ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis in Operations Research and Quantitative Logistics

Managing (Q, R) Inventory System with Priority-Differentiated Customer Class and Demand Lead Times

Charalampos Doxopoulos (593158)

Supervisor: Dr. Oguzhan Vicil

Second assessor: Prof. Dr. Ir. Rommert Dekker

Date: October 13, 2022

Abstract

We study a static inventory rationing policy of a single product, which serves two priority-differentiated customer classes (i.e. high-priority and low-priority). Demands arrive according to independent Poisson processes, with a fixed size of one unit, and are categorized according to their priority into critical or non-critical. The demand classes also differ in their advance demand information (ADI) structure. One of them provides perfect ADI, and its orders are due after a deterministic demand lead time (DLT), whereas the other one is due immediately. Inability to satisfy demand at its due time results in backorders. Lost sales and early deliveries are not allowed, and the replenishment lead time is deterministic.

We propose a continuous (Q, r, K) review policy. A common inventory pool is stored to satisfy both demand classes, as long as the physical stock is above a threshold level. When the physical stock drops to the threshold level, only critical demand is satisfied, while non-critical is backlogged at its due time. The backorders are cleared after the next replenishment arrival, according to the priority clearing mechanism.

Despite the complexity of the examined model, we derive expressions to determine the steady-state class-specific fill rates and backorders, and the expected on-hand stock, for given inventory control policy parameters. Moreover, we propose two algorithms to optimize the inventory system. The first one optimizes the policy parameters to minimize the expected inventory holding cost, while the class-specific service level requirements are met. The second one optimizes the policy parameters to minimize the total inventory cost. A numerical study is held to establish the quality of our proposed heuristics. Finally, we highlight the benefits of incorporating inventory rationing and ADI/DLT into a continuous (Q, r) model. We show that inventory rationing and ADI/DLT can save independently up to approximately 20% and 10% of inventory cost, respectively.

Contents

1 Introduction

Inventory management has interested researchers and managers for many years. Finding new ways to improve the efficiency and performance of an inventory system can result in profit increase and/or customer satisfaction. Two ways to achieve that are inventory rationing, in case of customer differentiation, and advance demand information (ADI), in case of appropriately shared information among the members of a supply chain.

Customer differentiation occurs often in production/inventory systems, where the requirements of service level or the stock-out costs vary. Customers can be divided into classes of different priority levels, where each class can have different treatment, according to some of its characteristics. In case of production, such a characteristic could be the shortage cost of an item. For example, consider a spare part that can be used in the production of two different finished goods. Choosing to use it for one of those goods means inability to produce the other one. In that case, based on the shortage costs, the two goods are two customers, with different criticality. In case of retail purchases, a characteristic to differentiate the retailers could be their importance to the company. Clients that have a significant role in company profits may be considered as important and thus, have different treatment than the rest.

To deal with multiple demand classes (i.e. customer classes), a simple way could be to use separate stock for each class. The inventory parameters of each class would be independent of the others. Although this policy could be useful, it leads to high safety stock and thus high inventory cost. Another way could be to use the same stock pool for the whole demand, regardless of the classes, which is known as round-up policy. In that case, to meet the service level requirements, the inventory control policy parameters should be determined by the highest class-specific service level. Thus, most of the expected service levels will be higher than required, which will increase the inventory costs.

Rationing arises quite often in a variety of contexts. Some examples could include the seats on an airplane, the cars in a rental company, and the rooms in a hotel, where the customers are divided into business and economy class. It is a well-known and widely used tool that balances supply with demand, in case of different categories of customers. Kleijn and Dekker (1999) give an overview of problems with multiple demand classes that can be approached with inventory rationing. Its main advantage is the ability to meet the class-specific service level requirements, while maintaining relatively low inventory, thanks to demand pooling. It allows prioritization of demand classes, by providing different service levels, without using separate inventories. To be more precise, there is a common stock pool that serves all the customers, according to their criticality. Let us assume a single product, the demand of which can be classified

into two priority classes (critical and non-critical). A part of the stock pool is used to satisfy the demand, regardless of the priority. This happens until the stock reaches a threshold level, after which the stock pool is used to satisfy exclusively the demand of the critical demand. If unsatisfied demand can be backlogged, then, when the physical stock drops to the threshold level, non-critical demand can be backlogged, but critical demand can continue being satisfied. When the stock level reaches zero, then neither critical orders can be satisfied and they are also backlogged.

One should decide on the backorder clearing mechanism. Deshpande et al. (2003) provide several clearing mechanisms that can be used in similar cases. Obviously, when a replenishment order arrives, satisfying critical backorders should be the priority. However, when satisfying the critical backorders the stock pool might drop to or below the threshold level. Then, we can either use the remaining inventory to satisfy the non-critical backorders or increase our reserved stock.

Another aspect that has become more relevant over the years is incorporating advance demand information (ADI) into inventory decisions. When used effectively, ADI can improve the performance of production/inventory systems (Karaesmen et al. (2004)). Access to more information about future demand can lead to better demand forecast and more effective inventory control and planning. Thus, ADI can lead to lower stock levels, while maintaining a satisfying customer service level or even increasing it.

ADI is often categorized as perfect or imperfect, in terms of the information about the time and the size of the future demand. If ADI is perfect, we have complete knowledge of the orders that will be placed, in advance. We know, beforehand, the exact quantity of them and their due time, whereas canceling them is not an option. In contrast, if ADI is imperfect, our knowledge is limited. Both quantity and due time are unknown; they are based on estimations, whereas cancellations are possible. Note that demand lead time (DLT) (i.e. the amount of time between the placement of an order and its due time) is deterministic in the perfect case, and stochastic in the imperfect.

In perfect ADI, customers do not allow delivery before the due date and thus, early shipment is forbidden. "This assumption is realistic in many though not all situations" (Hariharan and Zipkin, 1995). Some examples are:

1. The inventory capacity of a customer may be limited and thus, there is no room for stock replenishment before the expected delivery time. Moreover, receiving a delivery early results in a higher stock level, which increases the holding costs of the customer. Therefore, early shipment is not desired.

2. To prevent the machines of a production line from failure, maintenance is required, and some parts that decay need to be replaced. The managers set a usage time at which these parts need to be replaced with new ones. Thus, spare parts should be delivered at the end of the usage time and based on a just-in-time

(JIT) system.

