sciendo

PRODUCTION ENGINEERING ARCHIVES 2024, 30(3), 377-393

PRODUCTION ENGINEERING ARCHIVES

ISSN 2353-5156 (print)
ISSN 2353-7779 (online)

Exist since 4th quarter 2013
Available online at https://pea-journal.eu

A Model for Decision-making to Parameterizing Demand Driven Material Requirement Planning Using Deep Reinforcement Learning

Mustapha El Marzougui[1*], Najat Messaoudi[1], Wafaa Dachry[1,2], Bahloul Bensassi[1]

[1] Laboratory of Electrical and Industrial Engineering, Information Processing, Computing and Logistics, Faculty of Sciences Ain Chock, Hassan II University, Casablanca, Morocco; mustapha.elmarzougui@gmail.com (MEM); najatm2013@gmail.com (NM); bahloul_bensassi@yahoo.fr (BB)
[2] Laboratory of Engineering, Industrial Management and Innovation, FST, Hassan I University, Settat, Morocco; wafaa.dachry@uhp.ac.ma
*Correspondence: mustapha.elmarzougui@gmail.com

Article history
Received 20.12.2023
Accepted 10.08.2024
Available online 09.09.2024

Keywords
DDMRP,
Deep Reinforcement Learning,
Parameterization,
on-hand inventory,
on-time delivery.

Abstract
Demand-Driven Material Requirements Planning (DDMRP) is an emerging inventory management approach that has garnered significant attention from academia and industry. Numerous recent studies have highlighted the advantages of DDMRP compared to traditional methods such as material requirement planning (MRP), Theory of constraint (TOC), and Kanban. However, the performance of DDMRP relies on several parameters that affect its effectiveness. Parameterization models and the optimization of control variables have significantly contributed to the field of inventory management and have proven to be effective and practical in addressing challenges by providing a structured approach to handling complex variables and constraints. This paper introduces an innovative parameterization model that leverages deep reinforcement learning (DRL) to parameterize a DDMRP system in the face of uncertain demand. The main objective is to dynamically determine the optimal values for the variability and lead time factors within the DDMRP framework, to maximize customer service levels and optimize inventory efficiency. The results of this study emphasize the effectiveness of DRL as an automated decision-making approach for controlling DDMRP parameters. Additionally, the findings highlight the potential for enhancing the performance of the DDMRP approach, particularly in terms of on-time delivery (OTD) and average on-hand inventory (AOHI) by adjusting the variability and lead-time factors.

DOI: 10.30657/pea.2024.30.37

1. Introduction

The production environment has become increasingly complex and dynamic in today's rapidly changing industrial landscape. This complexity arises from global market fluctuations, rapid technological advancements, changing customer demands, and increasing competition. The intricate and dynamic production environment is commonly described by the acronym VUCA (Volatility, Uncertainty, Complexity, and Ambiguity; Bennett and Lemoine, 2014). The continuous evolution of these elements poses significant challenges for companies striving to maintain high performance and operational efficiency. The complexity of the production environment introduces various uncertainties and risks that can disrupt supply chains, impact production schedules, and affect customer satisfaction. Companies must navigate through unpredictable

market conditions, fluctuations in demand patterns, and intricate supply networks to ensure the timely delivery of products while optimizing inventory and maintaining high service levels.

As a response to these evolving challenges, the concept of DDMRP has emerged. DDMRP is an efficient inventory management. This system effectively addresses the issue of the bullwhip effect by strategically placing buffers along the supply chain (Ptak and Smith, 2016). DDMRP utilizes buffers, also known as decoupling points, to establish separation between the supply chain, material usage, and demand (Ptak and Smith, 2016). The purpose of these zones is to maintain optimal inventory levels at the decoupling points, ensuring that any fluctuations within the system do not propagate throughout the entire supply chain (Ptak and Smith, 2016). Based on a comprehensive literature review, multiple authors (Marzougui et al., 2021; Azzamouri et al., 2021; Benavente et al., 2023) have provided evidence of DDMRP's effectiveness across several key performance indicators (KPI). These indicators include service level or on-time delivery, inventory sizing measured through average on-hand inventory, and cost management (Miclo, 2016; Kortabarria et al., 2018; Velasco Acosta et al., 2020).

Nevertheless, this methodology faces a significant limitation as it relies on several parameters determined by the manager's judgment (Damand et al., 2022; Lahrichi et al., 2022; Duhem et al., 2023), which are based on rules defined by the authors of DDMRP (Ptak and Smith, 2016) and expressed as intervals. Such subjective selection of factors can lead to excessively large deviations, resulting in naturally inconsistent performance (Lee and Rim, 2019). Consequently, even slight deviations in the choice of factor values result in the buffer size effect within the supply chain. This effect causes distortions that compromise the integrity of the buffers. Indeed, parameterization plays a critical role in the dynamic adjustment component of DDMRP (Duhem et al., 2023), and it greatly impacts the performance of the system (Damand et al., 2022). The selection and configuration of DDMRP parameters have a tangible influence on the overall effectiveness of the methodology (Lahrichi et al., 2022), and these parameters must be continuously adjusted to effectively adapt to the actual demand (Duhem et al., 2023).

In ((El Marzougui et al., 2022a) ;(El Marzougui et al., 2022b) ; (El Marzougui et al., 2023)) research, they have made significant theoretical contributions by examining the feasibility and complementarity of integrating industry 4.0 technologies with the DDMRP system, highlighting the positive effects of integrating these technologies on various parameters of DDMRP and confirming the beneficial impact of artificial intelligence on DDMRP components.

In line with this context, (Boute et al., 2021) considers deep reinforcement learning as a valuable complementary tool in data-driven inventory control. This approach empowers the transformation of data into effective decision-making processes, further enhancing the overall performance of inventory management systems. To the best of our knowledge, there is only one article (Duhem et al., 2023) that suggests a reinforcement learning model in a DDOM (Demand-Driven Order Management) to parameterize two specific parameters, namely order spike threshold (OST) and order spike horizon (OSH). However, limited emphasis has been placed on determining DDMRP parameters such as variability factor (F_V) and lead-time factor (F_{LT}).

Therefore, this article introduces a decision-making approach that integrates DRL with the DDMRP methodology. The objective is to accurately determine the values for the variability factor and lead time factor, rather than providing them as intervals (Ptak and Smith, 2016). In this research, we propose a learning model based on the DRL algorithm to optimize DDMRP parameters. The application of a DRL algorithm in this context is motivated by the positive outcomes achieved through its utilization of various inventory optimization concepts (Cuartas and Aguilar, 2023; Duhem et al., 2023). DRL is an effective machine learning approach for solving problems in which an agent possesses no prior knowledge about the environment. Enhancing decision-making efficiency in supply chain operations is a critical competitive realm within today's uncertain business environments (Esteso et al., 2022).

Hence, the main contribution of this work is the implementation of a deep reinforcement learning model that dynamically sets more efficient parameters for the variability factor and lead time factor in an environment of unknown demand, improving performance metrics such as on-hand inventory and on-time delivery.

The paper is structured into seven sections. Section 2 provides an overview of fundamentals and includes a literature review. Section 3 focuses on describing the problem description, assumptions, and modelling. Section 4 presents the learning-optimization approach. Section 5 details the experimental design. In Section 6, the results and discussion are presented. The last section of the paper concludes with concluding remarks and highlights future research directions.

2. Fundamentals and Literature review

In this section, we begin by introducing the DDMRP component and providing a description of buffer sizing and order generation. Additionally, we present a literature review focused on the parameterization of the DDMRP approach. Next, we explore the relevant literature on reinforcement learning within the context of our specific scenario. Finally, we highlight the contribution of this article

2.1. DDMRP Model

Ptak and Smith, (2016) define the DDMRP process through five consecutive steps. The first step of the DDMRP model is to determine which parts should be buffered called decoupling points. The function of these buffers is to create independence between demand and supply in the supply chain. Hence, a new concept generated by this step is decoupled lead time (DLT). DLT is defined as a qualified cumulative lead time defined as the longest unprotected/ un-buffered sequence in a bill of material'. In other words, the 'DLT is calculated as the summation of the lead time on that longest unprotected sequence'. Once buffers are placed, the second step is sizing.. Ptak and

Smith (2016) recommend dividing the buffer into three zones, each with distinct sizes and purposes: green, yellow, and red. These buffers must be adapted and adjusted for a dynamic environment. The third step deals with the dynamic adjustment of buffer size. This latter takes into account some factors that vary by industry and company such as trends and seasonality, new products, new markets, etc.. (Ptak and Smith, 2016).

As described by (Ptak and Smith, 2016), the last two steps of DDMRP are defined as follows:

The fourth step involves the demand-driven planning phase. In this step, the decision maker decides if a replenishment order must be generated or not for a DDMRP buffer and determines the quantity to be ordered based on the value of the net flow equation.

The last step is visible and collaborative execution, where DDMRP calculates a buffer execution status based on the current stock instead of the net flow equation.

Based on (Ptak and Smith, 2016), this section intended to develop two components of the DDMRP model buffer sizing and order generation.

