Hybrid Algorithm based on Reinforcement Learning and DDMRP methodology for inventory management

Carlos Andrés Cuartas Murillo[1] **and Jose Lisandro Aguilar Castro**[1, 2, 3]

[1] GIDITIC, Universidad EAFIT, Medellín, Colombia
[2] Dpto de Automática, Universidad de Alcalá, España
[3] CEMISID, Universidad de Los Andes, Mérida, Venezuela
cacuartasm@eafit.edu.co, jlaguilarc@eafit.edu.co

Abstract. This article proposes a hybrid algorithm based on Reinforcement Learning and on the inventory management methodology called DDMRP (Demand Driven Material Requirement Planning) to determine the optimal time to buy a certain product, and how much quantity should be requested. For this, the inventory management problem is formulated as a Markov Decision Process where the environment with which the system interacts is designed from the concepts raised in the DDMRP methodology, and through the Reinforcement Learning algorithm – specifically, Q-Learning. The optimal policy is determined for making decisions about when and how much to buy. To determine the optimal policy, three approaches are proposed for the reward function: the first one is based on inventory levels; the second is an optimization function based on the distance of the inventory to its optimal level, and the third is a shaping function based on levels and distances to the optimal inventory. The results show that the proposed algorithm has promising results in scenarios with different characteristics, performing adequately in difficult case studies with a diversity of situations such as scenarios with discontinuous or continuous demand, seasonal and non-seasonal behavior with high demand peaks, multiple lead times, among others.

Keywords*:* Smart inventory; DDMRP; Inventory Management System; Reinforcement Learning, *Q-Learning*.

1. Introduction

An efficient inventory management requires a special interest in companies dedicated to commercialization or production. Thus, “inventory represents one of the most important investments of companies compared to the rest of their assets, being essential for sales and optimizing profits" (Durán, 2011). Hence, the relevance of an efficient inventory management, as well as production planning, are critical elements that represent a competitive advantage, and that constitute a determining factor for the long-term survival of the organization (Silver, Pyke, Thomas, 2017). Inventory management has traditionally been approached through the implementation of MRP (Material Requirement Planning) (Rossi et.al, 2017), a methodology introduced by Joseph Orlicky (1976), which aims to plan material requirements (Huq and Huq, 1994). However, and despite its popularity, this methodology has an important limitation since its precision is not suitable for dynamic environments. Therefore, small variations in the system lead to the bullwhip effect in the supply chain, which consists of distortions that are generated between the number of units demanded versus those purchased (Constantino et. al 2013). This effect has been widely studied in the literature (Steele (1975), Mather (1977) and Wemmerlov (1979)), and generates changes in work schedules, increases costs, among other things.

Given the above, the present work is carried out based on an alternative methodology: the DDMRP, developed by Ptak and Smith (2011), which allows a better adaptation in environments with high variability, and therefore, more efficient inventory management. The "Demand Driven" approach, called DDMRP, introduces the creation of decoupling to absorb variability, reduce lead times, and reduce overall capital investment.

Thus, in this article, a hybrid algorithm is developed based on Reinforcement Learning and on the DDMRP inventory management methodology, to determine the optimal time to buy a product, and the quantity requested on the purchase order. It is important to highlight with respect to this last aspect (quantity of units), that it should not be very high since the demand for more resources increases the costs; nor very low because it can cause unsatisfied demand, production delays, among other problems.

The main contribution is the definition of a hybrid algorithm based on Reinforcement Learning and on DDMRP to determine when and how much to buy a certain product. The hybrid algorithm is defined with three different reward functions based on the DDMRP theory, an optimization function or a shaping function. They are evaluated in multiple case studies, which differ from each other according to the next characteristics: discontinuous or continuous demand, seasonal and non-seasonal behaviors, with high or low demand peaks, with different lead times, among others. Thus, the main contribution of this work is the implementation of a hybrid reinforcement learning algorithm that allows a more efficient inventory management process than the one proposed in the DDMRP theory. Additionally, an alternative formula to the one defined in the DDMRP theory is also proposed, to calculate the optimal inventory level in a more efficient way.

The article is organized as follows: in session 2, a literature review is presented; Section 3 describes the theoretical framework. In section 4, the experimentations are carried out; and in section 5, an analysis and discussion of the results is presented. Finally, in section 6, the conclusions of the study are described.

2. Literature review

The general trend of research on inventory management has been usually using the MRP, as stated Rossi et. al (2017), in which they remark that around 75% of manufacturing companies use MRP as the main method for planning production. Since the introduction of the MRP, a wide variety of investigations have been developed, such as the proposed by Pooya, Fakhlaei, Alizadeh-Zoeram (2021), in which dynamic systems are used to reduce the impact of the bullwhip effect produced by demand, and thus, reduce production costs.

As an alternative system to MRP has been developed DDMRP, a system that solves the problem of the bullwhip effect through the positioning of decoupling points or buffers located in the supply chain (Ptak and Smith, 2016). The main function of these buffers is to store a certain number of products to avoid the variability of demand or variability in the supply chain. Around the DDMRP, researches have been developed mainly focused on exploring the advantages of this methodology in organizations, such as the one proposed by Velasco et al. (2020), where the authors recreate a simulation environment of the system through Arena software and demonstrate the efficiency of the system in manufacturing environments, obtaining results such as a reduction in lead time of 41%, and a decrease of 18% in inventory levels.

On the other hand, authors such as Kortabarria et al. (2018) present a case study of a manufacturing company of home appliance components in which they compare an inventory management methodology based on MRP to one based on DDMRP. Their results have reduced the bullwhip effect and rush orders. Also, Shofa and Widyarto (2018) developed a case study for a company in the Indonesian automotive sector where their results show through simulation that the delivery times of the DDMRP method were reduced from 52 to 3 days, and additionally, the levels of inventory were lower than when the MRP approach was used.

But DDMRP and MRP are not the only models that have been studied in the literature. Mathematically, inventory management has been proposed as an optimization problem whose objective is to maximize profit and minimize costs. These models have been applied in various organizational areas; for example, authors such as Hubbs et al., (2020) and Karimi et al. (2017) developed an inventory management system aimed at human resource scheduling in production. Analogously, Paraschos et al. (2020) developed a model to optimize the tradeoff between machinery maintenance, equipment failures, and quality control.

In summary, although several studies propose inventory management systems from methodologies such as MRP and DDMRP, or like an optimization problem, no research was found in the literature using reinforcement learning techniques and DDMRP for inventory management.

3. Theoretical framework

3.1 Inventory Management

Inventories are all those items or stock used in production or commercialization in an organization (Durán, 2012, p.56). Some important aspects about how to obtain and maintain an adequate inventory are: absorbing fluctuations in demand, having protection against the lack of reliability of the supplier or a product that is difficult to ensure a constant supply, obtaining discounts when ordering with larger quantities, and reducing order costs if they are carried out less frequently (Muller, 2011). Regarding this last aspect, Peterson, Silver and Pyke (1998) point out that there are basically five categories of costs associated with inventory management: the unit cost of the value of the product, costs of maintaining the products, ordering costs, stockout costs, and those associated with control systems.

On the other hand, DDMRP combines relevant features of MRP, distribution resource planning (DRP) and Six Sigma. It is a system that allows the adaptation to dynamic demand environments and that avoids the amplification of the bullwhip effect in the supply chain through buffers. In general, these buffers act as decoupling points of fluctuations, not only in demand, but also those inherent or associated with the supply chain. Thus, the DDMRP implements buffers (also called decoupling points) whose function is to create independence between the supply chain, use of materials, and demand. This is achieved by establishing optimal inventory levels at the decoupling points, in such a way that if any variation is generated in the system, it is not transmitted through the entire supply chain. In the next subsections are presented some concepts related to DDMRP.

3.1.1 Buffer

The buffers are made up of three zones: red, yellow, and green, which will be described below.

Red Zone

It is the lower zone of the buffer and is associated with low inventory levels. The way to calculate its base (BZR) is:

$$BZR = ADU * DLT * LTF \quad (11)$$

Where:
ADU: is the average daily usage.
DLT: Lead Time between buffers or decoupling points.
LTF: variability factor that gives a greater threshold in delivery times.

Now, the upper limit of the red zone (TOR) is given by:

$$TOR = BZR\ X\ FV \quad (12)$$

*w*here,

FV: variability factor that gives a greater slack to the area in case the demand for the product is highly variable.

Yellow Zone

It corresponds to the intermediate level of the buffer. The lower limit of the yellow zone is TOR, and the upper limit (TOY) is calculated as:

$$TOY = TOR\ +\ (ADU\ X\ DLT) \quad (13)$$

Green Zone

It corresponds to the upper zone of the buffer and is associated with high inventory levels. The lower limit of the zone is given by TOY. To determine the upper limit of this zone (TOG), it is necessary to calculate the following three factors:
i) Order cycle (DOC): this factor represents the number of days between orders. It sets the imposed or desired number of days of inventory until a new replenishment order is made. The way to calculate it is:

$$ADU\ X\ Days\ between\ orders \quad (14)$$

ii) Base of the red zone (BZR), calculated according to equation (11)

iii) Minimum order quantity that can be made (MOQ).

Now, once the 3 factors have been calculated, the TOG is calculated as follows:

$$TOG = TOY\ + max\ (DOC, BZV, MOQ) \quad (15)$$

3.1.2 Qualified Demand

Qualified demand is made up of the sum of demand orders existing to date, and the sales orders that exceed the OST level (Order Spike Threshold) in a certain time horizon (OSH). This time horizon is equivalent to the DLT value. Note that the OST level represents the maximum demand threshold for it to be considered as a demand peak. This ensures that high levels of demand are identified, as well as the supply of materials necessary to satisfy them. This level is defined as the value of the ADU.

3.1.3 Net flow inventory

Net flow inventory position (NFP) is a concept defined in the DDMRP methodology associated with the amount of inventory available. This generates the signal to request a supply order; in other words, it defines the need to make a purchase. To calculate it, Ptak & Smith (2016) define the following equation:

$$NFP = OH + OP - QD \quad (1)$$

Where:
OH: Inventory available; quantity of stock available to be used.
OP: Quantity of stock ordered not received.
QD: Qualified demand orders.