1.1 Problem Statement

In this thesis, we consider a static rationing policy of a single product inventory, which serves two priority differentiated customer classes (high-priority and low-priority). The demand of the high-priority class is characterized as critical, and the demand of the low-priority class as non-critical. Those two classes result from the different service level requirements, since critical demands require higher service level than non-critical ones.

Apart from their service level, the two classes differ in terms of their ADI. One of them provides perfect ADI, where its orders are due after a deterministic lead time, whereas the orders are due immediately for the other class. Specifically, we consider two different cases. In the first case, critical demands are due immediately, while the non-critical demands are due after a fixed DLT. In the second case, critical demands are due after a fixed DLT, and non-critical demands are due immediately. The same ADI/DLT structure was also studied by Koçağa and Şen (2007) and Vicil (2021b). The demands are assumed to be independent Poisson processes, with a fixed size of one item, and no early deliveries are allowed.

To be more precise, the research is based on a continuous (Q, r, K) review policy. The distribution center keeps a common inventory pool to satisfy the demand of both customer classes. At their corresponding demand due times, as long as the common physical stock is above K (i.e. threshold or critical level), the orders of both demand classes are satisfied on a first-come, first-served (FCFS) basis. When the on-hand stock drops to or below K, we continue satisfying only the critical class demands while the non-critical class demands are backlogged. When the inventory position drops to or below r (i.e. reorder point), we order Q (i.e. replenishment quantity) number of units, which arrive after a deterministic replenishment lead time. The threshold level is assumed to be constant and independent of the arrival time of the next replenishment order.

In our model, there are no lost sales, but inability to satisfy the demand on time results in backorders, which are cleared through the priority clearing mechanism. This means that when the replenishment order arrives, at first, the inventory is used to clear the existing critical backorders. Then, only if the remaining inventory is above K, can the existing non-critical backorders be cleared.

Our objective is manifold. First, for given inventory control policy parameters (i.e. Q, r, K), we approximate the steady-state class-specific fill rates, the steady-state class-specific expected backorders (the expected time-weighted backorders), and the steady-state expected on-hand stock. Second, we optimize the inventory parameters,

such that the expected on-hand inventory is minimized, while meeting the class-specific service level requirements. Then, we optimize the inventory parameters, such that the expected total inventory cost is minimized. Finally, we identify the benefits of incorporating demand lead time and inventory rationing into a continuous (Q, r) review model.

1.2 Purpose and Contribution

Although many researchers studied the effect of rationing on inventory systems, few of them considered ADI, which is becoming more relevant over the years, thanks to continuous technological improvements. Our inspiration has been originated from those few researchers, and specifically from Koçağa and Şen (2007) and Vicil (2021b), who incorporated perfect advance demand information into a threshold inventory rationing problem, with a continuous $(S-1, S)$ review policy. On top of that, there is no literature that studies the effect of ADI on inventory rationing in a continuous (Q, R) policy, which policy is widely used in the industry. Thus, the purpose of this research is to fill this gap.

Our model is based on the inventory rationing problem of Deshpande et al. (2003). They provide an example of inventory rationing (Q, r, K) policy in the US military, where different military services (i.e. navy, army, and airforce) have different service level requirements for common parts. Thus, they store a common inventory pool to satisfy the demand of all the military services and use rationing to meet the different service level requirements. The main difference between their and our problem is DLT. They consider DLT $= 0$. Vicil (2021b) indicates the difficulty of incorporating DLT into a threshold rationing policy. Since demand arrivals and due times differ, our model requires a different analysis in terms of the steady-state probabilities. Therefore, we provide structural results and properties of the steady-state distribution, in the same way Vicil (2021b) did for the $(S-1, S)$ policy.

Thus, we are able to present the first study that incorporates ADI/DLT into an inventory rationing system that serves two priority-differentiated demand classes, in the framework of a continuous (Q, r, K) review policy. Our research answers questions such as:

- For given policy parameters (Q, r, K), what are the steady-state class-specific expected backorders and fill rates, and the expected on-hand stock?
- For given class-specific fill-rate constraints, what are the optimal policy parameters (Q, r, K) that minimize the expected holding cost rate?
- For given inventory cost parameters, what are the optimal policy parameters (Q, r, K) that minimize the expected total inventory cost rate?

- How beneficial is the incorporation of demand lead time and inventory rationing to a continuous (Q, r) review model?

While we study a research problem, we expect our solution approach to be used in multiple industry settings. This is due to the fact that inventory rationing is a widely used technique, and over the years, more and more companies obtain access to advance demand information. Thus, the ones which will be able to use this information effectively will gain a great advantage over the others.

1.3 Outline

The remainder of this paper is organized as follows. In the next section, we review the most relevant literature. In Section 3, we derive all the necessary expressions to analyze our model and develop two heuristic methods, which both aim to optimize the inventory control policy. The first one focuses on minimizing the expected inventory holding cost, while satisfying the service level requirements, whereas the second one focuses on minimizing the expected total inventory cost. The numerical results of our heuristics and the evaluation of their performance are presented in Section 4.

2 Literature Review

In this section, we present the literature, relevant to our research. At first, we refer to the papers that study the inventory rationing policy and multiple demand classes. Then, we focus on incorporating ADI/DLT into inventory systems, and lastly, on combining inventory rationing with ADI/DLT.

2.1 Inventory Rationing Literature

Among the first to introduce multiple demand classes is Veinott Jr (1965). He considers a dynamic inventory model, with several demand classes for a single product where the system is periodically reviewed, the critical levels are equal to zero, and unsatisfied demand can be either partially or completely backlogged. Evans (1968), Topkis (1968), and Kaplan (1969) work on a similar model and prove that rationing is substantial in inventory systems. In contrast to Evans (1968) and Kaplan (1969), who considers two demand classes of different importance, Topkis (1968) presents a more general problem with n demand classes. In detail, his model includes periodically replenishment, the periods between which are divided into sub-periods. For each sub-period, the demand is observed, and the critical level is found. Unfulfilled demand is allowed to be completely backlogged, partially backlogged, or completely lost. When the replenishment order arrives, satisfying the backorders is the priority.

