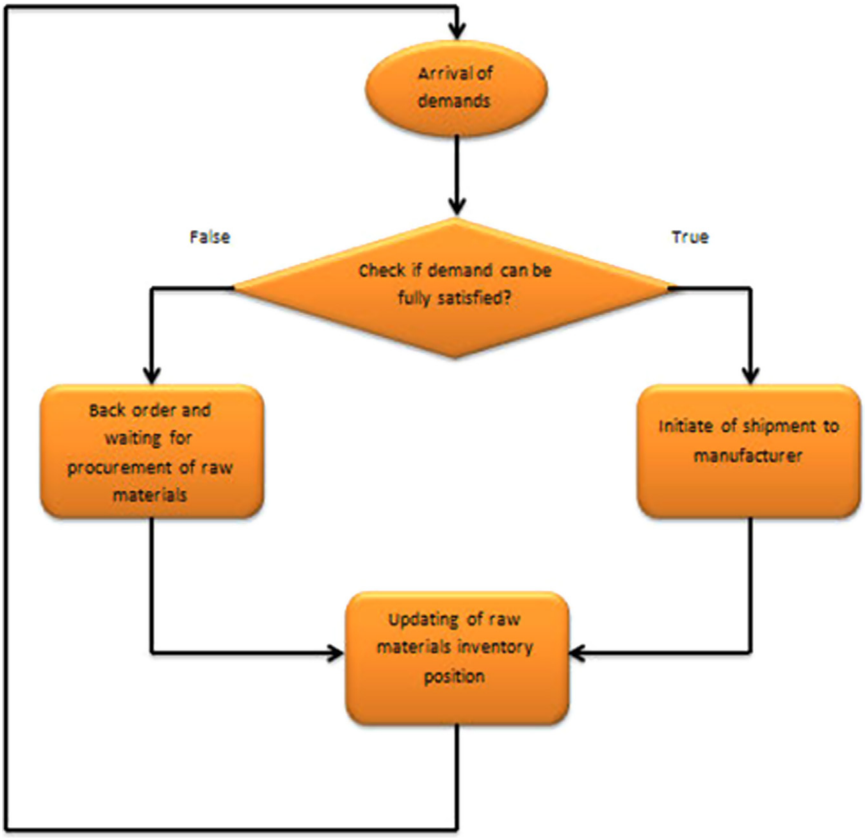

Arrival of demands
Check if demand can be fully satisfied?
False
True
Back order and waiting for procurement of raw materials
Initiate of shipment to manufacturer
Updating of raw materials inventory position

Fig. 5. Logic of supplier behavior.

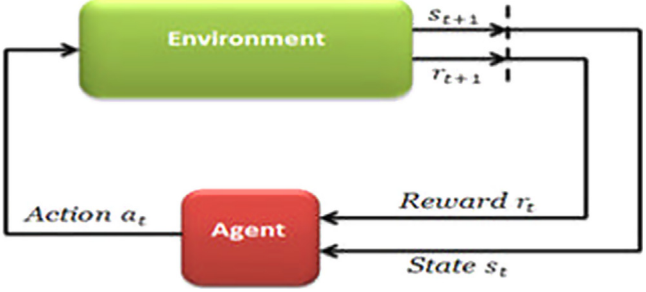


Fig. 6. Interaction of agent–environment based on RL algorithm.

The above expression implies that each state and the corresponding reward depend on the previous state and action, only.

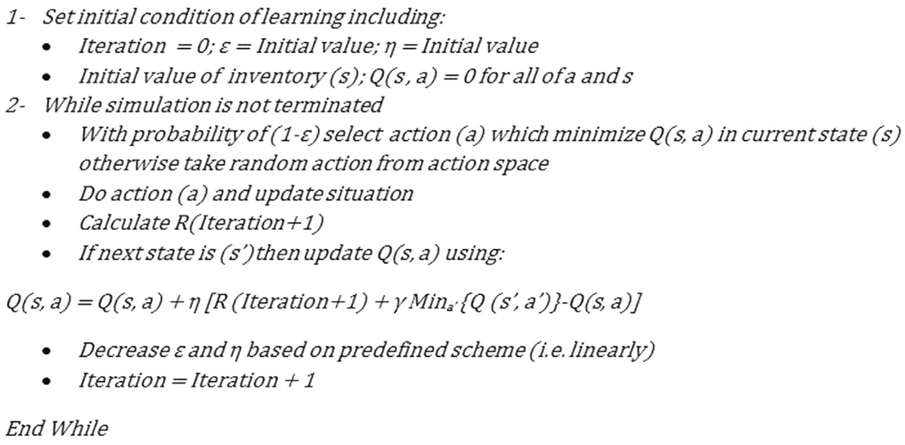
RL is a method for teaching the agent to learn what to do and how to map situations to actions for maximizing the numerical reward signal that can potentially help achieving an optimum policy. RL requires some mechanism and parameters for attaining this goal. In this framework, $P(s'|s,a)$ is the probability of arriving state s' by selecting action a in state s. In RL, the reward or punishment should be considered in a way that leads the agent toward its goal. Also, $Q(s,a)$ is the action-value function which accumulates the net reward or net punishment value to the future. Finally, an agent in each state takes actions subject to maximizing the net accumulated reward or minimizing the net accumulated punishment. Definition of the reward function (R_t) and $Q(s,a)$ are shown in the following equations:

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{2}$$

$$Q(s,a) = E\{R_t | s_t = s, a_t = a\} \tag{3}$$

The value of action a in state s is calculated based on expected value of future reward in trajectory which started in state s, action a is chosen in the first state, and the optimal policy is chosen thereafter for an infinite number of steps; γ is the discount coefficient between 0 and 1. Due to the difficulties in modeling of environment dynamics and calculation of expected value, the action-value function is estimated by approximation methods. In this research, we employed the Q-learning algorithm (Alpaydin, 2010) for estimation of $Q(s,a)$, and then the estimated values are used for agent action selection to achieve a desirable policy. In the implementation of RL we also considered γ equals to 1, because discounting is not considered in SCOM (Chaharsooghi et al., 2008).