3.1.4 Optimal level of inventory

Ptak and Smith (2016) define the optimal level of inventory from the following equation:

$$OH^* = TOR + \frac{(TOG - TOY)}{2} \quad (3)$$

3.1.5 Purchase order

The buy signal is generated when the NFP is less than or equal to the TOY level. The number of recommended units to request in the purchase order (SR) is calculated from:

$$SR = TOG - NFP$$

Otherwise, no purchase order is generated.

3.2 Reinforcement Learning

Reinforcement Learning (RL) is a type of learning where actions to take are not defined, rather than that, these are discovered based on experience (Sutton and Barto, 2018). In other words, learning takes place through trial and error, and the rewards obtained in each of those.

These interactions are generally modeled as a MDP, which is made up of the following elements: the agent, in charge of the learning and decision-making process; and the environment, which are all the objects with which the agent interacts. (Watkins, 1989). These are a formalization of a sequential decision-making process where actions are influenced not only by immediate actions, but also by those taken in future situations and states (Sutton and Barto, 2018). To do this, the agent selects an action, and the environment generates a new situation and a reward for the action chosen.

In general, the structure of an MDP consists of 4 parts: the possible states (s), the possible actions (a), a transition function and a reward function (R). If the actions are deterministic, then a transition function is defined to assigning each (s, a) a new state (s') as a result of the interaction between both. On the other hand, if the action is stochastic, then the transition function is defined as a probability function, where $P(s'|s,a)$ represents the probability of being in a state s' given the couple s and a. It should be noted that the final objective of the MDP is to find a policy: π: $s \rightarrow a$ that maximizes the expected value of the rewards associated with the states. Thus, we seek to maximize the expected profit given by the function (Sutton and Barto, 2018):

$$G_t = R_{t+1} + R_{t+2} \cdots + R_t \quad (4)$$

Where:

R_t: It is the reward obtained in episode t.

Which, defined in a recursive and generalized way, gives (Sutton and Barto, 2018):

$$G_{(t)} = \sum_{k=0}^{\infty} \gamma^k R_{\mathrm{t+k+1}} \quad (5)$$

Where:

$G_{(t)}$: It is the reward function obtained in episode t.
k: Interval of time.
γ: discount factor.
$R_{(s,t+\mathrm{k+1})}$: reward for action taken in the moment $t + k + 1$ by the state s.

Now, the agent's behavior in relation to the probability of selecting a certain action is defined based on the policies. In this way, it determines how desirable it is to take an action in a specific state. Under a certain policy, action-value functions are defined. The way to calculate the function is as follows (Sutton and Barto, 2018):

$$q_\pi(s, a) = E_\pi[R_t | S_t = s, A_t = a] \quad (6)$$

Where:

$q_\pi(s, a)$: Action value function of state s.
E_π: Expected value under policy π.
R: Reward.
S_t: State at time t.
A_t: Action at time t.

3.2.1 Q-Learning

Q Learning is an RL algorithm introduced by Watkins (1989). It is characterized by being an off-policy, a policy where the optimal policy is learned independently of the agent's actions. This, as stated by Sutton and Barto (2018), allows the convergence of the algorithm to be faster. Now, regarding the calculation of the Q values with which the stock value function is constructed, it is carried out as follows (Watkins, 1989):

$$Q(S, A) \leftarrow (1 - \alpha) * Q(S, A) + \alpha[R + \gamma \, max \, Q(S', A) - Q(S, A)] \quad (7)$$

Where:

$Q(S, A)$: expected reward value for action taken in state *S*
α: Learning rate.
R: $Rewards$.
γ: Discount factor.

3.2.2 Shaping function

The Shaping function was initially introduced by Skinner (1958) because of the effectiveness obtained by training an animal by giving it rewards during the learning process, once it performed behaviors similar to those desired. Similarly, the shaping function has been

implemented in RL algorithms by authors such as Ng, Harada and Russell (1999), proving to be a very efficient technique, and sometimes indispensable for a quick convergence of the learning algorithm (Sutton and Barto, 2018). Formally, the reward function is defined as:

$$R(s,a,s') = R(s,a,s') + F(\text{s,a,s'}) \quad (8)$$

Where:

R(s,a,s')= Reward function.
F(s,a,s')=Shaping function

For its part, F is defined as:

$$F(s,a,s') = \gamma\phi(s') - \gamma\phi(s) \quad (9)$$

Where:
γ: Discount factor.
ϕ: Function that defines how close or far the agent is from the target.

The inventory environment is described in Figure 1. It will consist mainly of three components: the first associated with the purchase orders (left part of Fig. 1: Supply Side); the second one to the buffer (in the center of Fig. 1), and the third is the demand side, associated with the demand in a given time horizon (OSH) and the OST.

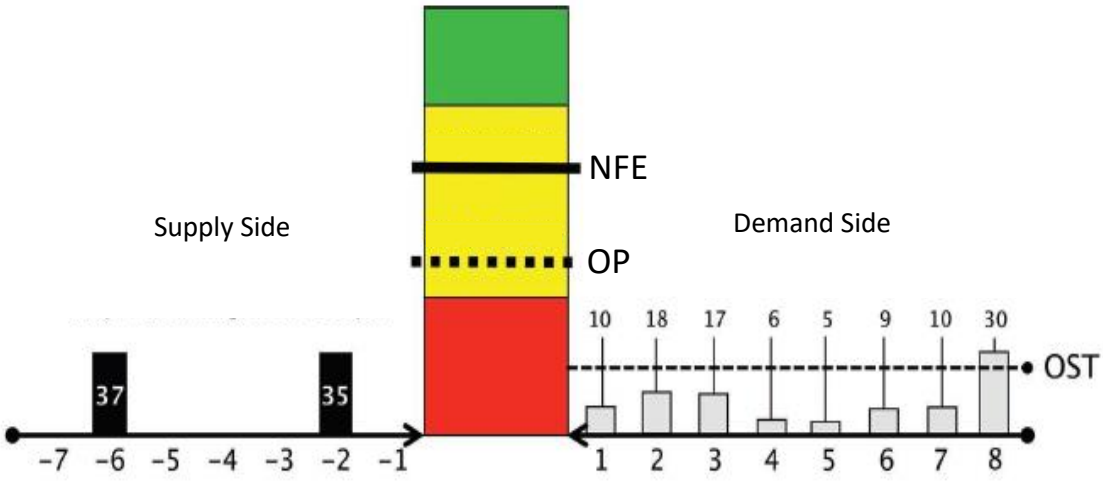


Figure 1 *Inventory Management. Ptak y Smith (2016)*

4. Proposed model

The proposed model in this study is the definition of an optimization problem (W) that aims to minimize the distance between the real inventory (OH) and the optimal inventory level (OH*) defined in Eq. (3):

$$W = \ min[\sum_{t=1}^{n}(OH^* - \ OH)] \quad (10)$$

Now, to solve this optimization problem, it is structured as an MDP where the environment is defined from the theoretical concepts of DDMRP (Figure 1), and the policy to optimize and learn is the request for orders of products. This last process is carried out through Q-Learning.

4.1 Purchase order

The time horizon of this component is given by the DLT, and represents the order of products that are pending to be received. Once a purchase order is placed, it is represented as an order in the period: $P_t - DLT$, where P_t: Period where the order was placed.

4.2 Our optimal inventory level

Although in Eq. (3) DDMRP theory provides a function to calculate the OH*, we propose an alternative function defined as:

$$OH^* = TOR + \frac{(TOG + TOY)}{2}$$

Note that both OH* will be used in each of the case studies to compare the performance between them

4.3 Markov decision process

Next, each of the MDP components of the algorithm are defined.

4.3.1 Actions

The action of agent A_t will be based on the number of units to buy at a certain time. Based on OH inventory, the agent must determine the optimal number of units to request in the order (if necessary).

4.3.2 Rewards

Three reward approaches were developed.

R_1 : Rewards based on DDMRP levels

The first approach is based on the state (S_t) of the inventory. Since the most desirable level for the DDMRP theory is yellow, a rewards function R_1 will then be defined such that:

$$R_1 = \begin{cases} -1, & TOY < S_t \leq TOG \\ 1, & TOR < S_t \leq TOY \\ 0, & 0 \leq S_t \leq TOR \end{cases} \quad (11)$$

This reward function seeks to optimize the desirable level of inventory, in this case, yellow.

R_2 : Rewards based on optimization

Given that the goal is to minimize the distance between the OH and the optimal OH*, the following reward function is defined based on equation (10):

$$R_2 = \frac{1}{W} \quad (12)$$

It should be noted that by maximizing the reward function in Eq. (12), the optimization defined in Eq. (10) is minimized. With this reward, it is sought to optimize the distance of the inventory to the optimal inventory value.

R_3 *: Rewards based on Shaping*

Finally, a Shaping approach based on equation (11) will be used, such that a new reward function will be defined:

$$R_3 = R_1 + \phi(s) \tag{13}$$

Where $\phi(s)$ is calculated as follows:

$$\phi\,(s) = \text{OH*}(s) - \text{OH}(s) \tag{14}$$

This reward allows optimizing both the inventory level and distance to the optimal value.

4.3.2 States

The model states are given by three components: OH, OH*, and the lead time (LT). This information will be stored at time t in a tuple with the following structure:

$$S = (OH, OH^*, LT\ )$$

For example, the state S (100, 120, 3) represents an inventory level of 100 units, an optimal inventory of 120, and a lead time of 3.

4.4 Variables and assumptions

Since the algorithm uses the inventory environment defined by the DDMRP theory as the environment, each of the variables explained in section 3.1 are used. The assumptions used to develop the model associated with the *real scenario* are defined below:

- Demand: given that the historical data of the demand was very limited in the proposed scenarios, it was decided to generate pseudo-random data for learning the model by means of the Mersenne Twister algorithm, from the maximum and minimum demand identified in the historical data. The Mersenne Twister algorithm was selected for two reasons: first, because it is one of the best generators of pseudo-random numbers (Matsumoto and Nishimura, 1998), and second, because its characteristics can significantly favor the convergence time of an algorithm (Bonato et al., 2013).
- ADU-OSH: For the real scenario, these variables were calculated from the demand of the previous number, based on a 60-day moving average.
- DLT-Lead Time-OSH: taken from the median of the Lead Time of the last year of the historical data.
- Initial OH: the initial OH was determined from the final inventory of the period prior to the testing of the historical data.
- MOQ: calculated as the minimum purchase order present in the historical data.