Nahmias and Demmy (1981) are the first that analyzed a rationing policy in a continuous (Q, R) model. Assuming static rationing, constant replenishment lead time, and independent Poisson demand processes with zero lead time, which can be divided into two classes, they are able to derive expressions for the expected backorders. To simplify their model, they assume existence of at most one outstanding order. Both Dekker et al. (1998) and Deshpande et al. (2003) present models that ignore this assumption. Assuming a continuous review policy with $Q = 1$, Dekker et al. (1998) are able to derive the exact service level of the non-critical demand but only to approximate the service level of the critical demand, since it depends on the incoming replenishment orders. Although they do not specify a clearing mechanism, they compare the approximated fill rates by assuming three different clearing mechanisms, through simulations. While Nahmias and Demmy (1981) focus only on calculating the fill rates of two customer classes, for given inventory parameters, Deshpande et al. (2003) deals with optimizing the parameters of a continuous review policy, based on some specific service levels. Another significant difference between these papers is that in the latter, the authors ignore the assumption of the outstanding order limitation. They compare four different control policies, one of which is the priority clearing mechanism, which we also use in our analysis. Ha (1997a) analyzes inventory rationing of a single item production system with multiple demand classes and lost sales. He shows optimality of a stationary critical level policy in a $M/M/1$ queuing system. Ha (1997b) examines also the same problem with the exception of dividing customers into two demand classes, and allowing demand to be backlogged but not lost. Later, he also extends the model he studied in the paper Ha (1997a), by assuming Erlang distribution of the processing time, and thus a model of $M/E_k/1$ (Ha (2000)). Similar to Deshpande et al. (2003), Arslan et al. (2007) study a model, with the extension to incorporate multiple demand classes, with different shortage costs or service level requirements. They develop a cost optimization model and a heuristic solution approach that finds the optimal threshold level that meets the service level requirements with respect to the lowest possible inventory. Assuming multiple demand classes and time-independent penalty costs in a multi-period system, Wang and Tang (2014) investigate a dynamic inventory rationing system with a mixture of backorders and lost sales types. To overcome the complexity of their model, they develop a heuristic method. Their numerical study shows that in case of both backorders and lost sales, improvements in the system performance is significant only if one class dominates in priority most of the period.

Investigating an inventory allocation problem, with two demand classes in a continuous review $(S-1, S)$ framework, Vicil and Jackson (2016) introduce a novel approach to estimate service levels and optimize stock levels. They provide a heuristic to determine the steady-state probabilities, when lead time is generally distributed, and a stock minimization algorithm, less complex than the already existing, since it requires

computing the steady-state distribution only once. In addition to the service level constraints in their model, Vicil and Jackson (2018) use waiting time constraints. Their optimization algorithm is exact and their methodology decreases the computational complexity of the optimization routine. Relying on the approach of Vicil and Jackson (2016), for the steady-state analysis, Vicil (2021a) develop an optimization algorithm to minimize the average expected cost rate in an inventory system with two priority-differentiated demand classes, where inability to satisfy the orders of one class leads to backorders, whereas inability to satisfy the orders of the other class leads to lost sales. He finds that under certain circumstances, the steady-state distributions of a system with generally distributed lead times are identical to the steady-state distributions of a Continuous-Time Markov Chain system that have the same mean. Moreover, Vicil (2022) provides an optimization algorithm to minimize the expected cost rate per unit time, for an inventory system, in which unmet demand is backlogged and then cleared according to the priority clearing mechanism.

2.2 ADI/DLT Literature

Trying to answer the question of whether two stages of a manufacturing process should be considered as unite or separate, Simpson Jr (1958) is the first to introduce DLT to inventory systems. He presents a model which optimizes the service times of each stage of the process, with respect to the prior determined service level criteria. Hariharan and Zipkin (1995) study a continuous review system with a fixed DLT for all the customers (perfect ADI). They conclude that either increasing the demand lead time or decreasing the replenishment lead time have the same results in the expected backorders and inventory levels, at each stage. Both cases can reduce the future demand uncertainty and thus the inventory. Gallego and Özer (2001) use a portfolio of customers with different positive demand lead times. They show that, in case of zero setup costs, a state-dependent base stock policy is optimal, whereas in case of positive setup costs, a state-dependent (s, S) policy becomes optimal. Their numerical study proves that ADI improves the system performance. Considering discrete time system, Karaesmen et al. (2002) examine the structure of optimal inventory policies. Since there are difficulties in applying their exact optimal policy in real life, they also propose a heuristic which is based on the base stock control policy with demand lead time information. Özer and Wei (2004) establish optimal policies for a capacitated inventory system, in which the manufacturer has access to advance demand information. They analyze a periodic threshold review policy with positive setup costs, where they order full capacity only if inventory falls below a threshold level. Tan et al. (2007) study the effect of imperfect ADI to ordering decisions. They find that the system performs better for lower level of ADI imperfectness and higher level of demand variability.

Similar to Gallego and Özer (2001), Wang and Toktay (2008) incorporate ADI with a periodic review policy, with the exception that early shipment is allowed. They conclude that the statement of Gallego and Özer (2001), that increasing demand lead time and decreasing replenishment lead time is equally beneficial, does not hold for the case of flexible deliveries. Thus, they suggest aiming to decrease the DLT, in these cases. Benjaafar et al. (2011) consider a production system with finite capacity and stochastic production times. In their study they assume that the time between two consecutive updates of demand information is random and the customers are allowed to request for order fulfillment either earlier or later than the expected date (imperfect ADI). They conclude that ADI is always beneficial for the supplier but might not be for the customers. Allowing the exceeding stock to be returned to the upstream supplier, Topan et al. (2018) investigate a single item periodic review policy, with lost sales and imperfect ADI. They show that the quality of ADI affects the inventory costs and allowing returning the exceeding stock increases significantly the benefit of ADI.

2.3 Combining Inventory Rationing with ADI/DLT

Using imperfect ADI, where the order due dates can alter and the orders can be canceled, Gayon et al. (2009) formulate a problem of a production/inventory system with multiple customer classes, as a continuous time Markov decision process. They conclude that incorporating ADI can reduce the costs of the supplier, complementary to the cost reduction incurred by inventory rationing without ADI. In contrast, Tan et al. (2009) model a similar problem by using discrete time. They propose solution methods based on Monte Carlo simulation.

