SCIENTIFIC OASIS

Decision Making: Applications in Management and Engineering

Journal homepage: www.dmame-journal.org
ISSN: 2560-6018, eISSN: 2620-0104

Agent-based Decision Making and Control of Manufacturing System Considering the Joint Production, Maintenance, and Quality by Reinforcement Learning

Mohammad Reza Nazabadi[1], Seyed Esmaeil Najafi[1,*], Ali Mohaghar[2], Farzad Movahedi Sobhani[1]

[1] Department of Industrial Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
[2] Faculty of Management, University of Tehran, Tehran, Iran

ARTICLE INFO

Article history:
Received 19 July 2023
Received in revised form 11 December 2023
Accepted 16 December 2023
Available online 20 December 2023

Keywords: Reinforcement learning; Agent-based modeling; Production Planning; Maintenance; Quality control; Real-time decision making.

ABSTRACT

Taking an integrated approach towards production, maintenance, and control in manufacturing systems is crucial due to the profound impact of their interconnections. Investigating these aspects in isolation may lead to infeasible solutions. This research focuses on the real-time and autonomous decision-making process concerning joint production planning, maintenance, and quality problem in a stochastic deteriorating production system with limited maintenance activities. Formulating the problem as a continuous semi-Markov decision process accounts for the complexities of the real production system and the occurrence of events over an uneven and continuous period. While dynamic programming is a common tool for addressing joint optimization problems, it has limitations, such as the curse of dimensionality. In this study, the optimal policy of the decision-maker agent is obtained by the goal-directed machine learning method called (R-SMART) and agent-based modeling. To the author's knowledge, the proposed approach is novel, and there is little research on such an implementation of the joint optimization problem. The quality of the optimal policy is evaluated through heuristic and simulation-optimization methods in various scenarios. The results demonstrate that the proposed RL-based method outperforms others in most scenarios, achieving a stable, integrated optimal policy.

1. Introduction

Production planning, maintenance, and ensuring the quality of finished products pose significant challenges in manufacturing systems. The primary objective of production planning is to efficiently schedule tasks and allocate resources, aiming to achieve operational and economic goals such as cost minimization, tardiness minimization, and production maximization [1]. In addressing maintenance issues, the primary focus is optimizing production machine availability while minimizing associated costs.

* *Corresponding author.*
E-mail address: enajafi1515@gmail.com

https://doi.org/10.31181/dmame712024885

Maintenance activities utilize time that could otherwise be allocated to the production process. However, delaying maintenance and repair activities to boost production may elevate the risk of system malfunction. Consequently, maintenance planning and production planning often find themselves in conflict [2].

The quality of finished products is also affected by the depreciation of the production system. As the production system experiences more depreciation, the likelihood of producing lower-quality finished products increases.

In modern manufacturing systems, it is crucial to consider the mutual interaction between production planning, maintenance, and quality. However, only a limited number of researchers have explored this integrated perspective [3].

While dynamic programming (DP) is a common tool employed by most researchers in this field, it comes with limitations:

i. Detailed system models are required, including transition probabilities for each state-action pair.
ii. Computational requirements grow exponentially with an increase in the number of state variables (Curse of Dimensionality).

As a result, DP methods prove inefficient when dealing with complex and large-scale problems, such as those encountered in many real manufacturing systems.

Artificial intelligence (AI) has made much progress in recent years and encourages researchers to adopt AI in various fields, including manufacturing. The combination of AI and manufacturing systems leads to the term "smart manufacturing," that is, the incorporation of intelligence systems such as the Internet of Things and Machine Learning techniques into manufacturing processes for accurate measurement and inspection of indicators, e.g., inventory level and quality of products [4]. In addition, the manufacturing systems offer many opportunities to apply AI techniques to support decision-making [5].

In recent years, a goal-directed learning method called reinforcement learning (RL) has been applied in many areas, including manufacturing. RL, one of the three machine learning methods, focuses on learning how to map situations to actions to maximize numerical rewards [6,7]. However, as suggested by Sutton and Barto [6], RL methods prove to be efficient alternatives for multi-state optimization.

This research aims to propose an integrated method for production planning, maintenance, and product quality within a production system. The objective is to derive an optimal policy using agent-based modeling (ABM) and reinforcement learning. Formulating the optimization problem as a semi-Markov decision process (SMDP) allows us to treat it as a continuous decision-making process. In our approach, a decision-maker agent interacts with the agent-based simulation model of the production system. The agent observes the system state, selects an action, and implements it into the simulation model. The resulting next state and acquired reward are then fed back to the agent. This iterative process continues until the agent attains the final optimal policy. To evaluate this policy, we implement a heuristic policy, comparing the results through a simulation-optimization approach. A comprehensive performance analysis will be conducted to assess the efficiency of the RL-based method in comparison to existing policies.

The main contributions of this paper are as follows:

i. Addressing the integrated production, maintenance, and quality control problem in the production system with a limited number of maintenance activities between consecutive repair activities. Previous research often assumed an unlimited number of

partial maintenance procedures between two consecutive repairs, which is unrealistic in many real-world scenarios.

ii. Developing an agent-based simulation model for the production system to facilitate the seamless integration of the decision-maker agent and the simulation model. While discrete event simulation has been widely applied in this field, the agent-based approach proves to be more consistent and scalable, particularly for large-scale problems.

iii. Formulating the integrated production planning, maintenance, and quality problem as a continuous semi-Markov decision process, and the average reward RL algorithm achieves the optimal or near-optimal policy. The use of SMDP enables real-time decision-making, a crucial aspect of Industry 4.0.

iv. Conducting a comprehensive efficiency comparison between the proposed RL+ABM approach and heuristic and simulation-optimization methods in various scenarios. These scenarios are designed based on the utilization rate of the production system, categorized as low, normal, and high.

The remainder of the paper is structured as follows: Section 2 presents the literature review. Section 3 introduces the production system and outlines the assumptions made. In Section 4, we provide the agent-based model of the production system, offer a brief review of the Markov and semi-Markov decision processes, and describe the implemented RL algorithm. Section 5 details the alternatives for evaluating the acquired policy. Numerical results of the integrated production, maintenance, and quality problem are presented in Section 6. Finally, Section 7 concludes the paper, identifies research limitations, and outlines potential future research directions.

2. Literature Review

In recent years, as manufacturing systems have become more intricate and customer demands continuously evolve, researchers have concentrated on the integrated optimization of production planning, maintenance, and product quality. The research can be categorized in terms of the problem definition, assumptions of the production system, and problem formulation. In the following, we cover the recent related papers that addressed the combined optimization in the domains of production systems.

Although some literature mentioned the economic effects of the combined view of production, maintenance, and quality [8, 9], the literature that covers the topic is rare [3].

Integrated production and maintenance planning in a single machine-single product system with PM policy has been investigated by Aghezzaf *et al.* [10] and Chen [11]. Chouikhi *et al.* [12] addressed the combined maintenance and quality problem in a single-machine production system. The system is subject to deterioration, which impacts the product quality, and condition-based maintenance has been considered. Khatab *et al.* [13] investigate optimally integrating production quality and condition-based maintenance problems in a single-product, single-machine production system. Hadian *et al.* [14] consider maintenance, buffer stock, and quality control problems in deteriorating single-machine production systems. The problem is formulated by mathematical modeling, and a genetic algorithm has been applied to find the joint planning. In a similar production system, Cheng *et al.* [15] applied a simulation-optimization (SO) approach to jointly optimize the inventory size, PM policy, and lot size. The integrated problem of the production strategy and quality control in a single-machine production system is considered by Bouslah *et al.* [16]. The SO has been applied to jointly optimize the production lot size, the inventory threshold, the maintenance and repair activities threshold, and the sampling plan. Another integrated policy is presented by Bouslah *et al.* [17]. They

jointly optimize a single product-multi machine production system subject to operation-dependent and quality-dependent failures, so increasing the degradation of upstream machines leads to the production of defective products in the preceding machines. SO has yielded the optimal policy to control the production thresholds, the quality control level, and the preventive maintenance thresholds. Fakher *et al.* [18] consider the multi-product, single-machine production system and develop an integrated optimization model to maximize the expected profit. The simultaneous production planning and quality control problem in a single machine-single product production system is optimized through the SO method by Rivera-Gómez *et al.,* [19]. A single machine-single product system subject to quality deterioration to find optimal control policy is presented by Rivera-Gómez *et al.* [3]. Preventive maintenance (PM) and quality control policies are suggested to increase system availability, and the SO approach has been used to acquire the optimal policy. Some research has also investigated the application of meta-heuristic methods in production and maintenance scheduling problems [20, 21]. They consider a single degrading production system, and genetic algorithm, simulated annealing, and teaching learning-based optimization have been used to solve the problem.

The application of machine learning in production lines has been reviewed by Kang *et al.* [22]. In this research, quality, and availability are the most important applications of machine learning in manufacturing systems. Reinforcement learning is an emerging machine learning method for optimization problems in manufacturing systems, specifically in the combined optimization problem of the topic.