One of the most challenging issues in the implementation of a RL algorithm is action selection in the learning phase. The agent needs to exploit its knowledge for selecting the best action in each specific state while it is also needed to explore a different action to evaluate new possibilities which may lead to the improvement of the agent policy. So, each agent needs a tradeoff between exploitation (action selection based on its knowledge) and exploration (random action selection). So in each step, with a probability of $P_{exploitation}$ agent select action according to its current knowledge about estimation of $Q(s,a)$ and with probability of $P_{exploration}$, selects an action randomly. While the summation of $P_{exploitation}$ and $P_{exploration}$ should be equal to 1 in each step, it is wise to get $P_{exploration}$ to be larger than $P_{exploitation}$ in the initial steps and gradually decrease $P_{exploration}$. One of the common methods for action selection based on exploration and exploitation, is the so-called ε-greedy policy (Alpaydin, 2010). The best action with respect to $Q(s,a)$ is called the greedy action. According to ε-greedy philosophy, in state s with a probability of ε ($P_{exploration}$), the agent selects any action except the best one (with the same probability for the non-greedy actions) and with a probability of $1-\varepsilon$ ($P_{exploitation}$) selects the greedy action. In this approach, ε is set to a value close to



```
1- Set initial condition of learning including:
    • Iteration $= 0$; $\varepsilon =$ Initial value; $\eta =$ Initial value
    • Initial value of inventory (s); $Q(s, a) = 0$ for all of a and s
2- While simulation is not terminated
    • With probability of $(1\text{-}\varepsilon)$ select action (a) which minimize $Q(s, a)$ in current state (s)
      otherwise take random action from action space
    • Do action (a) and update situation
    • Calculate $R(Iteration+1)$
    • If next state is (s') then update $Q(s, a)$ using:

$Q(s, a) = Q(s, a) + \eta\,[R\,(Iteration+1) + \gamma\,Min_{a'}\{Q\,(s', a')\}\text{-}Q(s, a)]$

    • Decrease $\varepsilon$ and $\eta$ based on predefined scheme (i.e. linearly)
    • Iteration $=$ Iteration $+ 1$

End While
```

Fig. 7. Developed algorithm based on Q-learning for solving SCOM problem.

Table 2
Coding of retailer states.

Inventory in hand	$(-\infty,-30]$	$(-30,-10]$	$(-10,0]$	(0,10]	(10,20]	(20,30]	(30,50]	$(50,\infty)$
State	1	2	3	4	5	6	7	8

Table 3
Coding of retailer actions.

Order quantity	10	20	30	40	80	100	160
Action	1	2	3	4	5	6	7

Table 4
Coding of distributer states.

Inventory in hand	$(-\infty,0]$	(0,30]	(30,50]	(50,80]	(80,110]	(110,140]	(140,160]	$(160,\infty)$
State	1	2	3	4	5	6	7	8

Table 5
Coding of distributer actions.

Order quantity	10	20	30	60	70	100	160
Action	1	2	3	4	5	6	7

Table 6
Coding of manufacturer states.

Inventory in hand	$(-\infty,0]$	(0,20]	(20,40]	(40,60]	(60,80]	(80,100]	(100,120]	$(120,\infty)$
State	1	2	3	4	5	6	7	8

Table 7
Coding of manufacturer actions.

Order quantity	20	30	40	60	80	100	120
Action	1	2	3	4	5	6	7

1 initially, and is gradually decreased to zero. The algorithm developed for the SCOM problem is shown in Fig. 7.

In this algorithm, the inventory and manufacturing costs are used for developing of punishment functions of $R(t)$. $R(t)$ is essentially the feedback which the agent receives from interaction with supply chain system in t-th time step. Hence, the agent should ideally achieve to the policy which minimizes accumulated punishment. As the actual searching space is infinite and it is impossible to

Table 8
Coding of supplier states.

Inventory in hand	$(-\infty,0]$	(0,20]	(20,40]	(40,60]	(60,80]	(80,100]	(100,120]	$(120,\infty)$
State	1	2	3	4	5	6	7	8

Table 9
Coding of supplier actions.

Order quantity	20	40	60	70	80	90	100
Action	1	2	3	4	5	6	7

Table 10
Action-value matrix for retailer agent.

	Action(1)	Action(2)	Action(3)	Action(4)	Action(5)	Action(6)	Action(7)
State(1)	578.9289	709.2339	721.497	613.3572	513.3801	469.6121	248.4729
State(2)	239.0474	215.8933	224.8062	195.4382	195.0247	201.9397	182.6577
State(3)	117.5855	93.7045	96.43031	85.28146	96.68859	90.10457	102.4766
State(4)	160.9015	141.2609	113.3442	118.5011	110.1733	128.3232	127.6145
State(5)	99.09235	129.8192	102.6289	100.6761	103.3072	106.2238	98.32939
State(6)	142.1675	124.7719	122.2973	135.2981	134.2852	141.8093	141.1105
State(7)	187.7201	195.0995	169.5437	182.3046	188.1912	184.9668	186.5244
State(8)	411.238	300.8895	381.6361	413.9192	436.705	430.6285	476.8614

Table 11
Action value-matrix for distributer agent.

