

A reinforcement learning agent for maintenance of deteriorating systems with increasingly imperfect repairs

Alberto Pliego Marugán, Jesús María Pinar-Pérez, and Fausto Pedro García Márquez

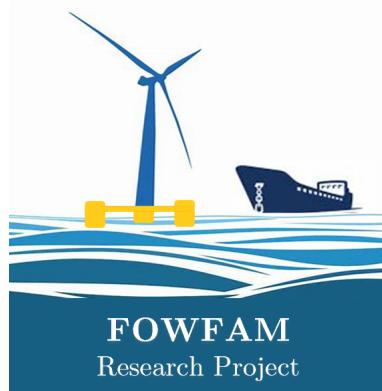
Published in **Reliability Engineering & System Safety - RESS** (ELSEVIER), December 2024.

Cite as: Marugán, A. P., Pinar-Pérez, J. M., and García Márquez, F. P. (2024). A reinforcement learning agent for maintenance of deteriorating systems with increasingly imperfect repairs. *Reliability Engineering & System Safety*, 252, 110466.

DOI: <https://doi.org/10.1016/j.ress.2024.110466>

Article available under the terms of the **CC-BY-NC-ND** licence

The work reported herewith has been financially supported by the Spanish Ministerio de Ciencia, Innovación y Universidades , under Research Grant **FOWFAM** project with reference: PID2022-140477OA-I00.



A reinforcement learning agent for maintenance of deteriorating systems with increasingly imperfect repairs

Alberto Pliego Marugán^{a,*}, Jesús M. Pinar-Pérez^a, Fausto Pedro García Márquez^b

^a CUNEF Universidad, Leonardo Prieto Castro 2, Madrid, Spain

^b Ingenium Research Group, Universidad de Castilla La-Mancha, Av. Camilo José Cela, Ciudad Real, Spain

Abstract

Efficient maintenance has always been essential for the successful application of engineering systems. However, the challenges to be overcome in the implementation of Industry 4.0 necessitate new paradigms of maintenance optimization. Machine learning techniques are becoming increasingly used in engineering and maintenance, with reinforcement learning being one of the most promising. In this paper, we propose a gamma degradation process together with a novel maintenance model in which repairs are increasingly imperfect, i.e., the beneficial effect of system repairs decreases as more repairs are performed, reflecting the degradational behavior of real-world systems. To generate maintenance policies for this system, we developed a reinforcement-learning-based agent using a Double Deep Q-Network architecture. This agent presents two important advantages: it works without a predefined preventive threshold, and it can operate in a continuous degradation state space. Our agent learns to behave in different scenarios, showing great flexibility. In addition, we performed an analysis of how changes in the main parameters of the environment affect the maintenance policy proposed by the agent. The proposed approach is demonstrated to be appropriate and to significantly improve long-run cost as compared with other common maintenance strategies.

Keywords: Maintenance management, Reinforcement learning, Gamma deterioration process

*Corresponding author. Email address: alberto.pliego@cunef.edu

1. Introduction

Globalization and the ultrahigh competitiveness of current and emerging markets necessitate the ongoing modernization and sophistication of engineering systems. However, the development of increasingly complex multicomponent systems introduces myriad—and often unprecedented—potential failure mechanisms, which work alongside normal wear-and-tear-related deterioration. Nevertheless, ongoing reliability in the face of increasing sophistication is of paramount importance if such systems are to benefit the industries, businesses, and commercial ventures for which their use is intended. This makes efficient and cost-effective maintenance management essential.

Maintenance costs are estimated to constitute between 15% and 70% of total production costs [1], with ongoing processes modernization and automation only serving to increase the importance of maintenance. Accordingly, comprehensive maintenance strategies and methodologies have evolved and/or been developed in every industrial and service sector, as exemplified by the automotive, food, energy, and pharmaceutical industries, as well as by social services such as education and healthcare [2].

Maintenance strategies can be divided into two major categories: corrective maintenance (CM) and preventive maintenance (PM). CM is reactive, being initiated after a component fails, while the purpose of PM is to prevent such component failures before they occur. PM further encompasses predictive maintenance (PdM) and condition-based maintenance (CBM), which differ in the way maintenance-need is assessed. PdM involves the use of precise formulas in conjunction with the accurate measurement of environmental factors, such as temperature, vibration, and noise, using sensors or inspections, and maintenance-need is assessed based on analysis of these factors. Accordingly, PdM has the ability to forecast forthcoming maintenance events, making it highly accurate and efficient. Conversely, CBM relies solely on real-time measurements, and maintenance actions are executed once a parameter surpasses a predefined threshold. This means that CBM systems engage in maintenance activities only when required. Furthermore, maintenance strategies are often applied in accordance with a policy having a specific set of characteristics, such as age-replacement, failure-limit, random-age-replacement, repair-cost-limit, and periodic-preventive-maintenance policies [3].

Improving these maintenance strategies is one of the main challenges facing the emergence of “industry 4.0”, a term for the next-generation developments envisaged for modern and future systems, typically encompassing three main directions, as outlined below [4]:

- The first direction concerns adaptability to changing conditions, which includes innovation capability, individualization of products, flexibility, and decentralization. In this field, the availability of all the productive resources of a company is essential to ensure adaptive capacity.
- The second direction concerns sustainability and ecological activities. Improving the efficiency of a productive processes implies a reduction of energy waste. Moreover, poor maintenance management can cause additional pollution from productive processes, for instance, leakages in natural gas or petroleum production [5], poor water quality [6], or noise pollution by cars [7].

- The third direction concerns the use of technologies for increasing mechanization, automation, digitalization, and networking. These characteristics depend on the use of electronics, information technologies, real-time data, mobile computing, cloud computing, big data, and the internet of things (IoT) [8].

The huge amount of data generated and made available by the third developmental direction will facilitate the creation of intelligent maintenance policies via machine learning techniques. Machine learning is a powerful tool for extracting useful information in this massive data environment. Current literature contains numerous algorithms for data-driven decision making in the field of maintenance, and research interest in machine learning for maintenance management is clearly increasing. This interest is strengthened by the necessity of data processing and the increasing importance of the maintenance of systems.

This paper is centered in one of the three major paradigms of machine learning: reinforcement learning (RL). RL seeks a set of optimal actions by an agent within a defined environment for maximizing rewards. With RL, the final reward is cumulative, since it is the result of progressive actions corresponding to a specific action policy. Accordingly, RL shows enormous promise for addressing computational problems in a way that achieves long-term goals [9].

Clearly, the use of machine learning techniques has significantly increased in recent years, but the increase in the use of RL is even more significant. **It should be noted that today** the number of publications mentioning RL in the field of maintenance is almost 20-times greater than **a decade ago**.

The objective of this study **is** to explore the capacity of RL agents to generate policies that improve the maintenance of deteriorating systems. Any improvement in maintenance policy will be assessed in terms of long-term costs. **The proposed model can be applied in industrial systems or components subjected to deterioration.** For instance, maintenance of renewable energy systems such as wind turbines or solar panels, maintenance of elevators in commercial buildings, conveyor belts in warehouses, irrigation systems in agriculture, office equipment such as printers or HVAC systems, public lighting systems, etc. It must be mentioned that our RL agent has been developed to minimize long run cost rates, i.e. to improve maintenance from a purely economic perspective. Hence, as the deterioration may increase at intolerable levels, this methodology is not applicable in its current formulation to critical safety systems such as maintenance of aircrafts, nuclear plants, etc, where failures can be catastrophic.

The main novelty of this study lies in the combination of a maintenance model in which each repair is less effective as more repairs are conducted, and a RL agent whose structure directly addresses the maintenance problem without the need to discretize the degradation state. This combination significantly aligns the model with reality, where the degradation process is continuous, repairs are imperfect, and systems are affected by consecutive repairs.

The remaining content of this paper is structured as follows: Section 2 reviews the most pertinent literature on deteriorating systems and maintenance models. Similar studies are presented to highlight the main contributions of our work. Section 3 briefly explains the main concepts of RL and the Double Deep Q-Network (DDQN) structure employed in this work. Section 4 presents the proposed system to be subjected to degradation and the possible maintenance actions. Section 5 describes the environment and the RL agent proposed in

this paper. Section 6 shows different scenarios to be analyzed, the main results, and a comparison of the proposed maintenance policy with other conventional policies. Finally, section 7 presents the main conclusions of our work.