Kuhnle *et al.* [23] applied reinforcement learning to find the optimal maintenance schedule for parallel working machines to reduce the system's downtime and increase production. Xanthopoulos *et al.* [24] formulate the combined production maintenance optimization problem in a single machine-single product system as a Markov decision process. They applied an average reward RL method called R-learning [25] to find the optimal policy. The Kanban and threshold-type inventory and production policy (s, S) have examined the quality of the RL-based policy. An extension of previous research has been proposed by Paraschos *et al.* [26]. They formulate combined production maintenance and quality control of the single machine-single product as a Markov decision process. Like the previous paper, the optimal policy has been obtained by the R-learning algorithm. The manufacturing system is affected by several deterioration failures, and the quality of the finished product is related to the system's deterioration level. Along the same line, the R-learning algorithm is applied to find the optimal preventive maintenance and production scheduling policy in a single-machine production system [27]. Wang *et al.* [28] studied the integrated production scheduling and maintenance optimization problem in a single-machine production system with deteriorating effects. They applied a Q-learning-based solution framework to find the optimal joint policy. In a multi-machine production system, preventive maintenance is acquired by the Deep reinforcement learning agent [29]. The problem is formulated as an MDP, and a Double Deep-Q-Network is applied to learn the policy.

A multi-agent deep reinforcement learning algorithm was developed to optimize the maintenance scheduling in a parallel production system by Rodríguez *et al.* [30]. The algorithm learns a maintenance policy that technicians perform in the stochastic multi-machine production system under the uncertainty of failures. The RL agents partially observe the state of each production machine to coordinate maintenance decisions, leading to the dynamic allocation of maintenance tasks to technicians (with different skills).

Lee and Mitici propose [31] a deep reinforcement learning algorithm for predictive aircraft maintenance planning and minimizing maintenance costs. Convolutional Neural Networks and

Monte Carlo estimate the remaining useful life of parts (RUL). They compare the efficiency of the policy obtained by DRL with the mean-estimated RUL.

Wesendrup and Hellingrath [32] investigate the production, spare parts, and maintenance planning for a single-machine system using RL. The research aims to maximize production revenue by meeting customer demands and minimizing costs. They apply Proximal Policy Optimization to post-prognostics production planning and control decision-making.

Zhenggeng Ye *et al.* [33] investigate a joint optimization of preventive maintenance and quality in manufacturing systems. They offer machine-level dynamic reliability and quality models to deal with complex interactions in manufacturing networks. In addition, they propose a Deep Deterministic Policy Gradient (DDPG) algorithm to obtain the joint quality and reliability policy in the manufacturing system.

Geurtsen *et al.* [34] investigate the joint optimization problem of production activities and maintenance in an assembly line. The assembly line consists of a serial production line with N machines and N-1 buffers, and the maintenance must be scheduled on the last machine of the production line. They employ average-reward deep reinforcement learning techniques to find the optimal joint policy.

Although the research mentioned investigates the joint optimization policy, it is often considered paired topics (e.g., production and maintenance), and little research has addressed the joint optimal policy of production, maintenance, and quality. Furthermore, in the limited research that has explored this tripartite optimization, to the author's knowledge, the assumption of the restricted number of maintenance activities between two consecutive major repair activities is not included, and the unlimited number of maintenance can be performed on the production system to return the system to the previous deterioration stage. However, maintenance activities are typically limited in the practical production system, and major repair actions are unavoidable in many instances to restore the system to its initial stage. This research aims to bridge this gap by considering these real-world constraints. Moreover, integrating an RL-based decision-maker agent with an agent-based production system to derive a joint optimal policy is seldom explored in the reviewed literature.

3. Problem Description

The problem investigated in this paper is a production system contains a single production machine and a storage facility with capacity I_{max}. The production system produces only one type of product stored in the storage facility to satisfy customer demands. The production times are exponentially distributed with parameter λ_p, and the cost of production is C_p.

During the production process, the machine deteriorates from d_0 (as-good-as-new) to d_n (malfunction). The deterioration level of the production machine is defined by d stages so that after the occurrence of deterioration failures, the stage transfers to $d+1$ until the breakdown level d_{max}. In each deterioration stage between $1, \cdots, d_{max}-1$, the machine can be maintained at a cost of C_m or repaired at a cost of C_r, and returned to the previous deterioration stage or the initial stage, respectively. The maintenance and repair activities are exponentially distributed with parameter λ_m and λ_r. In addition, the maximum number of authorized maintenance activities between two consecutive repair activities is limited to U_{max}. In d_{max}, the production system fails, and the repair activity must be carried out. By repair activity, the deterioration level of the system back to the stage d_0.

It should be noted that the production process can occur when the deterioration level of the system is between d_0 and d_{max-1} and the maintenance and repair activity will be the only option when the deterioration level of the system is d_{max}.

Moreover, in each deterioration level, the production machine may encounter an unexpected breakdown with probability B_d and cost C_b. Maintenance and repair activities can be conducted to prevent such breakdowns. The breakdown recovery duration is exponentially distributed with parameter λ_b. It is assumed that $\lambda_b > \lambda_r > \lambda_m$ and $C_b > C_r > C_m$.

The quality of the product is affected by the deterioration level. There is a probability of producing low-quality products Q_d according to the deterioration level with cost C_q . The production system is more likely to produce a low-quality product when the deterioration level of the system increases. The cost of the low-quality product is the difference between the acquired profit of high-quality and low-quality products.

The time interval between two successive customer demands is considered exponentially distributed with parameter λ_d. The amount of each demand follows Poisson distribution with parameter λ_n. At the time of demand arrival, the demand is immediately satisfied if the inventory is sufficient and the profit P is acquired. Otherwise, the customer demand is back-ordered by FCFS (first come – first served) policy, and the maximum allowed back-orders is B_{max}. It is clear that the customer demand is lost (missed orders) when the system has reached the maximum permitted back-orders and the cost C_l is acquired. The production system includes the holding cost C_h when the inventory of the finished product is available. Otherwise, the shortage cost C_s is considered.

In the following section, the agent-based model of the production system is developed as a manufacturer agent by implementing the characteristics mentioned above. Then, the state space, the set of actions, and the reward function associated with the actions are defined to specify the input data of the decision-maker agent. Finally, the decision-maker agent, the decision-making procedure by reinforcement learning techniques, and the interaction of the manufacturer agent and the decision-maker agent are described.

4. System Modeling

Dynamic and time-dependent processes can be modeled as agent-based models. As highlighted by Macal and North [35], the agent-based model comprises three fundamental elements:

i. Definition of the agents, attributes, and behaviors;
ii. The agents' relationship and interactions;
iii. The agents' environment;

Agents, defined as artificial individuals operating based on their behavior and collaborating or competing with other agents [36], must possess specific characteristics to be considered as agents. These characteristics include autonomy, modularity, sociality, and conditionality [35]. Additionally, agents can be equipped with learning or evolutionary capabilities, such as artificial intelligence, allowing them to adapt to changes in themselves or the environment [37].

Agent-based modeling is more general and consequential than traditional approaches, e.g., Discrete Event Simulation and System Dynamics, because it captures more complex structures and dynamics [38].

This paper employs the agent-based modeling approach to simulate the production system. Based on the characteristics outlined for agents, two entities are considered: the manufacturer agent and the decision-maker agent. Additionally, the production system and its finished product buffer are modeled as the environment.

4.1 Manufacturer Agent

The manufacturer agent is responsible for the production process. This agent has six states; the initial state is "ready for the action." In the initial state, the manufacturer agent can receive one of the following messages from the decision-maker agent:

i. Produce: the manufacturer agent with $0 \leq d < d_{max}$ transits to "produce" state with $1 - B_d$ probability and produce a high-quality product with $1 - Q_d$ probability after λ_p, or with B_d probability, the agent transits to the "breakdown" state, and after λ_b, the agent's state returns to the initial state. During the production process, with $f_{d,p}$ probability
($d =$ deterioration, $p =$ number of productions since the last deterioration failure), the deterioration level of the production machine increases by one unit.

ii. Maintain: the agent starts the maintenance activity, and after λ_m, the deterioration level decreases by one unit.

iii. Repair: the manufacturer agent starts the repair activity, and after λ_r, deterioration level will be as good-as-new (*d=0*), and the state will return to the initial state.

Idle: the agent transits to the "idle" state until the next customer demand arrival.

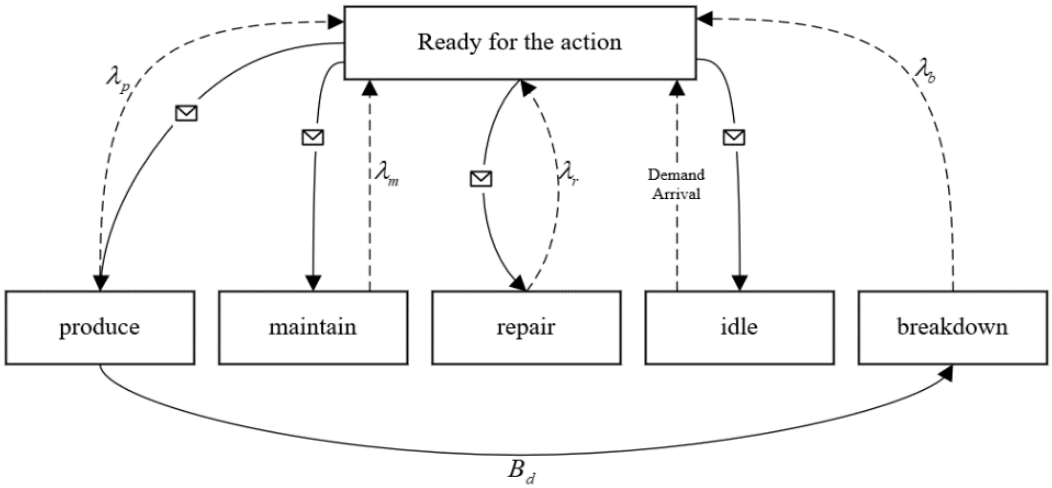

Fig. 1. The behavior of the manufacturer agent

The manufacturer agent operates without interruption until the completion of the process in each state. Interactions between agents are only permitted when the manufacturer agent is in the initial state. The behavior of the manufacturer agent is presented in Figure 1.