2.1.1. Buffer sizing

Buffer sizing (BS) determines the amount of inventory contained in the buffers. Buffer sizing is the shock absorber to mitigate both supply and demand variability and reduce or eliminate the bullwhip effect. The size of the buffer is the summation of three calculated zones (Ptak and Smith, 2016): red, yellow and green, which will be described in this section. To simplify matters, the relevant notations utilized in the learning model for DDMRP parameter optimization problems are presented in the following Table 1:

Table 1. Notations (Authors)

Notation	Description
H	Time horizon
DLT	Decoupled Lead Time.
D(t), t ∈ H	Demand quantity on day t
S_0	Initial on-hand inventory.
MOQ	Minimum order quantity
F_V	Variability factor
F_{LT}	Lead time factor
$H_{order.}$	Desired order cycle horizon
T_{spike}	Order spike threshold.
H_{spike}.	Order spike horizon
OH t , t ∈ H	On-hand inventory at the beginning of day t.
PO t , t ∈ H	On-order inventory at the beginning of dayt.
NFE t , t ∈ H ˜	Net flow equation at the beginning of day t.
QD t, t ∈ H	Qualified demand quantity on day t.
ROt, t ∈ H	Order amount launched on day t.
IO t, t ∈ H	Incoming order amount on day t.
ADU t, t ∈ H	Average daily usage calculated on day t.
TOR t, TOY t, TOG t, t ∈ H	Respectively the top of red, the top of yellow and the top of green levels on day t
OTD	On-Time Delivery.
AOHI	Average on-hand inventory

- **The red zone (RZ)** presents the security level of the inventory in the buffer. It is calculated according to the formula below:

Red Zone = Red base zone + Red safety zone

Red base zone = ADU*DLT*F_{LT} and
Red safety zone = ADU*DLT*F_{LT}*F_V
Red Zone (RZ) = TOR (Top Of Red)

$$\mathbf{RZ = (ADU*DLT*F_{LT}) + (ADU*DLT* F_{LT}*F_V)} \quad (1)$$

- ***The yellow zone (YZ)*** encompasses the demand for the days that extend beyond the reach of a replenishment order. It is calculated according to the formula below:

*Yellow Zone = ADU*DLT*
TOY (Top Of Yellow) = TOR + YZ

$$\mathbf{YZ = ADU*DLT} \quad (2)$$

- ***The green zone (GZ)*** presents the stock level that is considered to be comfortable and no replenishment order is generated. It is calculated according to the formula below:

*Green Zone = Max {MOQ; ADU*DLT*F_{LT}; ADU*H_{order}}*

The minimum order quantity (MOQ) and the desired order cycle (Horder) are not considered in the calculation of the green zone. Hence, it will be calculated by the formula

*Green Zone = ADU*DLT*F_{LT}*
TOG (Top Of Green) = TOY + GZ

$$\mathbf{GZ = ADU*DLT*F_{LT}} \quad (3)$$

- ***Buffer size (BS):*** The buffers contain three zones as shown in Figure 1:

Buffer size = RZ + YZ + GZ
Buffer size = TOG

$$\mathbf{BS = ADU*DLT*F_{LT} + ADU*DLT*F_{LT}*F_V + ADU*DLT + ADU*DLT*F_{LT}} \quad (4)$$

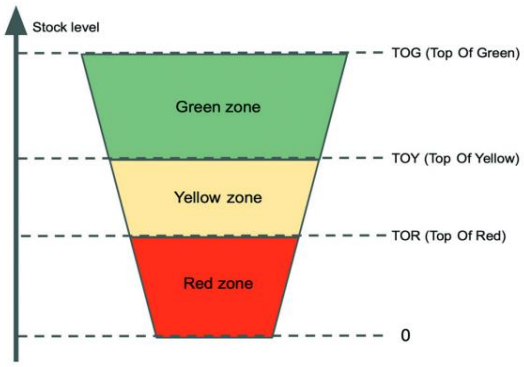

Fig. 1. DDMRP zone buffers

The determination of buffer size is influenced by two critical factors, which are under the control of the manager. These factors, namely the lead-time factor (**F_{LT}**) and the variability factor (**F_V**), are established based on the guidelines provided in Table 2 and Table 3 as referenced in (Ptak and Smith, 2016). These rules do not provide exact values of the optimal factors.

Table 2. The percentage used by F_{LT} (Ptak and Smith, 2016)

Lead time	Lead-time factor range
Long	20-40% of ADU over DLT
Medium	41-60% of ADU over DLT
Short	61-100% of ADU over DLT

Table 3. The percentage used by F_V (Ptak and Smith, 2016)

Variability	Variability factor range
High	61-100% of Red-Base
Medium	41-60% of Red –Base
Low	0-40% of Red-Base

2.1.2. Order generation/ Buffer's replenishment

Order generation is the process by which supply orders are generated by evaluating actual inventory, a stock that has been ordered but not received, and qualified sales order demand. In this step, the decision maker decides if a buffer replenishment order must be generated and calculates the quantity to be ordered. Supply order generation is created by the use of the net flow equation (NFE).

- **Net Flow Equation (NFE)** is a new concept of the DDMRP approach that defines an amount of available inventory that generates a signal of supply order. It is calculated according to the formula below (Ptak and Smith, 2016):

$$\textbf{Net Flow Equation (NFE)} = \textbf{OH} + \textbf{OP} - \textbf{QD} \quad (5)$$

On-hand inventory **(**OH) is inventory physically available i.e. quantity of stock available to be used.

$$\textbf{OH (Day t+ 1)} = \textbf{OH (t)} - \textbf{Demand (t)} + \textbf{Incoming Supply (t)} \quad (6)$$

Pending Orders (OP) is the quantity of stock that has been ordered but not received;

$$\textbf{OP (Day t+ 1)} = \textbf{OP (t)} - \textbf{Incoming Supply (t)} + \textbf{Order amount (t)} \quad (7)$$

- Qualified Demand (QD) is constituted by the sum of demand orders existing on the day, with the future demand considered as a spike.Whether an order is considered a spike depends on two factors: the order spike threshold (OST) and the order spike horizon (OSH). The demand orders that exceed the qualified order spike threshold level over a given order spike horizon are considered spikes.

$$\textbf{QD (day t)} = \textbf{demand (day t)} + \textbf{future demands considered as a spike} \quad (8)$$

Where:
OST: represents the maximum demand threshold for it to be considered as a demand peak which would qualify for net flow equation inclusion;
OSH: is the length of time in the future in which spikes are considered;
When the total number of orders received for an item exceeds a certain threshold on a given date, this is considered an 'order spike', as illustrated on the fourth day in Figure 2.

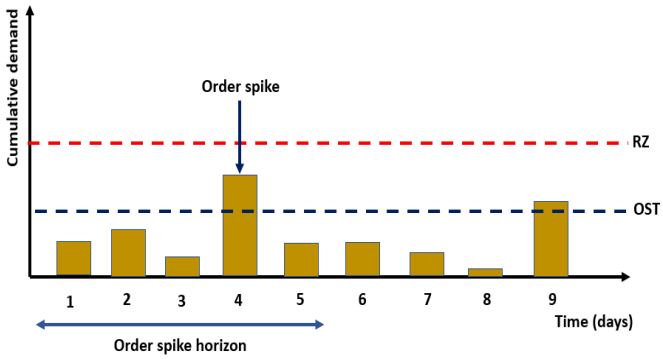

Fig 2. An example of an order spike on day 4

(Ptak and Smith, 2016) suggest three different methods for establishing the OST, this value should always be within the range of the red zone total:

- At 50 % of the TOR: OST = 50%*TOR;
- At the top of the red safe zone: OST = RBZ;
- ADU factor: OST = a*ADU

Regardless of the choice used for the determination of OST, this value must always be within the range of the total of the red zone (i.e., $\mathbf{a_{OST}} \leq 1$) to preserve the buffer's integrity (Ptak and Smith, 2016). Therefore, we can deduce that OST represents the percentage of red zone

$$\textbf{OST} = \mathbf{a_{OST}}\textbf{*ZR} \quad (9)$$

For OSH, it is recommended (Ptak and Smith, 2016) to set it to:

- At least one DLT: OSH = 1.0*DLT

Based on the net flow equation **(5)** value and top of yellow, a replenishment order (RO) is generated and calculated following these formulas equation **(10)**:

- *If* $NFE_t \leq TOY_t$ $RO_t = TOG - NFE$ (RO_t defines the quantity of the order that is to be launched)
- *Otherwise,* $RO_t = 0$ (No order is generated) **(10)**

2.1.3. Setting of DDMRP parameters / Parameterization of DDMRP

The purpose of this section is not to discuss the literature review of DDMRP in a general way but to present a specific DDMRP parameterization. Some recent comprehensive reviews can be found in studies (Marzougui et al., 2021), (Azzamouri et al., 2021), and (Benavente et al., 2023). As many authors and researchers (Bahu et al., 2019; Velasco Acosta et al., 2020) have underlined the DDMRP process depends on several parameters that affect its application and its performance. (Ptak and Smith, 2016) provide some approximate tools to fix the parameters based on the manager's experience and judgment. Several authors have highlighted the absence of automatic or semi-automatic rules to fix these parameters (Bahu et al., 2019) (Velasco Acosta et al., 2020) (Miclo, 2016) . Based on the literature review summarized in Table 4 (Appendix A), there is limited work dealing with DDMRP parameterization using some optimization methods.

The authors (Miclo, 2016) proposed an automatic optimization method utilizing the simulated annealing algorithm to determine parameter values. This metaheuristic approach was tested in the Kanban game. Furthermore, (Lee and Rim, 2019) introduced an alternative formula to tackle the issue of fixed parameters for the factor of variability and lead time. (Velasco Acosta et al., 2020) performed a statistical analysis to establish parameterization rules based on data from the case study. Similarly, in the (Martin, 2020) study, parameterization was approached statistically. The authors designed a series of experiments to test various parameterization scenarios and derived general rules based on the analyzed dataset. To gain a deeper understanding of the impact of DDMRP parameterization on KPI, specifically OH and OTD, we are interested in studies (Damand et al., 2022) and (Lahrichi et al., 2022). These studies propose automatic methods to optimize some parameters of a DDMRP with two different approaches Genetic Algorithm and Mixed Integer Linear Programming. Nevertheless, these studies are subject to certain assumptions and limitations. In the study conducted by (Damand et al., 2022), a comprehensive analysis was performed, considering all eight possible parameters of DDMRP. A metaheuristic algorithm was employed. However, it should be noted that these approaches do not guarantee an optimal solution. The study assumed known demand, obtained through either forecasting or historical data, over a planning horizon of one year. To obtain fronts of non-dominated solutions, the authors utilized a multi-objective algorithm called NSGA-II. This algorithm was utilized to determine and fix all parameters once a year based on the historical data from the previous year. (Lahrichi et al., 2022) proposed a mathematical model utilizing Mixed Integer Linear Programming (MILP) for parameterizing three factors of DDMRP. The objective was to minimize the average on-hand inventory while ensuring 100% of OTD. The study successfully obtained optimal solutions. However, it is important to note that this study assumed known demand and generated 51 replenishment orders over 100 days**.** These studies lack a learning method and exhibit limited responsiveness when confronted with uncertain demand (Duhem et al., 2023).

To our knowledge, only (Duhem et al., 2023) have successfully integrated a Demand-Driven Operating Model (DDOM) and an RL agent. They proposed an automatic method to adjust DDOM parameters using reinforcement learning. They attempted to adjust OST and OSH by using the Branching Dueling Q-Network algorithm to solve problems with unknown demand. They adapted their algorithm to a multiproduct environment and various experimentations. This demonstrates that their approach is well-suited for industrial applications where demand is unknown or difficult to predict.