	Action(1)	Action(2)	Action(3)	Action(4)	Action(5)	Action(6)	Action(7)
State(1)	54,537.6	53,114.83	53,484.32	53,157.31	53,386.63	53160.74	52,684.66
State(2)	65,701.69	66,248.92	59,593.88	65,550.95	66,197.58	64,182.42	67,252.23
State(3)	37,645.67	38,428.32	37,152.98	37,754.94	36,620.02	36,096.83	39,476.85
State(4)	49,421.98	48,068.84	57,738.81	47,721.3	47,675.09	48,740.18	47,919.58
State(5)	51,158.7	50,598.06	49,961.04	50,413.26	49,334.22	49,253.39	53,322.98
State(6)	45,940.12	44,849.9	45,531.2	40,235.98	44,900.75	45,966.6	45,779.96
State(7)	19,363.61	20,745.73	18,881.73	19,512.39	19,806.52	19,520.68	17,942.48
State(8)	5220.265	5370.955	5240.724	5173.607	5285.571	5236.23	5174.607

Table 12
Action-value matrix for manufacturer agent.

	Action(1)	Action(2)	Action(3)	Action(4)	Action(5)	Action(6)	Action(7)
State(1)	23,027.82	22,030.81	22,711.43	22,148.95	20,485.43	22,323.57	21,848.42
State(2)	28,384.15	28,372.42	28,123.75	24,408.93	25,379.38	28,495.07	28,694.38
State(3)	28,710.38	29,470.14	29,373.42	27,637.7	28,699.86	29,096.55	22,768.86
State(4)	22,406.73	22,084.46	24,747.68	23,038.39	17,611.55	22,941.38	22,447.46
State(5)	5199.444	7220.198	4885.826	4441.954	8437.157	9480.728	4331.759
State(6)	5719.183	6333.241	6848.294	8180.545	5576.718	7146.562	4678.928
State(7)	2388.405	2402.4	4663.871	2393.403	5628.425	3699.539	2367.315
State(8)	8661.735	4413.84	4443.653	6703.814	7889.036	5841.031	10328.51

search infinite space, we used matrix with eight rows and seven columns to discretize the problem state space into a finite one which can be handled by Q-learning. Based on the literature (Giannoccaro and Pontrandolfo, 2002) in this matrix, the rows are representative of states (inventory position) and columns are representative of action (quantity of order). Hence the arrays m and n represent value of action n in state m. Further, η is the learning coefficient (learning rate) set to a value between 0 and 1, which is decreased gradually to zero. This scheme causes the learning to influence the agent and less as time passes in the simulator.

4.1. Implementation of RL for retailer agent

As mentioned before, the Markov decision process for each agent includes eight states and in each state, the agent has seven actions to take. For the retailer agent, the simulation showed proper states and actions which are shown in Tables 2 and 3 respectively.

Table 13
Action-value matrix for supplier agent.

	Action(1)	Action(2)	Action(3)	Action(4)	Action(5)	Action(6)	Action(7)
State(1)	27,936.15	27,905.23	27,815.22	29,336.97	28,715.89	27,612.74	27,987.63
State(2)	32,009.59	31,047.3	32,002.52	31,536.22	32,746.76	31,723.1	31,575.18
State(3)	31,292.5	31,180.87	32,315.65	30,827.55	26,683.61	31,076.47	31,441.19
State(4)	24,770.74	26,437.97	22,267.29	26,229.04	27,170.49	25,698.89	28,403.42
State(5)	8458.919	7599.912	11,182.6	11030.33	8316.997	15,376.53	6389.806
State(6)	9059.088	7568.501	6537.433	13972.88	11159.19	6423.871	8846.496
State(7)	5674.202	4492.777	8179.035	5385.62	3923.054	6383.123	5319.235
State(8)	4985.185	5092.984	5156.244	5140.385	5166.063	4990.229	5075.646

The retailer agent faces four costs of C_h, C_b, C_o, C_t, but as shown in Table 3, there is no action with zero quantity, so C_o (order setup cost) cannot influence the ordering policy. For this reason, although C_o is considered for calculation of the total supply chain cost, it is not included in the punishment function. The retailer punishment function is as shown in the following equation:

$$R(t) = C_t O_{12}(t) + \sum_{l=1}^{7} C_h I_{1l}^{h} + \sum_{l=1}^{7} C_b I_{1l}^{b} \tag{4}$$

The initial learning coefficient (η) for the retailer agent is 0.21 and in each step it is decreased by 2.55×10^{-4}; also the initial value of ε is 0.98 and is decreased by 1.88×10^{-3} in each step.

4.2. Implementation of RL for distributer agent

The state and action values are determined from running the simulation and are shown in Tables 4 and 5 respectively.

The distributor agent's holding, backlog and transportation costs are used in the distributer agent's punishment function, and the corresponding function is presented in the following equation:

$$R(t) = C_t O_{23}(t) + \sum_{l=1}^{7} C_h I_{2l}^{h} + \sum_{l=1}^{7} C_b I_{2l}^{b} \tag{5}$$

For this agent, the initial value of η is 0.19 and the initial value of ε is 0.95. The first parameter is decreased by 1.91×10^{-4} while the second parameter decreasing rate is 1.21×10^{-3} in each step.