4.2. Decision Maker Agent

The decision-making process can be formulated as a Markov-decision process that consists of four elements: state space, set of actions, transition probabilities, and rewards [6]. As shown in Figure 2, the transition times in MDPs are discrete and equally distributed. Therefore, there is no concept of time in MDP. However, in many decision-making problems, the next state of the system and the received reward can be affected by the duration of the transition time drawn from probability distributions. Such problems can be modeled as Semi-Markov decision processes and interpreted as continuous-time stochastic systems [7].

Fig. 2. The decision epochs in the Markov-decision processes

In this research, the decision-making process is formulated as a continuous semi-Markov decision process, and an average reward Reinforcement Learning method called R-Smart (Relaxed Semi-Markov Average Reward Technique) [39] has been applied for computing optimal or near-optimal policies. In the following, the decision-making process formulation has been described.

4.2.1. Decision epochs

The decision-maker agent can only send a message (make a decision) when the manufacturer agent is in the "Ready for the action" state. Because the transition time from one state to another is stochastic, the duration between decision epochs is unequal (Figure 3).

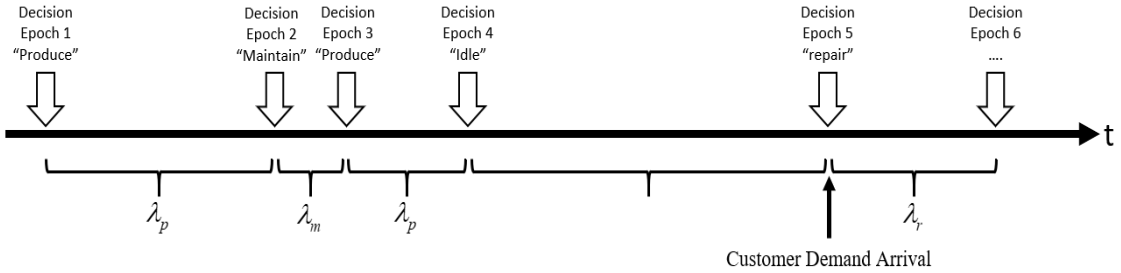

Fig. 3. The decision epochs in the SMDPs concept

4.2.2. State space

In every decision epoch, the decision-maker agent observes the state of the production system as the following vector:

$$S^t = (S_1^t, S_2^t, S_3^t) = (I(t), d(t), M(t)) \quad (1)$$

Where $I(t)$ is the inventory level of the production system at time t, $D(t)$ is the deterioration level of the production machine at time t, and $M(t)$ is the number of maintenance activities performed since the last repair activity. The entire state space of the production system can be written as:

$$S_1 = -B_{max}, \cdots, I_{max} \quad (2)$$

$$S_2 = 0, \cdots, D_{max} \quad (3)$$

$$S_3 = 0, \cdots, U_m \quad (4)$$

So, the total number of production system states is:

$$N(s) = (d+1) \times (I_{max} + B_{max} + 1) \times (U_m + 1) \quad (5)$$

4.2.3. Set of actions

In every decision-making epoch, the decision-maker agent can initiate one of the following actions:

i. Authorize the manufacturing agent to produce a new product
ii. Authorize maintenance activities
iii. Authorize repair activities
iv. Authorize the manufacturing agent to remain idle

Admissible actions in every state are defined by the action-state function $A(S_1, S_2, S_3)$ that is written as:

$$A(S_1,S_2,S_3) = \{produce, maintain, idle\}\ where: \begin{cases} S_1 = -B_{max}, \cdots, I_{max-1} \\ S_2 = 0, \cdots, d_{max} - 1 \\ S_3 = 0, \cdots, U_{max} - 1 \end{cases} \quad (6)$$

$$A(S_1,S_2,S_3) = \{maintain, repair, idle\}\ where: \begin{cases} S_1 = -B_{max}, \cdots, I_{max} \\ S_2 = d_{max} \\ S_3 = U_{max} - 1 \end{cases} \quad (7)$$

$$A(S_1,S_2,S_3) = \{repair, idle\}\ where: \begin{cases} S_1 = -B_{max}, \cdots, I_{max} \\ S_2 = d_{max} \\ S_3 = U_{max} \end{cases} \quad (8)$$

$$A(S_1,S_2,S_3) = \{produce, idle\}\ where: \begin{cases} S_1 = -B_{max}, \cdots, I_{max} - 1 \\ S_2 = 0 \\ S_3 = 0, \cdots, U_{max} \end{cases} \quad (9)$$

$$A(S_1,S_2,S_3) = \{idle\}\ where: \begin{cases} S_1 = I_{max} \\ S_2 = 0 \\ S_3 = 0, \cdots, U_{max} \end{cases} \quad (10)$$

4.2.4. Reward function

According to the state of the production system, every action of the decision-maker agent is associated with the relative reward:

$$R_{t+1} = C_h(t) + C_b(t) + C_l(t) + C_m(t) + C_r(t) + C_p(t) + C_q(t) - P(t) \quad (11)$$

Let t_i denotes the decision epochs. Then:

$$C_h(t) = (\int_{t_i}^{t_{i+1}} I(t)dt) \times C_h \quad (12)$$

$$C_b(t) = (\int_{t_i}^{t_{i+1}} B(t)dt) \times C_b \quad (13)$$

$C_l(t) = N(l) \times C_l$, where $N(l)$ denotes the number of missed orders between t_i and t_{i+1} (14)

$P(t) = N(P) \times P$, where $N(P)$ denotes the number of sales between t_i and t_{i+1}. (15)

Also, the cost of maintenance, repair, breakdown, and low-quality products is calculated by the relevant costs. The main goal of the decision-maker agent is to maximize the expected sum of profits and minimize the expected sum of costs listed in Eq. (11).

4.2.5. Average reward reinforcement learning

In the average reward reinforcement learning techniques for SMDPs, it is assumed that the time spent in each transition is not unity. The average reward (ρ) calculation can be mathematically expressed as:

$$\rho_\pi = \lim_{n\to\infty} \frac{E\left[\sum_{t=1}^{n} r(s_t, \pi, s_{t+1})\right]}{E\left[\sum_{t=1}^{n} t(s_t, \pi, s_{t+1})\right]} \quad (16)$$

Where π is the policy, r is the acquired reward, and t is the transition time. Finding the policy that returns the highest average reward is the primary goal of the average reward RL techniques. R-learning [25], variants of R-learning [40], and R-SMART [39,41] are some of the average reward RL algorithms.

This research uses the R-SMART algorithm to find the optimal policy for the combined production, maintenance, and quality problem. The algorithm estimates the action value of each state as $Q(s, a)$ and attempts to find the optimum policy by maximizing the average reward. The agent updates the action value of each state as follows:

$$Q(s,a) \leftarrow (1-\alpha^t)Q(s,a) + \alpha^t \left[r(s,a,s') - \rho^t t(s,a,s') + \eta max_{a' \in A(s')} Q(s',a')\right] \quad (17)$$

Where ρ is the average reward, *r(s, a, s')* is the transition reward from the current state to the next state, and *t(s, a, s')* denotes the transition time. The average reward is updated when the agent chooses the greedy action in the current state by the following equations:

$$R(t) \leftarrow R(t-1) + r(s,a,s') \quad (18)$$

$$T(t) \leftarrow T(t-1) + t(s,a,s') \quad (19)$$

$$\rho^{t+1} \leftarrow (1-\beta^t)\rho^t + \beta^k \left[\frac{R(t)}{T(t)}\right] \quad (20)$$

In the above equations, α and β are the learning rates, and η is the positive scalar. The interaction between the decision-maker agent, the manufacturer agent, and the environment is shown in Figure 4. The $Q(s,a)$ is updated after each action by Eq. (17), and this process continues until the combined optimal policy is achieved.

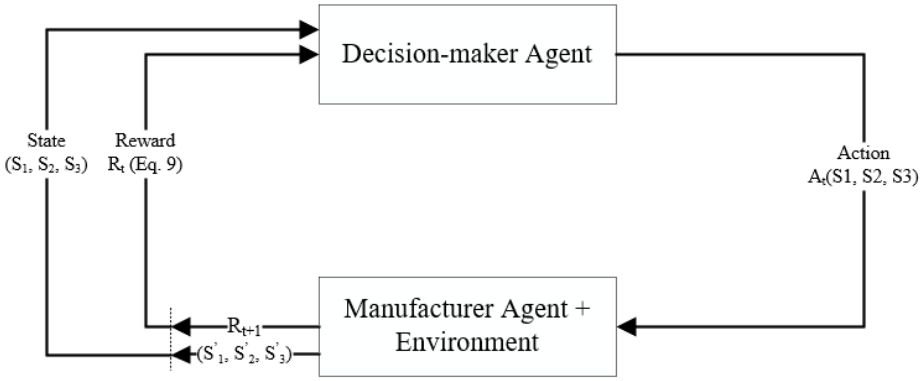

Fig. 4. The RL-based interaction of decision-maker agent

5. Evaluating the Control Policy

To assess the quality of the control policy acquired by the decision-maker agent, three alternatives are considered: The heuristic approach called (s, S)-CBM, simulation-optimization technique, and random decision-maker agent. Additionally, the Monte Carlo method is employed to compare the results of the proposed method with those of the alternatives.

5.1. (s, S)-CBM Policy

Push production systems usually use a well-known threshold-type control policy called (s, S). In this policy, when the inventory level drops below level s, the production is authorized to increase the inventory position to level S. In addition, the production and inventory control policies can be integrated with maintenance policies such as corrective maintenance (CM), preventive maintenance (PM), or condition-based maintenance (CBM) [42].