In the first published paper that combines inventory rationing and ADI/DLT, Koçağa and Şen (2007) consider an inventory system with two demand classes, where the orders of the first one need to be satisfied immediately, and those of the second one are due after a determined lead time (perfect ADI). The two demand classes also differ in terms of their priority. Koçağa and Şen (2007) derive expressions for the customer service level of both classes, and find the exact service level of the non-critical class and an approximation of the service level of the critical class. Their model results in significant cost savings, when compared to simulation study, proving the importance of ADI. Moreover, they prove that rationing is more beneficial when ADI/DLT provides information about the critical class than the non-critical one. The same problem was also studied by Vicil (2021b), but he proposes a new method for estimating the service levels, the quality of which significantly outperforms the heuristic given by Koçağa and Şen (2007). To be more precise, he proposes a model that works for a continuous review $(S-1, S)$ policy, and two customer classes. The contribution of his model is twofold. First, it is able to find the optimal service level, for given inventory policy

and replenishment lead time. Second, it can optimize the inventory policy parameters, for specific service levels. To estimate the steady-state distribution, he uses a similar approach as presented by Vicil and Jackson (2016) and thus he is able to analyze the steady-state probabilities under the certain approximation assumption. Finally, he indicates the benefits of studying the limiting behavior of an infinitesimal probabilistic analysis to continuous review policies.

3 Model

To formulate our model, we need to introduce some appropriate notations. Those notations are presented in Table 1. In Subsection 3.1 we further explain these notations and use them to build our model framework. In Subsections 3.2 and 3.3 we analyze and formulate our service level optimization and cost optimization models, respectively. To implement these models, we derive some necessary expressions for the fill rates in Subsection 3.4 and the performance measures in Subsection 3.5. Finally, in Subsections 3.6 and 3.7 we present and explain the algorithms used for the service level and the cost optimization.

Notation	**Definition**
Q	Order quantity
r	Reorder point
K	Threshold level
λ^c, (λ^n)	Demand arrival rate of (non-)critical demand class
β^c, (β^n)	(Non-)Critical demand class service level
L	Deterministic replenishment lead time
H	Deterministic demand lead time (DLT)
OH	On-hand stock
IL	Inventory level
IP	Inventory position
$B^c, (B^n)$	Number of outstanding (non-)critical backorders
$Y^c, (Y^n)$	Number of (non-)critical orders that have been placed but not yet due
OO	Number of outstanding replenishment orders
$\widehat{D}$	Demand with net impact on the inventory level
$D^c, (D^n)$	(Non-)Critical demand
$\beta^c, (\beta^n)$	(Non-)Critical demand fill rate
$\bar{\beta^c}, (\bar{\beta^n})$	Minimum required (non-)critical demand fill rate
$E[\cdot]$	Expected value of a random variable $[\cdot]$
A	Fixed ordering/setup cost incurred per replenishment
h	Holding cost per quantity unit per time unit
$b_c, (b_n)$	Shortage cost per (non-)critical unit short per time unit
C	Total inventory cost incurred per replenishment cycle
$P^c_{out}, (P^n_{out})$	Stock-out probability of (non-)critical demand

Table 1: Definitions of used notations

3.1 Model Framework

To specify our problem, we assume the inventory of an item is replenished in deterministic time L, in order to satisfy the demand of two classes (i.e. critical and non-critical). Demand arrivals are independent Poisson processes with rates λ^c and λ^n, for the critical and the non-critical demand classes, respectively. We assume unit demand sizes for both demand classes. The reason for the differentiation of demand is the service level requirements. In our model, to measure the service level of each class, we use class-specific fill rates, which are defined as the percentage of the class-specific orders that are satisfied from on-hand stock at their due time. Critical demand needs a higher service level than non-critical. Therefore, if β^c and β^n denote the fill rates of the critical and non-critical demand, respectively, it must hold $\beta^c > \beta^n$.

We assume a continuous review (Q, r, K) inventory control policy, where we order a constant replenishment quantity Q when the inventory position drops to r. There are no lost sales, and thus, every order is either satisfied on their corresponding due time or backlogged. Inventory and replenishment orders are allocated according to the following mechanism: At first, we specify a fixed threshold level K. When the on-hand stock is above K, the incoming demands will be satisfied according to a FCFS basis at their due times, regardless of their criticality. When the on-hand stock drops at or below K, we will only continue satisfying the critical demand. This means that if the non-critical demand is due immediately, then it will be backlogged. Otherwise, if the non-critical demand is due after a lead time, then it will be backlogged after its lead time, only if the on-hand stock at that moment will be at or below K. Critical class orders are backlogged only if the on-hand stock reaches zero. Backorders can be satisfied after the next replenishment arrival. When a replenishment order of size Q is received from the resupply, first the critical backorders will be satisfied, and then any remaining inventory that exceeds the threshold level K is used to clear the non-critical backorders on a first-come, first-served basis. This is due to the fact that the non-critical class backorders are cleared only after the reserved stock K is fully restored. Since the demand lead times are deterministic, arrival times of due times also follow independent Poisson processes with rates λ^c and λ^n. Hence, the depletion rate is

$$\text{(depletion rate)} = \begin{cases} \lambda^c + \lambda^n & OH > K \\ \lambda^c & 0 < OH \leq K \\ 0 & \text{otherwise} \end{cases}$$

where OH denotes the on-hand stock that is available at a specific time.

In the first case of our analysis (i.e *Model 1*), we assume that critical demand is due immediately and non-critical demand after a deterministic DLT of H, whereas in the second (i.e *Model 2*), the opposite. For the remaining of this section, we focus

on *Model 1*. Moreover, as Koçağa and Şen (2007) and Vicil (2021b), to simplify our model, we also assume that $H \leq L$, so that the DLT is not allowed to be quoted longer than the replenishment lead time.

3.2 Service Level Optimization Model

In this subsection, we first derive some important expressions, based on which we define our optimization problem. In a continuous review model with all stock-outs backlogged, and constant lead times, the inventory level (IL) and inventory position (IP) have limiting distributions and can be connected with the following formula:

$$IL(t+L) = IP(t) - \widehat{D}(t, t+L) \tag{1}$$

where $\widehat{D}(t, t+L)$ is the demand that impacts the inventory level from an arbitrary time t, until L time units later.

