

RESEARCH

Open Access



Research on supply chain efficiency optimization algorithm based on reinforcement learning

Tao Zhou¹, Lihua Xie¹, Chunbin Zou¹ and Yong Tian^{1*}

*Correspondence:

yong_tian34@outlook.com

¹Sichuan China Tobacco Industry Co., Ltd, Chengdu 610016, Sichuan, China

Abstract

Supply chain efficiency is critical to enterprises and can affect their competitiveness. The supply chain faces an uncertain and complex external market environment, facing the problem of supply chain efficiency optimization; the traditional optimization method is ineffective, which can better face the current environment and deal with problems. It has advantages in optimizing supply chain efficiency and has been widely used. This paper first expounds on the importance of supply chain management status, the limitations of traditional supply chain management methods, and reinforcement learning in the application of supply chain optimization. Then, through experiments, reinforcement learning, supply chain optimization problems, and the analysis of related algorithm design, the optimal algorithm focuses on inventory management optimization. Finally, this paper points out the future research directions and development trend of the supply chain efficiency optimization algorithm based on reinforcement learning.

Keywords: Supply chain efficiency optimization; Reinforcement learning; Q-learning algorithm; Inventory management optimization

1 Introduction

1.1 Importance of supply chain management

Supply chain management is very significant for enterprise operations. It can be explained from the following aspects: improving the market competitiveness of enterprises; high efficiency of supply chain management, which can directly reduce the enterprise production, storage and transportation costs, improve the product quality and product supply ability, effectively meet the customer demand and customer experience, improve customer satisfaction and enterprise brand value, and let enterprises get enough market competitive advantage in the industry; improving supply chain management by integrating sophisticated technologies, optimizing procedures, encouraging cooperation, enhancing demand forecasting, implementing sustainable principles, and promoting agility and continuous improvement. These tactics increase efficiency and competitiveness. He et al. [1] studied agent reinforcement learning based on the depth of q network. Efficient supply chain

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

management reduces costs by avoiding waste and optimizing logistics while improving product quality via better coordination and monitoring. This results in faster market responses, higher consumer satisfaction, and a competitive edge.

A multiobjective optimization system was developed, which relies on reinforcement learning (RL) and random forest (RF) algorithms and is used for textile manufacturing quality, productivity and cost control, maximizing the rewards of all the subjects, the target optimization problem as a Markov game paradigm, realizing the textile process-related equilibrium optimal solution, help enterprises to optimize the product and overall process performance. Random forest algorithms boost productivity in textile manufacturing by enabling predictive analytics for demand forecasting, improving quality control by identifying defect patterns, optimizing processes through variable evaluation, facilitating data-driven decision-making, and adapting to changing data patterns, resulting in increased efficiency and reduced waste.

Coordinate supply chain management and control of inventory save enterprise funds, promote the sustainable development of enterprises, avoid excessive or insufficient inventory, reduce waste, improve the ability of resource sharing and resource utilization, and promote the green development of our enterprises.

There is a need to enhance the response-ability of enterprises to make decisions and the ability to resist risks. Through the integration of the supply chain, enterprises can timely perceive and respond to the changes in supply and demand in the market, reasonable supply chain layout, disperse risks, form a mutual assistance mechanism between various links, and improve the flexibility of market decision-making and the overall risk resistance. Predictive modeling in supply chain management has substantial benefits, including increased efficiency and responsiveness to market swings. It enables precise demand forecasting by evaluating previous data and recognizing trends, allowing firms to align inventory levels with anticipated customer demand, eliminating stockouts and excess inventory. Achamrah et al. [2] also proposed a new deep reinforcement learning algorithm based on an artificial immune system to promote the storage, transportation, processing, and protection of products in the supply chain. To resolve the problem of returnable transportation goods, Q-learning and Deep Q-Learning algorithms are described in detail, and CPLEX is used to solve three models developed for SM, DM, and IRPPDS modes utilizing CPLEX. The results show that the method promotes mutual assistance between various links and the algorithm reduces cost and improves productivity.

1.2 Limitations and improvement methods of the traditional supply chain management methods

The traditional supply chain management method is not mature; the lack of adequate information sharing and communication mechanisms results in the information between different participants in the supply chain not being synchronized, and the work coordination is not strong. Supply chain management (SCM) improves corporate competitiveness by managing the flow of commodities, information, and funds along the supply chain. Effective SCM lowers production, storage, and transportation costs, increasing operational efficiency. In addition, the traditional method is mainly for manual operation and paper documents, which is inefficient and prone to errors. So, big data, artificial intelligence, and new technologies such as blockchain must be introduced to improve efficiency and accuracy. Ali et al. [3] expounds on the meaning of supply chain collaboration and the

use of machine learning technology in supply chain collaboration, establishing an intelligent simulation model of the supply chain, making the schedule, process route, truck loading, and delivery quotation intelligent automation. It describes all the participants in the supply chain, such as suppliers, manufacturers, wholesalers, retailers, customers, etc., enabling the subjects in the supply chain to realize information sharing.