In relation to the *theoretical scenario*, the assumption used for the construction of the model was the use of the Mersenne Twister random number generator to simulate the demand in the learning process. Also, for the rest of the variables, this scenario has established the parameters for each of them.

5. Experimentation

5.1 Case studies

The algorithm will be implemented in two test scenarios: the first one will be a theoretical scenario made from the data presented in chapter 9 of the book *Demand Driven Material Requirements Planning*, by the authors Ptak and Smith (2016), used to simulate the behavior of DDMRP for a given product. This first scenario will be used to compare, from a theoretical point of view, the behavior of our algorithm proposed with the behavior of the DDMRP in the simulation environment that the authors of the book proposed.

On the other hand, in the second case, a real scenario will be carried out in an organization of the logistics sector, in which the behavior of 3 products with different distribution centers and demand behavior will be evaluated. Below we define each of the case studies. Table 1 shows the products that will be analyzed in the real and theoretical cases.

Case study	Min demand	Mean demand	Median demand	Max demand	Standard dev. demand	LT	MOQ
P1	2	10.53	9	30	6.03	7	20
P2	0	2.52	0	72	8.20	9	40
P3	0	3.61	0	256	14.78	15	16
P4	0	1.05	0	26	2.61	7	4

***Table* 1** *Case study characteristics features.*

5.1.1 P1: Theoretical case study

This first case study is developed from the simulation data in chapter 9 of the book *Demand Driven Material Requirements Planning*, by the authors Ptak and Smith (2016). This case study is tested in 21 days with a product whose demand is continuous (see Figure 2). Figure 2 shows that there´s only one demand peak in day 7, the maximum demand requested is 30 units, the median demand is 9 units (and 10.53 in mean), has a lead time of 7 days and a MOQ of 20 units. Notice that the period of time is so short that there´s no evidence to conclude the demand has any kind of stationary or trend behavior.

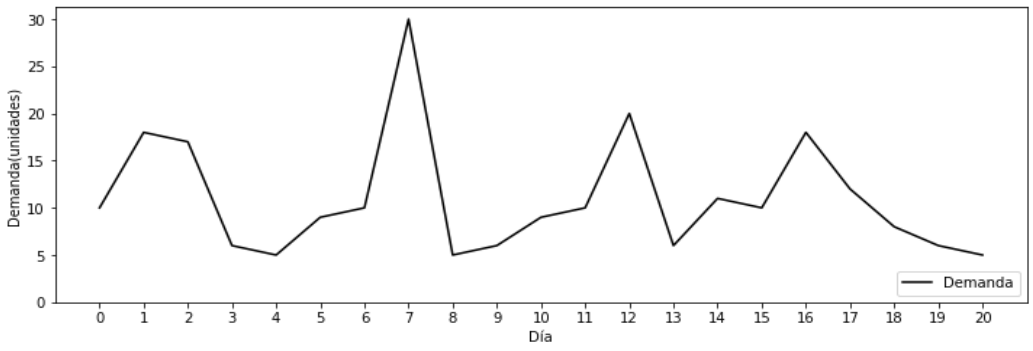


Figure 2 *Case T1 Demand*

5.1.2 P2: case study of product 39933

This case study was taken from the historical behavior of product 39933 from the operation center 11 of the logistics company. This case has a discontinuous demand, and it is characterized by having a stationary time series in mean and in variance (see Figure 3). Also, there's neither a noticeable trend or a significant change in variance over time. The median demand of this product is 0 (and mean 2.52), the maximum demand is 72 units, has a MOQ of 40 units, and an LT of 9 days (see Table 1).

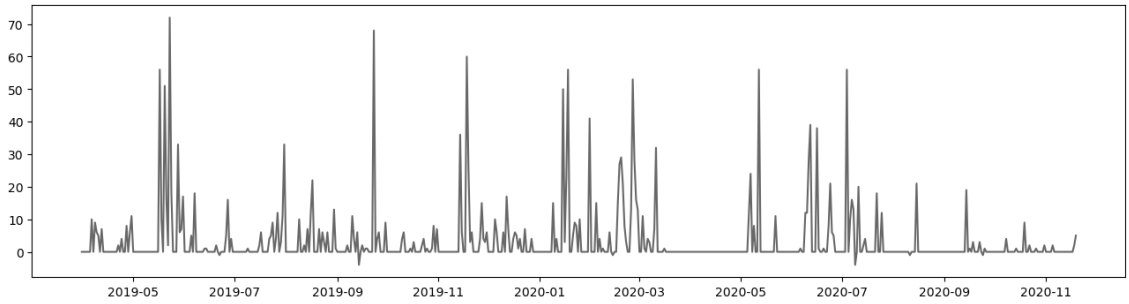


Figure 3 *Case P2 Demand*

5.1.3 P3: case study of product 28440

This case study was taken from the historical behavior of product 28440 of the operation center 12 of the company in the logistics sector under study. The demand for this product is discontinuous, and it is characterized by having a trend time series with non-stationary variance in mean or variance (see Figure 4). Notice how the variation of the variance, in terms of dispersion of data, is lower in the time window of the first half of the time series. and higher at the end of it. Additionally, it can be seen that the quantity demanded increases overtime. Also, notice that this case study has the highest standard deviation and demand quantities over the case studies (see Table 1). The maximum demand requested is 256 units, has a median of 0 units (mean of 3.61), has a MOQ of 16 units, and an LT of 16 days, being the case that takes the higher time to be delivered to the operations center.

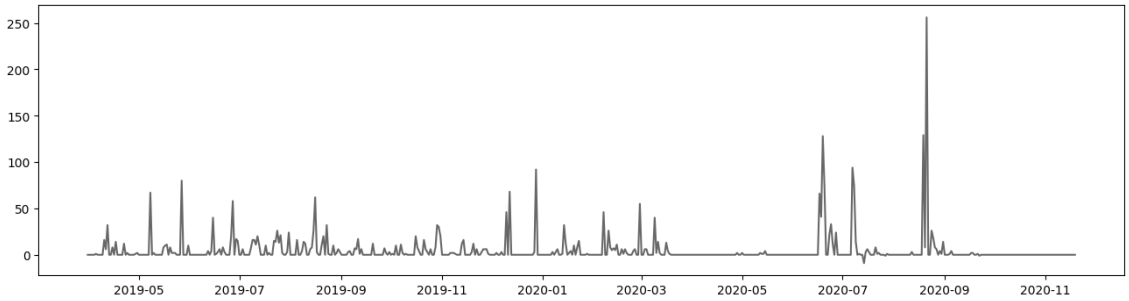


Figure 4 *Case P2 Demand*

5.1.4 P4: case study of product 43387

This case study was taken from the historical behavior of product 43387 from the operation center 14 of the company in the logistics sector under study. The demand for this product is discontinuous and characterized by having a seasonal time series with non-stationary variance in mean or variance (see figure 5). Notice how there´s seasonality around the midterm of both

years and how the variance increases overtime. Overall, of the case studies, this product has the lower standard deviation, it has a maximum demand of 26 units and a median of 0 (1.05 in mean). This product has an MOQ of 4 units and a LT of 7 days (see Table 1).

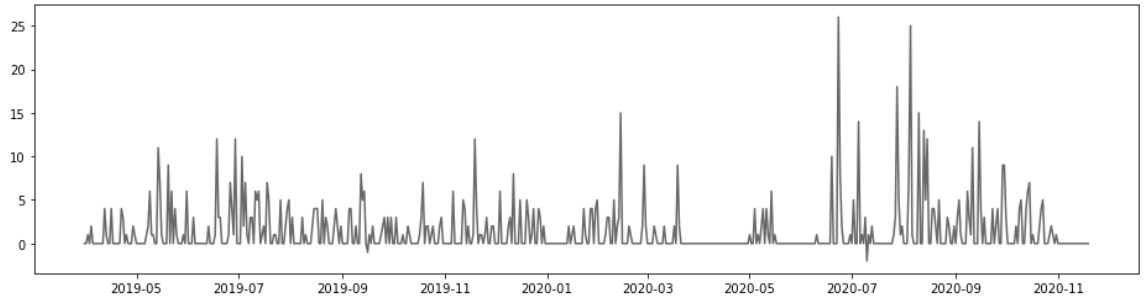


Figure 5 *Demand-P4 scenario*

5.2 Evaluation Metrics

The evaluation metrics will be classified into 2 categories, RL metrics and logistic metrics, they will be described below:

5.2.1 Logistic metrics

Below the logistics metrics are presented

Bullwhip effect ratio (REL): this metric is used to evaluate the ability to avoid spreading distortions between the orders purchased and demand of the product (Romero et al., 2016). In our algorithm, it will be compared the orders purchased by the optimal policy learned by our RL and the demand of the test period. The expression to calculate it is defined as:

$$REL = \frac{\sigma^2_{orders\ purchased}}{\sigma^2_{demand}}$$

where:

σ^2: $variance$.

The closer the ratio is to 1, the less is the distortion of the bullwhip effect. Note that a result equal to one means there is no distortion, thus, there´s no bullwhip effect.

Number of stockouts (BS): This metric is proposed to evaluate the number of times the stock is broken. This is a very relevant event because it can cause an increased risk of lost sales as well as it leads to reduced customer satisfaction and lowered loyalty levels (Merrad et al., 2020). Note that the lower the number of stockouts, the better the inventory policy was learned.

Average OH* distance (AOHD): This metric is used with the objective of evaluating the performance of the closeness of the inventory of our RL algorithm to its optimal level. Therefore, the closer to zero, the better. Mathematically, it represents the Euclidean distance between OH and OH*. Its formulation will then be:

$$d(OH, OH*) = |OH - OH*|, \forall\ OH, OH* \in \mathbb{R} \qquad (19)$$

Now, to evaluate the general behavior throughout the episodes, the median of these distances will be calculated.

5.2.2 RL metrics

Below, the reinforcement learning metrics are presented.