Now, let us assume that at any time t, $B^c(t)$, $B^n(t)$, $Y^n(t)$, and $OO(t)$ represent the number of outstanding critical backorders, the number of outstanding non-critical backorders, the number of non-critical class orders that have been placed but not yet due, and the number of outstanding replenishment orders, respectively. Then, it holds:

$$IL(t) = OH(t) - B^c(t) - B^n(t) \tag{2}$$

$$IP(t) = IL(t) + OO(t) - Y^n(t) \tag{3}$$

A noteworthy phenomenon that may appear in the examined system is that, contrary to the regular (Q, r, K) inventory model, the on-hand stock can be larger than the inventory position. This may happen as IP drops when demand arrives, whereas OH drops when demand is due. For instance, consider the case where at a random time t, it holds $IP(t) = OH(t) = r + Q$. Now, consider that until H time units after t, there are no critical demand arrivals and a positive x number of non-critical arrivals ($D^c(t, t+H) = 0, D^n(t, t+H) = x$). Then, at time $t+H$, it holds $IP(t+H) = r+Q-x$, while $OH(t+H) = r+Q$. Hence, $OH(t+H) > IP(t+H)$. This phenomenon occurs due to the fact that the non-critical demand is due after a demand lead time, and it may make our analysis more complicated.

Furthermore, we can split the demand, according to its class. Consider that D^c and D^n denote the critical and non-critical demand, respectively. Then, for a random time t, it holds:

$$\widehat{D}(t, t+L) = D^c(t, t+L) + D^n(t, t-H+L) \tag{4}$$

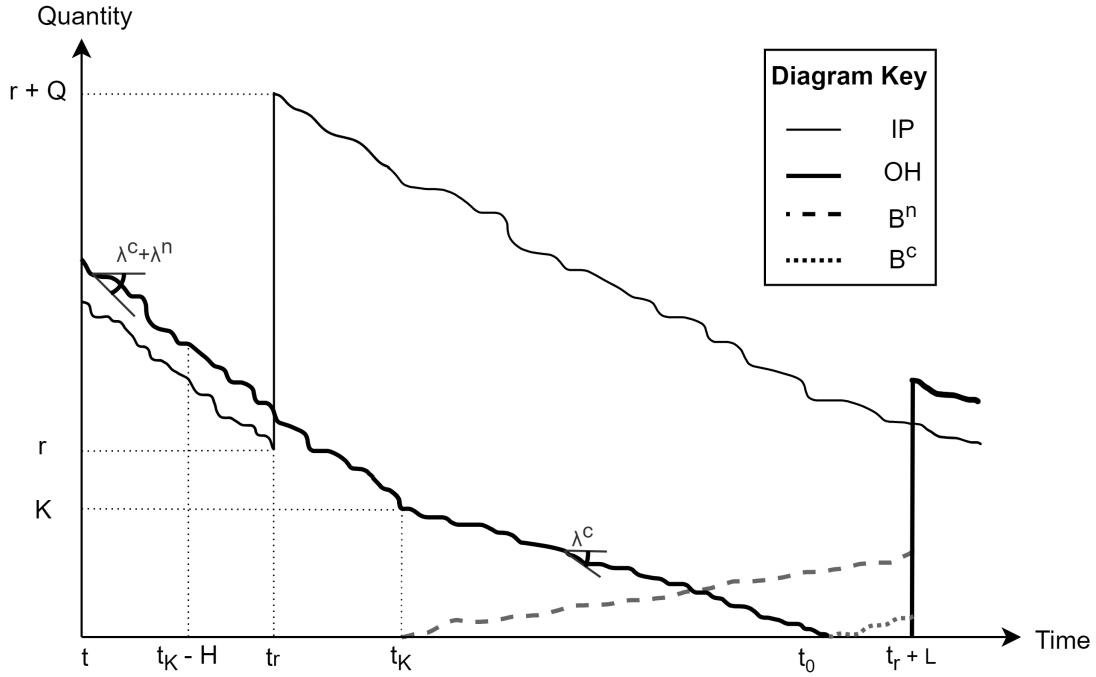
$$Y^n(t) = D^n(t-H, t) \tag{5}$$

Observe that the non-critical orders that are placed within the interval $(t-H+L, t+L)$ are not due before $t+L$ and thus, they do not have a net impact on the inventory level. Moreover, although the non-critical orders that are placed within $(t-H, t)$ are due within the interval $(t, t+H)$, which is in $(t, t+L)$, they do not affect the inventory level at time $t+L$. The reason is that, $IP(t)$ has already taken into account $D^n(t-H.t)$ by Equations 3 and 5. Due to Equation 2, we can conclude that $IL(t+L)$ is given by $IP(t) - \hat{D}(t, t+L)$.

Figure 1 illustrates a typical cycle of a (Q, r, K) review policy of *Model 1*. It depicts the inventory position, the on-hand stock, and the critical and non-critical backorders, which are denoted as B^c and B^n, respectively. The average inclination of the OH line shows the depletion rate, as defined in Subsection 3.1. As already mentioned, the on-hand can be greater, smaller or equal to the inventory position. In the illustrated example, OH happens to be larger than IP, from time t until t_r. At time t_r, IP drops to r units. Thus, we place an order of Q and IP increases to $r+Q$, instantly. At time t_k, OH hits the threshold level K, which means that the non-critical backorders will no longer be satisfied, until the next replenishment order arrives. Note that since we are considering *Model 1*, non-critical demands are due H time units after their arrival and thus backlogged at their due time. Therefore, non-critical backorders are the non-critical orders received after $t_k - H$. As a result, at t_k, the depletion rate of the on-hand stock decreases from $\lambda^c + \lambda^n$ to λ^c. The critical orders are still being satisfied until a stock-out, at time t_0. All the demand orders that arrive after t_0 are backlogged, regardless of their class. At time $t_r + L$, the replenishment order placed at t_r arrives, and the backorders are cleared according to the priority clearing mechanism. Therefore, if $IP(t_r + L) - \widehat{D}(t_r, t_r + L) \geq K + B^c(t_r + L) + B^n(t_r + L)$, all the backorders will be satisfied. In other words, if $\widehat{D}(t_r, t_r + L) \leq r + Q - K$, all the backorders are cleared and the level of on-hand stock becomes $r + Q - \widehat{D}(t_r, t_r + L)$, which is equivalent to $IP(t_r) - \widehat{D}(t_r, t_r + L)$. Otherwise, the clearing mechanism will not clear all the backorders but instead some backorders will remain unsatisfied until the arrival of the next replenishment order. In that occasion, there are two sub-cases: The first one appears when $D^c(t_k, t_r + L) \geq Q$. Then, even by satisfying only the critical backorders the on-hand stock will drop to or below K. Thus, we will satisfy only the critical backorders on a first-come, first-served basis, until all of them are satisfied or no physical stock is left. The second sub-case appears when $D^c(t_k, t_r + L) < Q$. Then, all the critical backorders will be satisfied, but the on-hand stock will drop to

K before satisfying all the non-critical backorders. Thus, only some of the non-critical backorders will be satisfied, while the remaining ones may be satisfied only after the next replenishment arrival.