Supply chain However, traditional supply chain management methods lack measures to deal with risks. Mahmud et al. [4] propose a multiobjective optimization model. In an integrated SC scheduling problem (ISCSP), manufacturers, suppliers, and batch decisions are optimized simultaneously to respond to customer needs, which ensures higher flexibility in process routing, reduces inventory cost, and gains a competitive advantage in the market. Cost efficiency, quality control, customer responsiveness, optimal inventory, resource utilization, effective coordination, risk management, technology integration, adaptability, and sustainability practices are essential aspects of supply chain management that provide a competitive edge. Together, these components boost an organization's operational efficiency and market competitiveness.

Lack of consideration for sustainable development. Traditional supply chain management methods often focus on the short-term economic benefits within the enterprise while ignoring the impact on society and the environment. This may lead to the waste of social resources, environmental harm, and the violation of rights and interests of employees.

1.3 Application of reinforcement learning in supply chain optimization

The application of reinforcement learning in supply chain optimization mainly includes inventory management, logistics distribution, production planning, risk management, and supply chain coordination. Reinforcement learning optimizes supply chain processes by improving inventory management through dynamic replacement and precise demand forecasting while lowering costs. In logistics, it enhances delivery routes, maximizes vehicle use, and responds to disturbances. Overall, it improves decision-making, efficiency, and inventory and logistics management responsiveness.

Inventory management and production planning Through learning the history of the warehouse and real-time data information, the reinforcement learning algorithm can calculate the optimal inventory level of a single product and, at the same time, can, according to the change in market demand, adjust the inventory strategy, optimize the inventory level, save the cost of inventory, and improve the level of customer service. Reinforcement learning can also determine the best production plan for each product and adjust it dynamically so as to optimize the production plan and improve production efficiency. Estes et al. [5] introduced the application of reinforcement learning (RL) in production planning and control, solved the problems of production scheduling, procurement, inventory and supply chain, and elaborated Q-learning and its variants. Zhao et al. [6] proposed a reinforcement learning-driven brainstorm optimization algorithm (RLBSO) to balance the allocation of resources, reduce the calculation of TEC, and solve the problem of workshop scheduling.

Supply chain coordination Learning the history of the warehouse and real-time data can determine the optimal distribution route and dynamically adjust the distribution route according to the demand for orders and traffic conditions to reduce transportation costs. Reinforcement learning can optimize collaborative decision-making in the supply chain to improve the overall efficiency of the supply chain. Multiagent reinforcement learning, Q-learning, SARSA, and deep reinforcement learning (DRL) are reinforcement learning approaches that will enhance collaborative supply chain decision-making. These strategies offer optimal strategy learning, flexible decision-making, and consistent policy changes, allowing supply chain partners to improve their efficiency and responsiveness. Oroojlooy-jadid et al. [7] proposed a deep reinforcement learning (RL) algorithm to seek the optimal solution in the constantly changing and unpredictable state of a supply chain.

Risk management aspects Learning the history of the warehouse and real-time data can identify potential risk factors and determine the optimal risk response strategies to reduce the risk of supply chain interruption. Aboutorab et al. [8] studied the use of the reinforcement learning (RL-PRI) method to assist risk managers in actively identifying their operational risks, explained the working principles and steps of the process in detail, and demonstrated the excellent performance of the RL-PRI method by comparing with the manual identification of the professional risk managers.

In addition, reinforcement learning can also be used to optimize other aspects of the supply chain, such as procurement, quality control, and customer service. This process can be done by enabling adaptive learning, real-time decision-making, and process optimization. It personalizes client encounters, anticipates challenges, and optimizes resource allocation, resulting in increased efficiency, lower costs, and more customer happiness.

2 Goals, basic principles, and practical application of reinforcement learning

The goal of reinforcement learning is to train an agent to allow him to make optimal decisions in a dynamic environment. It defines both state and action spaces. By taking an action the agent moves from state to state, obtaining reward signals from the environment as feedback. It requires learning a strategy to select the optimal action in any state to maximize the long-term cumulative reward. The primary objective is to train an agent to maximize long-term cumulative rewards by selecting optimal behaviors in a dynamic environment, utilizing defined state and action spaces and value functions to evaluate state-action combinations.

The rationale is that the value function evaluates the value of each state–action pair. Commonly used value functions include action and state value functions. Learning algorithms, such as Q-learning algorithms, can constantly update these value functions through experiments and learning, making the strategy more accurate. In the learning process, we have to use the existing knowledge and explore new movements. Using the current knowledge to obtain the maximum return, we can explore better solutions and finally find the optimal strategy in a dynamic environment.