2.1.4. Deep Reinforcement Learning (DRL)

Reinforcement learning (RL) is one of the three types of machine learning, along with supervised and unsupervised learning. It involves an agent actively acquiring knowledge by engaging and interacting with its environment (Sutton and Barto, 2017). RL draws inspiration from psychological theories, of animal learning and achieves optimal decision-making strategies by learning from experience. In RL, an agent is defined as the decision maker, while the surrounding environment encompasses everything outside the agent. RL focuses on determining the optimal action for an agent to maximize the total reward it receives over time. In the field of artificial intelligence, a key objective is to develop fully autonomous agents that engage with their environments, learning optimal behaviours and continuously improving through a process of trial and error.

As illustrated in Figure 3, RL involves an **agent** engaging with an **environment**, where at each time step t, the agent perceives the current **state** of the system, represented as $st \in S$ (where S denotes the set of possible states). The agent then selects an **action** at $\in A(st)$ (where A(st) represents the set of feasible actions given the state st) and receives a **reward** $rt \in R$. Subsequently, the system transitions randomly to a new state $st+1 \in S$. The sequence of actions performed by the agent in response to the environment's states is known as the **policy**. RL algorithms require a balance between **exploration** and **exploitation**. Initially, agents should explore the environment to learn, gradually transitioning to exploiting their acquired knowledge for optimal decision-making. If agents solely rely on exploitation, they may miss out on better policies. Therefore, exploration is necessary to discover optimal solutions. This process follows the principles of a **Markov decision process** (MDP) (Sutton and Barto, 2017), which provides a mathematical framework for modelling the decision-making process in Markov processes (S, A, T, R, γ), where:

- A set of states S defines the environment. S is the set of all possible states of the environment observed in period t;
- A set of actions A (s) to be taken by the agent when it is in state s. A is the set of all possible actions selected by the agent;
- A transition model T (s', (s, a)) where a ε A (s); represents transition probabilities from one combination of state and action to the next
- A reward function R (s) that denotes the rewards based on the cost minimization objective
- γ is the discount factor for future rewards
- The value function

$$V(s)=R(s) + \gamma\Sigma s'\epsilon SP(s'(s,\pi(s))V(s')).$$

The objective of the MDP is to find a policy that maximizes the expected value of the rewards associated with the states:
$\pi : s \rightarrow a$
Thus, we seek to maximize the expected profit given by the function (Sutton and Barto, 2017):

$$Gt = Rt+1 + Rt+2 \cdots + Rt$$

where: Rt: It is the reward obtained in episode t

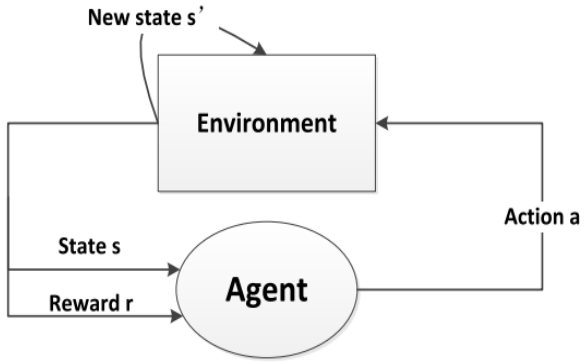

Fig 3. The general framework in RL

In recent years, the success and popularity of RL have surged, primarily driven by its remarkable achievements in solving complex problems that involve making sequential decisions. RL agents naturally consider not only immediate rewards or costs but also future ones, allowing them to learn the most advantageous sequence of actions within a given environment instead of treating each action independently (Sutton and Barto, 2017). In addition, RL has shown the potential to surpass human-level performance, as demonstrated by (Mnih et al., 2013). Unlike supervised learning, RL is not limited by the need for human-supervised data. The remarkable progress in the field of RL can be attributed to the invaluable achievement resulting from the integration of RL and deep learning (DL) or deep neural networks, known as DRL (Sutton and Barto, 2017). Deep neural networks can be used as function approximators (neural network (NN)) in reinforcement learning to represent value functions or policy functions, enabling more complex and high-dimensional representations. DRL combines the fields of RL and DL to tackle decision-making problems that were previously difficult to handle, such as chess, Go, and Atari games (Silver et al., 2016) . It offers the potential to address challenging environments characterized by high-dimensional state and action spaces. The integration of RL and DL techniques empowers RL agents with enhanced environmental perception capabilities. DRL algorithms have been successfully employed in diverse domains, including robotics, video games, medicine and healthcare, finance, transportation systems, and Industry 4.0 (Mousavi et al., 2018).

Several DRL algorithms commonly used for continuous environments have been developed in the literature. The most relevant algorithms are:

- Deep Deterministic Policy Gradient (DDPG): is an off-policy algorithm that combines deep Q-learning with policy gradients, enabling the learning of continuous control policies in high-dimensional action spaces.
- Twin Delayed Deep Deterministic Policy Gradient (TD3): is an improvement over DDPG that introduces twin value networks and delayed policy updates, leading to better stability and performance.
- Proximal Policy Optimization (PPO): a policy optimization algorithm that aims to find an optimal policy by iteratively updating the policy parameters. It has gained popularity for its stability and sample efficiency.

Article contribution and motivation

The parameterization of DDMRP in the literature has certain gaps that require attention. Previous studies (Damand et al., 2022) and (Lahrichi et al., 2022), have proposed mathematical and algorithmic optimization approaches and utilized demand history to optimize key parameters of DDMRP. However, these methods assume prior knowledge of demand by relying on forecasting techniques to predict demand over the planning horizon. In contrast, the fundamental principle of DDMRP is to respond to actual demand in real time. Duhem's work (Duhem et al., 2023), on the other hand, focused solely on studying two factors (OST and OSH) within a discrete environment using a learning approach. Such an excessive inconsistency of the existing guideline for parameterization of these key factors is the core motivation of this study.

To address these limitations, this research endeavour focuses on integrating DRL with DDMRP to dynamically optimize additional parameters, specifically the variability factor (F_V) and lead time factor (F_{LT}). The main objective is to maximize the completion of these parameters by leveraging DRL techniques within a continuous environment. By employing a DRL algorithm capable of operating in a continuous environment, it is implied that both the state and action spaces are continuous rather than discrete. This means that any agent trained within this environment must possess the ability to enable continuous control. It should be noted that in this study, the demand is not known in advance and will only be available within a predetermined horizon (H), aligning with the principles of DDMRP. The purpose of this work is to explore the potential of DRL in facilitating more effective decision-making within the DDMRP framework. Through the dynamic optimization of parameters and the ability to adapt to evolving conditions, the proposed approach seeks to determine the optimal values for the variability and lead time factors. These optimized factors aim to minimize on-hand inventory while maximizing the on-time delivery over a defined time horizon.

3. Problem description, assumption and modelling

In this work, the parameterization of DDMRP is considered an optimization problem to define a set of DDMRP parameters (DDMRP Parameters Optimization Problem: DDMRP-POP) (Damand et al., 2022). Since the DDMRP is based on current demand steering, we will assume that demand is known only during the demand visibility horizon. We are trying to provide a decision support system that will allow real-time decision-making for DDMRP parameter optimization. We define the DDMRP parameter optimization problem (DDMRP-POP) as follows:

• *Data:*

1) Time step is fixed to DLT+ H_{order} denoted by H ; where H = [1; 2; .; DLT + H_{order}]
2) The demand over a future planning horizon is known to day t+ H.
3) The average daily usage is the average demand.
4) The DLT of the considered buffered product.
5) The initial on-hand inventory.

• *Parameters to be optimized:*
  1) F_V: The variability factor
  2) F_{LT}: The lead time factor
• *Objective:*
  1) Minimizing the average on-hand inventory
  2) Maximizing the on-time delivery
• *Constraints:*
  1) Respect a DDMRP buffer sizing and replenishment policy

Assumptions and justifications

We examine a single buffered reference and address the challenge of determining the variability and lead-time factors to optimize the size of the buffer. The choice of these data characteristics and assumptions should be justified for the following reasons:

- F_V and F_{LT} have an impact on the dimensions of the red and green zones. This should be addressed by a manager since there is no predefined automatic rule for optimal parameterization;
- All the demand quantities are known during the time horizon H.(H_{order}=4 days and DLT=7 days)
- The core principle of DDMRP is to synchronize supply chain planning and execution with actual demand signals to improve operational performance (inventory, customer service levels, OTD..). For that, the historical data of the demand was not used in the proposed scenario and it was decided to generate pseudo-random data for learning the model. This data presents an actual demand identified in the time step H to avoid historical and prediction demand. However, the agent did not learn demand policies on their history but on their actual demand in limited horizon to avoid forecast and to have a good knowledge about future demand that can obtained from customers.
- ADU is updated each time step (ADU = the mean of demand).
- The buffer zone sizes (TOG, TOY, and TOR) are updated each day;
- The buffers have already been positioned. The DLT is known for all products.
- The initial on-hand inventory is known. q_t on-hand $t \in H$; The on-hand inventory at the beginning of day t is set to be equal to the demand during the lead time (DLT*ADU). This ensures sufficient inventory to cover consumption over the lead time and prevent stockouts on the initial days.
- On the first day of planning, on-order quantity and incoming supply are set at 0;
- There are no capacity constraints, when a replenishment order is launched at day t, it is delivered without delay at day t+ DLT.
- OSH is equal to DLT and OST is equal to 50% of the red zone.
- AOHI presents a KPI for right-sizing inventory as promised by DDMRP. We seek to optimize DDMRP parameters to minimize the stock and increase DDMRP performance.
- OTD represents a KPI for customer service as promised by DDMRP, measuring the rate of satisfied demands. All demands must be satisfied, and no unsatisfied demands (stock-outs) are allowed. In other words, we seek to optimize DDMRP parameters to improve customer service level and therefore increase DDMRP performance.
- Two objectives will be reached simultaneously: minimizing average on-hand inventory and maximizing on-time delivery.

4. Learning-Optimization Methodology

This section explains how the DDMRP-POP has been mapped into a Markov decision process (MDP) and solved using the proposed approach. We propose a learning approach to resolve (DDMRP-POP) because the DDMRP system must decide to optimize the parameters. DRL is a learning method that focuses on decision-making in cases where demand is known only in the determined horizon and cannot be predicted. The learning model aims to use the available demand information over the future horizon to enable optimization of DDMRP parameters. Therefore, the description problem of dynamic parameterization of DDMRP can be regarded and modelled as a Markov decision process. RL is characterized by interaction with the environment. An agent must learn from its interaction with the environment and its own experiences. Therefore, a learning data generation process for to present environment is required. Considering an agent that observes the customer demands variation over the horizon H. the agent makes its decision as a function of the customer's demand. The data collected by the agent over a horizon H represents a stochastic environment. The random variable is demand over horizon H. We need to define the time step called the future horizon over which the demand is known.