2. Stochastic degradation processes and RL maintenance

Most systems employed in production processes are subject to degradation. A deteriorating system can be defined as a system with an increasing probability of the occurrence of failures [10], i.e., a decreasing reliability over time. However, most of these systems can be maintained or repaired. Constructing accurate models that define degradation processes is essential for operations and maintenance purposes and product design. Such models provide valuable information of the reliability, [remaining useful life \(RUL\)](#), and actual conditional state of a product during its lifecycle.

An interesting classification of the main degradation models was proposed by Kang et al. [11]. In terms of this classification regime, this paper is focused on monotonical stochastic degradation processes (SDPs) with single-mechanism degradation. The term "monotonical" indicates that the degradation is irreversible, i.e., the state of the system worsens over time unless a maintenance activity is carried out. This situation corresponds to most actual degradation phenomena. According to Peng and Tseng [12], a good stochastic model should satisfy three main properties: clear physical explanation; easy formulation; and adaptability to exogenous events. Within this field, the most common stochastic processes satisfying these properties are gamma, inverse Gaussian, and Wiener processes for continuous degradation, and Markov chains for discrete degradation modelling.

In this paper, we propose a continuous monotonic degradation model based on the gamma stochastic process. Gamma-process-based models were introduced in 1975 by Abdel-Hameed [13] and have since been widely used to model deterioration. An extensive review of gamma degradation processes is provided by van Noortwijk [14].

The increasing importance of maintenance has led to the development of policies and algorithms to obtain optimal maintenance policies [15] considering SDP. However, it is not possible to define an optimal maintenance for all systems since their maintenance does not always have the same goals and must be adapted to each type of system. There are numerous reported methodologies for the maintenance of systems subject to SDP, including value iteration algorithms [16], stochastic filtering [17], multi-objective optimization [18], stochastic programming formulation [19, 20], and others [21, 22, 23]. In addition to these algorithms and methods, some researchers have recently employed the capacities of RL to improve different aspects of maintenance management. Some RL-based approaches are employed to aid the maintenance tasks on safety-critical systems, i.e., those systems whose failure or fault entails catastrophic consequences [24]. Therefore, the main objective in the maintenance of these types of system is to maximize the system's reliability. For instance, Aissani et al. [25] developed a multi-agent approach for effective maintenance scheduling in a petroleum refinery. They achieved a continuous improvement of solution quality by employing a SARSA algorithm. Mattila and Virtalen [26] proposed two formulations for scheduling the maintenance of fighter aircraft via RL techniques, i.e., λ -SMART and SARSA algorithms, and achieved improved results with respect to heuristic

baseline policies. However, RL algorithms are mostly employed in non-safety-critical systems where the main goal of maintenance is to maximize profit, which does not always coincide with maximizing reliability. In this field, RL has been employed for several system types, including manufacturing and production systems used in flow line manufacturing [27]; civil infrastructure systems used for bridges [28], pavements [29], and roads [30]; transportation systems used in the maintenance of ships [31]; power and energy systems used in offshore wind farms [32], power grids [33, 34], and energy storage systems [35]; and other more specific systems such as those used in medical equipment [36] and Mobile Edge Computing systems [37]. An exhaustive review of the use of RL for maintenance of different types of systems is provided by Marugán [38].

In this paper, we are mainly interested in RL-based models for deteriorating systems. Several approaches can be found in this field, for instance, Andriopoulos and Papakonstantinou [39] proposed a stochastic optimal control framework for the maintenance of deteriorating systems with incomplete information. They considered stochastic, non-stationary, and partially observable ten-component deteriorating systems in four possible degradation states. They employed a DDMAC structure, which was compared with several baseline maintenance polices, such as fail replacement (FR), age-periodic maintenance (APM), age-periodic inspections with CBM (API-CBM), time-periodic inspections with CMB (TPI-CBM), and risk-based inspections with CBM (RBI-CBM). Their proposed agent clearly outperformed all the baselines. Peng and Feng [40] introduced a study addressing the decision-making problem of CBM for lithium-ion batteries, representing their capacity degradation with a Wiener process. To tackle this problem, they employed an algorithm known as Gaussian process with reinforcement learning (GPRL). Unlike the prevailing approaches, which primarily focus on maximizing discounted rewards, the GPRL algorithm aims to minimize long-term average costs. This alternative approach demonstrated superior performance in comparison with the conventional methodology. Wang et al. [41] employed a Q-Learning-based solution in a multi-state single machine with deteriorating effects. They developed a PM strategy that combined time-based PM and CBM, and they employed a discrete deterioration model using a Markov chain with four possible states, which was used to demonstrate the high performance and flexibility of the proposed RL approach. Zhang et al. [42] proposed a customized Q-Learning method called Dyna-Q to deal with a system with a large number of degradation levels and where the degradation formula is unknown. Due to the number of possible states, this model can be considered halfway between a discrete and continuous degradation model. Adsule et al. [43] studied degradation in terms of the wear of a component. They considered a Gaussian model for the stochastic degradation, and a SMART RL algorithm was employed. The agent was able to obtain an optimal or near optimal policy to determine maintenance actions and inspection scheduling. Zhao and Smidts [44] proposed a case study of a pump system used in nuclear power plants with a Gamma deterioration process. The problem was presented as a partially observable Markov decision problem where knowledge of the system is improved with Bayesian inference. Zhang et al. [45] modelled the SDP for a multi-component system based on the compound Poisson and gamma processes. They employed a DQN algorithm to optimize the CBM policy under different scenarios. The gamma process is also employed by Yousefi et al. [46] who proposed a Q-Learning algorithm to find policies in a repairable multi-

component system being subjected to two failure processes - degradation and random shocks. Despite considering a continuous SDP, they discretized the deterioration into four levels, allowing them to describe a discrete MDP.

Compared with these previous studies, the main contributions of this study are:

- A deteriorating system and maintenance model that consider imperfect maintenance with an important novelty with respect to the literature found. The maintenance model considers imperfect maintenance, and repairs become increasingly imperfect as more repairs are undertaken. This behavior is represented by a truncated normal distribution whose mean depends on the number of previous reparments done over the system.
- The implementation of a RL agent, which allows for the improvement of maintenance policies compared to conventional maintenance strategies. The proposed methodology allows generation of maintenance policies without the need for setting preventive maintenance thresholds.
- A study of the RL agent performance in different scenarios and a numerical analysis of the effect of changing key parameters (costs of maintenance activities, inspection intervals, degradation rate) on the maintenance policies generated by the agent. This article aims to demonstrate not only that RL techniques are suitable for generating maintenance policies in deteriorating systems, but also that they can be extremely flexible facing parameter changes.
- Our RL agent can operate in a continuous deterioration space without the need for a discretization process.

3. RL framework

RL is a computational strategy that proposes an iterative trial-and-error interaction between an agent and its environment. This process leads the agent to generate a maintenance policy aimed at maximizing a specific reward. Key components of an RL system encompass the agent, the available actions, the associated rewards, and the environmental context. The interaction between the agent and environment is often depicted as illustrated in [Figure 1](#).

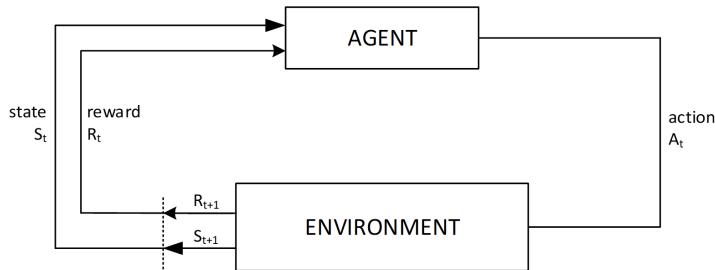


Figure 1: General RL structure. Adapted from [9]

Interaction between agent and environment is typically explained within the formal framework of Markov decision processes (MDPs) [9]. A MDP problem is

formed by the pertinent tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R})$, where \mathcal{S} denotes **the state space**, \mathcal{A} stands for **the action space**, $\mathcal{T}: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is the transition probability function providing the probability of transitioning from state s to s' due to action a , and $\mathcal{R}: \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ stands as the reward function, stipulating the reward due to a transition from state s to s' [9].