Trial and error and feedback to achieve the goal of optimization It is used in financial transactions, robot control, and game strategy. For example, Google's AlphaGo team used a reinforcement learning algorithm to train a Go AI program, beating the human champion in the competition, which attracted wide attention; in the control of autonomous

cars, they can learn how to navigate and avoid obstacles in complex traffic environment, and DeepMind team used a reinforcement learning algorithm to train an AI program that performs in Atari games. The DeepMind team's reinforcement learning algorithms are practical due to their adaptive learning capabilities, which allow AI programs to adjust methods based on contextual feedback. The algorithms may extend learned strategies to new settings, making them applicable in sectors other than gaming, such as robotics and healthcare. Additionally, DeepMind's use of deep neural networks allows for efficient high-dimensional data processing, facilitating learning from raw sensory inputs.

3 Analysis of supply chain optimization problems

3.1 Optimization of inventory management

In the process of supply chain optimization, inventory management optimization is essential. Gijbrecchts et al. [9] introduced the depth of reinforcement learning in inventory management, the double purchasing, sales loss, and three inventory management, each inventory problem modeling as a process of Markov decision and then using the effective asynchronous advantage participants-criticism (A3C) DRL algorithm. The A3C DRL algorithm enhances inventory management by accurately anticipating demand, reducing holding and stockout costs, and allowing for dynamic adjustments. It also improves decision-making, automates procedures, improves demand forecasting, and optimizes resource allocation, resulting in a more efficient and cost-effective inventory system. Inventory management optimization is the most common driver of the supply chain, and Q-learning is the most common algorithm [10]. The goal is to determine the best inventory level of each product to meet customer needs and minimize inventory costs. Inventory cost mainly includes two parts, holding and out-of-stock costs:

$$\begin{aligned} \text{Holding cost} = & \text{storage fee} + \text{insurance premium} \\ & + \text{capital occupation cost} + \text{other;} \end{aligned} \quad (1)$$

$$\begin{aligned} \text{shortage cost} = & \text{sales loss} + \text{customer dissatisfaction} \\ & + \text{emergency purchase cost} + \text{other.} \end{aligned} \quad (2)$$

There are many ways to optimize inventory management, such as the reorder point model (ROP) and economic batch model (EOQ), as well as methods based on optimization theory and artificial intelligence technology, such as nonlinear planning, linear planning, dynamic planning, and reinforcement learning.

Inventory management optimization method based on reinforcement learning Reinforcement learning is an important part of machine learning and interacting with the environment. Alves et al. [11] use deep reinforcement learning to consider uncertain needs and advance time and solve the problems of production planning and logistics.

Distribution in the multiechelon supply chain Deep reinforcement learning improves multitier supply chains by increasing adaptability, decision-making, efficiency, and customization. However, it necessitates a large amount of data, has computational hurdles, risks overfitting, complicates implementation, and lacks interpretability; therefore these advantages and disadvantages must be carefully considered before adoption. To solve the

problem of low supply chain performance (SCP) of (IRS) policy, Wang et al. [12] established a multiagent simulation model based on reinforcement learning and dynamic inventory replenishment strategy. This method is applied to aerospace manufacturing, can adapt to changes in supply and demand, and is very practical.

The application of reinforcement learning in inventory management optimization mainly has the following steps:

1) Define the state and action spaces. The state space refers to all the possible states of the inventory management system, and the action space refers to all the actions the inventory manager can take.

2) Define the reward function. The inventory manager will be rewarded for taking action in multiple states.

3) Initialize the reinforcement learning algorithm. Reinforcement learning algorithms usually represent the optimal policy using value or policy functions.

4) Interaction with the environment. The reinforcement learning algorithm interacts with the environment and learns the optimal strategy. In the optimization problem of inventory management, the environment refers to the inventory management system and the reinforcement learning algorithm, which observes the state and reward of the environment by taking different actions.

5) Update the value or policy functions. Reinforcement learning algorithms update value functions or policy functions according to the results of interaction with the environment.

After multiple interactions with the environment, the reinforcement learning algorithm can learn the optimal inventory management strategy. Reinforcement learning improves inventory management by allowing algorithms to learn ideal inventory levels from contextual interactions, defining state (inventory levels) and action (order amounts) spaces and minimizing costs with a reward function. The company sells a product whose demand is random and follows a normal distribution. The holding cost of the product is 1 yuan per unit per day, and the cost of the shortage is 12 yuan per unit per day – use of reinforcement learning to optimize the company's inventory management strategy. The state space is defined as the inventory level, and the action space is defined as the order quantity. The reward function is defined as