4.3. Implementation of RL for manufacturer agent

Result from the manufacturer's states and actions are presented in Tables 6 and 7

Despite the similarities between a manufacturer's states and actions with the other agents, costs and inventory structure are different. The manufacturer agent deals with two kinds of inventories: raw material and finished goods. It is also responsible for manufacturing goods that complicates the behavior of the manufacturer agent. With consideration of such differences, its punishment function is presented in Eq. (6) as follows:

$$R(t) = C_t O_{34}(t) + C_m m(t) + \sum_{l=1}^{7} C_{hr} I_{3l}^{hr} + \sum_{l=1}^{7} C_{hf} I_{3l}^{hf} + \sum_{l=1}^{7} C_b I_{3l}^{b} \tag{6}$$

In this agent, the initial values for η and ε are 0.19 and 0.95 respectively and they are decreased by 1.91×10^{-4} and 1.21×10^{-3} in each step.

4.4. Implementation of RL for supplier agent

Similar to other four agents, suitable values of states and actions are determined for the supplier agent – based on simulation results, and the results are shown in Tables 8 and 9.

The supplier punishment function is formed based on holding, backlog and transportation costs and is shown via the following equation:

$$R(t) = C_t O_{45}(t) + \sum_{l=1}^{7} C_h I_{4l}^{h} + \sum_{l=1}^{7} C_b I_{4l}^{b} \tag{7}$$

In Eq. (7), $O_{45}(t)$ is the representative of the size of order in the t-th time step for the supplier, which is satisfied by sources that are out of the designed model, and when the index j equals to 5, the supplier is virtual.

Initial values for η and ε are 0.19 and 0.95 and also they are decreased by 1.91×10^{-4} and 1.21×10^{-3} in each step.

5. Performance analysis

The any logic software was used to simulate the problem. In the learning phase, the developed model was simulated for 15 years to estimate the action-value function. Then estimated values were used for action selection based on the policy generated to solve the SCOM problem. For statistical performance analysis of the supply chain, the simulation model of the supply chain was run for 5 years and also the simulation outputs were used for risk evaluation purpose.

5.1. Simulation output

The simulation output in this research has two components: the first for the training phase of the RL algorithm and the second for statistical analysis and risk evaluation. For the training phase, we considered one specific matrix with eight rows and seven columns for each agent. In the beginning of the simulation, all elements of the matrix is set to zero, and then agents experience a variety of scenarios through the trial and error in the simulation and update their knowledge of values of different actions in each state. After running the 15-year simulation training phase, the estimation of action-value functions are available for each agent. Action-value matrices for each agent are shown in Tables 10–13.

In the above-mentioned matrices, the rows denote states and columns denote actions, so for each row, the column with the least value is the preferred action for the state concerned. For instance, in Table 10 action(7) is preferred in state(1) and state(2) while actions 6, 5, 7, 3, 4 and 2 are suitable for states 3, 4, 5, 6, 7 and 8 respectively. According to Tables 11–13, the action-value matrices for the other agents are similar to the retailer agent, and there is one optimum action with least value for each state.

In the second simulation experiment, conducted for a five-year duration, the action-value matrixes are used for action selection by agents and then the output information is gathered for analyzing the supply chain performance. The pie charts in Fig. 8 show the cost distribution in different agents of the supply chain.

According to Fig. 8, most of expenses in retailer are dedicated to ordering while the minority of costs is related to shortage of goods. This fact implies that the learned ordering policy is effective in dealing with fluctuating demand of customers. Also most expenses

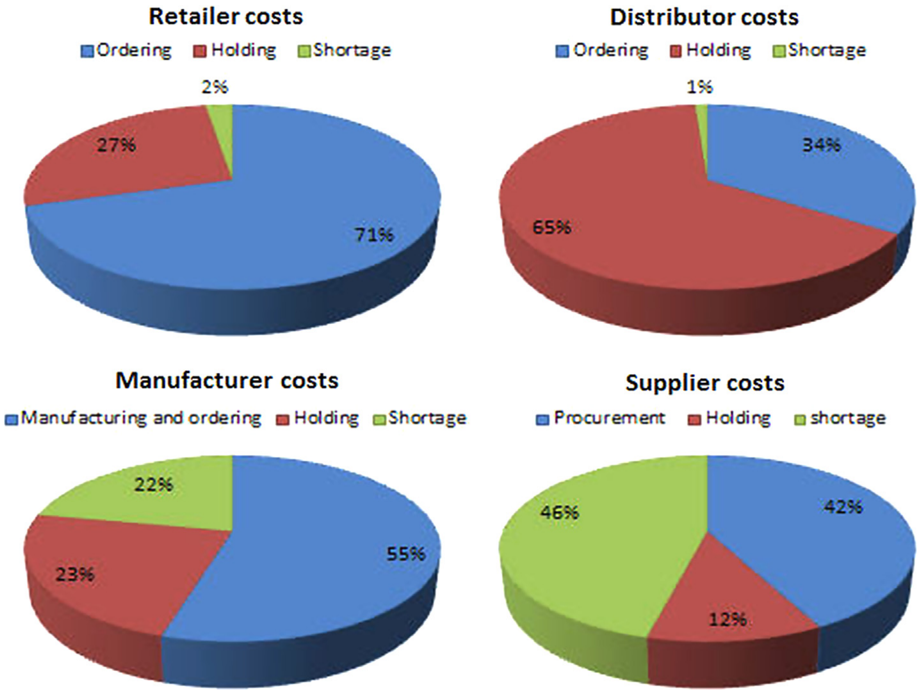


Fig. 8. Cost distribution in different agents of supply chain.