Average Accumulated Reward (AAR): The average accumulated reward metric is used to evaluate the performance of the policy learning process (Sutton and Barto, 2018), in our proposed model, the purchase order policy. A higher AAR value is better since the algorithm has obtained a higher reward on average. It is calculated as follows:

$$AAR = \sum_{1}^{t} \frac{(r_1 + r_2 + \ldots + r_t)}{N}$$

Where: r_t represents the rewards at episode t, and N: the number of episodes.

Percentage of Best-Accumulated Reward (PBAR): The best accumulated reward (BAR) is defined as the global maximum sum of rewards obtained in an episode in the whole run of the learning process. Thus, PBAR represents the proportion of BAR achieved. PBAR at time t can be calculated as:

$$PBAR_t = \frac{max\ (r_1 + r_2 + \ldots + r_t)}{BAR}$$

Where BAR is calculated as:

$$BAR = max\ (r_1 + r_2 + \ldots + r_N)$$

This metric is important because it shows how long it takes in terms of episodes to achieve the best-accumulated reward.

Rate of Convergence of the Algorithm (AC): This metric is widely used in the RL context by authors like Sutton (1988), and Watkins and Dayan (1992) to prove the ability of an algorithm to find an optimal value. Although the capabilities of Q learning algorithm to converge are proved mathematically by Watkins and Dayan (1992), convergence is used in this paper to prove that our algorithm is working as it should in terms of finding an optimal policy, and to visualize the speed of convergence. To show the convergence of our algorithm, the average accumulated rewards (AAR) and the number of episodes is compared in the learning process. It is a way to view the convergence of an algorithm in practice (Sutton and Barto, 2018).

5.3 Learning and evaluation periods

In all the proposed scenarios, a learning process was carried out in a simulation environment composed of a time window of 800 days. Once the learning process was carried out, on the non-theoretical case studies, the evaluation process was carried out from April 1, 2019, to April 1, 2020. In the theoretical case study, the evaluation process was carried out in the 21 days of simulation proposed in the book.

5.4 Results Analysis

In each of the next case studies, we compare the results of our models (R1-R2-R3) versus the results obtained by DDMRP´s theory (named “DDMRP” in the results tables). Note that the DDMRP is not a reinforcement learning model, it is obtained by calculating the units to be requested in the purchase order (SR) as defined in Section 3.5.1. Additionally, for each case study are shown the results obtained by our RL approach using the optimal inventory level defined by the DDMRP theory (see Eq. (3), named in the results as "DDMRP OH*") and the results obtained

based on the optimal inventory level that we propose (see Section 4.2, named in the results as "our proposed RL OH*").

5.4.1 P1: Theoretical case study.

In Table 2.a, the model with the fastest learning process was the one with the reward function R1 (P1R1). The PBAR reaches 93% of the BAR within 100 episodes. The slowest learning model was P1R2, in 100 episodes it reaches only 80% of the BAR, and it takes 20,000 to reach 100%. Additionally, P1R1 and P1R3 models are the ones that obtain the best performance, in 30,000 episodes their AAR were 1 and 0.99, respectively. According to Table 2.b, in terms of learning, our proposed RL OH* has a very similar performance in the different models, which also is similar to *DDMRP* OH*.

With respect to convergence, the models based on DDMRP's OH* (P1R1-P1R2-P1R3) and based on our proposed RL OH* (P1R1P-P1R2P-P1R3P) converge properly, evidencing a successful learning process by all the algorithms (see Figure 7).

Test results can be observed in Table 3.a. For each of the products (P1-P2-P3-P4), four models are compared against each of the metrics. R1-R2-R3 are models related to each of the reward function defined in section 4.3.1, and the fourth model, DDMRP, is based on the purchase order policy defined in DDMRP´s theory.

The results of the metric AAR show that the model with the best performances is R3. According to the results of both, DDMRP´s OH* (Table 3.a) and our proposed RL OH* (Table 3.b), the highest result is obtained by this last model. This means that model R3 accumulated more rewards on average, indeed, it behaves better in terms of being at the desirable level (yellow). Particularly, in Table 3.a, the AAR of model R3 is 0.88, meaning that it accumulated 0.88 rewards per day on average (being the value 1 the maximum possible). Continuing with the metric AOHD, the best results are obtained, in the case of DDMRP´s OH* by R2 and, in our proposed RL OH* by R1 and R3. Note that in this case study all of our models proposed outperformed the purchased order policy of DDMRP theory. In Table 3.a, this model has a median of 16 units away from the OH * level, this being the closest distance between the compared models (the visual behavior of AOHD can be seen in Figures 6 and 8). In relation to the BS metric, it can be observed that none of the models performs any number of stockouts. In this sense, all the models behave well since at no time do they run out of inventory.

Finally, in terms of REL, in Table 3.a, it can be seen that all of our models outperformed the order purchase policy of DDMRP. DDMRP has a rate of 10.05, meaning it was the most affected in terms of bullwhip effect ratio.

*a) Results based on DDMRP OH**

	P1R1		P1R2		P1R3	
Episodes	**AAR**	**PBAR**	**AAR**	**PBAR**	**AAR**	**PBAR**
100	0.06	0.93	0.21	0.80	0.10	0.91
200	0.02	1.00	0.19	0.80	0.17	0.98
500	0.07	1.00	0.15	0.80	0.33	0.99
1000	0.05	1.00	0.08	0.85	0.39	0.99
2000	0.04	1.00	0.04	0.85	0.51	1.00
5000	0.20	1.00	0.23	0.92	0.44	1.00
10000	0.58	1.00	0.48	0.94	0.61	1.00
20000	0.89	1.00	0.84	1.00	0.90	1.00
30000	1.00	1.00	0.99	1.00	1.00	1.00

*b) Results based on our proposed RL OH**

	P1R1P		P1R2P		P1R3P	
Episodes	**AAR**	**PBAR**	**AAR**	**PBAR**	**AAR**	**PBAR**
100	0.06	0.93	0.21	0.80	0.42	0.97
200	0.02	1.00	0.19	0.80	0.42	0.97
500	0.07	1.00	0.15	0.80	0.39	0.97
1000	0.05	1.00	0.08	0.85	0.37	1.00
2000	0.04	1.00	0.04	0.85	0.39	1.00
5000	0.20	1.00	0.23	0.92	0.60	1.00
10000	0.58	1.00	0.48	0.94	0.64	1.00
20000	0.89	1.00	0.84	1.00	0.83	1.00
30000	1.00	1.00	0.99	1.00	0.92	1.00

Table 2 *P1 Training results*

a) Results based on DDMRP OH*

PRODUCT	REWARD FUNCTION	AAR	AOHD	BS	REL
P1	R1	0.15	15	0	8.96
	R2	0.16	**13**	0	**5.82**
	R3	**0.88**	15	0	**5.82**
	DDMRP	N/A	16	0	10.05

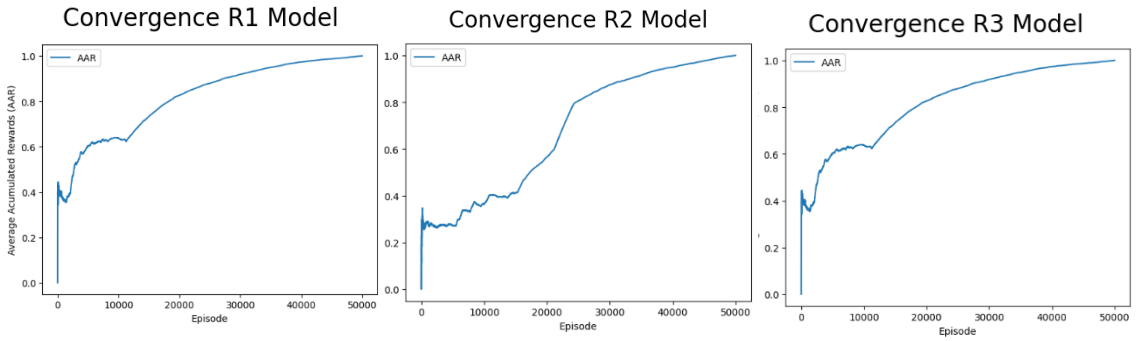
*b) Results based on our proposed RL OH**

PRODUCT	REWARD FUNCTION	AAR	AOHD	BS	REL
P1	R1P	0.69	**33**	0	27.31
	R2P	0.09	48	0	**25.42**
	R3P	**0.70**	**33**	0	**25.42**

Table 3 *P1 Test results*

Figure 6 *Model results behavior based on our proposed RL OH* behavior- P1 Case*

a) Results based on DDMRP OH*



b) Results based on our proposed RL OH*

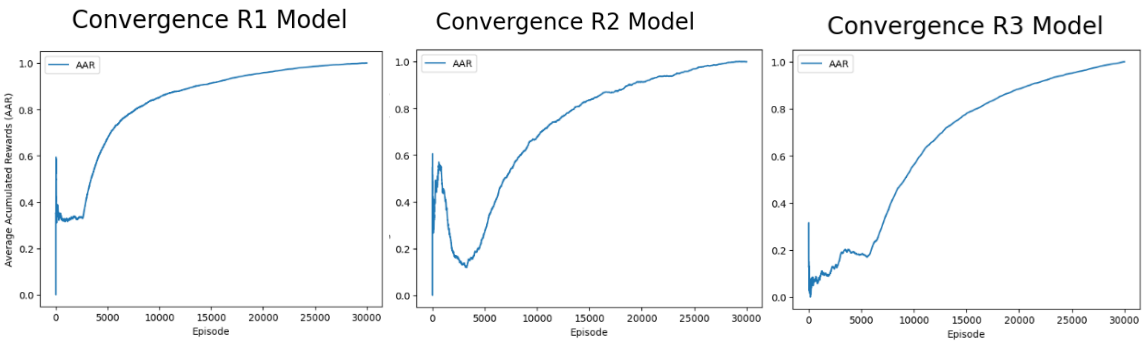


Figure 7 *P1 Convergence of models*

Figure 8 *Model results behavior based on our proposed RL OH* behavior- P1 Case*

5.4.2 P2: Product 39933

According to Table 4.a, the model with the fastest learning process uses the reward function R1. It reaches a PBAR of 32% in 100 episodes. It is observed that the learning time increases considerably in this case study with respect to P1 because the complexity of the variables of P2 case study are much higher. It can be observed that while the best model of P1 (P1R1) obtains 93% of the PBAR in 100 episodes, the best model of P2 (P2R3) requires around 2000 episodes to reach 92%, in other words, it takes around 20 times longer to reach nearly the same level of PBAR. In Table 4.b are shown the results of our approach, and the results are similar.