Figure 1: Typical Cycle of Model 1



Although the reorder point is larger than the threshold level in the example illustrated in Figure 1, there are no restrictions on the range of its value.

As already mentioned, the class-specific service levels are defined as class-specific order fill rates. Therefore, their values depend on the probability that they will be satisfied at their due times. We know that non-critical demand can be satisfied, if and only if on-hand stock is at least K, and critical class demand can be satisfied, if and only if on-hand stock is above zero. Thus, as Vicil (2021b) did, we assume $P_\infty(\cdot)$, the steady-state probability distribution of a random process, and then the expected service levels can be expressed, using the PASTA principle.

$$\beta^n(Q,r,K) = 1 - P_\infty\left(OH \leq K | (Q,r,K)\right) \tag{6}$$

$$\beta^c(Q,r,K) = 1 - P_\infty\left(OH = 0 | (Q,r,K)\right) \tag{7}$$

The objective of the service level optimization problem is to find the optimal inventory parameters that minimize the expected holding cost rate subject to class-specific service level requirements. Thus, if we represent the minimum required service level of the critical and non-critical demand class as $\bar{\beta}^c$ and $\bar{\beta}^n$, respectively, we can formulate our problem as following.

$$\begin{aligned} \min_{Q,r,K} \quad & hE[OH] \\ \text{s.t.} \quad & \beta^c(Q,r,K) \geq \bar{\beta}^c \\ & \beta^n(Q,r,K) \geq \bar{\beta}^n \\ & Q, K \geq 0 \\ & \bar{\beta}^c > \bar{\beta}^n > 0 \end{aligned} \tag{8}$$

Note that by $E[\cdot]$, we symbolize the expected value of a parameter and by h the inventory holding cost per unit per time. Since in our case h is assumed to be constant, then minimization of the expected holding cost rate means minimization of the expected on-hand stock. Thus, our objective function can be written as following.

$$\min_{Q,r,K} \quad E[OH] \tag{9}$$

3.3 Cost Optimization Model

To optimize the total inventory cost, we should consider the ordering/setup, holding, and shortage cost which comprise the total cost. If A denotes the fixed ordering/setup cost incurred per replenishment cycle (i.e. the time between the placement and the arrival of a replenishment order), then the ordering/setup cost per time unit is $A\frac{\lambda^c+\lambda^n}{Q}$. Moreover, if h denotes the holding cost per quantity unit per time unit, the holding cost incurred per time unit is equal to $hE[OH]$. Finally, denoting the fixed shortage cost per critical or non-critical unit short per time unit by b_c and b_n, respectively, the shortage cost per time unit can be expressed as $b_cE[B^c] + b_nE[B^n]$. Hence, the expected total inventory cost incurred per time unit, $E[C(Q,r,K)]$, can be found through Equation 10.

$$E[C(Q,r,K)] = A\frac{\lambda^c+\lambda^n}{Q} + hE[OH] + b_cE(B^c) + b_nE(B^n) \tag{10}$$

The objective of our cost optimization model is to find the optimal inventory parameters that minimize the total expected cost. Thus, we can formulate our problem as following.

$$\begin{aligned} \min_{Q,r,K} \quad & E[C(Q,r,K)] \\ \text{s.t.} \quad & Q, K \geq 0 \end{aligned} \tag{11}$$

3.4 Deriving Fill Rates

In this subsection, we derive expressions for the expected service levels for given inventory position and threshold level. Our analysis is inspired by the derivation of

the service levels by Koçağa and Şen (2007), which is given for the $(S-1, S)$ inventory model in a similar setting. The notations $\beta^n(t)$ and $\beta^c(t)$ denote the expected fill rates of non-critical and critical demand, respectively, at a specified time t. In other words, they denote the expected probability of not hitting a stock-out of class-specific demand, during the next replenishment cycle, considering the state of the inventory system at time t . Note, that we are examining the case in which non-critical demand is due after a fixed DLT, whereas the case in which critical demand is due after a DLT is presented in the Appendix A.

First, we derive an exact expression for the expected service level of the non-critical demand. Consider the interval $(t, t+L]$. We know that if the on-hand stock level drops to or below the threshold level, we no longer satisfy the non-critical demand, until the next replenishment arrival. Hence, to satisfy a non-critical demand order that is due at $t+L$, the on-hand stock must be at least K, at that moment, and there must be no backorders. The service level of the non-critical demand is the percentage of non-critical demand orders that are satisfied at their due time, which, on the long run, equals the probability that a non-critical demand order is satisfied at its due time. Without loss of generality, let us suppose that an arbitrary non-critical demand due time is $t+L$. In order to be able to satisfy the non-critical demand, it should hold $OH(t+L) > K$, which implies $B^c(t+L) = B^n(t+L) = 0$. Hence, we can calculate the non-critical fill rate as following.

$$\beta^n(K, IP(t)) = P\big[OH(t+L) > K\big] = P\big[(\widehat{D}(t, t+L) < IP(t) - K\big] \tag{12}$$

The above formula holds, as from Equation 2, we know that:

$$IL(t+L) = OH(t+L) - B^c(t+L) - B^n(t+L)$$

Moreover, since in our case $B^c(t+L) = B^n(t+L) = 0$, then $IL(t+L) = OH(t+L)$, and according to Equation 3:

$$IP(t) - \widehat{D}(t, t+L) = IL(t+L) = OH(t+L)$$

We know that demand arrivals follow a Poisson distribution with rates λ^c and λ^n, according to their criticality. We also know that the probability mass function of a discrete random variable X that follows a Poisson distribution with parameter $\lambda \geq 0$ is $f(k;\lambda) = P[X = k] = \frac{\lambda^k e^{-\lambda}}{k!}$, whereas for a given average rate r (per time unit), at which arrivals occur, we can calculate the probability that k arrivals occur in the interval of length T_0 time units, as $f(k; rT_0) = \frac{(rT_0)^k e^{-rT_0}}{k!}$.