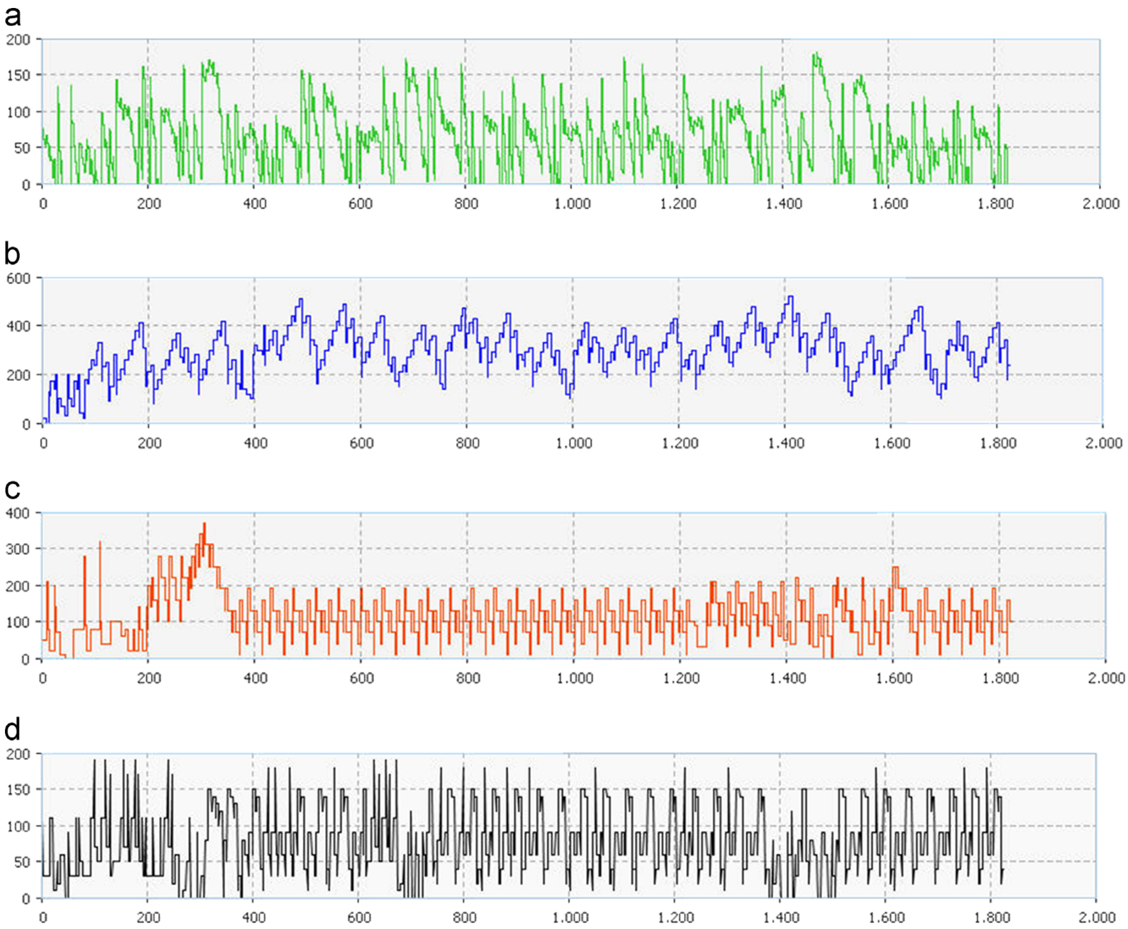


Fig. 9. (a) Daily inventory level of retailer agent in a five-year time interval, (b) daily inventory level of distributer agent in five year-time interval, (c) daily inventory level of manufacturer agent in five year-time interval, and (d) daily inventory level of supplier agent in five year-time interval.

in the distributor, the manufacturer and the supplier agents related to holding, manufacturing and shortage, respectively. Although the supplier agent faces a high ratio of shortage, it is not propagated and the supply chain and customer demands are satisfied with a low ratio of shortage. Daily inventory levels for each agent during a five-year interval are shown in Fig. 9(a–d).

Though non-stationary demand of customers leads to different fluctuating patterns of the inventory levels in supply chain agents, Fig. 9 shows that these patterns are stable. In this situation, stability of inventory level patterns can imply rationality of agents in decision making about ordering management.

Table 14
Result of distribution fitting for supply chain total cost.

Function	Sq error
Lognormal	0.0109
Weibull	0.0158
Beta	0.0193
Erlang	0.0415
Exponential	0.0415
Gamma	0.0415
Normal	0.331
Triangular	0.405
Uniform	0.479

Table 15
Result of distribution for average of customer waiting time.

Function	Sq error
Lognormal	0.0465
Erlang	0.0769
Gamma	0.0777
Weibull	0.127
Beta	0.14
Normal	0.161
Triangular	0.178
Exponential	0.293
Uniform	0.308

5.2. Statistical analysis of the supply chain performance

For statistical analysis of the supply chain performance, 50 simulation experiments with a five-year duration are conducted. In each simulation run, the total cost of the supply chain for a five-year time interval and the average customer waiting time are recorded. Result of distribution fitting for total costs and average waiting time are presented in Tables 14 and 15

As shown via Tables 14 and 15, criterion for ranking of the probability distribution function is square of errors. In both tables, the lognormal distribution is the best choice according to this criterion. The lognormal distribution has two parameters i.e. mean (μ) and standard deviation (σ) which determine shape and scale of the fitted distribution. In the process of distribution fitting, estimation of μ and σ were 8.9558 and 0.4822, respectively for the supply chain total cost distribution. Corresponding values for average customer waiting time are 2.1726 and 0.2655. Via their estimated values, the lognormal distribution functions for supply chain performance criteria are presented by Eqs. (8) and (9). Here, x_1 is the representative variable for supply chain total cost and x_2 represents average customer waiting time.