To explain the method used, it is essential to detail the methodological procedures of the study. Here's a structured explanation:

a) *Problem Description and Data Collection*: Identify specific DDMRP parameters needing optimization and define objectives. Collect relevant data.
b) *DRL Framework Selection:* Choose an appropriate DRL algorithm based on problem's complexity and data nature.
c) *Define the State, Action, and Reward*: Specify the state and action spaces. Design a reward function aligned with objectives.
d) *Simulation Environment Setup*: Develop a simulation environment that accurately replicates the real-world DDMRP system. Ensure it interacts with the DRL agent and provides feedback on actions taken.
e) *Model Evaluation and Validation*: Evaluate the DRL model to ensure it generalizes well to unseen data.
f) *Iterative Approach and Continuous Learning*: Implement an iterative approach for continuous optimization, balancing exploration and exploitation. Enable the DRL model to adapt to changing supply chain conditions and ensure continuous learning for dynamic environment adaptation.

Sections 3 and 2.3 provide detailed explanations for steps a) and b). The remaining steps will be detailed in the following sections

4.1. State

We define the model state across several sub-states that are given by the following variables:

- **State definition 1**: Demand variability $S_V(t)$

According to the APICS dictionary "Mathematically, variability or uncertainty in demand can be calculated by standard deviation, absolute mean deviation or variance of forecast errors".

The coefficient of variability (CV) is calculated by the following equation:

$$CV = Standard\ Deviation/Sample\ Mean$$

$S_V(t)$ presents the state of demand variability during the horizon to encode Fv in time steps to gain a basic comprehension of its fluctuations

$$S_V(t) = Standard\ deviation\ of\ demand\ over\ H\ /\ (ADU + Standard\ deviation\ of\ demand\ over\ H) \quad (11)$$

- **State definition 2**: Lead Time $S_{LT}(t)$

For this state, we choose an indicator that represents the relative availability of the products against the expected demand during the DLT. This state represents an indicator known as the inventory days of supply (IDS) or days of inventory. The IDS is a measure that provides an estimate of how many days an inventory level can sustain or cover the average daily demand over the DLT. It indicates the number of days the available or expected inventory can last based on the average daily usage and the lead time.

$S_{LT}(t)$ presents the lead time state to encode the lead time factor F_{LT} in time step t;

$S_{LT}(t)$= overall inventory level/inventory needed over DLT

$$S_{LT}(t) = IDS/DLT \quad (12)$$

Where: IDS = (OH + OP)/ ADU

So, we define the state vector at the time step from equations (11) and (12) as:

$$S(t) = [S_V(t), S_{LT}(t)] \quad (13)$$

Actions

The action of agent A_t will be based on the percentage of lead time factor, and variability factor to select at a certain time. The agent must determine the optimal value of the lead-time factor, and variability factor to take or to choose based on the coefficient of variability and lead time.

Regarding the action vector, we chose to implement a continuous action space.

The set of variability factor **(F_V)** actions available to an agent are fixed $A_V = \{0, 1\}$

The set of lead time factor **(F_{LT})** actions available to an agent is fixed $A_{LT} = \{0, 1\}$

This information will be stored at time t in a tuple with the following structure:

$$A(t) = [F_V(t); F_{LT}(t)] \quad (14)$$

4.2. Reward

In the RL context, the agent selects an action, and the environment generates a new situation and a reward for the action chosen. Hence, the major objective of the reward is to maximize the cumulative future reward. In this study, the goal of the DDMRP-RL approach is to minimize the on-hand inventory and maximize the OTD.

- **Reward definition R1**

R1: Rewards based on DDMRP levels. This reward aims to optimize the quantity of inventory available, which ideally falls within the range of TOR and TOY. Therefore, R1 will be formulated to:

$$R1 = \begin{cases} 5 & \text{if } TOR < OH \leq TOY \\ -5 & \text{Else} \end{cases} \quad (15)$$

- **Reward definition R2**

R2: Rewards based on service level. This reward system aims to maximize OTD of orders, ensuring a high rate of satisfied customer demands, while also preventing stockouts. A demand is considered satisfied when the available on-hand inventory, including incoming orders for the day, exceeds the demand quantity.

$$R2 = \begin{cases} 5 & \text{if } OH >= D(t) \\ -5 & \text{else} \end{cases} \quad (16)$$

By combining the two reward functions equations (15) and (16) that consider minimizing inventory level and maximizing on-time delivery at the same time, the primary reward function will be:

$$R = R1 + R2 \quad (17)$$

- **Rewards definition R3, R4**

R3, R4: Rewards based on shaping

To enhance the efficiency of the learning process, we have added a reward system based on shaping for the agent, along with supplementary information concerning its progress towards the goal. By providing these intermediate rewards, the agent can acquire knowledge incrementally and concentrate on the crucial steps or actions that lead to the desired outcome. When we mention rewards based on shaping, we mean that the agent receives intermediate rewards or feedback throughout its learning journey to steer its behaviour toward achieving the desired outcome. These rewards relatively rely on lower and upper bounds that the optimization algorithm needs to take into consideration

$$R3 = \begin{cases} 0 & \text{If } 0{,}05 < F_V \leq 0{,}95 \\ -5 & \text{Else} \end{cases} \quad (18)$$

$$R4 = \begin{cases} 0 & \text{If } 0{,}2 < F_{LT} \leq 0{,}95 \\ -10 & \text{Else} \end{cases} \quad (19)$$

So, the reward R will:

$$R = R1 + R2 + R3 + R4 \quad (20)$$

5. Experimental design

This section presents a series of numerical experiments conducted to assess the efficacy of the proposed DLR algorithm. The experiments aim to validate the effectiveness of the solutions derived from the algorithm. We implemented the deep reinforcement learning approach on a computer equipped with a M1 chip (16Gb in RAM). JAVA 8 programming language was used. In this initial trial, we generated a dataset that captures demand information spanning a planning horizon of one year. The dataset used for this experiment is available for reference at the provided link. The idea of demand generation is inspired by Lahrich's (2022) model (Figure 4); (Demands for one year following a normal distribution).

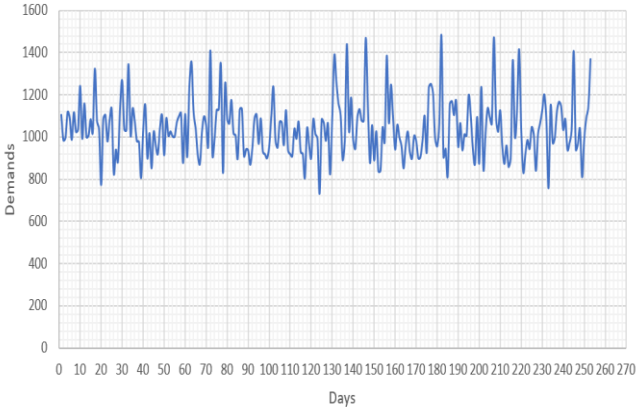

Fig. 4. Customers' demands

In our approach, on day t, the agent is provided with demand information exclusively for the planning horizon H, without relying on forecasts for the upcoming year to mitigate potential inaccuracies. We represent the demand values for a specific planning horizon using nested lists, where each inner list corresponds to the demand values for a particular period. The length of each internal list appears to be ($DLT+H_{order}$). Our objective is to equip our agents with precise real-time data sources regarding fluctuations in demand and lead times. Based on the state of real-time data of demands, we aim to enhance the performance of DDMRP and improve decision-making capabilities by determining the most suitable actions regarding variability and lead time factors. The environment defines two continuous states S1 and S2. The first state S1 represents the standard deviation of the sample over its mean, normalized between 0 and 1. S2 represents the ratio of overall inventory level to inventory needed over DLT. The environment provides a continuous action space with two actions, a_1 (Fv) and a_2 (F_{LT}), both ranging from 0 to 100. These actions are used to compute the buffer size. The model focuses on a single finished product, and each episode represents 308 days (Table 5 (Appendix A)) . Furthermore, a replication consists of 1200 episodes. We propose to calculate two factors Fv and F_{LT}, with Fv ∈ { 0; 1} and F_{LT} ∈ {0;1}.

5.1. Experiment 1: DDPG, TD3, and PPO Performance

After defining the actions (a1-a2) and states (S1-S2) as described in the previous section, we proceeded to test and compare the performance of the agents using three distinct DRL algorithms: DDPG, TD3, and PPO. These algorithms were briefly introduced in section 3 of the document. In our setting, we aim to define a sequence of actions that maximize the on-time delivery while minimizing the average inventory. These actions are designed to optimize the trade-off between timely delivery and inventory management. To evaluate the performance of the DRL algorithm in learning from experience, the algorithm was executed on the same instance for 1200 consecutive episodes. Each episode comprised 274 time steps, and we calculated the average cumulative profit achieved by summing the per-step profit at the final time step T. The decision to employ a sample size of 1,200 episodes is grounded in the identification of a stable reward curve. This stability indicates that the system has achieved consistent behaviour, making additional episodes unlikely to yield significant new insights. Selecting this stable point ensures efficient resource utilization, avoiding unnecessary computational costs. The convergence of the reward curve toward a constant behaviour implies repeatability and reliability in observations.

Figures 5 illustrate the evolution of each algorithm's ability to maximize reward as the episodes progress. This experiment aims at identifying the best DRL algorithm that can achieve the best reward function. This confirms that the encoding employed is capable of representing all the states of the system. As illustrated in Figure 4, the episodic reward plot typically shows the reward or return obtained by the agent over each episode during the learning process. By analyzing the episodic reward plot, we can observe the trend and patterns in the agent's performance. In our evaluation, we found that the DDPG model suffered from high fluctuations and irregularities indicating potential instability or suboptimal learning. On the other hand, TD3 demonstrated better performance in terms of episodic reward, although it displayed instability and began converging around the 1000th episode with a reward of [1400,1600]. PPO exhibits convergence around the 600th episode and maintains stability within the reward range of [2000, 2200]. Based on the observations depicted in the Figures, it is evident that the reward gradually stabilizes with an increasing number of episodes. This indicates that the learning process has been accomplished .On the other hand, PPO outperforms DDPG, converging around the 1400th episode. However, the PPO model stood out as the best-performing model, consistently delivering excellent results surpassing both DDPG and TD3 in performance. Based on the comparison results obtained, we have decided to exclude DDPG and TD3 algorithms from the remaining experiments and solely utilize the PPO algorithm. These results validate our decision to employ a DRL algorithm in a continuous environment for various reasons. We specifically chose an algorithm called Proximal Policy Optimization (PPO), known for its ability to deliver satisfactory results using default parameters or minimal adjustments to hyperparameters (Schulman et al., 2017).