In figure 9 can be seen that models based on DDMRP's OH* (P2R1-P2R2-P2R3) and models based on our proposed RL OH* (P2R1P-P2R2P-P2R3P) converge properly, evidencing a successful learning process by all the algorithms.

Table 5 shows the results for product P2. The results on the metric AAR show that the model with the best performances is model R1. Note that the results on both, DDMRP´s OH* (Table 5.a) and our proposed RL OH* (Table 5.b), the highest result is obtained by this model. This means that R1 accumulated more rewards on average, indeed, it behaves better in terms of being at the desirable level (yellow). Particularly, in Table 5.a the AAR of R3 is 0.71, meaning that it accumulated 0.71 rewards per day on average (being the value 1 the maximum possible).

Continuing with the metric AOHD, the best results are obtained, on both in DDMRP´s OH* and in our proposed RL OH*, by R2. It can be seen that this model is a median of 17 units away from the OH * level, this being the closest distance between the compared models (the visual behavior of AOHD can be seen in Figures 10 and 11). In relation to the BS metric, it can be observed that all the models of DDMRP´s OH* have one day of stockout excepting R2 that has 14 days, this significantly changes with our proposed RL OH* in which the worst model, R1, has only 4 days of stockouts.

Finally, in terms of REL, in Table 5 is shown that the worst performances were obtained by R3 with our proposed RL OH* (P2R3P), and R1 based on DDRMP´s OH* (P2R1P), meaning they were affected the most in terms of the bullwhip effect ratio.

*a) Results based on DDMRP OH**

	P2R1		P2R2		P2R3	
Episodes	**AAR**	**PBAR**	**AAR**	**PBAR**	**AAR**	**PBAR**
100	0.09	0.32	0.01	0.06	0.07	0.29
200	0.11	0.49	0.01	0.07	0.11	0.34
500	0.36	0.49	0.01	0.07	0.24	0.70
1000	0.39	0.51	0.01	0.07	0.56	0.87
2000	0.43	0.54	0.01	0.10	0.79	0.92
5000	0.53	0.79	0.14	0.73	0.92	0.94
10000	0.69	0.88	0.56	0.87	0.96	0.99
20000	0.87	0.97	0.89	1.00	0.99	1.00
30000	0.95	1.00	0.95	1.00	0.99	1.00

*b) Results based on our proposed RL OH**

	P2R1P		P2R2P		P2R3P	
Episodes	**AAR**	**PBAR**	**AAR**	**PBAR**	**AAR**	**PBAR**
100	0.09	0.23	0.01	0.04	0.01	0.10
200	0.13	0.25	0.01	0.04	0.01	0.10
500	0.25	0.42	0.02	0.24	0.02	0.11
1000	0.44	0.56	0.20	0.48	0.03	0.18
2000	0.63	0.73	0.43	0.60	0.09	0.42
5000	0.83	0.88	0.62	0.70	0.59	0.81
10000	0.92	0.88	0.75	0.89	0.83	0.91
20000	0.97	1.00	0.89	0.96	0.92	0.98
30000	0.99	1.00	0.95	0.98	1.00	1.00

Table 4 *P2 Training results*

*a) Results based on DDMRP OH**

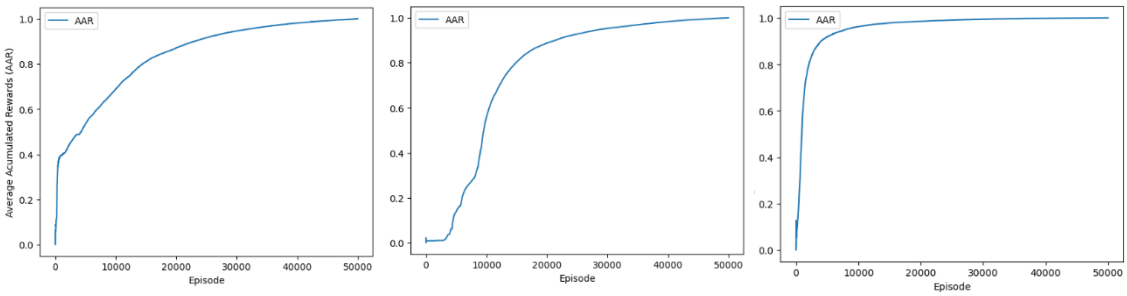
PRODUCT	REWARD FUNCTION	AAR	AOHD	BS	REL
P2	R1	**0.71**	24	**1**	1.47
	R2	0.14	**17**	14	**1.38**
	R3	0.23	50	**1**	**1.38**
	DDMRP	N/A	30	**1**	1.42

*b) Results based on our proposed RL OH**

PRODUCT	REWARD FUNCTION	AAR	AOHD	BS	REL
P2	R1P	**0.71**	20	4	**1.38**
	R2P	0.15	**17**	**0**	**1.38**
	R3P	0.51	34	1	1.8

***Table 5** P2 Test results*

*a) Results based on DDMRP OH**



*b) Results based on our proposed RL OH**

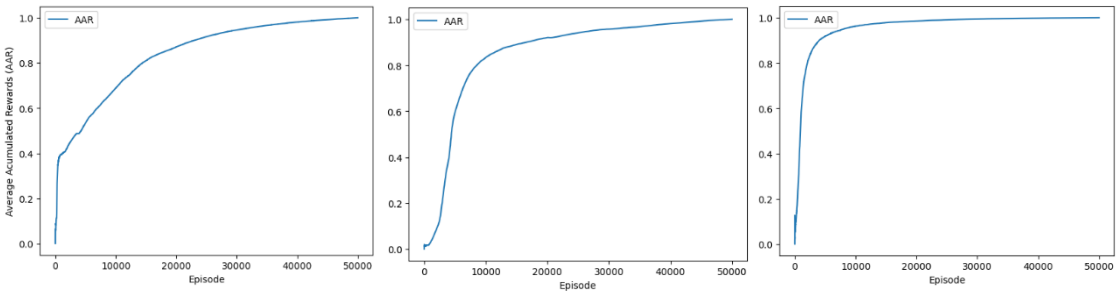


***Figure 9** P2 Convergence of models*

Figure 10 *Model results behavior based on DDMRP´s OH* - P2 Case*

Figure 11 *Model results behavior based on our proposed RL OH* behavior- P2 Case*

5.4.3 P3: Product 28440

With respect to Table 6, to analyze the performance of the DDMRP´s OH*, the model with the fastest learning process was R3 (P3R3 and P3R3P). It reaches a PBAR of 39% in 100 episodes and 97% in 10,000 episodes. In Figure 12 is observed that models based on DDMRP's OH* (P3R1-P3R2-P3R3) and based on our proposed RL OH*(P3R1P-P3R2P-P3R3P) converge properly, evidencing a successful learning process by all the algorithms.

The test results on the metric AAR for P3, are shown in Table 7. The model with the best performance is R1, for both, DDMRP´s OH* (Table 7.a) and our proposed RL OH* (Table 7.b). This means that R1 accumulated more rewards on average, indeed, it behaves better in terms of being at the desirable level (yellow). Particularly, in Table 7.a the AAR of R3 is 0.82, meaning that it accumulated 0.82 rewards per day on average.

With respect to the metric AOHD, the best results are obtained by models R1 and R2 of our proposed RL OH*. It can be seen that these models had a closeness to the OH* of 23 and 17 units respectively (the visual behavior of AOHD can be seen in figures 13 and 14). In relation to the BS metric, it can be observed that none of the models performs any number of stockouts, in this sense, all the models behave well since at no time, they run out of inventory.

Finally, in terms of REL, in Table 7 the worst performance was obtained by DDMRP´s purchase order policy, with a ratio of 2.13, meaning it was affected the most in terms of bullwhip effect ratio.

*a) Results based on DDMRP OH**

	P3R1		P3R2		P3R3	
Episodes	**AAR**	**PBAR**	**AAR**	**PBAR**	**AAR**	**PBAR**
100	0.09	0.23	0.01	0.04	0.04	0.39
200	0.13	0.25	0.01	0.04	0.04	0.39
500	0.25	0.42	0.02	0.24	0.04	0.39
1000	0.44	0.56	0.20	0.48	0.09	0.47
2000	0.63	0.73	0.43	0.60	0.27	0.72
5000	0.83	0.88	0.62	0.70	0.57	0.89
10000	0.92	0.88	0.75	0.89	0.76	0.97
20000	0.97	1.00	0.89	0.96	0.94	1.00
30000	0.99	1.00	0.95	0.98	0.98	1.00

*b) Results based on our proposed RL OH**

	P3R1P		P3R2P		P3R3P	
Episodes	**AAR**	**PBAR**	**AAR**	**PBAR**	**AAR**	**PBAR**
100	0.01	0.11	0.02	0.08	0.04	0.39
200	0.01	0.13	0.02	0.08	0.04	0.39
500	0.07	0.31	0.02	0.09	0.04	0.39
1000	0.20	0.34	0.05	0.39	0.09	0.47
2000	0.39	0.56	0.20	0.52	0.27	0.72
5000	0.70	0.69	0.50	0.63	0.57	0.89
10000	0.84	0.83	0.73	0.84	0.76	0.97
20000	0.95	0.90	0.92	0.91	0.94	1.00
30000	1.00	1.00	1.00	1.00	0.98	1.00

Table 6 *P3 Training results*

*a) Results based on DDMRP OH**

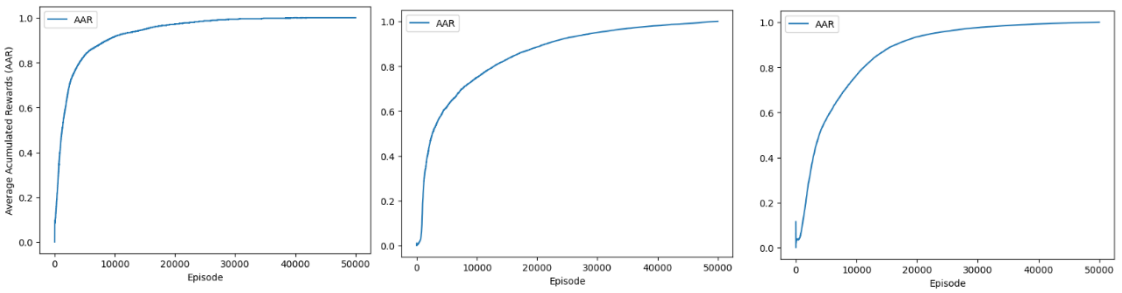
PRODUCT	REWARD FUNCTION	AAR	AOHD	BS	REL
P3	R1	**0.82**	**50**	0	0.96
	R2	0.48	67	0	**0.86**
	R3	0.48	67	0	**0.86**
	DDMRP	N/A	68	0	2.13