Considering Equation 4, the demands that have net impact value on the inventory level arrive with rates $\lambda_c + \lambda_n$ in the interval $(t, t-H+L]$ and λ_c in the interval $(t-H+L, t+L]$. Therefore we can compute the non-critical fill rate exactly, by the

following proposition.

Proposition.

$$\beta^n(K, IP(t)) = \sum_{i=0}^{IP(t)-K-1} e^{-[\lambda_c L+\lambda_n(L-H)]} \frac{\left[\lambda_c L + \lambda_n(L-H)\right]^i}{i!} \tag{13}$$

Since, Equation 13 holds for an arbitrary time t, it also holds for the steady state distribution:

$$\beta^n(K, IP) = \sum_{i=0}^{IP-K-1} e^{-[\lambda_c L+\lambda_n(L-H)]} \frac{\left[\lambda_c L + \lambda_n(L-H)\right]^i}{i!} \tag{14}$$

Now, we derive an approximate expression for the critical demand fill rate. Consider again the interval $(t, t+L]$ and assume that the on-hand stock at time $t+L$ is $OH(t+L)$. Then, the inventory system will satisfy an incoming critical demand that arrives and is due at time $t + L$ if and only if the on-hand stock is above zero at $t + L$.

$$\beta^c(K, IP(t)) = P\big[OH(t + L) > 0\big] \tag{15}$$

Although the above equation seems to be similar to Equation 12, we cannot use the same method to calculate it exactly, and thus give an approximation.

Contrary to the inventory position, the on-hand stock does not have a constant depletion rate and it is also affected by the demand lead time. Therefore, analyzing its distribution is very complicated. To simplify our analysis, we make the following assumptions.

- The order quantity Q is large enough with respect to the reorder point r.
- The reorder point r is larger than the threshold level K.

Since $r > K$, the inventory system continues to satisfy the demand of both priority classes even after the on-hand drops to or below the reorder point. Thus, until r, the on-hand stock has the same depletion rate with the inventory position $(\lambda_c + \lambda_n)$. We assume Q is large enough ($Q >> r$), so that in the majority of the cycle, OH is greater than 0 and after the replenishment order is received from the resupply, it is sufficient to clear all the existing backorders and restore the inventory level above r. We will use this assumption in our heuristic.

Since it is very difficult to determine $P[OH(t + L) > 0]$ from the knowledge of $IP(t)$ and total demand process over $(t, t+L]$, we will use the following approximation assumption to analytically derive the critical class fill rate level.

Approximation Assumption 1. At a random point t in time, $OH(t) = IP(t)$

We will condition on the on-hand stock level at time t to estimate the OH stock level at time $t+L$. Note that $IP(t)$ is always between $r+1$ and $r+Q$. Hence, the approximation assumption implies that $OH(t)$ is also between $r+1$ and $r+Q$.

To further analyze Equation 15, we split the corresponding probability according to two cases. The first one is the case where the on-hand stock remains above the threshold level throughout the interval $(t, t+L)$. Then, all the critical demand orders that arrive within $(t, t+L)$ are satisfied. To be more precise, in that case we satisfy all demand orders that are due in this interval, regardless their priority. The second case applies when the on-hand stock drops to the threshold level between t and $t+L$. Then, the critical demand orders that arrive between t_k and $t+L$ should be less than K, in order to avoid a stock-out situation. For this case, we need to consider the fact that after the on-hand stock reaches the threshold level, according to the inventory policy, only the critical orders can be satisfied. Thus, as shown in Figure 1, the depletion rate of the on-hand stock is $\lambda^c + \lambda^n$ until t_k, and then it reduces to λ^c. This means that the on-hand stock must first reach the threshold level when the depletion rate is $\lambda^c + \lambda^n$ and then from the threshold level to zero when the depletion rate is λ^c.

To calculate the above probabilities, we introduce the hitting time T, the first time that OH drops to the threshold level K. Hence, $T = t_k - t$. Moreover, we split the interval $(t, t+L]$ to $(t, t+L-H]$ and $(t+L-H, t+L]$. Then the following holds.

$$\begin{aligned}\beta^c(K, IP(t)) = {} & P[T > L] \\ & + P[\widehat{D}^c(t+T, t+L) < K \mid T \leq L-H] \cdot P[T \leq L-H] \\ & + P[\widehat{D}^c(t+T, t+L) < K \mid L-H \leq T \leq L] \cdot P[L-H \leq T \leq L]\end{aligned} \tag{16}$$

where,

$$P[T > L] = P[\widehat{D}(t, t+L) < IP(t) - K] \tag{17}$$

Since by Approximation Assumption 1, we assumed $OH(t) = IP(t)$, then $IL(t+L)$ will be a function of $\hat{D}(t, t+L)$, which is given by Equation 4. Hence, from a random point t in time, until $L-H$ time units later, the total demand process is governed by the total demand arrival rate $(\lambda^c + \lambda^n)$. From $L-H$ until L, the total demand is governed only by the critical demand arrival rate (λ^c).

Approximation Assumption 2. In a replenishment cycle with time interval $(t, t+L]$, non-critical demand has no impact on the inventory level between $t+L-H$ and $t+L$.

If T is in the interval $(t, t+L-H]$, our inventory system satisfies all the demand until T in a first-come, first-served basis, regardless of the priority. Therefore, due to Approximation Assumptions 1 and 2, the cumulative distribution function of T is:

$$F_1(K, IP(t), y) = P[T \leq y] = P[D^c(t, t+y) + D^n(t-H, t-H+y) \geq IP(t) - K] \quad (18)$$

In the above formula, we consider the non-critical demand orders that arrive from $t - H$ until $t - H + y$, because those arrivals affect the inventory position within the interval $(t, t + y]$.