In reinforcement learning, the agent's objective is defined by a special signal known as the reward, which is transmitted from the environment to the agent. At each time step, the reward is a single numerical value, denoted as $r_t \in \mathbb{R}$. The sequence of rewards after the time step t is $r_{t+1}, r_{t+2}, r_{t+3}, \dots$. The cumulative reward (G_t) represents the discounted reward or the sum of future rewards from the time t . For a trajectory of finite length K within the environment, G_t is defined by equation (1).

$$G_t = \sum_{k=0}^K \gamma^k r_{t+k} \quad (1)$$

where $\gamma \in [0, 1]$ is a discount factor that determines the relevance of the future rewards and forces the convergence for infinite-horizon returns. $k \in [0, K]$ is a subindex, being K the total number of future recompenses that the agent will receive until the end of the current episode. Rewards are used by the agent to generate a policy $\pi: \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$, i.e., a function providing the probability distribution of each action $a \in \mathcal{A}$ and each possible state $s \in \mathcal{S}$. Following a given policy π , a value function and an action-value function can be defined as:

$$V^\pi(s) = \mathbb{E}_\pi \left[\sum_{k=0}^K \gamma^k r_{t+k} \mid s_t = s \right] \quad (2)$$

$$Q^\pi(s, a) = \mathbb{E}_\pi \left[\sum_{k=0}^K \gamma^k r_{t+k} \mid s_t = s, a_t = a \right] \quad (3)$$

The policy π maps states to the probability of selecting each possible action. Therefore, if the agent follows the policy π at a certain time t , then $\pi(a|s)$ represents the probability of choosing the action a given a state s . Therefore, this policy depends only on the current state and not on the sequence of states and actions that preceded it, being aligned with the principles of MDP.

The main goal of the RL agent is to find the policy π^* that maximizes the expected reward, satisfying the Bellman optimality equations (4) and (5).

$$Q^*(s, a) = \sum_{s', r} p(s', r \mid s, a) \left[r + \gamma \max_{a'} Q^*(s', a') \right] \quad (4)$$

$$V^*(s) = \max_{a \in \mathcal{A}(s)} Q^{\pi^*}(s, a) = \sum_{s', r} p(s', r \mid s, a) [r + \gamma V^*(s')] \quad (5)$$

Therefore, π^* being the policy that maximizes the value functions, equations (6) and (7) will provide the optimal policy:

$$\pi^* = \arg \max_{\pi} V^\pi(s) \quad (6)$$

$$\pi^* = \arg \max_{\pi} Q^{\pi}(s, a) \quad (7)$$

These optimal policies can be attained by following different strategies. Depending on the characteristics of environment, different algorithms can be employed. A review of RL algorithms can be found in Shakya et al. [47].

In this paper, we employ the DDQN algorithm, proposed originally by Hasselt [48]. This algorithm, which is derived from the Deep Q-Network (DQN) algorithm, addresses the problem of Q-value overestimation, which is frequently provided by the standard DQN algorithm proposed by Mnih et al [49]. A DQN consists of a neural network that, given a state s , produces a vector of action values $Q(s; \theta)$, where θ represents the parameters of the neural network. The DQN algorithm incorporates three essential components: first, a neural network (main neural network) with parameters θ , which is employed to estimate Q-values of the current state s and a ; a second neural network (target neural network) with parameters θ' used to approximate the Q-values of the next state s' and next action a' ; a replay memory used to store the experiences for the learning process and the implementation of a target network with parameters θ [50]. The Bellman equation for a DQN is:

$$Q(s, a; \theta) = r + \gamma Q \left(s', \max_{a'} Q(s', a'; \theta') \right) \quad (8)$$

The main difference between a DDQN and a DQN is that the process of action selection and action evaluation are separate in a DDQN, as the target Q-values are determined by actions selected by the main network, while their Q-values are estimated using the target network. This adjustment effectively eliminates overestimation bias, leading to more precise Q-value estimates and enhanced training stability. Considering these changes, the Bellman equation for a DDQN results in:

$$Q(s, a; \theta) = r + \gamma Q \left(s', \arg \max_{a'} Q(s', a'; \theta); \theta' \right) \quad (9)$$

The main goal of DQNs and DDQNs is to estimate Q-values through deep neural networks, which is especially useful when the state space is too large to be collected in a table (as a Q-learning algorithm does). The architecture of a DDQN algorithm is illustrated in [Figure 2](#).

The decision to use a DDQN in this study was not arbitrary since it has been demonstrated that DDQN agents outperform other algorithms when dealing with very large state spaces. In this paper, we do not discretize the degradation level, so the state space is continuous while the action space is discrete. This features make DDQNs highly suited to work in this environment, as demonstrated in other studies [51, 52, 53]. Other suitable architectures, such as proximal policy optimization (PPO) and trust region policy optimization (TRPO) have been assessed for our environment, but they provided inferior results.

4. Proposed degradation process and maintenance model

This paper proposes a new approach to optimize the CBM policy for a gradually deteriorating single-unit system subjected to SDP. Degradation is modelled by a homogeneous gamma process. The proposed model is based on Marugan et al. [54].

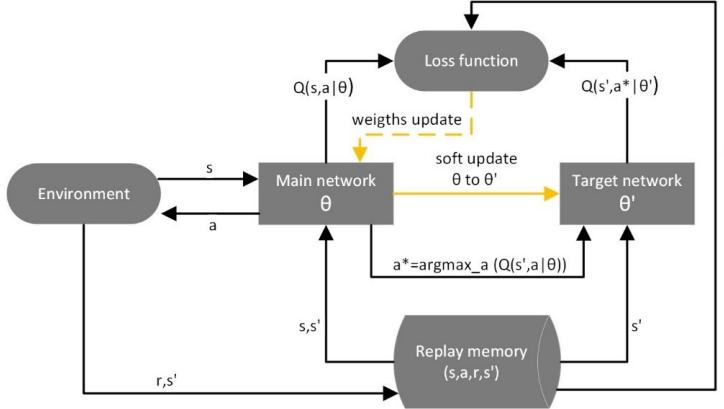


Figure 2: Double Deep Q-Network architecture

The gamma process, which is assumed to be strictly increasing over time if no maintenance action is carried out, can be formulated as $(X_t)_{t \geq 0}$. Let the random variable X_t stand for the deterioration state of the system at time t , where $X_0 = 0$ and $t \geq 0$. The degradation increment $\Delta X(t, \Delta t) = X_{t+\Delta t} - X_t$ is a continuous random variable following a gamma distribution with shape parameter $v(t, \Delta t)$ and scale rate β . Therefore, $\Delta X \sim \Gamma(v(t, \Delta t), \beta)$ and its probability density function (pdf) is:

$$f(t, \Delta t, x) = Pr(\Delta X = x) = \frac{x^{v(t, \Delta t)-1}}{\Gamma(v(t, \Delta t))} \beta^{v(t, \Delta t)} e^{-\beta x}, \quad \forall x \geq 0 \quad (10)$$

If $v(t, \Delta t)$ is a linear function, the model results in a stationary gamma process; otherwise, the process becomes non-stationary.

The cumulative density function is:

$$F(t, \Delta t, x) = \frac{\gamma(v(t, \Delta t), (\beta x))}{\Gamma(v(t, \Delta t))} \quad (11)$$

where $\gamma(\cdot)$ is the lower incomplete gamma function.

The survival function can be defined by:

$$\bar{F}(t, \Delta t, x) = 1 - F(t, \Delta t, x) = \frac{\Gamma(v(t, \Delta t), (\beta x))}{\Gamma(v(t, \Delta t))} \quad (12)$$

Besides the deterioration model employed to describe stochastically the state of the system, it is essential to define the way such states are obtained. In this field, continuous monitoring, which provides the system condition in real time, is the most accurate method. Continuous monitoring allows anomalies to be detected at initial stages, allowing maintenance actions to be performed immediately [55]. However, factors such as costs, technological limitations, legal issues, or other limitations make continuous monitoring inadequate for some systems. In such cases where continuous monitoring is not suitable, the deterioration state is often obtained via planned inspections. In this paper, we propose planned inspections to determine the state of the system. We consider

perfect inspection, i.e., the system state is revealed with certainty. Additionally, we assume that these inspections are instantaneous, so that the duration of the inspection is negligible. Inspections are executed at times $(T_n)_{n \in \mathbb{N}}$ with $(T_0) = 0$. Let T_n^- and $X_{T_n^-}$ be the time and the state of the system just before the inspection at time T_n , respectively.