$$f(x_1) = \frac{1}{x_1 0.4822\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln x_1 - 8.9558}{0.4822}\right)^2\right) \tag{8}$$

$$f(x_2) = \frac{1}{x_2 0.2655\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln x_2 - 2.1726}{0.2655}\right)^2\right) \tag{9}$$

For detailed investigation of recorded samples, scattering pattern of simulation records are shown in Fig. 10

According to Fig. 10, the scattering pattern of samples implies correlation between the supply chain total cost and the average customer waiting time. The correlation parameter (ρ) is 0.6841. Considering ρ, Eqs. (8) and (9) the bivariate lognormal distribution can be applied to address the stochastic behaviors of the supply chain. Eq. (10) describes the mathematical function of the fitted bivariate lognormal distribution and Fig. 11 shows its graphical form. In Eq. (10), index 1 refers to parameter/variable of supply chain total cost while 2 refers to parameter/variable of the average customer waiting time.

The coefficients of variation (CV) are calculated for both the supply chain total cost and the average customer waiting time. The CV value for supply chain total cost is 74.19 while corresponding

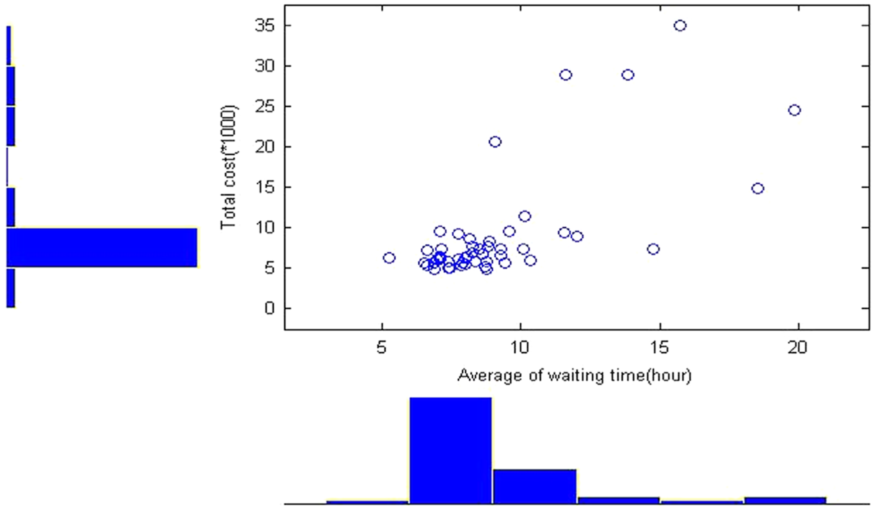


Fig. 10. Scattering of recorded samples according to supply chain total cost and average of customer waiting time.

value for the average customer waiting time is 32.00. As CV is a scale-less criterion for comparison of variation, it shows lower variation of x_2 in comparison with x_1. So a conditional bi-variate distribution and a given rough estimation of customer waiting time can improve estimation of the financial risk associated with a supply chain.

$$f(x_1,x_2)=\frac{1}{2\pi x_1 x_2 \sigma_1 \sigma_2 \sqrt{1-\rho^2}}\left\{\exp\frac{-1\times\left[(\ln(x_1-\mu_1)/\sigma_1)^2-2\rho(\ln(x_1-\mu_1)/\sigma_1)(\ln(x_2-\mu_2)/\sigma_2)+(\ln(x_2-\mu_2)/\sigma_2)^2\right]}{2(1-\rho^2)}\right\} \tag{10}$$

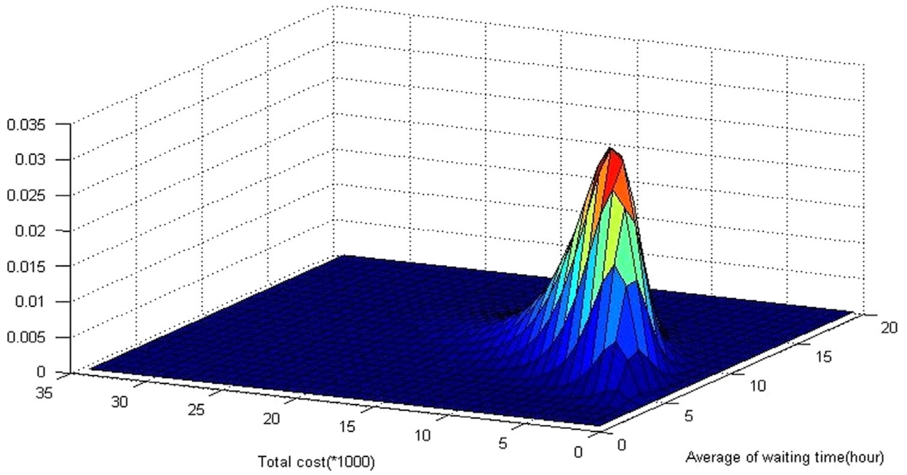


Fig. 11. Bi-variate distribution of supply chain performance.

Table 16
Financial risk of supply chain over five years interval.

CumProbability (%)	30	50	70	80	90	95	99
Total cost	6020.6	7752.7	9983.3	11633	14382	17136	23803

Table 17
Risk of customers waiting time over five years interval.