In this research, the (s, S)-CBM control policy has been used to evaluate the obtained control policy by the decision-maker agent. The algorithm of the policy is illustrated in Figure 5.

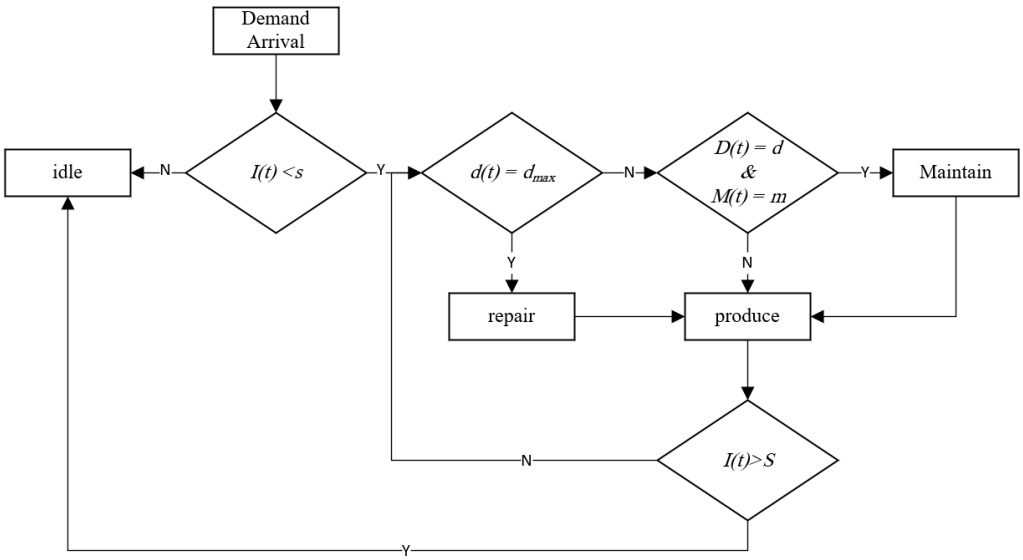

Fig. 1. The production system operation under (s, S)-CBM policy

The quality of the solution obtained from the heuristic approach is evidently reliant on the values of its parameters. To address this dependency, this paper employs a simulation-optimization technique to determine the optimal parameter value of the heuristic policy in each scenario.

5.2. Simulation – Optimization Techniques

The term simulation-optimization refers to the techniques applied to optimize the stochastic problems of parametric optimization [43]. SO involves searching for the value of the input parameters of the simulation model in a way that a specific objective is optimized.

The integrated production, maintenance, and quality control of this research can be formulated as discrete parametric optimization. The input parameters are the set of feasible actions in each state, and SO can be utilized to find the best action in each state such that the total reward is maximized.

This paper employs a commercial simulation optimization package to find the optimal parameter's value of the (s, S)-CBM policy and the optimal policy of the combined optimization problem. The SO package integrates metaheuristic approaches such as Scatter Search, Tabu Search, and Neural Networks into a single optimization procedure. Figure 6 illustrates the interaction between the SO package and the simulation model.

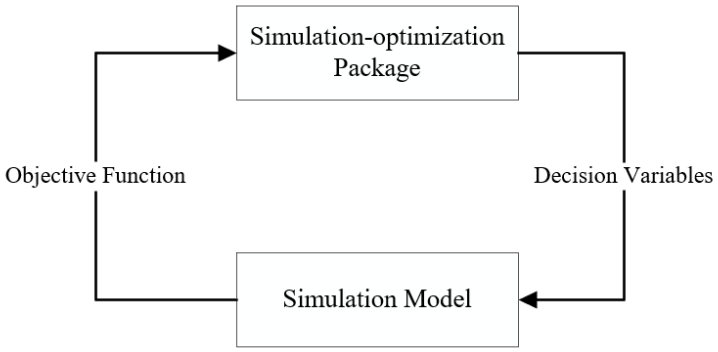

Fig. 2. The simulation-optimization process

As shown, the decision variables (which serve as input parameters for the simulation model) are determined by the SO package. The simulation model runs for a specific duration, and the cumulative reward is then returned to the package. This iterative process persists until the optimal policy is obtained. Table 1 outlines the decision variables and the objective functions for optimizing the (s, S)-CBM policy and the combined policy.

Table 1
The decision variables and the objective function of the simulation-optimization approach

Optimization problem	Decision Variables	Objective Function
(s, S)-CBM	s, S, l_r, l_m, U	Minimizing the cumulative value of Equation 11
Combined policy	$A(s_1, s_2, s_3) \forall s_1, s_2, s_3 \epsilon S$	

The SO process involves tuning parameters and hyperparameters. The number of iterations is determined to a point where the algorithm ceases to improve the objective function further. The number of replications is calculated based on the Central Limit Theorem (CLT), which suggests that a sample size equal to or greater than 30 is often considered sufficient. Here, the minimum of replications is set to 30, with a maximum of 100. The stop time for the simulation model fixed at 10,000 minutes, ensuring a sufficient number of events in the production system. Notably, the hyperparameters of the optimization process are automatically tuned by the commercial SO package.

5.3. Random Decision Maker Agent

As described in Section 4.2, the main goal of the decision-maker agent is to map situations into actions to maximize the reward. When the decision-maker agent always chooses the action with the maximum expected return, the action selection policy is "greedy." On the other hand, when the decision-maker agent chooses the action randomly, the action selection policy is "random." There is a well-known action selection policy in RL called "epsilon-greedy" that is illustrated by the following expression:

$$A_t = \begin{cases} argmax_{a \in A(s)} Q_t(s,a), & with\ probability\ 1-\varepsilon \\ any\ action\ in\ A(s) \quad , & with\ probability\ \epsilon \end{cases} \tag{21}$$

Where ε is the probability of taking a random action and $argmax_a$ denotes the value of action a at which the $Q(s,a)$ takes its maximal value. The parameter ε real-value is in the range (0,1). The $\varepsilon - greedy$ policy ensures that the agent explores the states of the system sufficiently.

In this research, the *ε-greedy* policy has been used to obtain the optimal policy by the decision-maker agent. However, for the random decision-maker agent (Random DMA) alternative, the ε is set to 1 to force the agent to select all actions randomly.

6. Numerical Result

In order to evaluate the efficiency of the proposed method, seven scenarios were conducted, as outlined in Table 3. These scenarios cover diverse system conditions and can be categorized as:

i. The base case (Scenario 1)
ii. The effect of increasing demand rate and quantity on the performance of the policies (Scenarios 2 and 3)
iii. The simultaneous effect of increasing demand rate and quantity, and the probabilities of breakdowns and producing low-quality productions on the performance of the policies (Scenarios 4 and 5)

iv. The simultaneous effect of increasing production rate, demand rate, and quantity, and the probabilities of breakdowns and producing low-quality productions on the performance of the policies (Scenarios 6 and 7)

All scenarios share the input parameters that are illustrated in Table 2.

Table 2
Input parameters of the agent-based simulation model

I_{max}	B_{max}	d_{max}	U_{max}	$f_{d,p}$	C_p	C_m	C_r	C_b	C_q	C_h	C_s	C_l	P
10	10	6	2	(0.04, 0.04, 0.05, 0.05, 0.06, 0.07)	0.5	50	150	300	1.5	0.3	0.6	100	2.1

The stress on the production system escalates from the second to the fifth scenario, gradually improving in the last two cases. Nevertheless, the system remains complex, marked by an elevated probability of breakdowns and low-quality productions. Consequently, the policies must authorize the production, maintenance, and repair activities properly to prevent increases in missed orders and breakdowns.

Table 3
Summary of scenarios

	$1/\lambda_p$	$1/\lambda_r$	$1/\lambda_m$	$1/\lambda_d$	λ_n	$1/\lambda_b$	B_d	Q_d
Scenario 1	1	20	2	10	2	25	(0, 0, 0, 0.001, 0.007, 0.01, 1)	(0, 0, 0.01, 0.05, 0.1, 0.15, 1)
Scenario 2	1	20	2	6	2	25	(0, 0, 0, 0.001, 0.007, 0.01, 1)	(0, 0, 0.01, 0.05, 0.1, 0.15, 1)
Scenario 3	1	20	2	5	3	25	(0, 0, 0, 0.001, 0.007, 0.01, 1)	(0, 0, 0.01, 0.05, 0.1, 0.15, 1)
Scenario 4	1	20	2	6	2	25	(0, 0, 0.001, 0.007, 0.015, 0.05, 1)	(0, 0.02, 0.05, 0.1, 0.15, 0.2, 1)
Scenario 5	1	20	2	5	3	25	(0, 0, 0.001, 0.007, 0.015, 0.05, 1)	(0, 0.02, 0.05, 0.1, 0.15, 0.2, 1)
Scenario 6	0.5	20	2	5	3	25	(0, 0, 0.001, 0.007, 0.015, 0.05, 1)	(0, 0.02, 0.05, 0.1, 0.15, 0.2, 1)
Scenario 7	0.5	20	2	6	2	25	(0, 0.01, 0.02, 0.05, 0.07, 0.1, 1)	(0, 0.05, 0.07, 0.1, 0.2, 0.25, 1)

The agent-based simulation model of each scenario was evaluated up to 1000 iterations by the Monte-Carlo method, and each iteration lasted over 10,000 minutes (7 days) to evaluate the obtained policy of the proposed method and the other alternatives.