Regarding the maintenance characteristics, we consider an imperfect maintenance for a repairable unit system. Two types of maintenance activities have been considered in this work: replacements and repairs. Like inspections, the maintenance interventions are assumed to be instantaneous, and the effect of these actions is observable immediately after the inspection, i.e., at time T_n^+ . The available maintenance actions are similar to the CBM model presented by Zhang et al. [42] but the behavior of the model presented herein is totally different. These maintenance activities are:

Replacements (R): A replacement leads the system to an "as good as new" (AGAN) state, i.e., the deterioration of the system after any replacement is $X_{T_n}^R = 0$. If this action is performed when the deterioration is above a failure threshold L , i.e., the deterioration reaches an unacceptable value above which the system will fail, the action is said to be a *corrective replacement*. However, if the action is carried out below the threshold, then the action is a *preventive replacement*, and no downtime costs are computed.

Repairs (P): This maintenance task is assumed to be imperfect. Several previous studies have modeled imperfect maintenance. We combine some characteristics from the models proposed in Huynh [56] and van Bérenguer [57] to determine the effect of an imperfect repair. We model the effect of imperfect repairs by subtracting a certain amount from the current deterioration level. This amount is sampled from a random distribution with a memory effect, i.e., it depends on the previous repairs. This memory is represented by assuming that, after a repair, the system cannot return to a deterioration state lower than that reached in the previous maintenance action. Note that we do not consider a corrective repairment; we assume that when the system has failed, it is necessary to reset the degradation to 0.

Let $X_{T_n}^P$ be the degradation state after a preventive repair action. We consider that $X_{T_n}^P = X_{T_n^-} - Z_n$, where Z_n , called the maintenance gain, is a continuous random variable distributed as a truncated normal distribution whose density is :

$$g_{\mu, \sigma, X_{T_n^-}}(x) = \frac{1}{\sigma} \frac{\phi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{X_{T_n^-}-\mu}{\sigma}\right) - \Phi\left(\frac{X^M-\mu}{\sigma}\right)} I_{[X^M, X_{T_n^-}]}(x) \quad (13)$$

where:

- $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density and cumulative distribution function of the standard normal distribution respectively;
- X^M is the deterioration value after the immediate previous maintenance activity;

- $I_{[X^M, X_{T_n^-}]}(x) = 1$ if $X^M \leq x \leq X_{T_n^-}$ and $I_{[X^M, X_{T_n^-}]}(x) = 0$, otherwise;
- μ and σ are the mean and standard deviation of the truncated normal distribution.

Similarly to [58], we assume that $\mu = \frac{X^M + X_{T_n^-}}{2}$ and $\sigma = \frac{X^M + X_{T_n^-}}{6}$.

The use of a truncated normal was proposed originally in reference [57] to model the maintenance gain. This model is appropriate because it captures the variability of the system deterioration after an imperfect repair and allows for considering practical limits in the model. In our approach, the system deterioration after a repair action is bounded by the current deterioration state and the deterioration state after the previous maintenance intervention.

An illustrative example of the proposed model is shown in Figure 3, which shows the increasing deterioration and the maintenance actions allowed by the model.

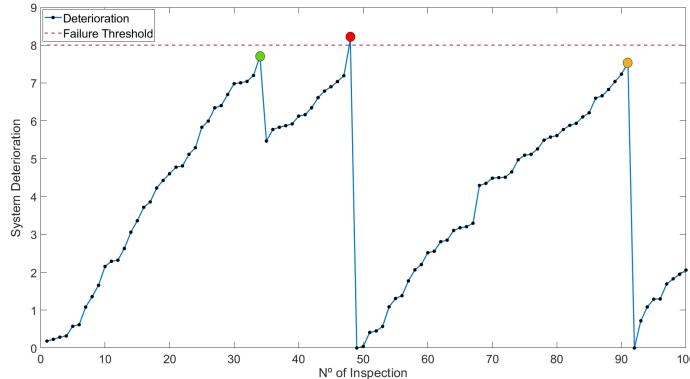


Figure 3: An example of the degradation process. (green circle: preventive repair; red circle: corrective replacement; orange circle: preventive replacement)

Note that we consider the working state of system to be binary: it is either functioning or not. The deterioration does not affect the performance of the system unless the failure thresholds is surpassed. Note that most literature on CBM consider a preventive maintenance threshold. One of the advantages of our approach is that this preventive threshold is not necessary since the RL agent will determine the best moment to perform either a corrective or a preventive maintenance. However, in our model, we define a corrective threshold L to determine a system failure.

5. Description of the environment and the RL Agent

5.1. State space representation

The RL agent needs to be input with the current state of the environment. This state will not only depend on the current deterioration of the system, but also on the number of preventive repair actions performed since the last reset to an AGAN state. Additionally, as deterioration is a time-dependent process, time

is essential to calculate the next state. The time period between two consecutive inspections (Δt) is defined to calculate the increment of deterioration between the state S^{T_n} and the state $S^{T_{n+1}}$. The definition of this parameter does not affect the Markovian properties of the process, since the value of Δt is constant and does not depend on any past event. The RL agent only needs to act for discrete times, i.e., an action is only ordered after a certain inspection; so the system state turns in a continuous degradation and discrete time state space. The degradation of the system at the inspection n is represented by X_{T_n} . As well as the system degradation, it is also necessary to input the degradation after the previous maintenance action, represented by X^M .

Therefore, this system state space will be given by:

$$S^{T_n} = \{X_{T_n}, X^M\} \quad \text{with } X_{T_n} \geq 0 \quad \text{and } 0 < X^M \leq X_{T_n} \quad (14)$$

5.2. Action space representation

According to the model description in Section 4, the action space for each inspection n is $\mathcal{A} = \{a_0, a_1, a_2\}$, where:

- a_0 corresponds to "no maintenance action" after the inspection n . System deterioration will continue according to the SDP defined in Section 4. After action a_0 at time T_n , the system state is $S^{T_n} = \{X_{T_n}, X^M\}$ with $X_{T_n} = X_{T_n}^-$.
- a_1 is a "preventive repair action" which leads the system deterioration to any state between $X_{T_n}^-$ and X^M according to the truncated normal model presented in Section 4. After action a_1 at time T_n , the system state is $S^{T_n} = \{X_{T_n}, X_{T_n}^-\}$ with $X_{T_n} \leq X_{T_n}^-$.
- a_2 refers to a "replacement action". Note that action a_2 encompasses both preventive and corrective replacements. Being preventive or corrective only depends on the state of the system when the action is performed. This action will provide different rewards regarding the state of the system; however, the consequence for the system state after the action is identical. Both actions set the system to an AGAN state. After action a_2 at time T_n , the system state is $S^{T_n} = \{0, 0\}$.

5.3. Rewards definition

The main purpose of this paper is to improve the maintenance strategy from an economic perspective. This objective is to minimize maintenance long-run cost. As aforementioned, the RL agent is created to maximize a long-term reward. These rewards will be defined in the function of both the deterioration state and the action selected by the agent.

Let C_P and C_R stand for the costs of preventive repair and replacement actions, respectively. As mentioned before, the inspections are assumed to be instantaneous, but if the system fails, we consider that for the time between consecutive inspections, the system is not functioning and therefore there is a loss of production due to the downtime, represented by C_{down} . Then, the reward at time T_n is defined as:

$$\mathbf{r}_{T_n}(a_{T_n}, X_{T_n}^-) = \begin{cases} 0 & \text{for } a_{T_n} = a_0 \text{ and } X_{T_n}^- < L \\ -C_p & \text{for } a_{T_n} = a_1 \text{ and } X_{T_n}^- < L \\ -C_R & \text{for } a_{T_n} = a_2 \text{ and } X_{T_n}^- < L \\ -C_R - C_{down} & \text{for } a_{T_n} = a_2 \text{ and } X_{T_n}^- \geq L \end{cases} \quad (15)$$

5.4. Agent definition and training

The DDQN algorithm was implemented in MATLAB with the following hyperparameters that correspond to the main hyperparameter configuration predefined in MATLAB.

- Exploration options: Epsilon decay = 0.005; Epsilon max = 1; Epsilon min = 0.01;
- Agent options: Sample time = 1, Discount factor = 0.99, Batch size = 64, Experience buffer length = 10000;
- Optimizer options: Optimizer: ADAM; Learn rate = 0.01; Gradient decay factor = 0.9.

Training options have been set as follows: Maximum episodes: 50000, and Maximum Episode Length = 500. The stopping criteria has been set to reach the maximum episodes number.