$$r(s, a) = -h * s - p * (d - s), \quad (3)$$

where $r(s, a)$ is the reward for achieving action a in state s , h is the holding cost, p is the out-of-stock cost, s is the inventory level, and d is the demand. Reinforcement learning algorithms use the Q-learning algorithm to learn the optimal strategy. After many interactions with the environment, the reinforcement learning algorithm learns the optimal inventory management strategy. Q-learning learns optimal methods based on essential characteristics, such as model-freeness, eliminate the need for knowledge of transition probabilities or reward functions. They assess the expected utility for acts utilizing a Q-value function, which is updated based on the incentives received. The algorithm balances exploration and exploitation, uses temporal difference learning to change Q-values depending on prediction mistakes, and ensures convergence to the optimal policy under specified conditions, making it a dependable choice for diverse applications. Q-learning is commonly used to optimize inventory management because of its capacity to successfully handle the complexities and uncertainties inherent in supply chain systems. Q-learning,

a model-free reinforcement learning algorithm, allows for calculating optimal inventory levels by learning from historical data and real-time information without the need for a preexisting environment model. Q-learning excels in supply chain optimization due to its model-free adaptability, effective exploration–exploitation balance, and continuous improvement through feedback. Its scalability, flexibility for real-time adjustments, simplicity, robustness under uncertainty, and integration potential with other methods enhance its effectiveness in optimizing inventory management and overall supply chain efficiency.

3.2 Optimization of transportation route planning

Transportation route planning and optimization determine the best route from multiple warehouses to multiple customers to meet customer needs, thus reducing transportation costs as much as possible. Ren et al. [13] studied the problem of multivehicle route planning in the supply chain in large-scale transportation tasks, pointed out the shortcomings of the current algorithm, proposed a multiagent reinforcement learning model, optimized the route length and arrival time, reduced computation time, and improved model performance. Route planning optimizes the delivery process by lowering travel time and costs and allowing for rapid modifications based on real-time information. It optimizes vehicle loads, increases customer communication, and encourages sustainability. Optimizing the route length and arrival time improves the computation time and model performance through improved efficiency and resource use. Shorter routes reduce trip lengths and computation times, allowing for faster decision-making. This simplification reduces computing complexity, allowing for speedier processing. Transportation costs mainly include two parts, fixed and variable costs:

$$\text{Fixed cost} = \text{vehicle depreciation fee} + \text{insurance premium} + \text{other}, \quad (4)$$

$$\text{Variable cost} = \text{fuel} + \text{labor} + \text{other}. \quad (5)$$

The methods used in transportation route planning optimization include the travel agent problem (TSP) and the vehicle path planning problem (VRP) and methods based on optimization theory and artificial intelligence technology, such as nonlinear and linear planning, dynamic planning, genetic algorithm, etc. The traveling salesman problem (TSP) facilitates transportation route design by giving a framework for determining the shortest delivery routes and reducing the travel distance and time. This optimization lowers operational expenses, fuel consumption, and vehicle wear while improving delivery schedules to increase customer satisfaction. Data analysis plays a crucial role in maximizing transportation routes by ensuring precise demand predictions, optimizing routes, analyzing costs, monitoring in real time, evaluating performance, simulating scenarios, grasping customer behavior, and improving supply chain integration.

3.3 Optimization of demand forecast

Accurate demand forecasting can help enterprises to better plan their production and enhance the overall competence of the supply chain. There are many demand-forecasting methods, including quantitative and qualitative methods. Qualitative methods mainly rely on market research and expert experience. In contrast, quantitative methods, such as the time series analysis method, mainly rely on historical data and statistical models to predict demand. To improve overall efficiency through accurate demand forecasting, several

measures or actions can be taken, including data collection and integration, advanced analytics, departmental collaboration, scenario planning, demand sensing techniques, and technology investment. Implementing these strategies can considerably improve the accuracy of demand forecasts, resulting in improved inventory management, lower costs, and increased overall supply chain efficiency.

4 Design of supply chain optimization algorithm based on reinforcement learning

4.1 Definition of action and state spaces

The state-space definition of the supply chain mainly includes inventory level, demand level, production cost, transportation cost, delivery time, and customer satisfaction. Reinforcement learning effectiveness in supply chain optimization is evaluated through cost reduction, service level improvement, inventory turnover, lead time reduction, customer satisfaction, resource utilization, flexibility, and risk mitigation. These metrics collectively assess the performance and efficiency of the algorithms in enhancing supply chain operations. The action space of a supply chain is defined as the adjustment of the elements in the state space of the supply chain. In supply chain optimization problems, reinforcement learning algorithms are mainly used to learn how to adjust the supply chain's state variables to maximize the supply chain's total benefits. Furthermore, reinforcement learning (RL) algorithms solve significant supply chain optimization challenges, such as inventory management, by dynamically altering stock levels to save costs while maintaining availability. They increase demand forecasting by analyzing historical data, production planning by allocating resources more efficiently, and logistics by altering routes in real time.