6.1. Obtaining the Optimal Integrated Policy by the RL-Based Decision-Maker Agent

As described in Section 4.2, the decision-maker agent has applied the average reward reinforcement learning algorithm called R-Smart to obtain the optimal integrated production, maintenance, and quality policy. The number of episodes and steps has been set to 20,000 and 10,000, respectively. Each episode starts with a unique random seed number to capture all the events. The positive scalar η is set to 0.999. The learning rate is set to be $\alpha = 0.01$ and $\beta = 0.009$. The parameter for $\varepsilon - greedy$ is initially set to be 0.1 and linearly reduced afterward.

The R-learning average reward ρ obtained by the decision-maker agent during the learning phase of Scenario One is depicted in Figure 7. Notably, as the number of learning episodes increases, the average reward received by the agent converges. This convergence behavior is consistent across all other scenarios as well.

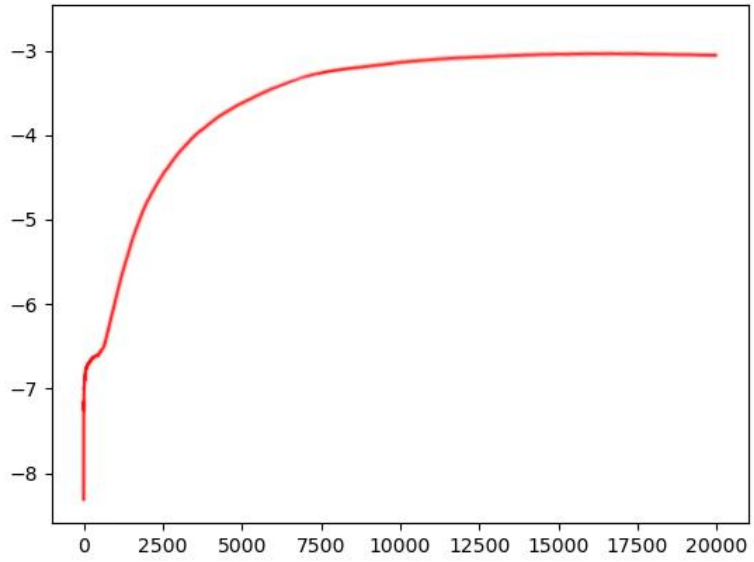

Fig. 7. The obtained average reward (ρ) by the RL-based DMA in Scenario 1

6.2. Optimal Parameter's Values of the (s, S)-CBM Policy

As shown in Figure 5, the (s, S)-CBM heuristic method has five parameters that directly affect the efficiency of the algorithm:

i. The value of lower inventory level *s*
ii. The value of upper inventory level *S*
iii. The limit of the repair activities l_r based on the deterioration level *d*
iv. The limit of the Maintenance activities l_m based on the deterioration level d and the number of authorized maintenance activities U

In order to compare the results of the proposed method and (s, S)-CBM policy on the same basis, the best values of the listed parameters need to be acquired. So, the agent-based simulation model of the production system is modified based on the (s, S)-CBM policy. Then, the input parameters are adjusted for each scenario. Finally, the simulation models are optimized by the SO package. The number of iterations is set to be 5000, and each iteration contains 30 to 100 replications. Each iteration's simulation model stop time is also set to 10,000 minutes. The linear constraints are defined to make the *S* value higher than the *s* value, and the l_m and l_r values are greater than zero. The optimal parameter's value yielded by the SO package is shown in Table 4.

Table 4
The optimal value of the (s, S)-CBM policy Parameters

	s	S	l_r	l_m	U
Scenario 1	1	2	5	2	0
Scenario 2	3	4	5	2	0
Scenario 3	8	9	5	2	0
Scenario 4	3	4	4	1	0
Scenario 5	7	9	4	2	0
Scenario 6	6	7	5	1	0
Scenario 7	2	5	2	1	0

6.3. Obtaining the Optimal Integrated Policy by the Simulation-Optimization

The objective of the simulation-optimization method is to minimize the cumulative value of Equation 11. In this regard, the number of iterations is set to be 20,000. Since the agent-based model

of the production system is stochastic, each iteration comprises 30 to 100 replications under similar conditions, enhancing the reliability and validity of the obtained objective function. Replications stop after minimum replications when a confidence level (90%) is reached or replications are continued so that the result falls within the confidence level. The simulation model stop time for each iteration is set to 10,000 minutes. The SO Package determines a set of state-action pairs in each iteration to minimize the objective function. Therefore, the number of decision variables equals the number of states of the production system, which is calculated by Equation 5. The decision variables represent the authorized action in each state.

At the end of the optimization process, the best set of state-action pairs is considered the optimal integrated policy.

6.4. Comparison between RL-Based Decision-Maker Agent and the Other Alternatives

Figure 8 illustrates the average cumulative reward of Equation 11 obtained by the proposed method and the other three alternatives. The results show that the proposed RL-based decision-maker agent (RL-based DMA) performed better in most scenarios than the other alternatives. However, in some scenarios, e.g., 3 and 5, the heuristic policy has performed slightly better than the RL-based policy. For comprehensive assessment, Figure 9 also displays the standard deviation of the results obtained by the control policies.

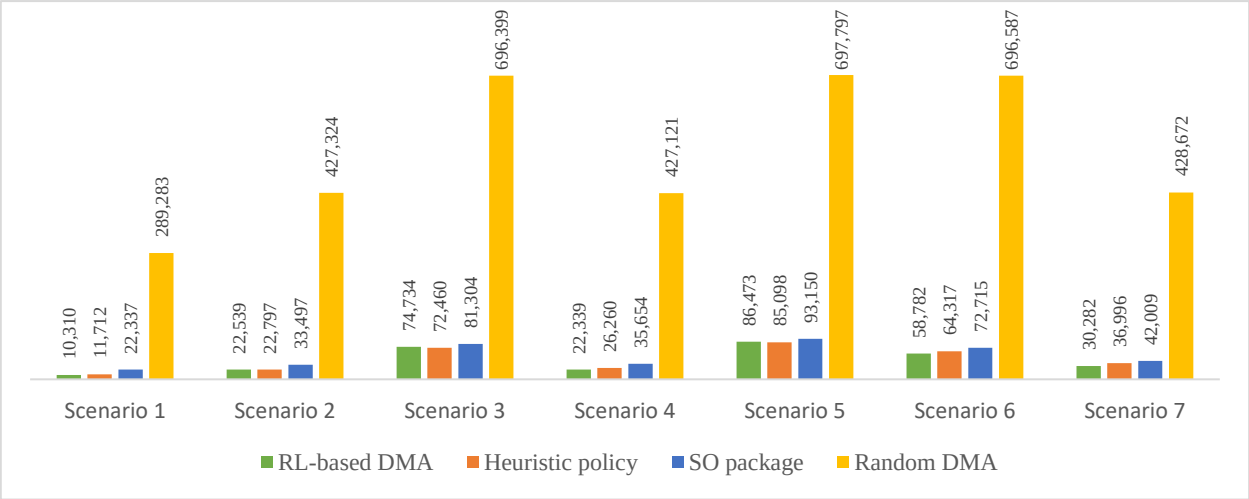

Fig. 8. The cumulative reward (Eq. (11)) of the control policies

According to the probabilistic nature of the simulation model, events vary in each model execution. Consequently, low variance indicates effective policy performance across different system states. The RL-based policy exhibited the minimum standard deviation. However, despite achieving commendable results in average cumulative reward, the heuristic policy consistently showed a higher standard deviation compared to other policies.

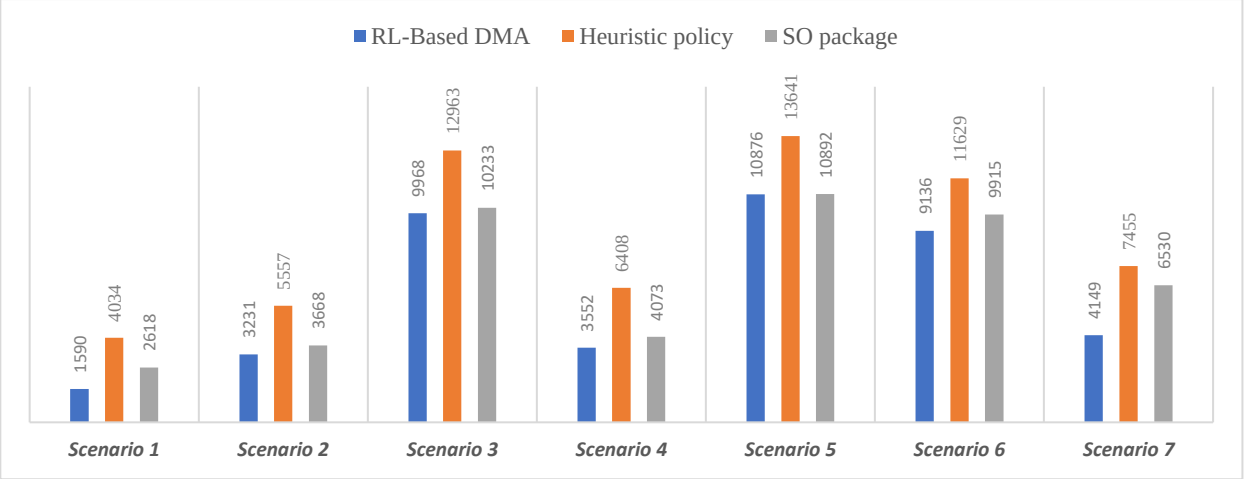

Fig. 9. The standard deviation of cumulative reward (Eq. (11)) of the control policies

For an in-depth understanding of each control policy, the following details outline their performance. The average inventory level, missed orders, authorized maintenance, authorized repair, machine breakdowns, and low-quality production achieved by the top three policies in each scenario are discussed below.