Note that the agent performance might be improved by tuning these parameters. However, our hyperparameter configuration is sufficient to demonstrate that a DDQN agent is able to generate successful maintenance policies under each scenario and to observe the behavior of the agent when the environment changes.

6. Case Studies

6.1. Numerical experiments

Seven different scenarios to obtain information of the performance of the proposed RL agent in different case studies have been considered. The data values employed in this study are based on the parameters of the case study presented by Zheng et al. [59]. Case 2* will be considered as the baseline case. It is worth mentioning that, for all cases, $C_R = 3500$ since we are interested in the effect of the ratio repair/replacement costs and $v(t) = 0.0115t$ since we assume the gamma deterioration process to be homogeneous. The rest of the parameters will vary between cases, as shown in Table 1 .

Table 1: Case studies parameters

Description	β	C_P	C_{down}	L	Δt
Case 1	Reduced repair costs	4.63	300	2000	8 100
Case 2*	Baseline	4.63	600	2000	8 100
Case 3	Increased repair costs	4.63	1500	2000	8 100
Case 4	Increased failure limit	4.63	600	2000	12 100
Case 5	Reduced downtimes cost	4.63	600	500	8 100
Case 6	Slower degradation	6.5	600	2000	8 100
Case 7	Longer inspection period	4.63	600	2000	8 150

Once the different case studies are analyzed, we calculate the long-run cost rate as a performance indicator of the proposed policy. The total costs of maintenance up to moment t can be defined by:

$$C(t) = C_P(t) + C_R(t) + C_{down}(t) = C_P N_P(t) + C_R N_{PR}(t) + (C_R + C_{down}) N_{CR}(t) \quad (16)$$

where $C_P(t)$, $C_R(t)$, and $C_{down}(t)$ are respectively the cumulative costs of preventive repair, corrective replacement, and production-loss costs due to unavailability of the system. These cumulative costs are given by the fix costs (C_P , C_R , C_{loss}) and $N_P(t)$, $N_{PR}(t)$, and $N_{CR}(t)$, which are the random number of repairs, preventive replacements, and corrective replacements, respectively, in the period $[0, t]$.

The long-run cost rate can be calculated by:

$$EC_\infty = \lim_{t \rightarrow \infty} \left[\frac{E[C(t)]}{t} \right] = \frac{E[C(S_i)]}{E[S_i]} \quad (17)$$

where S_i refers to the i^{th} renewal cycle (see Figure 3).

It is known that the system has regenerative properties since corrective replacements are forced when the deterioration surpasses the failure threshold, i.e., the deterioration will come to state 0 in a finite time. Due to such regenerative properties, the long-run cost rate can be calculated through the expected values in a single renewal cycle, a renewal cycle being the period from one replacement to the time just before the next replacement. Therefore, the long-run cost rate can be approximated by:

$$EC_\infty = \frac{C_P E[N_P(S_1)] + C_R E[N_{PR}(S_1)] + (C_R + C_{down}) E[N_{CR}(S_1)]}{E[S_1]} \quad (18)$$

The required expected values will be numerically obtained using the Monte Carlo method. Additionally, to be sure that results are statistically significant, the long-run costs rates will be estimated through confidence intervals.

6.2. Results

The proposed agent has been trained for the scenarios in Table 1, providing specific maintenance policies for each case. Figure 4 shows the system deterioration for the baseline case (Case 2*) and the rest of the cases using the policies proposed by the agent. Each scenario comprises a maintenance period of 1000 inspections, but only 250 inspections are shown for the sake of clear visualization.

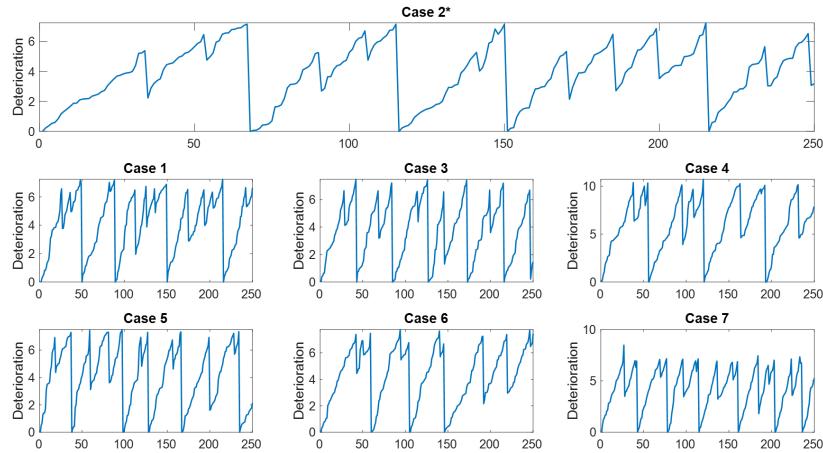


Figure 4: RL-based maintenance for all case studies

[Figure 4](#) allows us to conduct an initial analysis of how the agent performs in each scenario. For instance, it is clear that more repairs are carried out in Case 1, unlike Case 3, where fewer repairs are conducted. Therefore, the policy in Case 1 provides longer renewal cycles. It can be observed that in Case 6, due to slower degradation, the system is pushed closer to the failure threshold before a repair is executed, meaning that the agent allows greater risks of degradation exceeding such threshold. These and other observations are quantitatively analyzed in [Figure 5](#) and [Figure 6](#). These figures are based on complete periods of 1000 inspections and 200 Monte Carlo iterations. Therefore, the results are based on a total of 200,000 inspections.

[Figure 5](#) shows the total amount of maintenance actions that the RL agent proposes for each case. Moreover, the black line represents the average costs of maintenance for the Monte Carlo iterations.

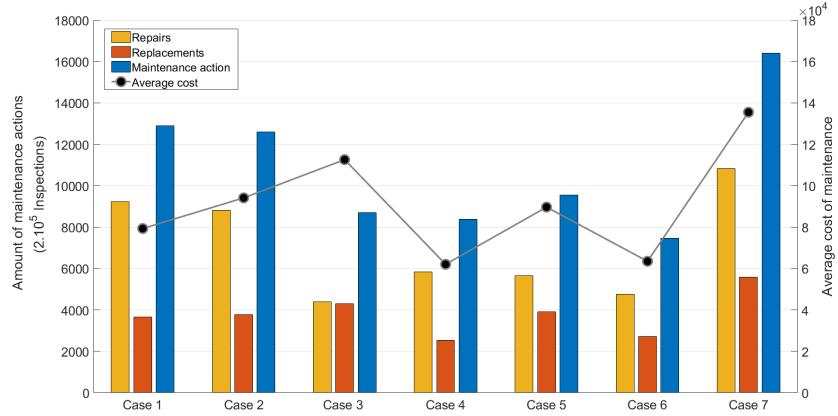


Figure 5: Amount of maintenance actions

[Figure 6](#) shows the percentage changes for each type of maintenance action with respect to Case 2*.

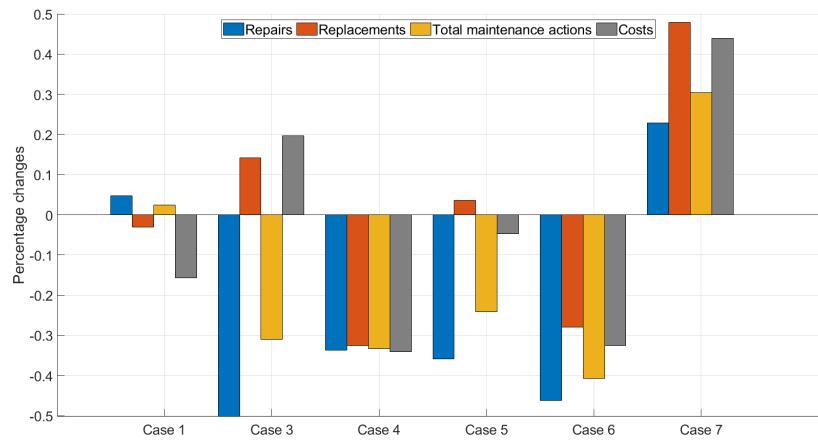


Figure 6: Percentage changes in the number of maintenance actions and costs with respect to Case 2*

[Figure 5](#) and [Figure 6](#) provide some important information regarding the maintenance policy generated for each case study. With respect to Case 2* we observe that:

- Repairs are cheaper in Case 1 and therefore the policy increases 4.7% the number of repairs and decreases 3.7% the number of replacements. These variations are rather small, but they lead to a reduction of 15.7% in the average costs.
- In Case 3, repairs are more expensive, and the policy drastically reduces the number of repairs by 50%. This forces the agent to carry out 14.1% more replacements. It is worth mentioning that although there is a reduction of more than 30% in the total number maintenance actions, the average maintenance costs increase significantly.
- In Case 4, the failure threshold is higher and therefore the number of maintenance actions and the average costs in the same time period are reduced. However, the ratio between repairs and replacements remains similar since the costs of maintenance actions have not changed. Therefore, a change in the failure threshold will affect the maintenance policy in terms of "when" but not "which" maintenance actions should be performed.
- Case 5 presents lower costs of corrective replacements through a reduction of downtime costs. We observe that the agent assumes more risk to perform a preventive replacement since, upon surpassing the maximum threshold L , the penalization is less significant. Therefore, the number of replacements increases by 3.6%, leading to a significant reduction of repairs (35.9%), which causes a slight reduction in costs.
- Case 6 considers that the degradation process is slower. Therefore, both the total number of maintenance actions and the mean costs of maintenance decrease. In general, this scenario is favorable in every way. It is clear that if the system degradation is slower, the number of maintenance activities of any type is reduced, and consequently the costs also decrease.
- Case 7 involves a variation in the time between inspections. If the period between inspections is longer, it is more likely that the maximum threshold L will be reached and therefore more corrective actions must be taken. In addition, preventive actions are proposed at lower deterioration states in order to avoid the risk of surpassing the threshold. It is worth mentioning that we are not considering costs of inspections, which would lead to savings in this case study.

In general, we can observe how the RL agent is able to learn from each case study and adapt the maintenance policy to the specificities of each case. Table 2 shows some results (mean and standard deviation) obtained from the analysis.

Using these statistical results, confidence intervals are used to estimate the expected values required to calculate long-run cost rates. By performing Anderson-Darling and Kolmogorov-Smirnov tests, we have verified that all the collected parameters are normally distributed in the Monte Carlo iterations. Therefore, the 95% confidence intervals for the long-run cost rate are given in Table 3 .

Table 2: Summary of case-study results

	N. of Repairs (N_P)		Number of Pre- ventive Replace- ments (N_{PR})		Number of Cor- rective Replace- ments (N_{CR})		Renewal Cycles Duration (S)	
	Mean	sd	Mean	sd	Mean	sd	Mean	sd
Case 1	46.19	3.17	17.99	1.22	0.28	0.53	53.33	3.25
Case 2	44.12	1.87	18.54	1.14	0.31	0.55	51.70	2.87
Case 3	21.95	0.79	20.71	1.15	0.80	0.88	45.37	1.57
Case 4	29.23	1.46	12.72	0.98	0.00	0.00	76.26	5.72
Case 5	28.27	1.57	18.43	1.48	1.10	1.05	49.99	2.89
Case 6	23.75	1.23	13.25	1.04	0.32	0.56	71.36	4.96
Case 7	54.22	2.05	27.55	1.42	0.32	0.56	35.27	1.67

Table 3: Relevant confidence intervals

	Interval for N_P		Interval for N_{PR}		Interval for N_{CR}		Interval for (S)		Long Run Cost Rate	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Case 1	45.75	46.63	17.82	18.16	0.21	0.35	52.88	53.78	1436	1503
Case 2	43.86	44.38	18.38	18.70	0.23	0.39	51.30	52.10	1764	1836
Case 3	21.84	22.06	20.55	20.87	0.68	0.92	45.15	45.59	2378	2462
Case 4	29.03	29.43	12.58	12.86	0.00	0.00	75.47	77.05	797	830
Case 5	28.05	28.49	18.22	18.64	0.95	1.25	49.59	50.39	1675	1760
Case 6	23.58	23.92	13.11	13.39	0.24	0.40	70.67	72.05	851	897
Case 7	53.94	54.50	27.35	27.75	0.24	0.40	35.04	35.50	3645	3767

By considering the central point of each interval, we calculate which part of the long-run cost rate corresponds to each type of maintenance action, obtaining the results presented in [Figure 7](#).

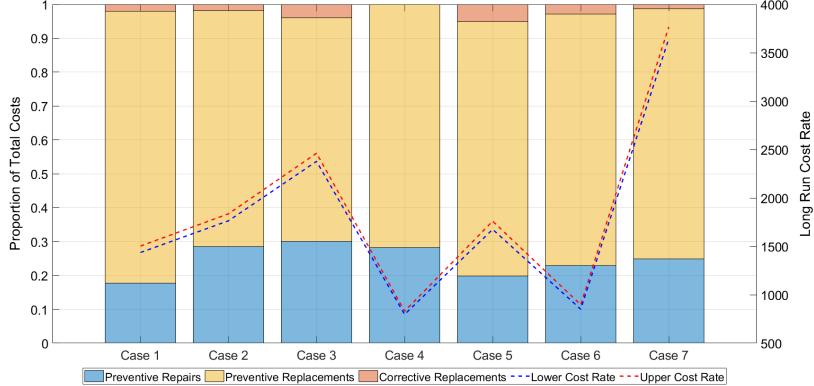


Figure 7: Long-run cost rates and distribution of costs for Cases 1–7

The long-run cost rate has a very similar shape to the average costs shown in [Figure 5](#). It is worth mentioning that average cost corresponds to a complete period of 1000 inspections and 200 iterations; however, the long-run cost rates correspond to the average costs during a single renewal cycle, i.e., from a replacement to the moment just before the next replacement. In general, we observe that, independently to the parameters configuration, around 70%, 27%, and 3% of long-run cost is due to preventive replacements, repairs, and corrective replacements, respectively.

It is worth mentioning that maintenance plays a crucial role in both availability and productivity. In our model, maintenance activities are assumed to be

instantaneous and therefore the system is available unless deterioration reaches the failure threshold. An important factor which would affect the system availability is the duration of maintenance activities, however, since this model is not described for a specific system, it is not convenient to provide such durations. However, the impact of maintenance on availability and productivity is partially captured by the parameter C_{down} . Consequently, the product of the total number of corrective replacements (N_{CR}) and C_{down} can be used as relative measure of unavailability to compare between the different cases of study as shown in Table 4 .

	Table 4: Impact on Availability			
	$E[N_{CR}]$	C_{down}	$E[N_{CR}] \cdot C_{down}$	Availability Ranking
Case 1	0.28	2000	560	<i>4th</i>
Case 2	0.31	2000	620	<i>3rd</i>
Case 3	0.80	2000	1600	<i>7th</i>
Case 4	0.00	2000	0	<i>1st</i>
Case 5	1.10	500	550	<i>2nd</i>
Case 6	0.32	2000	640	<i>5th</i>
Case 7	0.32	2000	640	<i>6th</i>

As can be observed in Table 4 , Case 4 presents the highest availability since it has the highest failure threshold. Cases 6 and 7 presents the result for the product $E[N_{CR}] \cdot C_{down}$, however, Case 6 has been positioned before since the total number of maintenance actions is three times less than in Case 7. Finally, in Case 3, the elevated cost of preventive activities impacts negatively on the availability.

6.3. Comparison to conventional maintenance policies

To determine the validity of the proposed procedure, we performed a comparison between the performance of the proposed RL agent and other conventional CBM strategies for Case 2*. In this paper, the RL Agent has the goal of improving maintenance from an economic perspective, i.e. the only objective is to minimize maintenance long run cost rate considering that when deterioration is above the failure threshold, a corrective maintenance action must be immediately done. Therefore, our maintenance model is defined in such way that failures do not cause safety problems or environmental risks but only economic losses. Additionally, system reliability is not considered to be an objective function in this paper. These are the reasons why some maintenance strategies such as risk-based or reliability centered maintenance are not included in this comparison.

Similarly to Andriotis and Papakonstantinou [39], we consider the following policies:

- *Fail Replacement (FR) policy*: Only corrective replacements are permitted. In this policy a corrective replacement is performed when the deterioration of the system is above the failure threshold L .
- *Age-based Periodic Maintenance policy*: This policy assumes that repairs and replacements are done periodically. Therefore, two important parameters must be defined: the time period between consecutive repairs and

the time period between consecutive replacements. In order to compare with the proposed RL based policy, both time periods have been optimized numerically with Monte Carlo iterations.