*b) Results based on our proposed RL OH**

PRODUCT	REWARD FUNCTION	AAR	AOHD	BS	REL
P3	R1P	**0.84**	23	0	**0.96**
	R2P	0.26	**17**	0	1.05
	R3P	0.48	67	0	0.86

Table 7 *P3 Test results*

*a) Results based on DDMRP OH**



*b) Results based on our proposed RL OH**

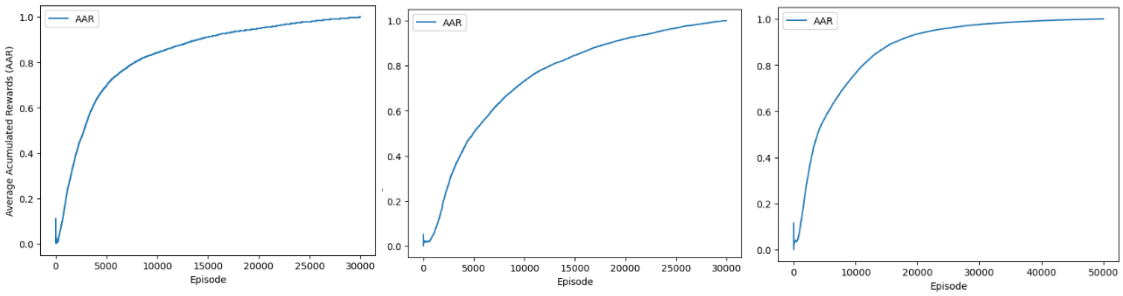


***Figure 12**: P3 Convergence of models*

***Figure 13** Model results behavior based on DDMRP´s OH* - P3 Case*

Figure 14 *Model results behavior based on our proposed RL OH* - P3 Case*

5.4.4 P4: Product 43387

In Table, for the training performance of the DDMRP´s OH* the model with the fastest learning process was R3 (P4R3 and P4R3P). Note that the PBAR reaches 88% in 100 episodes. The performance of our proposed RL OH* is very similar. In Figure 15 is observed that models based on DDMRP's OH* (P4R1-P4R2-P4R3) and based on our proposed RL OH*(P4R1P-P4R2P-P4R3P) converge properly, evidencing a successful learning process by all the algorithms.

Table 9 shows the general result for P4. The test results on the metric AAR show that the model with the best performances is R1 for DDMRP´s OH* (Table 9.a) and R3 for our proposed RL OH* (Table 9.b). The highest results obtained by these models mean that they accumulated more rewards on average. Particularly, in Table 9 the AAR of model R1 is 0.85, meaning that it accumulated 0.85 rewards per day on average.

Continuing with the metric AOHD, the best results are obtained by model R2 in the case of DDMRP´s OH*. In general, in this case study, all the models outperformed DDMRP´s ordering policy, this model (DDMRP) has the furthest distance to OH* (6 units in median) (the visual behavior of AOHD can be seen in figures 16 and 17). In relation to the BS metric, there´s a significant improvement between our proposed RL OH* and DDMRP´s OH*, this is due to the decrease in the number of stockouts in the test period. Note in Table 9 that with our proposed RL OH* only model R3 has one stockout.

Finally in terms of REL, in Table 9 is shown that the worst performance was obtained by DDMRP´s purchase order policy, with a ratio of 1.86, meaning it was affected the most in terms of bullwhip effect ratio, note that the closer it is to value one, the better.

*a) Results based on DDMRP OH**

	P4R1		P4R2		P4R3	
Episodes	**AAR**	**PBAR**	**AAR**	**PBAR**	**AAR**	**PBAR**
100	0.56	0.78	0.07	0.29	0.98	0.88
200	0.76	0.84	0.11	0.34	0.99	0.96
500	0.92	0.94	0.24	0.70	0.98	0.96
1000	0.98	0.98	0.56	0.87	0.97	0.96
2000	0.99	0.98	0.79	0.92	0.97	0.96
5000	1.00	0.98	0.92	0.94	0.96	0.98
10000	1.00	0.98	0.96	0.99	0.96	0.98
20000	0.99	1.00	0.99	1.00	0.96	1.00
30000	0.99	1.00	0.99	1.00	0.96	1.00

*b) Results based on our proposed RL OH**

	P4R1P		P4R2P		P4R3P	
Episodes	**AAR**	**PBAR**	**AAR**	**PBAR**	**AAR**	**PBAR**
100	0.56	0.78	0.64	0.64	0.73	0.80
200	0.76	0.84	0.80	0.70	0.88	0.82
500	0.92	0.94	0.91	0.85	0.97	0.88
1000	0.98	0.98	0.96	0.87	0.98	0.88
2000	0.99	0.98	0.98	0.90	0.99	0.92
5000	1.00	0.98	0.99	0.92	1.00	0.92
10000	1.00	0.98	1.00	0.92	1.00	0.95
20000	0.99	1.00	1.00	0.93	1.00	0.95
30000	0.99	1.00	1.00	1.00	1.00	1.00

Table 8 *P4 Training results*

a) Results based on DDMRP OH*

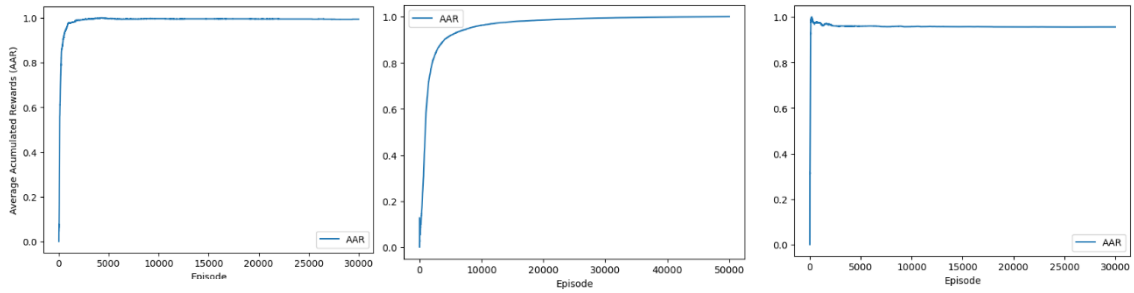
PRODUCT	REWARD FUNCTION	AAR	AOHD	BS	REL
P4	R1	**0.85**	4	5	1.25
	R2	0.52	**3**	14	1.54
	R3	0.82	4	5	**1.21**
	DDMRP	N/A	6	**2**	1.86

b) Results based on our proposed RL OH*

PRODUCT	REWARD FUNCTION	AAR	AOHD	BS	REL
P4	R1P	0.66	**1.5**	**0**	**1.44**
	R2P	0.66	**1.5**	**0**	**1.44**
	R3P	**0.90**	**1.5**	1	1.47