CumProbability (%)	30	50	70	80	90	95	99
Average of waiting time	7.6398	8.7811	10.0929	10.9797	12.3401	13.5897	16.285

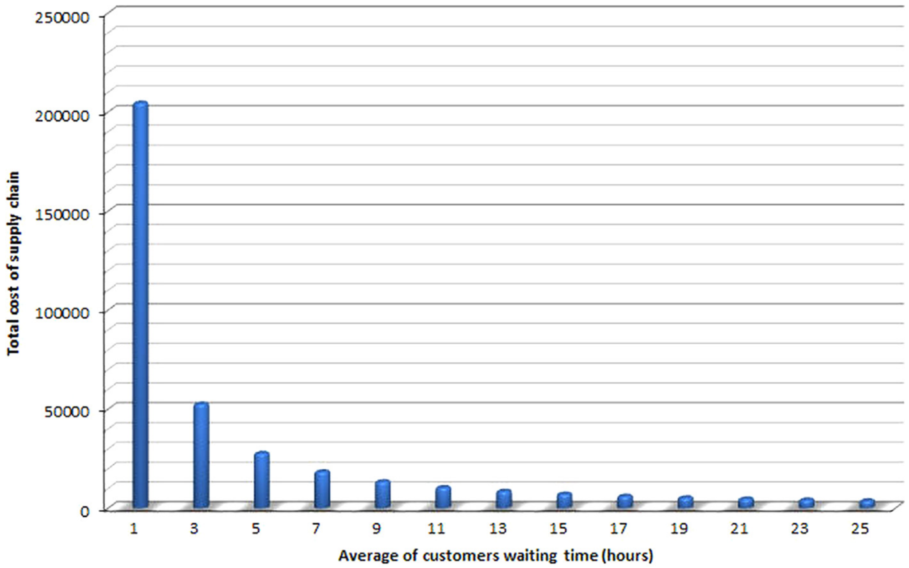


Fig. 12. Conditional financial risk of supply chain for different customer waiting time with 95% confidence interval.

5.3. Risk evaluation of supply chain performance

Risk evaluation is one of the most important tools in decision making under uncertainty and it also improves insight of decision makers to analyze expected outcomes. In this section, risk evaluation of supply chain performance is applied for the supply chain cost and customers' waiting time. Also according to the correlation of the supply chain cost and customers' waiting time, conditional risk evaluation has been performed.

Considering the probability distribution function of the supply chain total cost, which is presented by Eq. (8), and is based on the value at risk (Var) technique, the financial risk of the supply chain for a five-year time period is shown in Table 16.

Table 16 shows the risks of different costs. For instance, with a chance of 30%, the maximum possible cost in five years does not exceed of 6020.6 units of cost, while there is only 1% of risk that the total supply chain cost in five years exceeds 23,803 units.

In a similar manner, the risk of customers' waiting time is presented in Table 17.

Table 17 shows that there is only 1% of risk that customers' wait more than 16 h for their order delivery. This result implies that the considered supply chain is quite responsive with high service level.

For conditional risk evaluation of the supply chain with a given customer waiting time, the conditional distribution function is needed. In this case, we need to calculate $\mu_{x_1|x_2}$ and also $\sigma_{x_1|x_2}$ which are driven by the following equations:

$$\mu_{x_1|x_2} = \mu_1 - \left(\rho\frac{\sigma_1}{\sigma_2}[\ln x_2 - \mu_2]\right) \tag{11}$$

$$\sigma_{x_1|x_2} = \sigma_1\sqrt{(1-\rho^2)} \tag{12}$$

Finally, the conditional distribution of the supply chain total cost, with given customer waiting time is as follows:

$$f(x_1|x_2) = \frac{1}{x_1\sigma_{x_1|x_2}\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{\ln x_1 - \mu_{x_1|x_2}}{\sigma_{x_1|x_2}}\right)^2\right) \tag{13}$$

Based on Eq. (13) and for different given values of the customer waiting time, the financial risk of supply chain for five-year time interval with 95% confidence interval is as shown in Fig. 12.

Fig. 12 shows, with a 95% of confidence interval, the different financial risks for different waiting times. For example, for 1 h waiting time, the total cost of supply chain is supposed to be around 200,000, but, when we consider 3 h of waiting time, the total cost will exceed more than 50,000 units in only 5% of times. In this case, with acceptance of two more hours of waiting time, it is possible to decrease 150,000 units of the total cost.

6. Conclusions and future studies

In this research, an agent based modeling (ABM) paradigm is applied for modeling of a 4-echelon supply chain which faces fluctuations and stochastic customer demands. Then the supply chain ordering management (SCOM) problem was modeled as a Markov decision process. Due to the dynamic and time-varying nature of the system, the reinforcement learning (RL) algorithm was embedded within the simulation model of the supply chain to form an intelligent learning model, which is the main contribution of the paper. In the proposed model, several components of the supply chain, including the retailer, the distributer, the manufacturer and the supplier that learned a near-optimal policy of the inventory ordering system, were examined by developing and testing a variety of scenarios and the interactions among the agents. We also applied statistical analysis for a detailed evaluation of the supply chain performance criterion including supply chain total costs and the average customers' waiting time. Further, risk evaluation was conducted to obtain additional insights on the performance of the supply chain.

We believe that there are numerous avenues for the future researches. For instance, using a hybrid model, combining an agent based and the system dynamics, could be considered for evaluation of word-of-mouth impacts on the performance of the supply chain. On the other hand, the RL algorithm can also be applied for pricing strategies in a supply chain. Furthermore, there are many variants of RL algorithms, which may potentially lead to additional improvements of the supply chain behaviors in a dynamic situation.

Acknowledgment

Authors are grateful to anonymous referees for their valuable comments that improved quality of this research.

References

Alpaydin, E., 2010. Introduction to Machine Learning, 2nd ed. MIT press, Massachusetts.

Aissani, N., Bekrar, A., Trentesaux, D., Beldjilali, B., 2012. Dynamic scheduling for multi-site companies: a decisional approach based on reinforcement multi-agent learning. J. Intell. Manuf. 23, 2513–2529.