- *Threshold-based Maintenance (TBM) policy:* Maintenance actions are taken depending on the current state of deterioration of the system at the inspection time. Two thresholds are set and optimized, i.e., a preventive threshold, to determine when a preventive replacement is performed, and a corrective threshold, to define when a corrective replacement is required. In order to compare with the proposed RL based policy, both thresholds have been previously optimized numerically.
- *Age and Threshold-based Maintenance (ATBM) policy:* Maintenance actions are taken depending on both the current state of deterioration of the system and a certain time period between consecutive repairs and replacements. Four parameters have been considered in this strategy, i.e. two thresholds to determine if a preventive action or a corrective action must be done and two time periods to determine when a repair and replacement must be done. In order to compare with the proposed RL based policy, the four parameters (thresholds and time periods) have been previously optimized numerically.

Figure 8 shows the costs of maintenance for each policy in a total of 200 iterations.

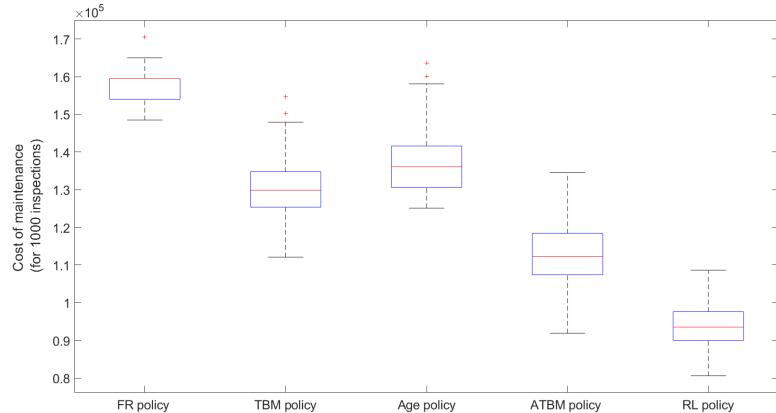


Figure 8: Comparison with other policies

Figure 8 shows that the agent is able to reduce the long-run cost rate by around 41%, 28%, 31% and 17% compared with the FR policy, TBM policy, Age policy, and ATBM policy, respectively. Therefore, the RL Agent proposed in this paper clearly outperforms other conventional maintenance policies.

7. Conclusions

This study successfully developed a homogeneous gamma degradation model whose maintenance framework is based on periodic and perfect inspections, i.e., inspections reveal the real degradation level of the system. Two type of

maintenance actions were considered: repairs or replacements. These actions are categorized as either corrective or preventive depending on the state of system at the time the action is carried out.

A model has been proposed wherein repair actions enhance the degradation following a probability distribution, representing imperfect maintenance subject to uncontrollable conditions. A novel feature of this model is that each repair action negatively affects the effectiveness of the subsequent repair by affecting the parameters of the probability distribution.

To optimize maintenance tasks, we implemented an RL agent with a DDQN structure, demonstrating its capability to decide when and what maintenance activities are advisable in different scenarios. One of the main advantages of this approach is that there is no requirement to define a preventive threshold. The RL-based agent discerns the ideal timing for executing corrective or preventive maintenance autonomously. In addition, this RL architecture was demonstrated to be highly effective when facing large or continuous state space. Another novelty of this study is the capacity of our RL agent to make decisions without discretizing the degradation variable.

Additionally, an analysis has been conducted to understand how each parameter influences the long-term maintenance costs based on the adopted policy. This study has demonstrated that the RL agent is able to create flexible policies adapted to changing environments.

Finally, the model was validated, revealing that our agent significantly improves long-term costs compared to [other maintenance](#) policies.

Acronyms

ADAM: Adaptive Learning Rates

AGAN: As Good as New

API-CBM: Age-Periodic Inspections with Condition-Based Maintenance

APM: Age-Periodic Maintenance

ATBM: Age and Threshold-based Maintenance

CBM: Condition Based Maintenance

CM: Corrective Maintenance

DDQN: Double Deep Q-Network

DDMAC: Deep Centralized Multi-Agent Actor Critic

DQN: Deep Q-Network

FR: Fail Replacement

GPRl: Gaussian Process with Reinforcement Learning

MDP: Markov Decision Process

O&M: Operation and Maintenance

PdM: Predictive Maintenance

PM: Preventive Maintenance

PPO: Proximal Policy Optimization

RBI-CBM: Risk-Based Inspections with Condition-Based Maintenance

RL: Reinforcement Learning

RUL: Remaining Useful Life

SDP: Stochastic Deterioration Process

TPI-CBM: Time-Periodic Inspections with Condition-Based Maintenance

TRPO: Trust Region Policy Optimization

Acknowledgement

The work reported herewith has been financially supported by the Spanish Ministerio de Ciencia, Innovación y Universidades, under Research Grant FOWFAM project with reference: PID2022-140477OA-I00.

References

- [1] D. S. Thomas, D. S. Thomas, The costs and benefits of advanced maintenance in manufacturing, US Department of Commerce, National Institute of Standards and Technology . . . , 2018.
- [2] R. Manzini, A. Regattieri, H. Pham, E. Ferrari, et al., Maintenance for industrial systems, Vol. 1, Springer, 2010.
- [3] H. Wang, A survey of maintenance policies of deteriorating systems, European journal of operational research 139 (3) (2002) 469–489.
- [4] H. Lasi, P. Fettke, H.-G. Kemper, T. Feld, M. Hoffmann, Industry 4.0, Business & information systems engineering 6 (2014) 239–242.
- [5] K.-M. Wollin, G. Damm, H. Foth, A. Freyberger, T. Gebel, A. Mangerich, U. Gundert-Remy, F. Partosch, C. Röhl, T. Schupp, et al., Critical evaluation of human health risks due to hydraulic fracturing in natural gas and petroleum production, Archives of Toxicology 94 (2020) 967–1016.
- [6] M. A. Vanshkar, A. P. M. R. Bhatia, Upcoming longest elevated flyover corridor of the state of madhya pradesh in the city of jabalpur is control the noise pollution, International Research Journal of Engineering and Technology (IRJET) 6 (9) (2019) 1406–1411.
- [7] C. Dierkes, L. Kuhlmann, J. Kandasamy, G. Angelis, Pollution retention capability and maintenance of permeable pavements, in: Global Solutions for Urban Drainage, 2002, pp. 1–13.
- [8] Y. Lu, Industry 4.0: A survey on technologies, applications and open research issues, Journal of industrial information integration 6 (2017) 1–10.

- [9] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, MIT press, 2018.
- [10] M. Kaminskiy, V. Krivtsov, A gini-type index for aging/rejuvenating objects, Mathematical and Statistical Models and Methods in Reliability: Applications to Medicine, Finance, and Quality Control (2010) 133–140.
- [11] K. Rui, G. Wenjun, C. Yunxia, Model-driven degradation modeling approaches: Investigation and review, Chinese Journal of Aeronautics 33 (4) (2020) 1137–1153.
- [12] C.-Y. Peng, S.-T. Tseng, Mis-specification analysis of linear degradation models, IEEE Transactions on Reliability 58 (3) (2009) 444–455.
- [13] M. Abdel-Hameed, A gamma wear process, IEEE transactions on Reliability 24 (2) (1975) 152–153.
- [14] J. M. Van Noortwijk, A survey of the application of gamma processes in maintenance, Reliability Engineering & System Safety 94 (1) (2009) 2–21.
- [15] A. Pliego Marugán, F. P. García Márquez, J. M. Pinar Perez, Optimal maintenance management of offshore wind farms, Energies 9 (1) (2016) 46.
- [16] W. Cheng, X. Zhao, Maintenance optimization for dependent two-component degrading systems subject to imperfect repair, Reliability Engineering & System Safety 240 (2023) 109581.
- [17] M. Zhang, O. Gaudoin, M. Xie, Degradation-based maintenance decision using stochastic filtering for systems under imperfect maintenance, European Journal of Operational Research 245 (2) (2015) 531–541.
- [18] A. F. Shahraki, O. P. Yadav, C. Vogiatzis, Selective maintenance optimization for multi-state systems considering stochastically dependent components and stochastic imperfect maintenance actions, Reliability Engineering & System Safety 196 (2020) 106738.
- [19] E. Leo, S. Engell, Condition-based maintenance optimization via stochastic programming with endogenous uncertainty, Computers & Chemical Engineering 156 (2022) 107550.
- [20] D. Ruiz-Hernández, J. M. Pinar-Pérez, D. Delgado-Gómez, Multi-machine preventive maintenance scheduling with imperfect interventions: A restless bandit approach, Computers & Operations Research 119 (2020) 104927.
- [21] A. Khatab, D. Ait-Kadi, N. Rezg, Availability optimisation for stochastic degrading systems under imperfect preventive maintenance, International Journal of Production Research 52 (14) (2014) 4132–4141.
- [22] C. Chuang, L. Ningyun, J. Bin, X. Yin, Condition-based maintenance optimization for continuously monitored degrading systems under imperfect maintenance actions, Journal of Systems Engineering and Electronics 31 (4) (2020) 841–851.