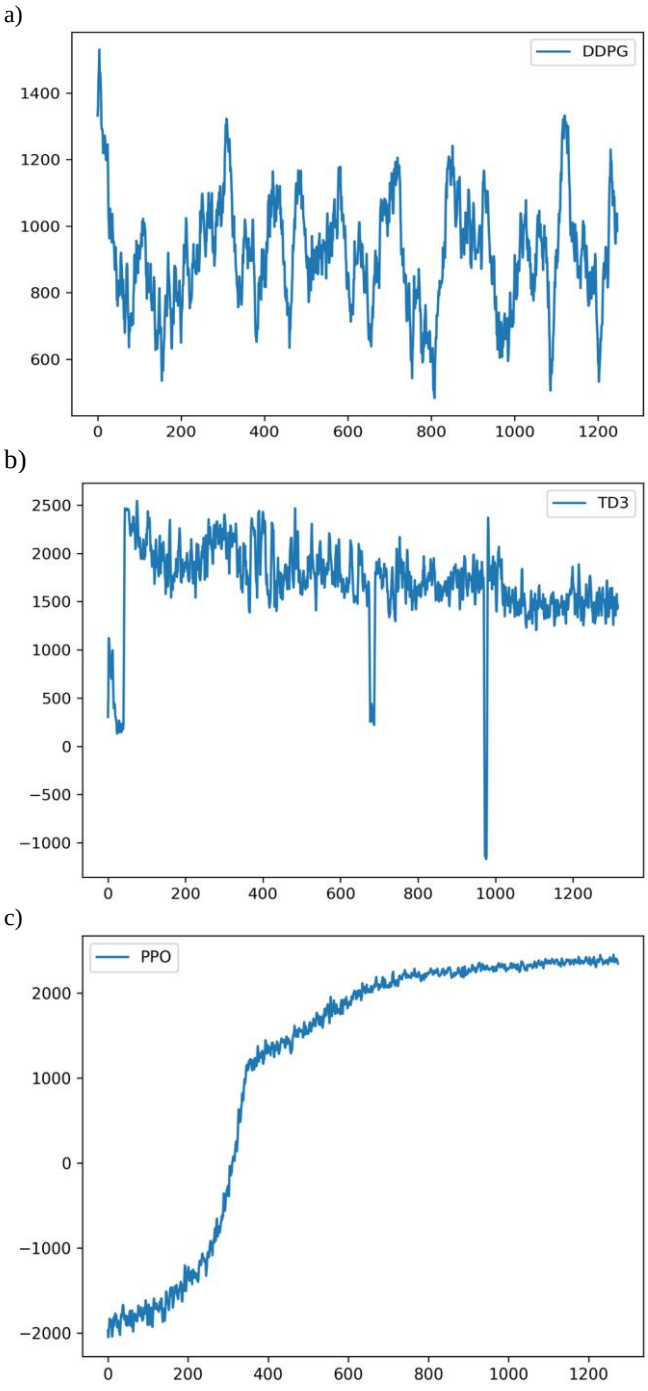

Fig. 5. Episodic reward for: a) DDPG, b) TD3, c) PPO

In this context, sample complexity pertains to the smallest dataset size necessary for the deep RL agent to acquire knowledge of a target function. PPO represents an actor-critic approach within DRL. It operates through two distinct deep artificial neural networks: one serves as the 'actor,' responsible for generating actions at each time step, while the other functions as the 'critic,' tasked with predicting the corresponding value. The actor's primary goal is to learn a policy that produces a probability distribution of feasible actions. Conversely, the critic aims to acquire the capability to estimate the reward function for specific states and actions. The divergence between the reward anticipated by the critic, characterized by parameters denoted as 'q,' and the actual reward obtained from the environment is captured within the loss function, denoted as 'Loss(q).

5.2. Experiment 2: PPO Performance against different DLT

In this experiment, we have introduced higher DLT values to explore more constrained versions of the task. These new DLT values can be seen as a broadened scope, allowing us to evaluate the agent's performance under increasingly challenging conditions. The primary objective of this experiment is to assess the agent's capacity to adapt to varying DLT values and document its performance. Figure 6 illustrates the rewards achieved in three scenarios with DLT values of 7 days (1DLT), 14 days (2DLT), and 21 days (3DLT), representing short, medium, and long DLT ranges, respectively.

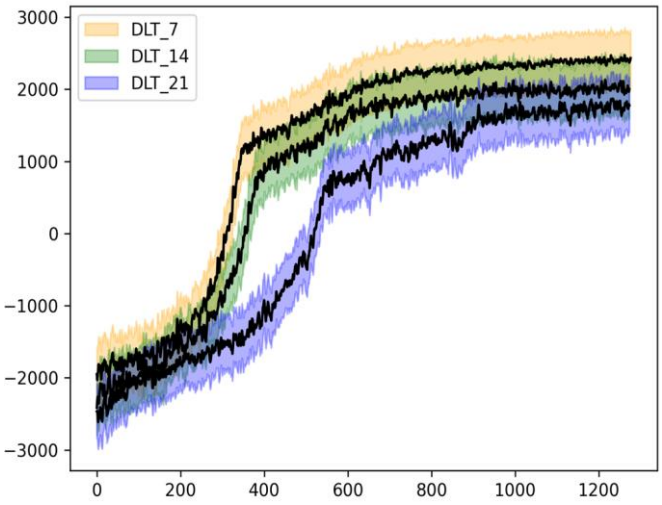

Fig. 6. Episodic reward for different DLT

It is evident that as DLT values increase, signifying greater environmental stochasticity, the task of finding an optimal policy becomes more challenging for the agent, resulting in higher variance in performance. Additionally, we observe that a higher DLT necessitates more steps for the PPO model to converge to a favourable policy. The results underscore the effectiveness of the PPO model when subjected to varying DLT values, especially those surpassing the originally defined parameters. We diligently monitored and recorded the agent's performance throughout the experiment.

5.3. Experiment 3: PPO Performance against high variability and short DLT

In this section, we explored how introducing unique data characteristics can affect the performance of our model. Specifically, we created data (Table 6 - Appendix A) that displayed a significant amount of variation, representing 77% of customer demands, along with a short DLT of just 5 days. The

goal of this experiment was to visually demonstrate how different data features could affect our model. The presence of high demand variability introduces a level of unpredictability, rendering it challenging for the model to rely on consistent demand patterns. Meanwhile, the short DLT imposes a sense of urgency, necessitating quick and precise decision-making with minimal margin for error. In this context, the model must rapidly adapt to fluctuating demands, optimizing allocation and delivery decisions within a constrained timeframe.

Figure 7 displays the episodic rewards and the policies determined by the agent over 100 days. Through this analysis, we assessed how the introduction of high variability in data, combined with a short DLT, affected the agent's performance. The agent requires additional episodes (1500) before it begins to converge towards an optimal policy, resulting in a lower episodic reward compared to experiments 1 and 2 (400 instead of 2000 previously recorded). The model requires approximately 1500 episodes to achieve only 20% (400) of the reward that was recorded before (2000). This confirms that the larger standard deviation has an impact on the agent's rewards.

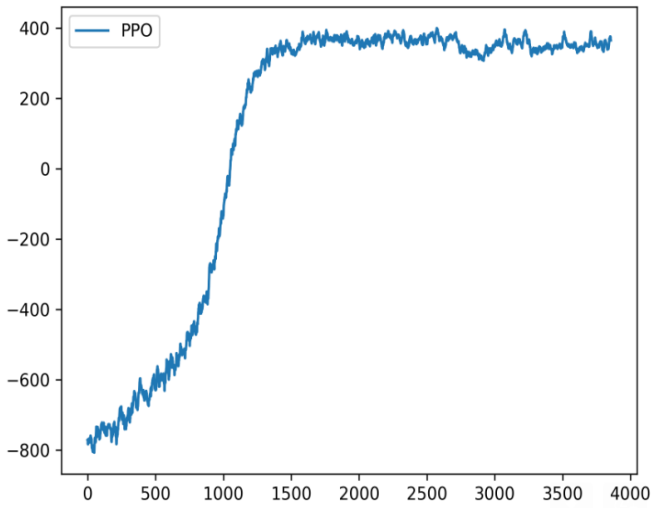

Fig. 7. Episodic reward

6. Results and discussion

Achieving the optimal parameterization of DDMRP poses a complex challenge, which proves to be demanding for both humans and machine learning algorithms, as demonstrated earlier. The task becomes even more difficult when there are additional constraints to fulfil and when incorporating realistic aspects of supply chain dynamics. To tackle these challenges, a parameterization approach for DDMRP is essential, striking a delicate balance between computational efficiency and complexity. This approach facilitates the effective configuration of DDMRP parameters, resulting in optimal inventory management and enhanced responsiveness within the supply chain.

In Experiment 1, three state-of-the-art deep reinforcement learning models were trained: DDPG, TD3, and PPO. The PPO algorithm exhibited remarkable time efficiency, prompting us to delve deeper into its reward performance. To evaluate its capabilities, it was observed that PPO required fewer time steps to converge towards an optimal policy. It's important to note, however, that shorter convergence time did not necessarily translate into similar episodic rewards when compared to TD3 and DDPG. In these experiments, PPO achieved a reward of 2000 within 600 episodes, while TD3 and DDPG failed to converge and remained unstable even after 1200 episodes. Additionally, it's worth mentioning that despite these differences in reward convergence, PPO remained the fastest among the three models in terms of training speed. DDPG and TD3 took approximately 240 minutes to complete 1200 episodes of training, whereas PPO accomplished the same number of steps in just 30 minutes.

In experiment 2, further investigations were conducted by modifying the DLT to introduce additional constraints to the agent's policy. To achieve this, we fixed the DLT values at 2DLT (14 days) and 3DLT (21 days), which represented more difficult scenarios compared to the initial settings. By imposing these stricter constraints, the aim was to test the agent's ability to adapt and find optimal control strategies even under challenging conditions. As a benchmarking result, it was confirmed that the PPO model demonstrated success in all tested scenarios when additional constraints were introduced to our environment. This accomplishment showcases the adaptability and resilience of the parameterization model to handle complex decision-making. Despite the increased difficulty level, the PPO model consistently performed well performance.