Table 9 *P4 Test results*

a) Results based on DDMRP OH*



b) Results based on our proposed RL OH*

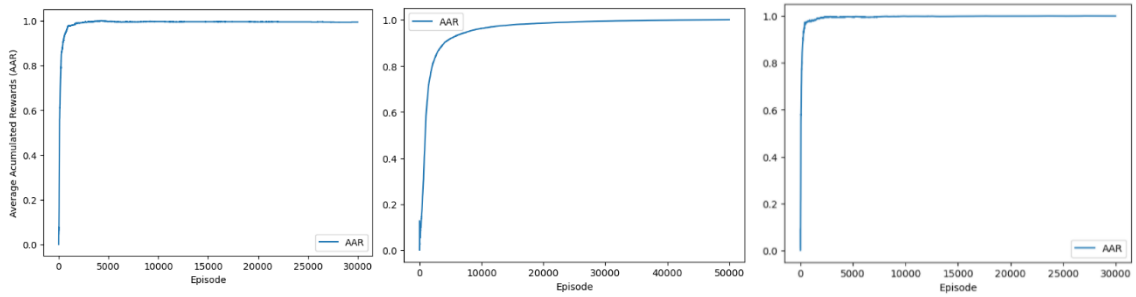


Figure 15 *P4 Convergence of models*

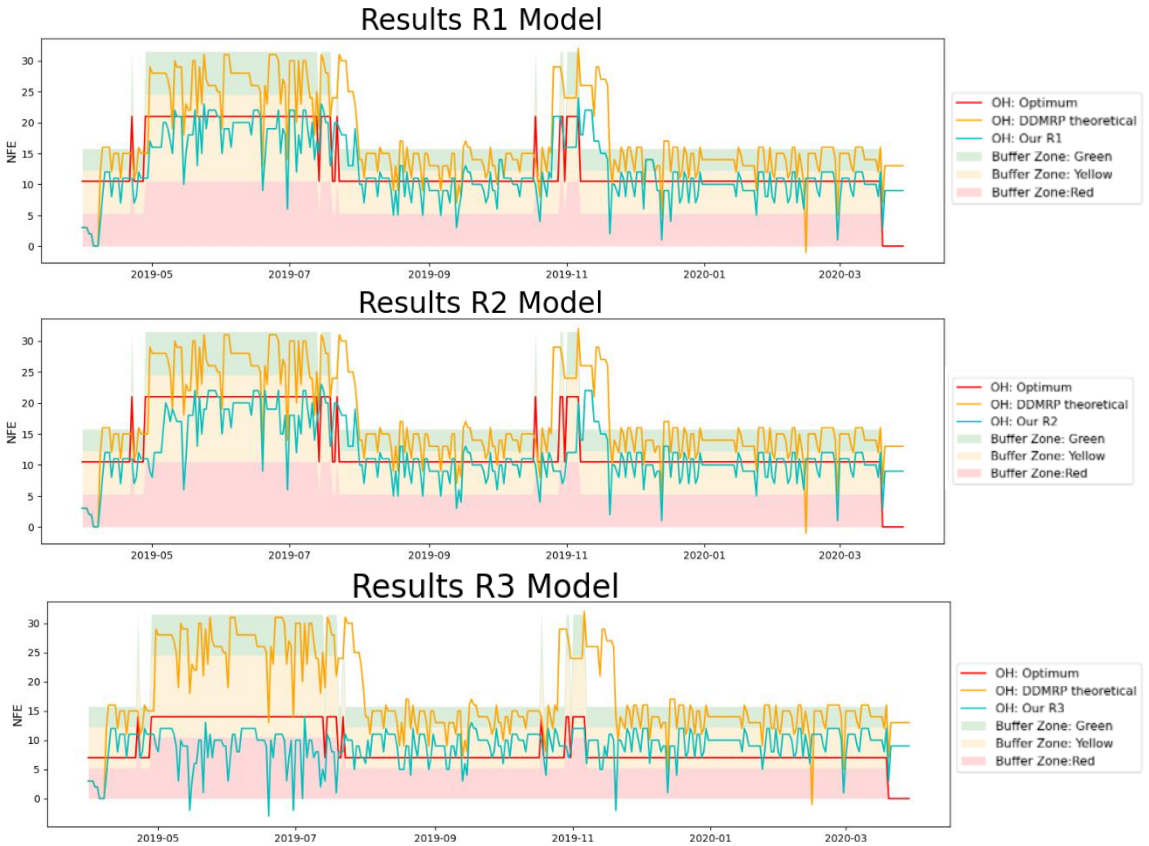


Figure 16 *Model results behavior based on DDMRP´s OH* - P4 Case*

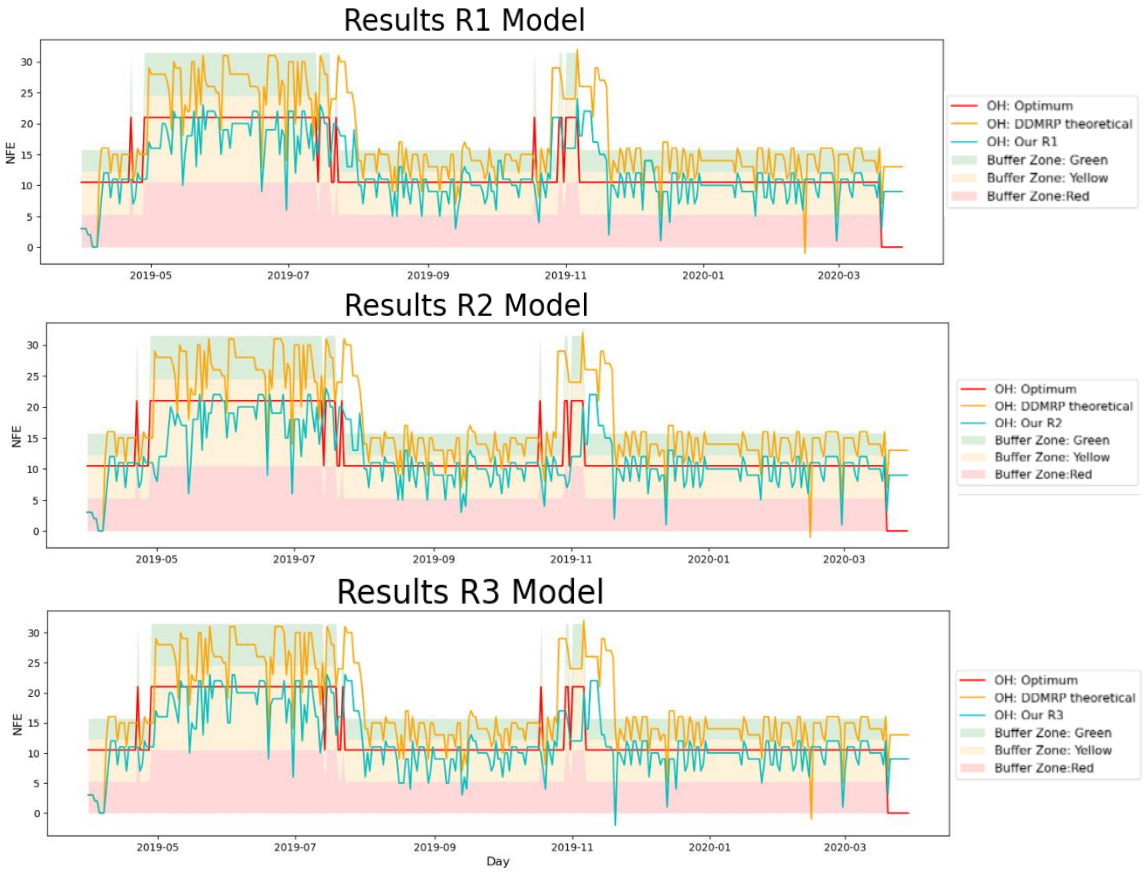


Figure 17 *Model results behavior based on our proposed RL OH* - P4 Case*

In Table 10, we compare the learning process performance, on average, between each of the reward function models that we proposed. In this table, we can see that the reward function that has the best learning performance is R3. In other words, R3 is the function that learns the fastest, in average, across all the case studies.

	PBAR		
Episodes	R1	R2	R3
100	0.55	0.31	0.60
200	0.57	0.34	0.61
500	0.72	0.53	0.79
1000	0.77	0.67	0.88
2000	0.86	0.73	0.92
5000	0.93	0.89	0.93
10000	0.95	0.93	0.98
20000	1.00	1.00	1.00
30000	1.00	1.00	1.00

Table 10 *Summarized learning performance of our proposed models*

In table 11 is summarized the results of the BS metric. In this table is shown how the performance of the algorithms is improved with our OH *. Note that in the P2 scenario the best performance is given with R2 with no inventory break. In the case of P4 the same happens, again, our OH* outperforms the BS metric compared to DDMRP's OH*. Note that the worst of the results obtained was a breakout while that the worst DDMRP's OH* scenario was 14 days of breakouts. Now, in scenarios P1 and P3 it can be observed that the performance is the same in both cases; however, in general, it can be observed that our proposed RL OH* has a higher performance than that proposed by the DDMRP theory. The result is highly significant given that in this way the risks of lost sales and customer satisfaction are being significantly reduced. Finally, in relation to the analysis of our proposed models, the best performance was obtained with R2 based on our proposed RL OH*. Note that in each of the study cases it was better or equal to the best result.

	P1		**P2**		**P3**		**P4**	
	DDMRP´s OH*	Our OH*	DDMRP´s OH*	Our OH*	DDMRP´s OH*	Our OH*	DDMRP´s OH*	Our OH*
R1	0	0	1	4	0	0	5	**0**
R2	0	0	14	**0**	0	0	14	**0**
R3	0	0	1	1	0	0	5	1
DDMRP	0		1		0		2	

Table 11 *BS results comparison*

In table 12 is summarized the results of the REL metric. it can be observed that even though in the theoretical scenario, the performance of the DDMRP's OH* is significantly higher; however, in the cases of real studies (P2-P3-P4), this difference is not so significant. Particularly, in the P2 case the best results were a ratio of 1.38, obtained by R2 and R3 of DDMRP's OH*, and it was also the same result for R1 and R2 of our proposed RL OH*. In the P3 scenario, the results of our OH* are also improved. Note that although the best result for both is given with R1 with a ratio of 0.96, R2 improves significantly, going from 0.86 to 1.05. Finally, in the P4 scenario, it can be observed that the performance of the DDMRP's OH* is higher than our OH* for R1 and R3 models; however, in R2 it improves. Note that the closer this value is to one, the less it is affected by the whip effect.

In relation to the performance of the proposed models, it can be observed that in this particular metric it was very diverse, R2 and R3 were the best in P1, R2 was the best in P2, R1 in P3 and R3 in P4. Observe that although they were diverse, in general, all the models proposed outperformed the order purchase model defined in the theory (DDMRP).

	P1		**P2**		**P3**		**P4**	
	DDMRP´s OH*	Our OH*	DDMRP´s OH*	Our OH*	DDMRP´s OH*	Our OH*	DDMRP´s OH*	Our OH*
R1	8.96	27.31	1.47	1.38	**0.96**	**0.96**	1.25	1.44
R2	**5.82**	25.42	1.38	1.38	0.86	1.05	1.54	1.44
R3	**5.82**	25.42	1.38	1.80	0.86	0.86	**1.21**	1.47
DDMRP	10.05		1.42		2.13		1.86	

Table 12 *REL results comparison*

In general, our proposed OH* in most cases has a better performance compared to the OH* defined by the DDMRP theory (see Eq. (3)), being significantly better in contexts of demand with high variability. This is given that it avoids the breakdown of inventory and is more robust against the bullwhip effect. Additionally, according to Tables 11 and 12, our proposed models, in

general, also outperformed in terms of efficiency, the purchase orders policy defined in the DDMRP theory. Now, to define which of the models is better logistically, for our criteria it is the R2 model. The above given the superiority in terms of BS and the good performance obtained in REL. We recommend this model (R2) even though it was not the most efficient in terms of time required in the learning process (see Table 10). Although a policy that has better performance in results is learned, it is not the fastest in the learning process.

However, if the case study has a high level of complexity or computational limitations, we recommend using R3 since it obtains good results in terms of learning (see Table 10) and in terms of results (see Table 11 and 12). The selection of the best model must be a tradeoff between whether what is sought is efficiency in terms of learning or performance of results.

5.5 Comparison with other works

The comparison of our proposal with other studies was carried out in relation to the following 3 comparison criteria:

- Technique: the techniques used.
- Bullwhip effect: it evaluates if the proposed model has a strategy to avoid distortions associated with the bullwhip effect.
- Adaptability: it evaluates if the proposed method in the article can be applied in demanding scenarios with different seasonal and trend behaviors.