- [23] J. Wang, X. Zhu, Joint optimization of condition-based maintenance and inventory control for a k-out-of-n: F system of multi-state degrading components, *European Journal of Operational Research* 290 (2) (2021) 514–529.
- [24] J. Bowen, V. Stavridou, Safety-critical systems, formal methods and standards, *Software engineering journal* 8 (4) (1993) 189–209.
- [25] N. Aissani, B. Beldjilali, D. Trentesaux, Dynamic scheduling of maintenance tasks in the petroleum industry: A reinforcement approach, *Engineering Applications of Artificial Intelligence* 22 (7) (2009) 1089–1103.
- [26] V. Mattila, K. Virtanen, Scheduling fighter aircraft maintenance with reinforcement learning, in: *Proceedings of the 2011 Winter Simulation Conference (WSC)*, IEEE, 2011, pp. 2535–2546.
- [27] X. Wang, H. Wang, C. Qi, Multi-agent reinforcement learning based maintenance policy for a resource constrained flow line system, *Journal of Intelligent Manufacturing* 27 (2016) 325–333.
- [28] S. Wei, Y. Bao, H. Li, Optimal policy for structure maintenance: A deep reinforcement learning framework, *Structural Safety* 83 (2020) 101906.
- [29] L. Yao, Q. Dong, J. Jiang, F. Ni, Deep reinforcement learning for long-term pavement maintenance planning, *Computer-Aided Civil and Infrastructure Engineering* 35 (11) (2020) 1230–1245.
- [30] A. Tanimoto, Combinatorial q-learning for condition-based infrastructure maintenance, *IEEE access* 9 (2021) 46788–46799.
- [31] A. V. Le, P. T. Kyaw, P. Veerajagadheswar, M. V. J. Muthugala, M. R. Elara, M. Kumar, N. H. K. Nhan, Reinforcement learning-based optimal complete water-blasting for autonomous ship hull corrosion cleaning system, *Ocean Engineering* 220 (2021) 108477.
- [32] J. Chatterjee, N. Dethlefs, Deep learning with knowledge transfer for explainable anomaly prediction in wind turbines, *Wind Energy* 23 (8) (2020) 1693–1710.
- [33] R. Rocchetta, L. Bellani, M. Compare, E. Zio, E. Patelli, A reinforcement learning framework for optimal operation and maintenance of power grids, *Applied energy* 241 (2019) 291–301.
- [34] Y. Yang, L. Yao, Optimization method of power equipment maintenance plan decision-making based on deep reinforcement learning, *Mathematical Problems in Engineering* 2021 (1) (2021) 9372803.
- [35] Q. Wu, Q. Feng, Y. Ren, Q. Xia, Z. Wang, B. Cai, An intelligent preventive maintenance method based on reinforcement learning for battery energy storage systems, *IEEE Transactions on Industrial Informatics* 17 (12) (2021) 8254–8264.
- [36] Y. Ma, H. Qin, X. Yin, Research on self-perception and active warning model of medical equipment operation and maintenance status based on machine learning algorithm, *Zhongguo yi Liao qi xie za zhi= Chinese Journal of Medical Instrumentation* 45 (5) (2021) 580–584.

- [37] J. Wang, L. Zhao, J. Liu, N. Kato, Smart resource allocation for mobile edge computing: A deep reinforcement learning approach, *IEEE Transactions on emerging topics in computing* 9 (3) (2019) 1529–1541.
- [38] A. P. Marugán, Applications of reinforcement learning for maintenance of engineering systems: A review, *Advances in Engineering Software* 183 (2023) 103487.
- [39] C. P. Andriotis, K. G. Papakonstantinou, Deep reinforcement learning driven inspection and maintenance planning under incomplete information and constraints, *Reliability Engineering & System Safety* 212 (2021) 107551.
- [40] S. Peng, et al., Reinforcement learning with gaussian processes for condition-based maintenance, *Computers & Industrial Engineering* 158 (2021) 107321.
- [41] H. Wang, Q. Yan, S. Zhang, Integrated scheduling and flexible maintenance in deteriorating multi-state single machine system using a reinforcement learning approach, *Advanced Engineering Informatics* 49 (2021) 101339.
- [42] P. Zhang, X. Zhu, M. Xie, A model-based reinforcement learning approach for maintenance optimization of degrading systems in a large state space, *Computers & Industrial Engineering* 161 (2021) 107622.
- [43] A. Adsule, M. Kulkarni, A. Tewari, Reinforcement learning for optimal policy learning in condition-based maintenance, *IET Collaborative Intelligent Manufacturing* 2 (4) (2020) 182–188.
- [44] Y. Zhao, C. Smidts, Reinforcement learning for adaptive maintenance policy optimization under imperfect knowledge of the system degradation model and partial observability of system states, *Reliability engineering & system safety* 224 (2022) 108541.
- [45] N. Zhang, W. Si, Deep reinforcement learning for condition-based maintenance planning of multi-component systems under dependent competing risks, *Reliability Engineering & System Safety* 203 (2020) 107094.
- [46] N. Yousefi, S. Tsianikas, D. W. Coit, Reinforcement learning for dynamic condition-based maintenance of a system with individually repairable components, *Quality Engineering* 32 (3) (2020) 388–408.
- [47] A. K. Shakya, G. Pillai, S. Chakrabarty, Reinforcement learning algorithms: A brief survey, *Expert Systems with Applications* 231 (2023) 120495.
- [48] H. Hasselt, Double q-learning, *Advances in neural information processing systems* 23 (2010).
- [49] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al., Human-level control through deep reinforcement learning, *nature* 518 (7540) (2015) 529–533.

- [50] S. Mo, X. Pei, Z. Chen, Decision-making for oncoming traffic overtaking scenario using double dqn, in: 2019 3rd Conference on Vehicle Control and Intelligence (CVCI), IEEE, 2019, pp. 1–4.
- [51] Y. Li, H. He, Learning of emss in continuous state space-discrete action space, in: Deep Reinforcement Learning-Based Energy Management for Hybrid Electric Vehicles, Springer, 2022, pp. 23–49.
- [52] A. Raghu, M. Komorowski, L. A. Celi, P. Szolovits, M. Ghassemi, Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach, in: Machine Learning for Healthcare Conference, PMLR, 2017, pp. 147–163.
- [53] X. Zhang, X. Shi, Z. Zhang, Z. Wang, L. Zhang, A ddqn path planning algorithm based on experience classification and multi steps for mobile robots, *Electronics* 11 (14) (2022) 2120.
- [54] A. P. Marugan, F. P. G. Marquez, J. M. Pinar-Perez, A comparative study of preventive maintenance thresholds for deteriorating systems, in: E3S Web of Conferences, Vol. 409, EDP Sciences, 2023, p. 04015.
- [55] S. Hao, J. Yang, C. Bérenguer, Condition-based maintenance with imperfect inspections for continuous degradation processes, *Applied Mathematical Modelling* 86 (2020) 311–334.
- [56] K. T. Huynh, A hybrid condition-based maintenance model for deteriorating systems subject to nonmemoryless imperfect repairs and perfect replacements, *IEEE Transactions on Reliability* 69 (2) (2019) 781–815.
- [57] P. D. Van, C. Bérenguer, Condition-based maintenance with imperfect preventive repairs for a deteriorating production system, *Quality and Reliability Engineering International* 28 (6) (2012) 624–633.
- [58] P. Do, A. Voisin, E. Levrat, B. Iung, A proactive condition-based maintenance strategy with both perfect and imperfect maintenance actions, *Reliability Engineering & System Safety* 133 (2015) 22–32.
- [59] R. Zheng, V. Makis, Optimal condition-based maintenance with general repair and two dependent failure modes, *Computers & Industrial Engineering* 141 (2020) 106322.