Steps for the reinforcement learning algorithm:

- 1) Initialize the state variable of the supply chain;
- 2) In the current state, perform a random operation;
- 3) Observe the environmental feedback and calculate the rewards;
- 4) Update the status variable;
- 5) Repeat steps 2–4 until the termination conditions are reached;
- 6) Output the optimal strategy.

4.2 Design of the reward function

The reward function refers to the reward of taking a specific action in a given state. In the supply chain optimization problem, the reward function reflects the supply chain's overall performance, including production cost, service level, and inventory level. The reward function is critical in reinforcement learning for increasing supply chain efficiency, since it provides a feedback mechanism to direct the learning process. It is crucial in reinforcement learning for supply chain efficiency as it allows for feedback on the effectiveness of actions taken by the agent. It quantifies success, guiding the agent to learn optimal strategies.

Commonly used supply chain optimization reward functions include total cost, service level, inventory level, and comprehensive indicators. The total cost includes the procurement, transportation, inventory, and service costs in the supply chain. The service level includes the order integrity rate, on-time delivery rate, customer satisfaction, and other factors. Inventory levels reflect the high and low inventory levels. Comprehensive indicators are a combination of multiple indicators based on specific problems.

Table 1 Four points to pay attention to when designing a reward function

Key points	Details
Consistency	Should be consistent with the overall objectives of the supply chain
Simplicity	Should be simple and clear, easy to calculate
Robustness	Should be able to deal with different situations
Dynamic	Should be able to reflect the dynamic changes in the supply chain

Table 2 Comparison of reinforcement learning algorithms

Algorithms	Details
Q-learning	Q-learning is a model-free reinforcement learning algorithm that does not need to know the transition probability and reward function of the environment. The algorithm learns the optimal strategy by continuously exploring and updating the Q-value function.
SARSA	SARSA (state-action-reward-state-action) is a model-based reinforcement learning algorithm that requires knowing the environment's transition probability and reward function. The SARSA algorithm learns the optimal strategy by constantly updating the SARSA cripple.
Deep Q network (DQN)	DQN is a reinforcement learning algorithm based on a deep neural network, which can deal with states and action space in high dimensions. The algorithm learns the optimal strategy by continuously training the deep neural network.
Policy gradient (strategy gradient)	Policy gradient is a gradient-based reinforcement learning algorithm that directly optimizes policy functions. The strategy gradient algorithm learns the optimal strategy by continuously updating the policy functions.

When designing a reward function, use four-point Tables 1 and 2.

The following is an example of a reward function based on the total cost:

$$R(s, m) = -C(s, m). \quad (6)$$

$R(s, m)$ is the reward of states s and action m , and $C(s, m)$ is the total cost of states s and action m . This reward function simply and effectively reflects the overall performance of the supply chain. The lower the total cost, the higher the reward.

4.3 The selection of the reinforcement learning algorithm

In the supply chain optimization problems, the reinforcement learning algorithms can be selected (see Table 2).

When choosing a reinforcement learning algorithm, consider the three key points in Table 3.

Here is an example of a supply chain optimization algorithm based on Q-learning:

- 1) Initialize the Q-value function;
- 2) Repeat the following steps until termination:
 - * Select an action in the current state,
 - * Act and observe the next state and reward,
 - * Update the Q-value function,
 - * Go to the next state;
- 3) Return the optimal policy.

Table 3 Three key points

Key points	Details
Complexity of the environment	If the environment is complex, then a reinforcement learning algorithm that can deal with states and action space in high dimensions, such as DQN or policy gradient, must be chosen.
Availability of data	If the amount of data is small, then a reinforcement learning algorithm that does not require a large amount of data, such as Q-learning or SARSA, must be chosen.
Computational resources	If computational resources are limited, then a reinforcement learning algorithm, such as Q-learning or SARSA, must be chosen.

Table 4 Validation of the platform indicators

Parameter	Index
Operating system	Windows 10, 64 bit
RAM	8 G
CPU	Intel(R) Core (TM) i7-9750H CPU @2.60 GHz 2.59 GHz
GPU	NVIDIA GTX-1050
HDD	1T
Cores	4
Programming languages and frameworks	Python + numpy

This algorithm is simple but can effectively learn the optimal supply chain strategy in Table 4.

5 Experimental design and analysis of the results

5.1 Experimental conditions

Experimental indicators: inventory level, demand level, transportation cost, delivery time.

Production cost and customer satisfaction.

The experiment included suppliers, manufacturers, warehouses, retailers, and users.

5.2 The experimental procedures

The experimental steps are as follows: initializing the supply chain state variable, updating the supply chain state variable according to the operation, calculating the rewards according to the current state of the supply chain, updating the state of the supply chain according to the current actions and rewards, and running the algorithm's result output for the specified number of iterations (Fig. 1 and Figs. 2, 3, 4, 5, 6, 7).