Paper	Techniques	Bullwhip effect	Adaptability
Ours	DDMRP and *Q Learning*	Yes	High
Paraschos et al. (2020)	Q Learning.	No	Medium
Kara and Dogan (2018)	*Q-Learning* y Sarsa	NO	Medium
Wang et al. (2020).	Economic Order Quantity (EOQ), Optimization	NO	Low
Karimi et al. (2017)	Q Learning	NO	Low
Giannoccaro and Pontrandolfo (2002).	Q Learning	NO	Medium

Table 13 *Comparison with other works.*

Paraschos et al. (2020) and Kara and Dogan (2018) propose an inventory management system that allows to optimally evaluate the tradeoff between cost (associated with equipment failures) and benefit. Wang et al. (2020) develop an order generation system based on price discount

strategies. Giannoccaro and Pontrandolfo (2002) develop an inventory management system that allows making decisions in relation to supply, production, and distribution. Wang et al (2020) develop an optimal replenishment and stocking strategy based on price discounts of the supplier. Finally, Karimi et al (2017) propose a model to optimize the trade-off between productivity and the level of knowledge of the human resource of a production company to maximize the expected profit.

Based on Table 13, our proposal differs from the rest of the articles because is the only one that proposes a model that avoids the distortions provided by the bullwhip effect in the supply chain. Particularly Giannoccaro, I., & Pontrandolfo, P. (2002), conclude that their proposed model can adapt to "slight changes of demand", similarly Kara and Dogan (2018), Karimi et al. (2017), Paraschos et al. (2020) showed evidence that their given models can adapt to uncertain demand but none evaluated the bullwhip effect. Finally, Wang et al. (2020) assumed constant demand for their proposed model, being this a very strong assumption and far from reality.

In relation to adaptability, the articles propose solutions for a specific process or business sector. Particularly, Wang et al. (2020) propose a model for a business that has specific pricing policies by their suppliers. Karimi et al. (2017) develop a model for a human resource planning area with specific variables that could not be replicable to other businesses. Similarly, Paraschos et al. (2020) develop a quality control model for detecting failures, Kara and Dogan (2018) for perishable products. Finally, Giannoccaro and Pontrandolfo (2002), although their model can be replicated in multiple business sectors, it is not so clear in the work how it can be used in other contexts.

6. Conclusion

This article implements a hybrid reinforcement learning algorithm based on the DDMRP theory and RF algorithms for inventory management that allows a more efficient ordering process. Additionally, we develop an alternative optimal inventory level function that outperforms the function defined by DDMRP. This was concluded by comparing the performance of the algorithm in scenarios with different characteristics, performing adequately difficult case studies with a diversity of situations, such as scenarios with discontinuous or continuous demand, seasonal and non-seasonal behavior, with high demand peaks, multiple lead times, among others.

The results obtained in relation to the model with the best performance was R2. It provides a balanced purchasing policy that optimizes the distance to optimal inventory, REL and minimizes stockouts. Note that although this was the best model, the other models proposed in our case studies were also promising as they were in general terms more efficient in terms of purchase orders than the model proposed by DDMRP.

In terms of Inventory level, we show that in cases like P4 and P2, where the level is too close to zero, the inventory can be broken multiple times as the variability of the units demanded changes. In the results, there's evidence that our proposed inventory level significantly reduces the number of occurrences, which can avoid the associated risks and costs. Continuing with the results of the REL ratio, the results show that in the case studies our models outperformed the model of the DDMRP´s theory. This, our models are more robust and less affected by the bullwhip effect.

In terms of learning performance, it was shown that in general, the most efficient model is R1 and the least efficient R2. Depending on the computational resources available, one model may

be more suitable than another. In our case studies, R2 adapted well to the resources, and it was possible to take advantage of its good results in the evaluation metrics.

For future work, it is proposed to build inventory management systems based on the SARSA and Deep Q Network reinforcement learning algorithms. The SARSA model is proposed with the objective of comparing the effect that an on-policy type (as it is) to the one used in this article (off policy). The online policy could lead to better learning process performances. On the other hand, the model based on Deep Q Network is proposed since the neural networks replace the Q table, which in practice can translate into an increase in the performance of the learning process since it is not based on a predefined discrete space (Q table). Finally, for future works, it will be explored an alternative exploitation-exploration policy that reduces the exploration rate over time with the objective of increasing the efficiency in the learning times of the model. With the current exploitation-exploration policy, it continues exploring at the same rate from the start to the end of the episode.

References

Bonato, V., Mazzotti, B., Fernandes, M., & Marques, E. (2013). A Mersenne Twister Hardware Implementation for the Monte Carlo Localization Algorithm. Journal of Signal Processing Systems for Signal, Image & Video Technology, 70(1), 75–85.

Costantino, F., Gravio, G.D., Shaban, A., & Tronci, M. (2013). Exploring the Bullwhip Effect and Inventory Stability in a Seasonal Supply Chain. International Journal of Engineering Business Management, 5.

Durán, Y. (2012). Administración del inventario: elemento clave para la optimización de las utilidades en las empresas. Visión Gerencial, (1), 55-78.

Edward A. Silver. (1981). Operations Research in Inventory Management: A Review and Critique. Operations Research, 29(4), 628–645.

Giannoccaro, I., & Pontrandolfo, P. (2002). Inventory management in supply chains: a reinforcement learning approach. International Journal of Production Economics, 78(2), 153–161. https://doi-org.ezproxy.eafit.edu.co/10.1016/S0925-5273(00)00156-0

Huang, J., Chang, Q., & Arinez, J. (2020). Deep reinforcement learning based preventive maintenance policy for serial production lines. Expert Systems With Applications, 160. https://doi-org.ezproxy.eafit.edu.co/10.1016/j.eswa.2020.113701

Hubbs, C. D., Li, C., Sahinidis, N. V., Grossmann, I. E., & Wassick, J. M. (2020). A deep reinforcement learning approach for chemical production scheduling. Computers and Chemical Engineering, 141. https://doi-org.ezproxy.eafit.edu.co/10.1016/j.compchemeng.2020.106982

Huq, Z., Huq, F., 1994. Embedding JIT in MRP: The case of job shops. Journal of Manufacturing Systems 13 (3), 153-164.

Kara, A., & Dogan, I. (2018). Reinforcement learning approaches for specifying ordering policies of perishable inventory systems. Expert Systems with Applications, 91, 150–158. https://doi-org.ezproxy.eafit.edu.co/10.1016/j.eswa.2017.08.046

Karimi-Majd, A.-M., Mahootchi, M., & Zakery, A. (2017). A reinforcement learning methodology for a human resource planning problem considering knowledge-based promotion. Simulation Modelling Practice and Theory, 79, 87–99. https://doi-org.ezproxy.eafit.edu.co/10.1016/j.simpat.2015.07.004

Kortabarria, A., Apaolaza, U., Lizarralde, A., & Amorrortu, I. (2018). Material management without forecasting: From MRP to demand driven MRP. Journal of Industrial Engineering and Management, 11(4), 632-650.

Lee, C.-J., & Rim, S.-C. (2019). A Mathematical Safety Stock Model for DDMRP Inventory Replenishment. Mathematical Problems in Engineering, 1–10. https://doi.org/10.1155/2019/6496309.

Mather, H. 1977. "Reschedule the Reschedules You Just Rescheduled – Way of Life for MRP?" Production and Inventory Management 18 (1): 60–79.

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Transactions on Modeling and Computer Simulation, 8(1), 3–30.

Merrad, Y., Habaebi, M. H., Islam, M. R., & Gunawan, T. S. (2020). A real-time mobile notification system for inventory stock out detection using SIFT and RANSAC. International Journal of Interactive Mobile Technologies, 14(5), 32–46.

Muller, M. (2011). Essentials of Inventory Management. AMACOM. Ng, A. Y., Harada, D., & Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward *Shaping*. In ICML (Vol. 99, pp.
278-287).

Orlicky, J. A. (1975). Material requirements planning: The new way of life in production and inventory management. McGraw-Hill.

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.

Peterson, R., Silver, E. A., & Pyke, D. F. (1998). Inventory Management and Production Planning and Scheduling (3rd ed.) JOHN WILEY & SONS.

Pooya, A., Fakhlaei, N., & Alizadeh-Zoeram, A. (2021). Designing a dynamic model to evaluate lot-sizing policies in different scenarios of demand and lead times in order to reduce the nervousness of the MRP system. Journal of Industrial & Production Engineering, 38(2), 122–136.

Ptak, C.A., & Smith, C (2011) Orlicky's Material Requirements Planning, McGraw Hill.

Ptak, C.A., & Smith, C. (2016). Demand driven material requirements planning (DDMRP), Industrial Press INC.

Romero Rodríguez, D., Aguirre Acosta, R., Polo Obregón, S., Sierra Altamiranda, Á., & Daza-Escorcia, J. M. (2016). Medicion del efecto latigo en redes de suministro. Revista Ingeniare, (20), 13+.

Shofa, M.J., Moeis, A.O., & Restiana, N. (2018). Effective production planning for purchased part under long lead-time and uncertain demand: MRP Vs demand-driven MRP. IOP Conference Series: Materials Science and Engineering, 337.

Silver, E. A., Pyke, D. F., & Thomas, D. J. (2017). Inventory and Production Management in Supply Chains: Fourth Edition. CRC Press.

Skinner, B. F. (1958). "Reinforcement today". American Psychologist, 13(3):94–99.

Steele, D. 1975. “The nervous MRP System: How to do battle.” Production and Inventory Management 16 (4): 83–89.

Sutton, R. (1988). Learning to predict by the methods of temporal differences. Machine Learning, 3(1), 9–44.

Sutton, R. S., Barto, A. G. (2018). Reinforcement Learning: An Introduction. MIT Press.

Velasco Acosta, A. P., Mascle, C., & Baptiste, P. (2020). Applicability of Demand-Driven MRP in a complex manufacturing environment. International Journal of
Production Research, 58(14), 4233–4245.

Wang, Y., Xing, W., & Gao, H. (2020). Optimal ordering policy for inventory mechanism with a stochastic short-term price discount. Journal of Industrial & Management Optimization, 16(3), 1187–1202. https://doi-org.ezproxy.eafit.edu.co/10.3934/jimo.2018199

Watkins, Christopher. (1989). Learning From Delayed Rewards. Doctoral Thesis, King’s College.

Watkins, C. J. C. H., Dayan, P. (1992). Q learning. Machine Learning, 8:279-292.

Wemmerlov, U. 1979. Design factors in MRP systems: A limited survey. Production and Inventory Management 20 (4): 15–35