In this subsection, an example of a sequence generated by the agent, which is assumed to be well-trained, is presented. Table 7 (Appendix A) presents the outcomes within the action generation process that have achieved a perfect 100% OTD. Within this table, each row provides information regarding demands, the variability factor, lead-time factor, on-hand inventory levels, and OTD.

The following remarks can be made out from Tables 7 and 8:

- The variability factor F_V changed each DLT scenario, even when the demand remained the same across all scenarios. This implies that the agent adjusted its state with the S(t)= [$S_V(t)$, $S_{LT}(t)$] vector to transition to another state. This confirms that the state encoding effectively encompasses sufficient information for making informed decisions regarding parameterization.
- The lead-time factor F_{LT} is relatively increased when DLT increases; This can be explained by the fact that longer DLTs mean that the supply chain is exposed to potential disruptions, delays, and variability for an extended period. This prolonged exposure can amplify the impact of any variability, including lead times and demands that are present in the supply chain.
- In all scenarios, the maximum service level of 100% OTD is achieved.
- The average on-hand inventory is optimized overall. It closely aligns with stock consumption during DLT. As DLT increases, average inventory also increases but surpasses DLT consumption. This outcome is not surprising

given that a longer DLT naturally leads to longer delivery times for new replenishment orders.

- The average stock is impacted by the DLT. Longer DLTs often require higher safety stock levels to compensate for the uncertainty associated with extended lead times. These increases in safety stock can contribute to a relative increase in both the lead-time factor and on-hand inventory.
- In Table 7, horizon H represents the time step at which the demand is known. This confirms that the optimal planning horizon should generally be situated between DLT and 2DLT. Within this range, the agent has an ideal timeframe to observe the outcomes of its parameterization actions without a heavy reliance on demand forecasting data and generate an optimal performance regarding average on-hand inventory that is equivalent to consumption during DLT. When a horizon as long as 3DLT is considered, the agent encounters challenges in achieving optimal inventory levels because it becomes increasingly dependent on demand forecasting. In this scenario, the on-hand inventory presents 1.75 of DLT.

To better visualize the results for on-hand inventory, Figures 8, 9, and 10 (Appendix B) show the evolution of on-hand inventory concerning the three buffer zones. As a result, the on-hand inventory is always maintained within the desired level or range defined by TOR (lower bound) and TOY (the upper bound). Occasionally, the on-hand inventory reaches the upper limit (TOY), but it does not go beyond this threshold. This suggests that our model is designed to avoid excessive overstocking. Importantly, the on-hand inventory is never allowed to go negative. This indicates that the system consistently maintains sufficient inventory on hand to meet customer demand without experiencing stockouts. It is important to note that this solution achieved a perfect 100% on-time delivery (OTD) rate, an outcome of paramount significance from the perspective of decision-makers.

It aligns with the principles of the DDMRP philosophy, which emphasizes the use of buffers to safeguard against shortages and strives to ensure the highest possible service level. The model effectively prevents stockouts, ensuring that the on-hand inventory never falls below the level of demand.

Experiment 3 aims to investigate how our PPO model performs in a challenging environment characterized by two crucial factors: high demand variability and a short DLT. This experiment offered valuable insights into both the strengths and limitations of the model when confronted with a dynamic setting and intricate decision-making scenarios. We present the results (Table 9 - Appendix A) of a sequence generated by the agent. Table 8 presents the outcomes within the action generation process that have achieved a perfect 100% OTD. Within this table, each row provides information regarding demands, the variability factor, the lead-time factor, on-hand inventory levels, and OTD. We present the results from the 25th demand to minimize the influence of the initial inventory effect.

Dividing the 99 days into distinct sub-periods and calculating the ADU, average inventory in each of these sub-periods can provide valuable insights into how the inventory management strategy performs under different conditions (Table 10 - Appendix A). This approach allows assessing the adaptability and effectiveness of the model over time.

The following remarks can be made out from Tables 8 and 9:

- The OTD consistently achieves a perfect 100% rate with a 5-day lead-time and remains unchanged across all time steps. This ensures that the model meets all demand during the lead- time without any shortages.
- The variability factor (F_V)is relatively high over time. This is explained by the high variability in customer demand (77%), which is influenced by a larger standard deviation of 1084. This confirms the correlation between variability factor Fv and the standard deviation of customer variability.
- The lead-time factor (F_{LT}) remains consistently stable over time, consistently hovering at approximately 49%. This consistent value indicates a medium-level lead-time factor, which corresponds to a relatively short DLT. This observation confirms the inverse relationship between F_{LT} and DLT, which is a fundamental characteristic of the DDMRP philosophy.
- There is a notable improvement in the average inventory over time, accompanied by a decrease in variability from 87% to 68%. This can be attributed to the impact of customer demand variability on the average inventory. Despite the persistence of relatively high variability (68%), our model maintains an optimized stock level. This confirms the robustness and adaptability of the model in handling variability without increasing the stock.

Figure 11 (Appendix B) shows the evolution of on-hand inventory to the three buffer zones. The model exhibits strong resilience in accommodating significant variability and typically maintains itself within the optimal range. However, there were instances where we observed inventory spikes that surpassed the (TOG). This behaviour can be attributed to the model's proactive approach to ensuring an adequate stock level in anticipation of an expected surge in demand in the coming days. This anticipation aligns with our strategy to effectively address and prepare for upcoming demand fluctuations.

7. Conclusion and research perspectives

This article introduces a novel approach to address the challenge of DDMRP parameterization in the presence of unknown demand scenarios. To tackle this issue, a DRL agent based on the PPO algorithm is employed. This agent enables the DDMRP methodology to efficiently handle buffer sizing by dynamically adjusting the variability factor and lead-time factor parameters. This study is considered one of the early studies that have addressed the potential benefits of employing a DRL approach to dynamically adjust DDMRP parameters. By demonstrating the successful utilization of the PPO model in optimizing buffer sizing under uncertain demand conditions, this research opens new avenues for enhancing the efficiency and adaptability of supply chain management practices.

While the experiments have not confirmed all of Ptak's rules, the approach remains grounded in reality by considering a continuous environment, as opposed to a discrete one, and by addressing unknown demand rather than relying solely on historical and forecasted demand.. Moreover, optimal parameters are proposed, highlighting the significance of employing an optimization-learning model for the parameterization of DDMRP.

The study introduced a DRL model to optimize DDMRP systems, demonstrating its potential to improve on-time delivery and average on-hand inventory. The effectiveness of this approach depends on accurate parameter selection, which can be variable and challenging to predict. Future work will focus on developing robust methods for parameter selection, enhancing predictive analytics, and providing detailed guidelines and case studies for DRL interpretation and implementation. Additionally, extensive testing across various scenarios will be conducted to validate the model's stability and universality. Despite these challenges, the use of DRL in optimizing DDMRP parameters is promising, suggesting potential advancements in inventory management and customer service effectiveness.

The results obtained from this study serve as a reference point for evaluating the performance of the parameterization learning model and establish a solid foundation for future advancements, improvements, and research directions in this approach.. One potential limitation may be the reliance on certain assumptions in the parameterization learning model. Future research could focus on validating these assumptions in real-world supply chain scenarios to ensure the model's robustness across diverse conditions. This paper offers various research perspectives and opportunities

- By continuously pushing the boundaries and exploring the limits of our model, it can ensure its effectiveness and suitability for a wide range of practical applications in supply chain scenarios.
- Through a comparative analysis to assess how the proposed learning model performs in contrast to an alternative learning algorithm, our objective is to enhance the parameterization of DDMRP.
- By taking into account certain constraints such as multiple products, production capacity, Minimum Order Quantities (MQO), and certain priorities.
- Testing various scenarios, including internal variability, order spike horizon, and order spike threshold.

Reference

Azzamouri, A., Baptiste, P., Dessevre, G., Pellerin, R., 2021. Demand driven material requirements planning (Ddmrp): A systematic review and classification. Journal of Industrial Engineering and Management, 14(3), 439–456. DOI: 10.3926/jiem.3331

Bahu, B., Bironneau, L., Hovelaque, V., 2019. Compréhension du DDMRP et de son adoption : premiers éléments empiriques. Logistique & Management, 27(1), 20–32. DOI: 10.1080/12507970.2018.1547130

Benavente, D., Peralta, S., Quispe, G., Moguerza, J., Raymundo, C., 2023. The Demand Driven MRP Implementation in Complex Manufacturing Industries: A Systematic Literature Reviews. International Journal of Engineering Trends and Technology, 71(3), 33–45. DOI: 10.14445/22315381/IJETT-V71I3P205

Bennett, N., Lemoine, G. J., 2014. What a difference a word makes: Understanding threats to performance in a VUCA world. Business Horizons, 57(3), 311–317. DOI: 10.1016/j.bushor.2014.01.001

Boute, R. N., Gijsbrechts, J., van Jaarsveld, W., Vanvuchelen, N., 2021. Deep Reinforcement Learning for Inventory Control: A Roadmap. In SSRN Electronic Journal. DOI: 10.2139/ssrn.3861821

Cuartas, C., Aguilar, J., 2023. Hybrid algorithm based on reinforcement learning for smart inventory management. Journal of Intelligent Manufacturing, 34(1), 123–149. DOI: 10.1007/s10845-022-01982-5

Damand, D., Lahrichi, Y., Barth, M., 2022. Parameterisation of demand-driven material requirements planning: a multi-objective genetic algorithm. International Journal of Production Research. DOI: 10.1080/00207543.2022.2098074

Duhem, L., Benali, M., Martin, G., 2023. Parametrization of a demand-driven operating model using reinforcement learning. Computers in Industry, 147(September 2022), 103874. DOI: 10.1016/j.compind.2023.103874

El Marzougui, M., Messaoudi, N., Dachry, W., Bensassi, B., 2022a. Industry 4.0 Technologies on Demand Driven Material Requirement Planning: Theoretical Background and Impacts. 59–69.