Algorithm result: The algorithm's output is an optimal strategy to optimize the supply chain's total revenue. Optimizing the supply chain's overall revenue through an ideal approach is critical for cost reduction, improved customer satisfaction, precise demand forecasting, effective resource allocation, flexibility, technological integration, and a competitive edge. These factors combined generate revenue growth and ensure the organization's long-term viability. The strategy says that in the current state, the optimal operation is keeping the level of demand, delivery times, production costs, transportation costs, inventory levels, and customer satisfaction unchanged.

5.3 Experimental results and evaluation of algorithm effects

1) In all the experimental metrics, the Q-learning algorithm performed the best, followed by the SARSA, DQN, and strategy gradient algorithms.

```
import numpy as np
class SupplyChain:
    def __init__(self):
        # Initialize the supply chain state variables
        self.inventory_level = 0
        self.demand_level = 0
        self.transportation_cost = 0
        self.production_cost = 0
        self.delivery_time = 0
        self.customer_satisfaction = 0
```

Figure 1 Initializing the supply chain state variable

```
def take_action(self, action):
    # Update the supply chain state variables based on the action
    if action == "increase_inventory":
        self.inventory_level += 1
    elif action == "decrease_inventory":
        self.inventory_level -= 1
    elif action == "increase_demand":
        self.demand_level += 1
    elif action == "decrease_demand":
        self.demand_level -= 1
    elif action == "increase_transportation_cost":
        self.transportation_cost += 1
    elif action == "decrease_transportation_cost":
        self.transportation_cost -= 1
    elif action == "increase_production_cost":
        self.production_cost += 1
    elif action == "decrease_production_cost":
        self.production_cost -= 1
    elif action == "increase_delivery_time":
        self.delivery_time += 1
    elif action == "decrease_delivery_time":
        self.delivery_time -= 1
    elif action == "increase_customer_satisfaction":
        self.customer_satisfaction += 1
    elif action == "decrease_customer_satisfaction":
        self.customer_satisfaction -= 1
```

Figure 2 Update the supply chain status variables according to this operation

```
def calculate_reward(self):
    # Calculate the reward based on the current state of the supply chain
    reward = 0
    if self.inventory_level > 0:
        reward += 1
    if self.demand_level > 0:
        reward += 1
    if self.transportation_cost < 10:
        reward += 1
    if self.production_cost < 10:
        reward += 1
    if self.delivery_time < 10:
        reward += 1
    if self.customer_satisfaction > 0:
        reward += 1
    return reward
```

Figure 3 Calculate the rewards based on the current state of the supply chain

```
def update_state(self):
    # Update the state of the supply chain based on the current action and reward
    self.inventory_level += np.random.normal(0, 1)
    self.demand_level += np.random.normal(0, 1)
    self.transportation_cost += np.random.normal(0, 1)
    self.production_cost += np.random.normal(0, 1)
    self.delivery_time += np.random.normal(0, 1)
    self.customer_satisfaction += np.random.normal(0, 1)
```

Figure 4 Update the status of the supply chain based on current actions and rewards

```

def run(self):
    # Run the algorithm for a specified number of iterations
    for i in range(1000):
        # Take a random action
        action = np.random.choice(["increase_inventory", "decrease_inventory", "increase_demand", "decrease_demand",
                                   "increase_transportation_cost", "decrease_transportation_cost", "increase_production_cost",
                                   "decrease_production_cost", "increase_delivery_time", "decrease_delivery_time",
                                   "increase_customer_satisfaction", "decrease_customer_satisfaction"])

        # Take the action
        self.take_action(action)

        # Calculate the reward
        reward = self.calculate_reward()

        # Update the state
        self.update_state()

        # Print the current state and reward
        print("Iteration:", i, "State:", self.inventory_level, self.demand_level, self.transportation_cost,
              self.production_cost, self.delivery_time, self.customer_satisfaction, "Reward:", reward)

```

Figure 5 Run the algorithm

```

# Create a supply chain object
supply_chain = SupplyChain()
# Run the algorithm
supply_chain.run()

```

Figure 6 Create a provisioning object and run the algorithm

```

81         print("Iteration:", i, "State:", self.inventory_level, self.demand_level, self.transportation_cost,
82               self.production_cost, self.delivery_time, self.customer_satisfaction, "Reward:", reward)
83     # Create a supply chain object
84     supply_chain = SupplyChain()
85     # Run the algorithm
86     supply_chain.run()
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591
2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2620
2621
2622
26
```

existing knowledge to maximize immediate rewards, and exploration, which searches out new actions to uncover possibly better long-term methods.