Bertsekas, D.P., Tsitsiklis, J., 1996. Neuro-Dynamic Programming.

Brintrup, A., Ranasinghe, D., McFarlane, D., Parlikad, A.K.N., 2009. International Academy of Production Engineering Conference, Cranfield, UK, 323–331.

Brintrup, A., 2010. Behaviour adaptation in the multi-agent, multi- objective and multi-role supply chain. Comput. Ind. 61, 636–645.

Chaharsooghi, K., Heydari, J., Zegordi, H., 2008. A reinforcement learning model for supply chain ordering management: an application to the beer game. Decis. Support Syst. 45, 949–959.

Duan, Y., Liu, Q., Xu, X., 2007. Application of reinforcement learning in robot soccer. Eng. Appl. Artif. Intell. 20, 936–950.

Fox, M.S., Barbuceanu, M., Teigen, R., 2001. Agent-oriented supply-chain management. Inf.-Based Manuf. 1, 81–104.

Giannoccaro, I., Pontrandolfo, P., 2002. Inventory management in supply chains: a reinforcement learning approach. Int. J. Prod. Econ. 78, 153–161.

Gosavi, A., 2003. Simulation-based Optimization: Parametric Optimization Techniques and Reinforcement Learning. Springer, Dordrecht.

Galasso, F., Thierry, C., 2009. Design of cooperative processes in a customer–supplier relationship: an approach based on simulation and decision theory. Eng. Appl. Artif. Intell. 22, 865–881.

Hull, J., 2012. Options, Futures, and Other Derivatives, 8th ed. Prentice Hall PTR, New jersey.

Jiang, C., Sheng, Z., 2009. Case-based reinforcement learning for dynamic inventory control in a multi-agent supply-chain system. Expert Syst. Appl. 36, 6520–6526.

Krause, T., et al., 2006. A comparison of Nash equilibria analysis and agent-based modelling for power markets. Int. J. Electr. Power Energy Syst. 28, 599–607.

Kwak, C., Choi, J.S., Kim, C.O., Kwon, I.-H., 2009. Situation reactive approach to vendor managed inventory problem. Expert Syst. Appl. 36, 9039–9045.

Kim, C.O., Kwon, I.-H., Kwak, C., 2010. Multi-agent based distributed inventory control model. Expert Syst. Appl. 37, 5186–5191.

Lancioni, R.A., 2000. New developments in supply chain management for the millennium. Ind. Mark. Manag. 29, 1–6.

Leitão, P., 2009. Agent-based distributed manufacturing control: a state-of-the-art survey. Eng. Appl. Artif Intell. 22, 979–991.

Li, J., Sheng, Z., Liu, H., 2010. Multi-agent simulation for the dominant players' behavior in supply chains. Simul. Model. Pract. Theory 18, 850–859.

Macal, C.M., Michael, J.N., 2009. Agent-based modeling and simulation. In: Winter Simulation Conference, pp. 86–98.

Pontrandolfo, P., Gosavi, A., Okogbaa, O., Das, T., 2002. Global supply chain management: a reinforcement learning approach. Int. J. Prod. Res. 40, 1299–1317.

Pathak, S. D., Dilts, D. M., Biswas, G., 2004. Simulating growth dynamics in complex adaptive supply networks. Simulation Conference, 2004. In: Proceedings of the 2004 Winter, pp.774–782.

Pan, A., Leung, S., Moon, K., Yeung, K., 2009. Optimal reorder decision-making in the agent-based apparel supply chain. Expert Syst. Appl. 36, 8571–8581.

Sutton, R.S., Barto, A.G., 1998. Introduction to Reinforcement Learning. MIT Press, Massachusetts.

Strader, T.J., Lin, F.-R., Shaw, M.J., 1998. Simulation of order fulfillment in divergent assembly supply chains. J. Artif. Soc. Soc. Simul. 1, 36–37.

Swaminathan, J.M., Smith, S.F., Sadeh, N.M., 1998. Modeling supply chain dynamics: a multiagent approach. Decis. Sci. 29, 607–632.

Sinha, A.K., Aditya, H., Tiwari, M., Chan, F., 2011. Agent oriented petroleum supply chain coordination: co-evolutionary particle swarm optimization based approach. Expert Syst. Appl. 38, 6132–6145.

Sun, Y., Xu, X., Hua, Z., 2012. Mitigating bankruptcy propagation through contractual incentive schemes. Decis. Support Syst. 53, 634–645.

Trentesaux, D., 2009. Distributed control of production systems. Eng. Appl. Artif. Intell. 22, 971–978.

Tehrani Nik Nejad, H., Sugimura, N., Iwamura, K., Tanimizu, Y., 2010. Multiagent architecture for dynamic incremental process planning in the flexible manufacturing system. J. Intell. Manuf. 21, 487–499.

Wagner, T., Guralnik, V., Phelps, J., 2003. enabling dynamic distributed supply chain management. Electron. Commer. Res. Appl. 2, 114–132.

Wang, Y.-C., Usher, J.M., 2005. Application of reinforcement learning for agent-based production scheduling. Eng. Appl. Artif. Intell. 18, 73–82.

Wang, S.-J., Liu, S.-F., Wang, W.-L., 2008. The simulated impact of RFID-enabled supply chain on pull-based inventory replenishment in TFT-LCD industry. Int. J. Prod. Econ. 112, 570–586.

Wang, S.-J., Wang, W.-L., Huang, C.-T., Chen, S.-C., 2011. Improving inventory effectiveness in RFID-enabled global supply chain with Grey forecasting model. J. Strateg. Inf. Syst. 20, 307–322.