El Marzougui, M., Messaoudi, N., Dachry, W., Bensassi, B., 2022b. Integration Model for Demand-Driven Material Requirement Planning and Industry 4.0. SAE International Journal of Materials and Manufacturing, 16(1), 5–16. DOI: 10.4271/05-16-01-0001

El Marzougui, M., Messaoudi, N., Dachry, W., Bensassi, B., 2023. Demand Driven Material Requirement Planning and Industry 4.0 Integration: Conceptual Framework and Hypotheses. International Journal of Engineering Trends and Technology, 71(12), 201–216. DOI: 10.14445/22315381/IJETT-V71I12P220

Esteso, A., Peidro, D., Mula, J., Díaz-Madroñero, M., 2022. Reinforcement learning applied to production planning and control. International Journal of Production Research. DOI: 10.1080/00207543.2022.2104180

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. https://pubmed.ncbi.nlm.nih.gov/34103776/

Lahrichi, Y., Damand, D., Barth, M., 2022. A first MILP model for the parameterization of DDMRP.pdf. Computers & Industrial Engineering, 174, 108769.

Lee, C. J., Rim, S. C., 2019. A Mathematical Safety Stock Model for DDMRP Inventory Replenishment. Mathematical Problems in Engineering, 2019. DOI: 10.1155/2019/6496309

Martin, G., 2020. Contrôle dynamique du Demand Driven Sales and Operations Planning. https://tel.archives-ouvertes.fr/tel-03165839%0 Ahttps://tel.archives-ouvertes.fr/tel-03165839/document

Marzougui, M. El, Messaoudi, N., Dachry, W., Sarir, H., Demand, B. B., Mrp, D., 2021. Demand driven mrp : literature review and research issues M El Marzougui, N . Messaoudi, W Dachry, H Sarir, B Bensassi To cite this version : HAL Id : hal-03193163.

Miclo, R., 2016. Challenging the "Demand Driven MRP" Promises: a Discrete Event Simulation Approach. 186.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M., 2013. Playing Atari with Deep Reinforcement Learning. 1–9. http://arxiv.org/abs/1312.5602

Mousavi, S. S., Schukat, M., Howley, E., 2018. Deep Reinforcement Learning: An Overview. Lecture Notes in Networks and Systems, 16, 426–440. DOI: 10.1007/978-3-319-56991-8_32

Ptak, C., Smith, C., 2016. Demand driven material requirements planning (DDMRP).

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., 2017. Proximal Policy Optimization Algorithms. 1–12.

Silver, D., Schrittwieser, J., Simonyan, K., Nature, I. A.-, 2017, U., 2016. Mastering the game of Go without human knowledge. Nature, 550(7676), 354.

Sutton, R. S., Barto, A. G., 2017. Reinforcement learning: An introduction (T. M. P. 978-0262039246. (Ed.); 2nd ed. Ca).

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. DOI: 10.1080/00207543.2019.1650978

Appendix

Appendix A

Table 4. Literature Review

Authors	issue	Methods used	Contribution
The present paper (2023)	Parameterization	Deep Reinforcement Learning (DRL)	Use DRL (PPO algorithm) to optimize DDMRP parameters such as Fv and F_{LT}
Luis DUHEM (2023)	Parameterization	Dueling Q-Network	Optimizing DDMRP parameters OSH (Order Spike Horizon) and OST (Order Spike Threshold) using a Dueling Q-Network
Lahrichi (2022)	Parameterization	mathematical optimization technique: Mixed Integer linear Programming (MILP)	Parameterization of three parameters; minimization of the average on hand inventory while a 100% OTD is forced by using a constraint. MILP outperforms the GA in fewer CPU time
Damand (2022)	Parameterization	Optimization metaheuristic/ Multi-objective optimization algorithm :	The multi-objective metaheuristic algorithm NSGA-II (Non-dominated Sorting Genetic Algorithm) was employed to optimize the eight important parameters of DDMRP with respect to two objectives. These objectives were to minimize the average on-hand inventory and maximize on-time delivery (OTD) at a 100% rate
Velasco Acosta, Mascle (2020)	Parameterization	Statistical analysis (consideration of data case)	In this study, a statistical analysis was conducted to derive parameterization rules that rely on data from the case study.
Martin (2020)	Parameterization	Statistical analysis (consideration of data case)	This study performs a statistical analysis to obtain parameterization rules that depend on data from case study
Lee and Rim (2019)	Parameterization	Heuristic	This study proposes a heuristic formula to fix two parameters of DDMRP: the lead time factor and the variability factor
Miclo (2016)	Parameterization	Optimization metaheuristic	The author suggests an automatic optimization by using a simulated annealing algorithm to fix the parameters. this metaheuristic approach was tested on the Kanban game

Table 5. Data Characteristics

DLT	Planning horizon	Data of demands	Min demand	Mean demand	Median demand	Max demand	Std dev. demand	CV
7 days	11 days	308 days	732	1041	1026	1483	137	13%

Table 6. Data Characteristics

DLT	Planning horizon	Min demand	Mean demand	Median demand	Max demand	Std dev. demand	CV
5 days	9 days	30	1 416	1 335	5 711	1 084	77%

Table 7. The optimal solution of DLT (7, 14 and 21)

Days	Demands	DLT =7				DLT =14				DLT =21			
		Optimal Parameters		*Objective functions*		*Optimal Parameters*		*Objective functions*		*Optimal Parameters*		*Objective functions*	
		F_V	F_{LT}	OH	OTD	F_V	F_{LT}	OH	OTD	F_V	F_{LT}	OH	OTD
25	1 118	55%	55%	7 117	100%	34%	50%	15 803	100%	46%	68%	18 374	100%
26	1 093	38%	37%	5 999	100%	15%	43%	14 685	100%	60%	63%	36 238	100%
27	988	56%	33%	4 906	100%	66%	54%	13 592	100%	64%	84%	35 145	100%
28	1 117	38%	40%	8 809	100%	43%	53%	21 785	100%	73%	80%	34 157	100%
29	1 025	68%	50%	7 692	100%	27%	58%	20 668	100%	53%	68%	33 040	100%
30	1 036	61%	49%	6 667	100%	17%	54%	19 643	100%	28%	75%	32 015	100%
31	1 242	26%	32%	5 631	100%	63%	55%	18 607	100%	70%	63%	30 979	100%
32	993	53%	30%	4 389	100%	46%	60%	17 365	100%	89%	63%	50 061	100%

33	1 159	63%	35%	10 374	100%	50%	48%	16 372	100%	41%	74%	49 068	100%
34	1 001	36%	46%	9 215	100%	37%	46%	15 213	100%	100%	74%	47 909	100%
35	1 007	31%	41%	8 214	100%	39%	45%	14 212	100%	54%	65%	46 908	100%
36	1 084	30%	41%	7 207	100%	44%	51%	13 205	100%	65%	75%	45 901	100%
37	1 022	54%	45%	10 031	100%	38%	60%	12 121	100%	41%	78%	44 817	100%
38	1 324	36%	26%	9 009	100%	46%	58%	20 364	100%	77%	77%	43 795	100%
39	1 079	28%	24%	7 685	100%	34%	50%	19 040	100%	50%	70%	42 471	100%
40	1 035	58%	51%	6 606	100%	57%	45%	17 961	100%	53%	64%	41 392	100%
41	774	26%	39%	5 571	100%	42%	44%	16 926	100%	56%	73%	40 357	100%
42	1 086	41%	33%	8 436	100%	22%	52%	16 152	100%	57%	71%	39 583	100%
43	1 105	69%	26%	7 350	100%	38%	63%	15 066	100%	57%	79%	38 497	100%
44	980	0%	41%	6 245	100%	44%	49%	13 961	100%	56%	69%	37 392	100%
45	1 078	45%	48%	5 265	100%	40%	40%	12 981	100%	50%	77%	36 412	100%
46	1 133	30%	39%	4 187	100%	46%	59%	21 599	100%	67%	79%	35 334	100%
47	827	54%	49%	3 054	100%	45%	46%	20 466	100%	43%	77%	34 201	100%
48	940	55%	34%	9 977	100%	15%	45%	19 639	100%	50%	75%	33 374	100%
49	885	51%	43%	9 037	100%	41%	50%	18 699	100%	43%	75%	32 434	100%
50	1 136	39%	44%	8 152	100%	46%	44%	17 814	100%	66%	76%	31 549	100%

Table 9. Optimal solution

Days	Demands	Fv	F_{LT}	OH	OTD	Days	Demands	Fv	F_{LT}	OH	OTD	Days	Demands	Fv	F_{LT}	OH	OTD
25	1474	56%	56%	5 258	100%	50	588	49%	53%	11 198	100%	75	231	53%	51%	5 972	100%
26	72	57%	44%	12 139	100%	51	2835	29%	61%	10 610	100%	76	1863	57%	38%	9 241	100%
27	83	47%	54%	12 067	100%	52	1984	36%	47%	7 775	100%	77	644	80%	54%	7 378	100%
28	1571	57%	42%	11 984	100%	53	1868	39%	54%	5 791	100%	78	1459	47%	57%	6 734	100%
29	1028	60%	52%	10 413	100%	54	1401	70%	41%	8 200	100%	79	2134	47%	46%	5 275	100%
30	939	52%	48%	9 385	100%	55	2300	69%	54%	12 950	100%	80	279	65%	26%	7 551	100%
31	2188	64%	56%	14 920	100%	56	3058	48%	42%	10 650	100%	81	73	75%	42%	10 719	100%
32	1501	45%	43%	12 732	100%	57	5320	37%	48%	7 592	100%	82	1245	39%	40%	10 646	100%
33	4245	49%	44%	11 231	100%	58	754	71%	53%	2 272	100%	83	671	24%	48%	9 401	100%
34	1380	35%	51%	6 986	100%	59	789	55%	42%	7 053	100%	84	1461	87%	47%	8 730	100%
35	1107	35%	41%	13 434	100%	60	99	47%	47%	6 264	100%	85	1808	36%	46%	7 269	100%
36	563	50%	38%	12 327	100%	61	738	50%	40%	6 165	100%	86	2668	41%	50%	10 909	100%
37	491	77%	39%	11 764	100%	62	2088	64%	38%	10 573	100%	87	266	82%	46%	8 241	100%
38	399	52%	52%	11 273	100%	63	1617	69%	49%	13 225	100%	88	1500	52%	39%	7 975	100%
39	518	71%	49%	14 431	100%	64	1041	13%	43%	21 969	100%	89	518	45%	44%	6 475	100%
40	1345	11%	47%	21 855	100%	65	1382	56%	58%	20 928	100%	90	1615	64%	35%	5 957	100%
41	397	88%	49%	20 510	100%	66	1394	81%	57%	19 546	100%	91	450	78%	49%	4 342	100%
42	2239	81%	58%	20 113	100%	67	1811	48%	65%	18 152	100%	92	2668	96%	57%	9 145	100%
43	239	51%	61%	17 874	100%	68	705	69%	44%	16 341	100%	93	845	61%	40%	10 612	100%
44	1399	97%	50%	17 635	100%	69	2461	45%	47%	15 636	100%	94	1403	63%	50%	9 767	100%
45	419	71%	48%	16 236	100%	70	371	18%	41%	13 175	100%	95	223	79%	48%	8 364	100%
46	277	43%	59%	15 817	100%	71	1061	54%	80%	12 804	100%	96	3355	60%	45%	8 141	100%
47	519	52%	60%	15 540	100%	72	962	47%	47%	11 743	100%	97	1535	35%	55%	8 119	100%
48	3036	57%	45%	15 021	100%	73	2043	42%	58%	10 781	100%	98	1335	74%	46%	6 584	100%
49	787	54%	49%	11 985	100%	74	2766	59%	50%	8 738	100%	99	1595	76%	42%	12 645	100%