4) The Q-learning algorithm performs the best regarding the total supply chain cost, average delivery time, and customer satisfaction, so the Q-learning algorithm is the best choice for supply chain efficiency optimization problems. The SARSA algorithm performs very closely to the Q-learning algorithm regarding total supply chain cost and average delivery time. Still, it is slightly inferior in terms of customer satisfaction, because it is an off-policy algorithm. This means that it learns the value of an optimal policy regardless of the agent's actions. This might result in suboptimal decisions in dynamic contexts where customer preferences and behaviors often change. Therefore the SARSA algorithm is also a good choice.

5) The DQN algorithm and policy gradient algorithm performed slightly inferior to the Q-learning algorithm and SARSA algorithm in terms of the average delivery time and total supply chain cost but in terms of customer satisfaction. The DQN algorithm and strategy gradient algorithm performed better. Therefore, if customer satisfaction is the most important indicator, then the DQN and strategy gradient algorithms are also good choices. The DQN and strategy gradient algorithms prioritize satisfaction in their incentive structures, allowing for customer-centered optimization. They use adaptive learning to recognize trends in client behavior, allowing for responsive modifications to shifting demands. These algorithms discover novel solutions to improve service offerings by experimenting with diverse strategies.

Overall, using algorithms such as Q-learning, SARSA, DQN, or policy gradient approaches improves supply chain performance by lowering costs, shortening delivery times, and improving inventory management. These algorithms also impact scalability, multiagent coordination, risk management, and long-term planning, resulting in more efficient and resilient supply chain operations.

5.4 Experimental conclusions and further studies

The Q-learning algorithm is best for the supply chain efficiency optimization problem. The SARSA algorithm is also a good choice. If customer satisfaction is the most important indicator, then the DQN and the strategy gradient algorithms are also good choices.

To get better application results, further research can be conducted to study the application of other reinforcement learning algorithms in supply chain efficiency optimization problems, such as the trust region policy optimization algorithm. Combining reinforcement learning algorithms with different optimization algorithms, such as genetic algorithms, Achamrah et al. [14] proposed a new solution based on a mixture of genetic algorithm, mathematical modeling, and deep reinforcement learning to deal with the current supply chain problems. Therefore modeling the real-world problem as a random and dynamic inventory routing problem improves the supply chain performance. Muthu et al. [15] explore the implementation of an intelligent IoT model for analyzing supply chain management at Chokhi Dhani Village resort. This model was utilized to understand audience behavior intelligence and determine the necessary services to sustain cultural harmony. Five modes were developed based on users' attitudes, and the model examined the interconnectedness among various audiences. The findings revealed a 52% variance in the model, with the most notable variances observed in the areas of finding meaning, linking ideas, using evidence, showing interest in ideas, and evaluating effectiveness. One can also

study the application of reinforcement learning algorithms in other supply chain issues, such as inventory management and transportation issues.

Advanced signal processing techniques are essential for ensuring secure and efficient communication in the future. Marketing information systems (MISs) leverage digital signal processes, including visual images, sound waves, and seismic waves, to effectively engage with audiences. Collaborative MISs are suggested to connect professionals and businesses more effectively. This research aims to address transportation challenges such as the total order intensity ratio, increasing fuel prices, delays, shortages of skilled workers, and warehouse conditions [16]. Complex contexts, data quality difficulties, and resource constraints challenge implementing reinforcement learning (RL) in supply chains. It is vital to balance exploration and exploitation, as too much exploration might damage decision-making. Integrating existing systems is hard, and deep learning's "black box" nature poses trust concerns. Regulatory and ethical constraints hamper implementation and careful planning, and stakeholder participation are needed.

6 Conclusions and outlook

Conclusions Reinforcement learning is an effective method for supply chain efficiency optimization. The Q-learning algorithm is the best choice for these problems, and the SARSA algorithm is also good. If customer satisfaction is the most important indicator, then the DQN and strategy gradient algorithms are also good choices.

Research prospect:

- Study the application of other reinforcement learning algorithms, such as the actor-critic algorithm and the trust region policy optimization algorithm, in supply chain efficiency optimization problems.
- Study the combination of reinforcement learning algorithms with other optimization algorithms, such as simulated annealing and genetic algorithms.
- Study the application of reinforcement learning algorithms in other supply chain issues, such as inventory management and transportation problems.
- Study the application of reinforcement learning algorithms in uncertain environments, such as demand and price uncertainty.

Specific research directions include:

- Multiagent reinforcement learning. Study how to apply reinforcement learning to multi-agent systems, such as competition and cooperation between suppliers, manufacturers, and retailers.
- Reinforcement learning under uncertainty. Study how to apply reinforcement learning to uncertain environments, such as demand and price uncertainty.
- Combine the reinforcement learning algorithm with other optimization algorithms to improve the algorithm's performance.
- Application of reinforcement learning to other supply chain issues. Study how to apply reinforcement learning to other supply chain issues, such as optimizing demand and transportation.