In Scenario 1, as depicted in Figure 10, the RL-based DMA achieved a lower average inventory level. This reduction in holding inventory did not lead to an increase in total missed orders, positioning this agent with the minimum number of missed orders. Notably, the RL-based DMA prioritized maintenance activities over repair activities, earning it the second rank in machine breakdowns and low-quality production.

The heuristic policy, on the other hand, outperformed in both maintenance and repair activities and in producing high-quality products. Although the heuristic policy authorized nearly half the number of maintenance activities compared to the RL-based DMA and almost the same number of repair activities, its total count of low-quality productions was 25% lower than that of the RL-based DMA.

Contrarily, the SO package authorized more maintenance and repair activities than both the RL-based DMA and the heuristic policy. Consequently, the lowest number of low-quality productions was achieved. However, the emphasis on maintenance and repair activities led to an increase in missed orders.

Fig. 10. The performance of the policies in Scenario 1

In Scenario 2, where the demand rate has increased, the RL-based DMA and the heuristic policy achieved identical cumulative rewards (Figure 11). The RL-based DMA prioritized low inventory holding and authorized more maintenance activities, while the heuristic policy aimed to balance the number of maintenance and repair activities to mitigate low-quality productions.

The SO package, however, authorized more repair activities than maintenance activities, resulting in higher breakdowns and low-quality production compared to the other alternatives. Despite this, the SO package increased the inventory level to meet the elevated demand rate, leading to a reduction in missed orders.

In Scenario 3, the production system faced more challenging conditions compared to previous scenarios. Alongside an increased demand rate, the amount of each demand has also risen. In this context, the RL-based DMA performed slightly less effectively than the heuristic policy. The RL-based DMA aimed to reduce costs by maintaining lower inventory levels and authorizing more maintenance activities (Figure 11). However, this approach led to an increase in the number of missed orders.

On the other hand, the heuristic policy, by authorizing more repair activities and increasing the inventory level, succeeded in reducing the number of missed orders and low-quality productions. The SO package, while authorizing more repair activities compared to other policies, still had the highest number of breakdowns and low-quality productions. However, the SO's policy managed to yield a lower number of missed orders compared to the RL-based DMA.

In Scenario 4, accounting for the increased probability of machine breakdowns and the production of low-quality products observed in Scenario 2, the RL-based DMA effectively maintained a lower inventory level and reduced missed orders (Figure 12). The heuristic policy outperformed in authorizing maintenance and repair activities, resulting in lower breakdowns and low-quality productions. Despite the high inventory level, the SO package achieved fewer missed orders compared to the heuristic policy.

The conditions considered in Scenario 5 are more complex, involving an increased demand rate, higher demand amount, elevated probability of breakdowns, and low-quality production. Despite the complexity, the holding inventory level for all policies is nearly identical. However, the RL-based agent achieved a lower number of missed orders compared to other policies (Figure 12). Similar to previous scenarios, the heuristic policy authorized the minimum required maintenance and repair activities, resulting in the highest number of high-quality products.

In Scenario 6, where an increase in production rate is considered, the new system configuration has notably improved the performance of the RL-based DMA. This agent achieved superior results in all aspects compared to the other policies, maintaining low inventory levels and obtaining the minimum number of missed orders, machine breakdowns, and low-quality productions (Figure 13).

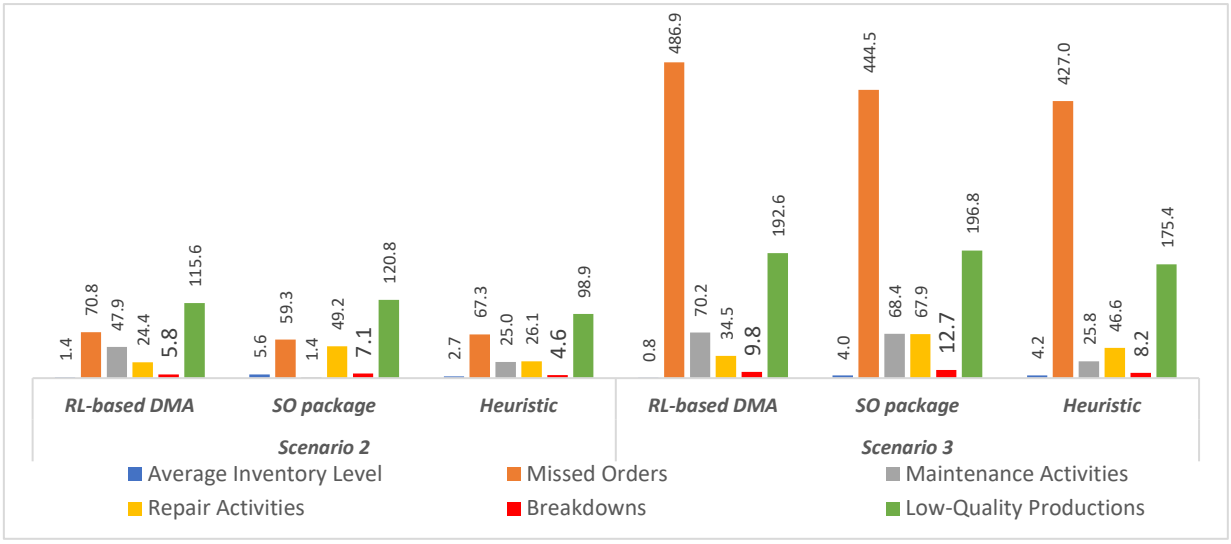

Fig. 11. The performance of the policies in Scenario 2 and Scenario 3

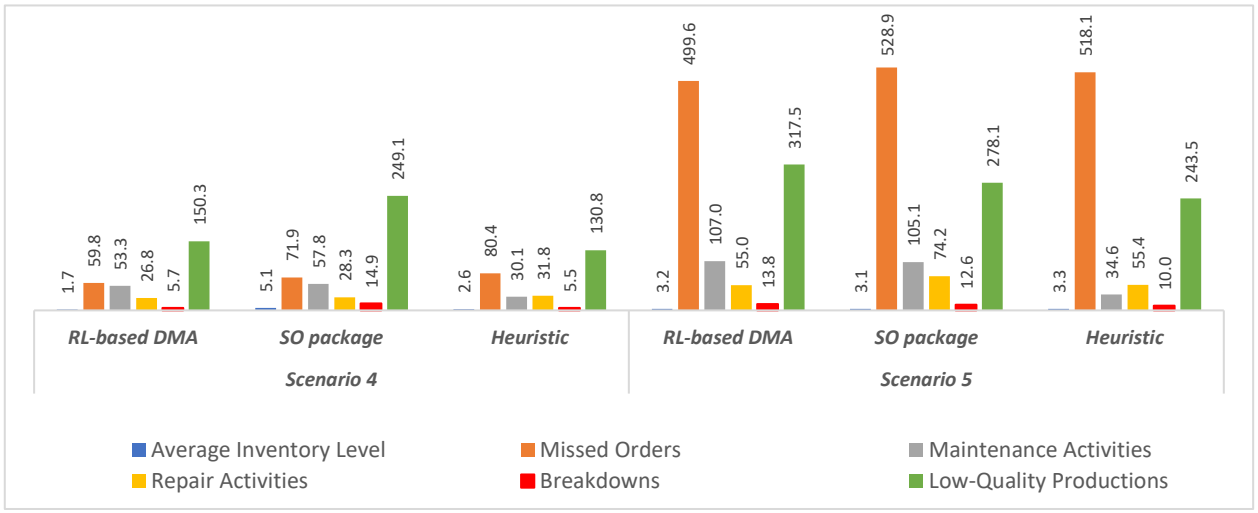

Fig. 12. The performance of the policies in Scenario 4 and Scenario 5

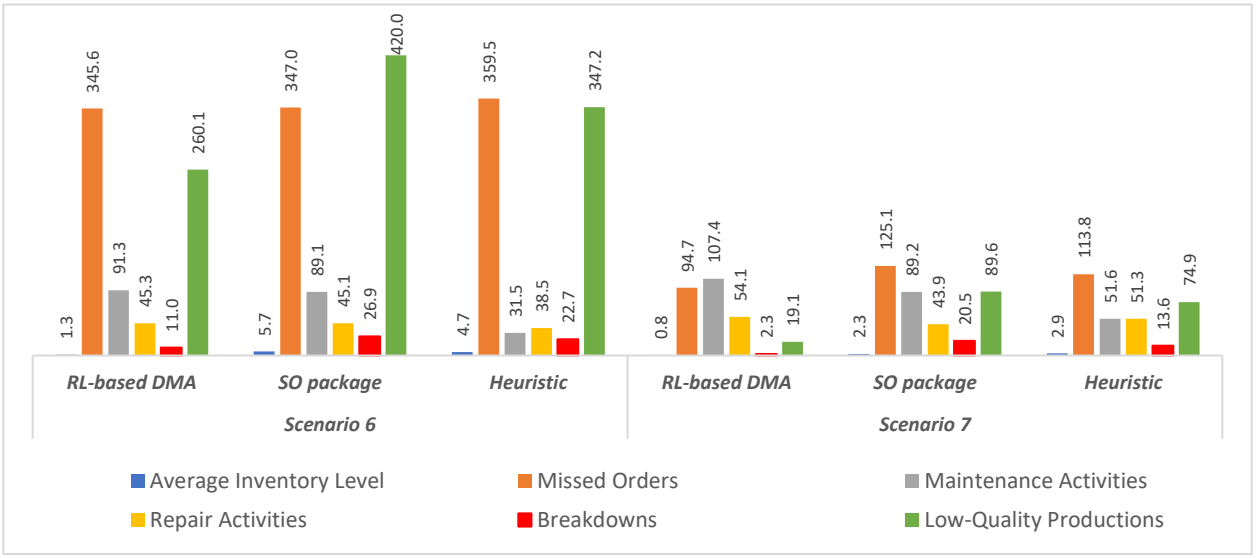

Fig. 13. The performance of the policies in Scenario 6 and Scenario 7

In the final scenario, an increased probability of machine breakdowns and low-quality production has been introduced compared to the previous scenario. As illustrated in Figure 13, the RL-based DMA demonstrated strong performance across all aspects. The agent achieved the minimum number of machine breakdowns and low-quality production by timely authorizing maintenance and repair activities. Similar to previous scenarios, the RL-based DMA prioritized maintenance activities over repair activities, maintaining a lower inventory level and minimizing the number of missed orders.