Table 10. Synthesis of optimal solutions (DLT =5)

Sub periods	ADU	Standard Deviation	CV	Average stock	Stock coverage	OTD
[25-49]	1 129	985	87 %	13 717	12,14 days (242 % of DLT)	100 %
[49-74]	1 657	1 109	67 %	11 605	7 days (140 % of DLT)	100 %
[75-99]	1 274	862	68 %	8 248	6,47 days (129% of DLT)	100 %

Appendix B

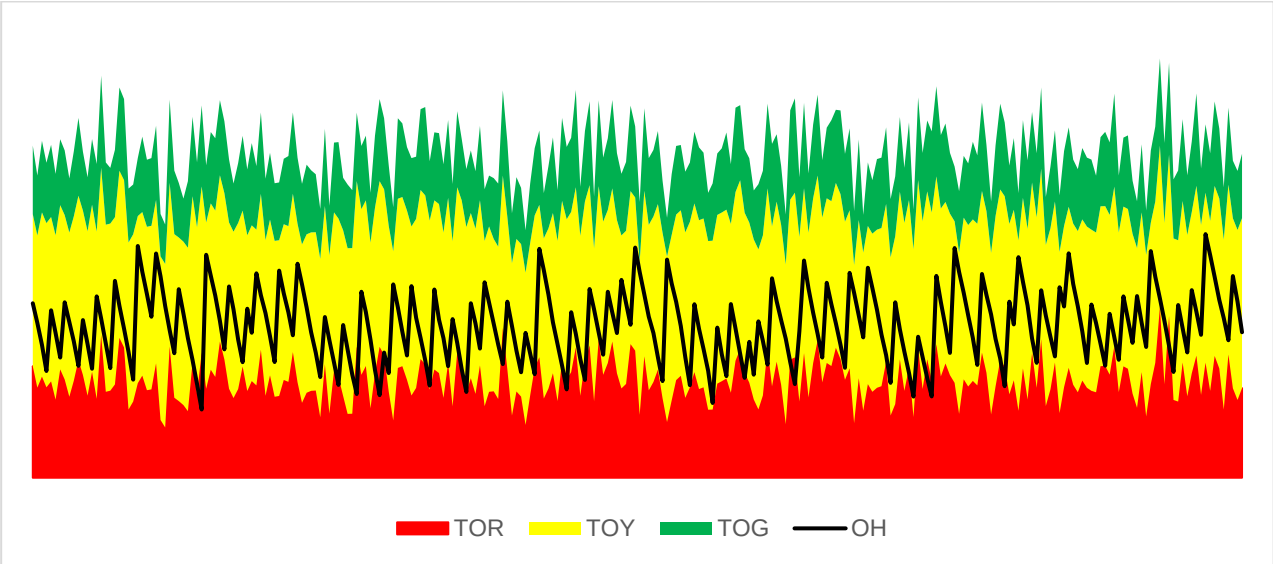

Fig. 8. Evolution of on-hand inventory DLT =7

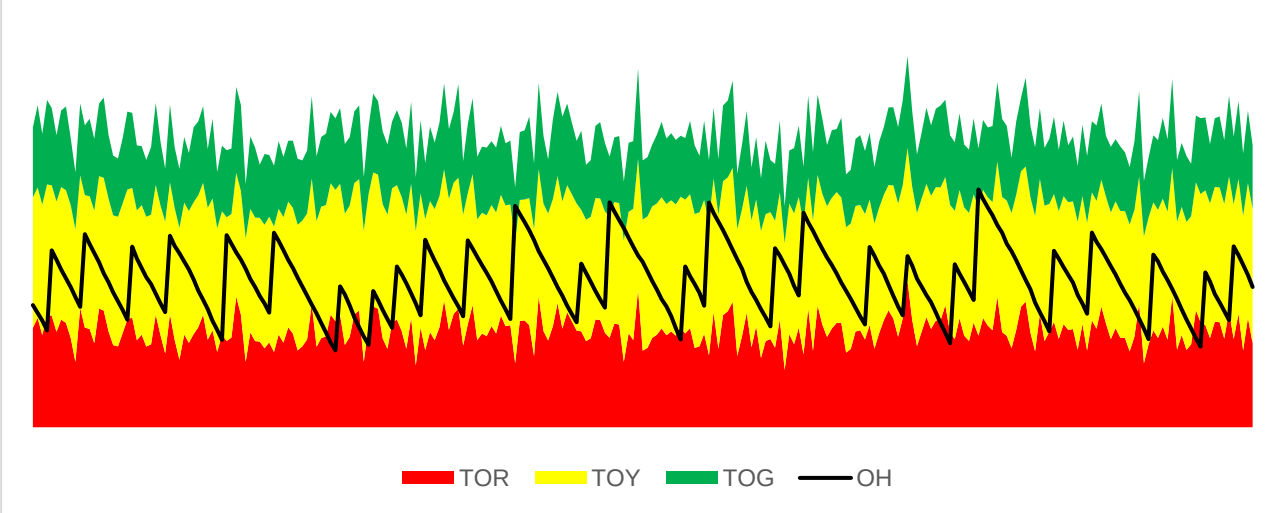

Fig. 9. Evolution of on-hand inventory DLT =14

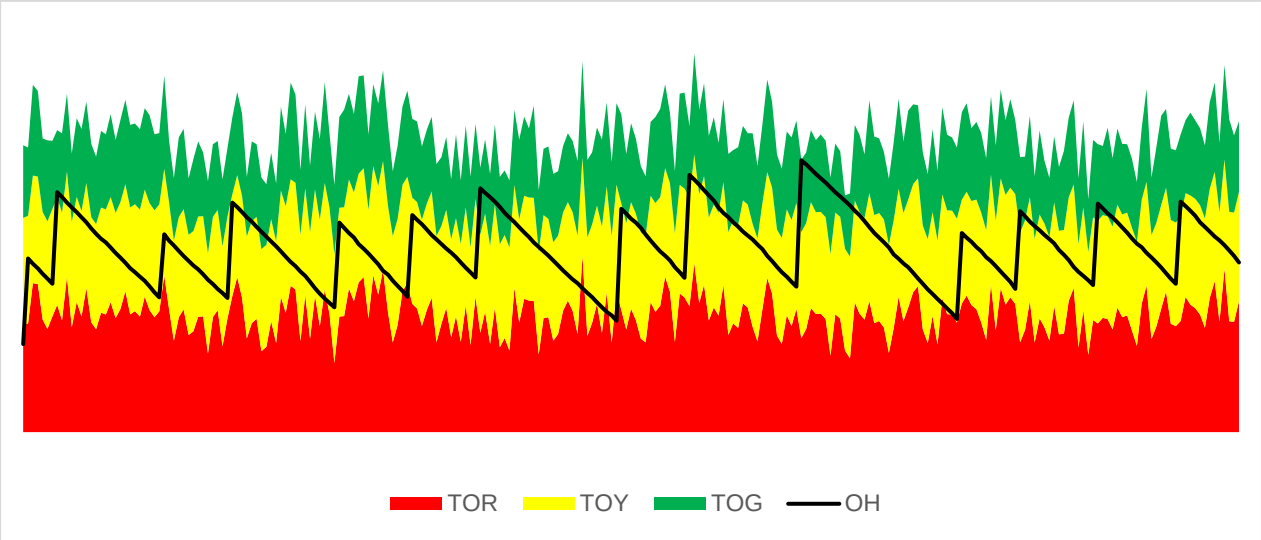

Fig. 10. Evolution of on-hand inventory DLT =21

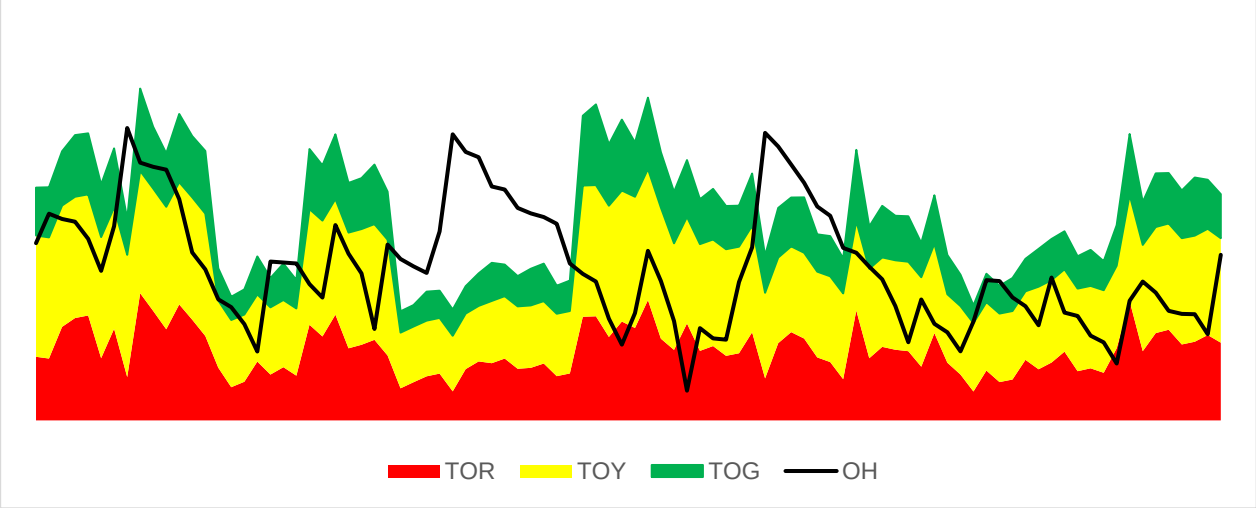

Fig. 11. Evolution of on-hand inventory