Author contributions

TZ and LX designed the framework, analyzed performance, validated results, and wrote the paper. CZ and YT collected the information required for the framework, provided software, performed a critical review, and administered the process. All authors read and approved the final manuscript.

Funding

The authors did not receive any funding.

Data availability

No datasets were generated or analyzed during the current study.

Code availability

Not applicable.

Declarations**Competing interests**

The authors declare no competing interests.

Received: 19 June 2024 Accepted: 28 August 2024 Published online: 06 November 2024

References

1. He, Z., Tran, K.P., Thomassey, S., Zeng, X., Xu, J., Yi, C.: Multi-objective optimization of the textile manufacturing process using deep-Q-network-based multi-agent reinforcement learning. *J. Manuf. Syst.* **62**, 939–949 (2022)
2. Achamrah, F.E., Riane, F., Sahin, E., Limbourg, S.: An artificial-immune-system-based algorithm enhanced with deep reinforcement learning for solving returnable transport item problems. *Sustainability* **14**(10), 5805 (2022)
3. Ali, N., Ghazal, T.M., Ahmed, A., Abbas, S., Khan, M.A., Alzoubi, H.M., et al.: Fusion-based supply chain collaboration using machine learning techniques. *Intell. Autom. Soft Comput.* **31**(3), 1671–1687 (2022)
4. Mahmud, S., Abbasi, A., Chakraborty, R.K., Ryan, M.J.: A self-adaptive hyper-heuristic based multi-objective optimisation approach for integrated supply chain scheduling problems. *Knowl.-Based Syst.* **251**, 109190 (2022)
5. Estes, A., Peidro, D., Mula, J., Díaz-Madroño, M.: Reinforcement learning applied to production planning and control. *Int. J. Prod. Res.* **61**(16), 5772–5789 (2023)
6. Zhao, F., Hu, X., Wang, L., Xu, T., Zhu, N., Jonrinaldi: A reinforcement learning-driven brainstorm optimisation algorithm for multi-objective energy-efficient distributed assembly no-wait flow shop scheduling problem. *Int. J. Prod. Res.* **61**(9), 2854–2872 (2023)
7. Oroojlooyjadid, A., Nazari, M., Snyder, L.V., Takáč, M.: A deep q -network for the beer game: deep reinforcement learning for inventory optimization. *Manuf. Serv. Oper. Manag.* **24**(1), 285–304 (2022)
8. Aboutorab, H., Hussain, O.K., Saberi, M., Hussain, F.K.: A reinforcement learning-based framework for disruption risk identification in supply chains. *Future Gener. Comput. Syst.* **126**, 110–122 (2022)
9. Gijsbrechts, J., Boute, R.N., Van Mieghem, J.A., Zhang, D.J.: Can deep reinforcement learning improve inventory management? Performance on lost sales, dual-sourcing, and multi-echelon problems. *Manuf. Serv. Oper. Manag.* **24**(3), 1349–1368 (2022)
10. Rolf, B., Jackson, I., Müller, M., Lang, S., Reggelin, T., Ivanov, D.: A review on reinforcement learning algorithms and applications in supply chain management. *Int. J. Prod. Res.* **61**(20), 7151–7179 (2023)
11. Alves, J.C., Mateus, G.R.: Multi-echelon supply chains with uncertain seasonal demands and lead times using deep reinforcement learning. *arXiv preprint* (2022). [arXiv:2201.04651](https://arxiv.org/abs/2201.04651)
12. Wang, H., Tao, J., Peng, T., Brintrup, A., Kosasih, E.E., Lu, Y., et al.: Dynamic inventory replenishment strategy for aerospace manufacturing supply chain: combining reinforcement learning and multi-agent simulation. *Int. J. Prod. Res.* **60**(13), 4117–4136 (2022)
13. Ren, L., Fan, X., Cui, J., Shen, Z., Lv, Y., Xiong, G.: A multi-agent reinforcement learning method with route recorders for vehicle routing in supply chain management. *IEEE Trans. Intell. Transp. Syst.* **23**(9), 16410–16420 (2022)
14. Achamrah, F.E., Riane, F., Limbourg, S.: Solving inventory routing with transshipment and substitution under dynamic and stochastic demands using genetic algorithm and deep reinforcement learning. *Int. J. Prod. Res.* **60**(20), 6187–6204 (2022)
15. Ferinia, R., Kumar, D.L.S., Kumar, B.S., Muthu, B.A., Asaad, R.R., Ramamoorthi, J.S., Daniel, J.A.: Factors determining customers desire to analyse supply chain management in intelligent IoT. *J. Comb. Optim.* **45**(2), 72 (2023)
16. Aggarwal, K., Khoa, B.T., Sagar, K.D., Agrawal, R., Dhingra, M., Dhingra, J., Kumar, R.L.: Marketing information system based on unsupervised visual data to manage transportation industry using signal processing. *Expert Syst.*, e13384 (2023)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.