Throughout the scenarios, The RL-based DMA consistently showed a tendency to hold less inventory and focus more on maintenance activities. Conversely, the heuristic policy tended to authorize the minimum required maintenance and repair activities. However, the behavior of the SO package varied across scenarios.

The RL-based DMA exhibits prominent performance, especially in the last two scenarios. The heuristic policy also demonstrated promising results in all scenarios due to parameter values obtained through SO. Despite this, the optimal integrated policy produced by the SO package did not outperform the other policies.

It's noteworthy that the RL-based DMA determines the policy in every decision epoch, with a limited feasible action set. The agent observes the reward of the corresponding action in the next step, leading to policy improvement with each visit to a specific state. In contrast, the SO algorithm must decide on all state-action pairs at every simulation model iteration. Consequently, the computational cost of acquiring the optimal integrated policy significantly increases with a large number of state spaces.

7. Conclusion and Discussion

This paper investigated the joint optimal production, maintenance, and quality policy in a deteriorating single-machine production system. While extensive research explores the joint optimal policy, limited attention is given to the simultaneous consideration of production, maintenance, and quality. Moreover, existing studies often overlook the crucial assumption of limited maintenance activities between two consecutive repair activities. In addition, the literature that applied agent-based modeling and reinforcement learning is rare. This paper aimed to investigate the production model with a more realistic assumption and evaluate the combination of agent-based modeling and reinforcement learning to obtain the joint optimal policy.

To cover the gap, we formulated the problem as a continuous semi-Markov decision process, facilitating real-time decision-making. The solution involved employing average-reward reinforcement learning in conjunction with agent-based modeling. To assess the quality of the acquired policy, we subjected it to evaluation by the heuristic policy, simulation-optimization method, and the random decision-maker agent.

The study encompassed seven scenarios, strategically designed to explore policy behavior under varying conditions, including increasing demand, probability of breakdowns, probability of producing low-quality products, and changes in the production rate of the system.

Policy evaluation was conducted based on the average and standard deviation of the cumulative reward. The average insights into the overall performance of the policy, while the standard deviation offered a measure of its stability amidst stochastic events within the system.

The results highlight the superior performance of the proposed RL-based method across most scenarios, consistently achieving better outcomes than the alternatives and displaying the lowest standard deviation of rewards. The heuristic policy also demonstrated an acceptable average cumulative reward, albeit with a higher standard deviation than the alternatives. SO package secured

a third-place ranking in cumulative average reward and second place in the standard deviation of rewards.

Notably, the heuristic policy slightly outperformed the RL-based policy in Scenarios 3 and 5. Scenario 3 incorporated an increased demand and probability of breakdowns, while Scenario 5 introduced an augmented demand, breakdown probability, and probability of producing low-quality products. It is important to note that the demand rate in both scenarios is less extreme than in Scenario 4 and 6. The results suggest that specific algorithms may yield better average cumulative rewards under certain conditions of the production system. However, as emphasized earlier, the comprehensive evaluation of policy performance should consider both cumulative reward and the standard deviation.

The reason for the higher standard deviation in the heuristic policy is that, regardless of the stochasticity of the production system, the heuristic policy consistently adopts a specific approach, which is considered a low performance in some system states. In contrast, the RL-based policy benefits from the decision-maker agent repeatedly observing each state, updating the appropriate state-action value based on acquired rewards. In the SO policy, the replication process in each iteration contributes to the reduction of the standard deviation.

The average cumulative reward of the proposed RL-based method closely aligned with the heuristic policy in scenarios where the production system was under pressure. This similarity arises because the learning agent tends to perform actions similar to the heuristic algorithm. However, The RL-based method demonstrated significantly superior performance in scenarios characterized by a balance between production and demand. In such cases, the RL-based decision-maker agent has greater flexibility in making different decisions.

In general, it can be asserted that the RL approach and the metaheuristics in connection with the simulation model (SO) approach will lead to a better optimal policy than the heuristics in the stochastic systems. Despite its effectiveness in some instances, the heuristics may have limitations in providing high-quality actions for all stochastic states due to its high variances. The heuristics approach is often based on approximations, rules of thumb, or expert knowledge, which may not be optimal or applicable in every scenario. In the SO approach, the simulation model is replicated varying number of times to converge to the real objective value, consequently reducing the variance of the policy. On the other hand, the RL agent can adapt and improve its policy over time based on feedback from the stochastic environment. This adaptive nature empowers the agent to discover better solutions, potentially surpassing heuristic approaches in terms of policy quality with low variances across a broader range of system states.

As the dimensions of the problem increase, and consequently, the number of possible states expands, RL, especially when leveraging neural network-based learning (Deep RL), could prove more efficient in obtaining the optimal joint policy. In contrast to SO methods, RL has the capability to explore a subset of states and provide an optimal joint policy for the problem.

Finally, it could be concluded that combining agent-based modeling with RL-based decision-maker agents eases the integration and interactions between agents. In addition, RL has excellent potential to solve multi-state optimization problems. However, there are some limitations as follows:

i. Building an agent-based model of the production system is more time-consuming than discrete event simulation.
ii. The algorithm is highly related and sensitive to parameters such as learning rates, scalars, number of steps and episodes, and exploration policy, making the calibration process time-consuming.

iii. Despite the convergence of the policy, the acquired policy may not necessarily be the best. However, utilizing a unique random seed number in each episode has been observed to lead to a better policy in stochastic systems.
iv. The selection process of different RL-based algorithms, along with the trial-and-error calibration of each algorithm, poses significant challenges in the implementation of reinforcement learning.

For future work, this research can be extended with the following suggestions:

i. Explore additional scenarios to evaluate policies using different combinations of demand, breakdown probabilities, probability of low-quality production, and production rate.
ii. Investigate the combination of meta-heuristic methods with RL algorithms for parameter calibration, reducing algorithm convergence time, and enhancing the quality of the acquired policy.
iii. Explore the application of other reinforcement learning algorithms, such as Q-learning, to compare their efficiency in achieving the joint optimal policy.
iv. Introduce greater complexity, e.g., multi-production machines and multi-products, and explore the application of multi-agent RL or Deep RL.

In addition to the aforementioned areas for future work, there is a proposal to evaluate the time required to obtain the joint optimal policy through SO and RL methods.

Author Contributions

Conceptualization, M.R.N, S.E.N, A.M, and F.M.S.; methodology, M.R.N, S.E.N, A.M, and F.M.S.; software, M.R.N, S.E.N, A.M, and F.M.S.; validation, M.R.N, S.E.N, A.M, and F.M.S.; formal analysis, M.R.N, S.E.N, A.M, and F.M.S.; investigation, M.R.N, S.E.N, A.M, and F.M.S.; resources, M.R.N, S.E.N, A.M, and F.M.S.; data curation, M.R.N, S.E.N, A.M, and F.M.S.; writing—original draft preparation, M.R.N, S.E.N, A.M, and F.M.S.; writing—review and editing, M.R.N, S.E.N, A.M, and F.M.S.; visualization, M.R.N, S.E.N, A.M, and F.M.S.; supervision, M.R.N, S.E.N, A.M, and F.M.S.; project administration, M.R.N, S.E.N, A.M, and F.M.S.; funding acquisition, M.R.N, S.E.N, A.M, and F.M.S. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

The data used to support the findings of this study are included in the article.

Conflicts of Interest

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

Acknowledgment

This research was not funded by any grant.

References

[1] Saidi-Mehrabad, M., Paydar, M. M., & Aalaei, A. (2013). Production planning and worker training in dynamic manufacturing systems. Journal of Manufacturing Systems, 32(2), 308–314. https://doi.org/10.1016/j.jmsy.2012.12.007

[2] Liu, Q., Dong, M., & Chen, F. F. (2018). Single-machine-based joint optimization of predictive maintenance planning and production scheduling. Robotics and Computer-Integrated Manufacturing, 51, 238–247. https://doi.org/10.1016/j.rcim.2018.01.002

[3] Rivera-Gómez, H., Gharbi, A., Kenné, J.-P., Montaño-Arango, O., & Corona-Armenta, J. R. (2020). Joint optimization of production and maintenance strategies considering a dynamic sampling strategy for a deteriorating system. Computers & Industrial Engineering, 140, 106273. https://doi.org/10.1016/j.cie.2020.106273

[4] Sharp, M., Ak, R., & Hedberg Jr, T. (2018). A survey of the advancing use and development of machine learning in smart manufacturing. Journal of Manufacturing Systems, 48, 170–179. https://doi.org/10.1016/j.jmsy.2018.02.004

[5] Sharma, A., Zhang, Z., & Rai, R. (2021). The interpretive model of manufacturing: a theoretical framework and research agenda for machine learning in manufacturing. International Journal of Production Research, 59(16), 4960-4994. https://doi.org/10.1080/00207543.2021.1930234

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

[7] Das, T. K., Gosavi, A., Mahadevan, S., & Marchalleck, N. (1999). Solving semi-Markov decision problems using average reward reinforcement learning. Management Science, 45(4), 560–574. https://doi.org/10.1287/mnsc.45.4.560

[8] Beheshti Fakher, H., Nourelfath, M., & Gendreau, M. (2017). A cost minimisation model for joint production and maintenance planning under quality constraints. International Journal of Production Research, 55(8), 2163–2176. https://doi.org/10.1080/00207543.2016.1201605

[9] Nourelfath, M., Nahas, N., & Ben-Daya, M. (2016). Integrated preventive maintenance and production decisions for imperfect processes. Reliability Engineering & System Safety, 148, 21–31. https://doi.org/10.1016/j.ress.2015.11.015

[10] Aghezzaf, E.-H., Khatab, A., & le Tam, P. (2016). Optimizing production and imperfect preventive maintenance planning' s integration in failure-prone manufacturing systems. Reliability Engineering & System Safety, 145, 190–198. https://doi.org/10.1016/j.ress.2015.09.017

[11] Chen, Y.-C. (2013). An optimal production and inspection strategy with preventive maintenance error and rework. Journal of Manufacturing Systems, 32(1), 99–106. https://doi.org/10.1016/j.jmsy.2012.07.010

[12] Chouikhi, H., Khatab, A., & Rezg, N. (2014). A condition-based maintenance policy for a production system under excessive environmental degradation. Journal of Intelligent Manufacturing, 25, 727–737. https://doi.org/10.1007/s10845-012-0715-9

[13] Khatab, A., Diallo, C., Aghezzaf, E.-H., & Venkatadri, U. (2019). Integrated production quality and condition-based maintenance optimisation for a stochastically deteriorating manufacturing system. International Journal of Production Research, 57(8), 2480–2497. https://doi.org/10.1080/00207543.2018.1521021

[14] Hadian, S. M., Farughi, H., & Rasay, H. (2021). Joint planning of maintenance, buffer stock and quality control for unreliable, imperfect manufacturing systems. Computers & Industrial Engineering, 157, 107304. https://doi.org/10.1016/j.cie.2021.107304

[15] Cheng, G. Q., Zhou, B. H., & Li, L. (2018). Integrated production, quality control and condition-based maintenance for imperfect production systems. Reliability Engineering & System Safety, 175, 251–264. https://doi.org/10.1016/j.ress.2018.03.025

[16] Bouslah, B., Gharbi, A., & Pellerin, R. (2016). Integrated production, sampling quality control and maintenance of deteriorating production systems with AOQL constraint. Omega, 61, 110–126. https://doi.org/10.1016/j.omega.2015.07.012

[17] Bouslah, B., Gharbi, A., & Pellerin, R. (2018). Joint production, quality and maintenance control of a two-machine line subject to operation-dependent and quality-dependent failures. International Journal of Production Economics, 195, 210–226. https://doi.org/10.1016/j.ijpe.2017.10.016

[18] Fakher, H. B., Nourelfath, M., & Gendreau, M. (2018). Integrating production, maintenance and quality: A multi-period multi-product profit-maximization model. Reliability Engineering & System Safety, 170, 191–201. https://doi.org/10.1016/j.ress.2017.10.024

[19] Rivera-Gómez, H., Gharbi, A., & Kenné, J. P. (2013). Joint production and major maintenance planning policy of a manufacturing system with deteriorating quality. International Journal of Production Economics, 146(2), 575–587. https://doi.org/10.1016/j.ijpe.2013.08.006

[20] Ghaleb, M., Taghipour, S., Sharifi, M., & Zolfagharinia, H. (2020). Integrated production and maintenance scheduling for a single degrading machine with deterioration-based failures. Computers & Industrial Engineering, 143, 106432. https://doi.org/10.1016/j.cie.2020.106432
[21] Sharifi, M., & Taghipour, S. (2021). Optimal production and maintenance scheduling for a degrading multi-failure modes single-machine production environment. Applied Soft Computing, 106, 107312. https://doi.org/10.1016/j.asoc.2021.107312
[22] Kang, Z., Catal, C., & Tekinerdogan, B. (2020). Machine learning applications in production lines: A systematic literature review. Computers & Industrial Engineering, 149, 106773. https://doi.org/10.1016/j.cie.2020.106773
[23] Kuhnle, A., Jakubik, J., & Lanza, G. (2019). Reinforcement learning for opportunistic maintenance optimization. Production Engineering, 13(1), 33–41. https://doi.org/10.1007/s11740-018-0855-7
[24] Xanthopoulos, A. S., Kiatipis, A., Koulouriotis, D. E., & Stieger, S. (2017). Reinforcement learning-based and parametric production-maintenance control policies for a deteriorating manufacturing system. IEEE Access, 6, 576–588. https://doi.org/10.1109/ACCESS.2017.2771827
[25] Schwartz, A. (1993). A reinforcement learning method for maximizing undiscounted rewards. Proceedings of the Tenth International Conference on International Conference on Machine Learning)pp. 298–305). Amherst MA USA : Morgan Kaufmann Publishers Inc. https://dl.acm.org/doi/10.5555/3091529.3091568
[26] Paraschos, P. D., Koulinas, G. K., & Koulouriotis, D. E. (2020). Reinforcement learning for combined production-maintenance and quality control of a manufacturing system with deterioration failures. Journal of Manufacturing Systems, 56, 470–483. https://doi.org/10.1016/j.jmsy.2020.07.004
[27] Yang, H., Li, W., & Wang, B. (2021). Joint optimization of preventive maintenance and production scheduling for multi-state production systems based on reinforcement learning. Reliability Engineering & System Safety, 214, 107713. https://doi.org/10.1016/j.ress.2021.107713
[28] Wang, H., Yan, Q., & Zhang, S. (2021). Integrated scheduling and flexible maintenance in deteriorating multi-state single machine system using a reinforcement learning approach. Advanced Engineering Informatics, 49, 101339. https://doi.org/10.1016/j.aei.2021.101339
[29] Huang, J., Chang, Q., & Arinez, J. (2020). Deep reinforcement learning based preventive maintenance policy for serial production lines. Expert Systems with Applications, 160, 113701. https://doi.org/10.1016/j.eswa.2020.113701
[30] Rodríguez, M. L. R., Kubler, S., de Giorgio, A., Cordy, M., Robert, J., & Le Traon, Y. (2022). Multi-agent deep reinforcement learning based Predictive Maintenance on parallel machines. Robotics and Computer-Integrated Manufacturing, 78, 102406. https://doi.org/10.1016/j.rcim.2022.102406
[31] Lee, J., & Mitici, M. (2023). Deep reinforcement learning for predictive aircraft maintenance using probabilistic remaining-useful-life prognostics. Reliability Engineering & System Safety, 230, 108908. https://doi.org/10.1016/j.ress.2022.108908
[32] Wesendrup, K., & Hellingrath, B. (2023). Post-prognostics demand management, production, spare parts and maintenance planning for a single-machine system using Reinforcement Learning. Computers & Industrial Engineering, 179, 109216. https://doi.org/10.1016/j.cie.2023.109216
[33] Ye, Z., Cai, Z., Yang, H., Si, S., & Zhou, F. (2023). Joint optimization of maintenance and quality inspection for manufacturing networks based on deep reinforcement learning. Reliability Engineering & System Safety, 236, 109290. https://doi.org/10.1016/j.ress.2023.109290
[34] Geurtsen, M., Adan, I., & Atan, Z. (2023). Deep reinforcement learning for optimal planning of assembly line maintenance. Journal of Manufacturing Systems, 69, 170-188. https://doi.org/10.1016/j.jmsy.2023.05.011
[35] Macal, C., & North, M. (2014). Introductory tutorial: Agent-based modeling and simulation. Proceedings of the 2014 Winter Simulation Conference. IEEE. https://doi.org/10.1109/WSC.2014.7019874
[36] Cuevas, E. (2020). An agent-based model to evaluate the COVID-19 transmission risks in facilities. Computers in Biology and Medicine, 121, 103827. https://doi.org/10.1016/j.compbiomed.2020.103827
[37] Santos, F., Nunes, I., & Bazzan, A. L. C. (2020). Quantitatively assessing the benefits of model-driven development in agent-based modeling and simulation. Simulation Modelling Practice and Theory, 104, 102126. https://doi.org/10.1016/j.simpat.2020.102126
[38] Borshchev, A., & Filippov, A. (2004). From system dynamics and discrete event to practical agent based modeling: reasons, techniques, tools. The 22nd International Conference of the System Dynamics Society. England: Oxford.
[39] Gosavi, A. (2004). Reinforcement learning for long-run average cost. European Journal of Operational Research, 155(3), 654–674. https://doi.org/10.1016/S0377-2217(02)00874-3
[40] Singh, S. P. (1994). Reinforcement learning algorithms for average-payoff Markovian decision processes. Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence (pp. 700-705). Seattle Washington: AAAI Press.

[41] Gosavi, A. (2011). Target-sensitive control of Markov and semi-Markov processes. International Journal of Control, Automation and Systems, 9, 941-951. https://doi.org/10.1007/s12555-011-0515-6
[42] van Horenbeek, A., Buré, J., Cattrysse, D., Pintelon, L., & Vansteenwegen, P. (2013). Joint maintenance and inventory optimization systems: A review. International Journal of Production Economics, 143(2), 499–508. https://doi.org/10.1016/j.ijpe.2012.04.001
[43] Gosavi, A. (2015). Simulation-based optimization. Berlin: Springer.