It is known that the probability that k arrivals, with rate λ, occur over a specified time y can be expressed by Erlang distribution with probability density function $f(y; k; \lambda) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}$. Therefore, we can express $f_1(t, y) = \frac{dF_1(t,y)}{dy}$ (i.e. the probability density function of T) as the Erlang-$\big(IP(t) - K\big)$ with rate $\lambda^c + \lambda^n$.

$$f_1(K, IP(t), y) = (\lambda^c + \lambda^n)^{IP(t)-K} e^{-(\lambda^c+\lambda^n)y} \frac{y^{IP(t)-K-1}}{(IP(t) - K - 1)!} \quad (19)$$

(see Appendix B for proof)

By Approximation Assumption 2, we consider the critical demand within the interval $(t, t + y]$ and the non-critical demand within the interval $(t, t + L - H]$.

If we denote the cumulative distribution function of T for the interval $(t+L-H, t+L]$ by $F_2(K, IP(t), y)$, then the following expression holds.

$$F_2(K, IP(t), y) = P[T \leq y] = P[D^c(t, t + y) + D^n(t, t + L - H) \geq IP(t) - K] \quad (20)$$

Therefore, we can express $f_2(t, y) = \frac{dF_2(t,y)}{dy}$ as:

$$f_2(K, IP(t), y) = \lambda^c e^{-(\lambda^c y + \lambda^n (L-H))} \frac{\big[\lambda^c y + \lambda^n (L - H)\big]^{IP(t)-K-1}}{(IP(t) - K - 1)!} \quad (21)$$

(see Appendix B for proof)

Using Equations 19 and 21, we can approximate the critical service level, as following:

$$\begin{aligned} \beta^c(K, IP(t)) = & \sum_{i=0}^{IP(t)-K-1} e^{-[\lambda_c L + \lambda_n (L-H)]} \frac{\big[\lambda_c L + \lambda_n (L - H)\big]^i}{i!} \\ & + \int_0^{L-H} f_1(K, IP(t), y) \cdot \sum_{i=0}^{K-1} e^{-\lambda^c (L-y)} \frac{[\lambda^c (L - y)]^i]}{i!} \, dy \\ & + \int_{L-H}^{L} f_2(K, IP(t), y) \cdot \sum_{i=0}^{K-1} e^{-\lambda^c (L-y)} \frac{[\lambda^c (L - y)]^i]}{i!} \, dy \end{aligned} \quad (22)$$

Since, Equation 22 holds for an arbitrary time t, it also holds for the steady state distribution:

$$\begin{aligned}\beta^c(K, IP) = &\sum_{i=0}^{IP-K-1} e^{-[\lambda_c L+\lambda_n(L-H)]} \frac{\left[\lambda_c L + \lambda_n(L-H)\right]^i}{i!} \\ &+ \int_0^{L-H} f_1(K, IP, y) \cdot \sum_{i=0}^{K-1} e^{-\lambda^c(L-y)} \frac{[\lambda^c(L-y)]^i]}{i!}\, dy \\ &+ \int_{L-H}^{L} f_2(K, IP, y) \cdot \sum_{i=0}^{K-1} e^{-\lambda^c(L-y)} \frac{[\lambda^c(L-y)]^i]}{i!}\, dy\end{aligned} \tag{23}$$

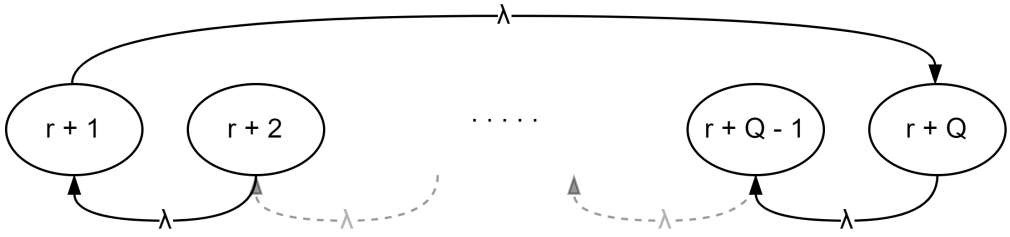
Solving Equations 12 and 22, we can compute the expected class-specific fill rates, with respect to the inventory position and the threshold level. However, the value of the inventory position depends on the current state in which the inventory system is. Hence, to compute the expected class-specific fill rates regardless the state of the inventory system (regardless time t), we analyze the steady-state probability of the inventory position. Although we consider the DLT for the non-critical customer class, the IP process is still governed by the demand arrival process, by the IP definition we made. Since the units are demanded one at a time and according to our inventory policy a replenishment order with quantity Q is placed when the inventory position drops at the reorder point r, it holds that $r + 1 \leq IP \leq r + Q$.

Lemma 1. Assume that a continuous review policy (r, Q) is applied. The unit orders arrive independently, according to a Poisson distribution and the orders that cannot be satisfied at their due time are backlogged. Then the steady-state probability of the inventory position follows Uniform distribution with bounds $[r + 1, r + Q]$.

Proof.

Suppose that demands are independent Poisson processes, with fixed size of one item and arrival rate λ. Then Figure 2 illustrates the inventory position as a continuous-time Markov chain with states $[r + 1, r + 2, ..., r + Q - 1, r + Q]$.

Figure 2: Continuous-Time Markov Chain of Inventory Position



If we denote by π_i the stationary probability that the inventory position at a random

time is $r + j$, then:

$$\pi_i = \frac{1}{Q} \tag{24}$$

Since the possible states of the model have equal stationary probabilities π_i, it holds that $IP \sim U[r + 1, r + Q]$.

Knowing that IP can get any value of the interval $[r + 1, ..., r + Q]$, with the same probability, we can now calculate the class-specific fill rates for given inventory control policy parameters, regardless time t.

$$\beta^i(Q, r, K) = \frac{\sum_{IP=r+1}^{r+Q} \beta^i(IP, K)}{Q} \qquad \forall i \in \{c, n\} \tag{25}$$

3.5 Estimating Performance Measures

Moreover, we can estimate some significant performance measures. Those are the expected on-hand stock, the class-specific backorders and the class-specific orders that are placed but not yet due.

It is possible that at a random time $t + L$ there are backorders in the system, which affect the level of the on-hand stock. Thus, we cannot ignore them, which make their analysis more complicated. From Equations 2 and 3, the on-hand stock at a random time $t + L$ is given by the following formula.

$$OH(t + L) = IP(t) - \widehat{D}(t, t + L) + B^c(t) + B^n(t) \tag{26}$$

Since 26 holds for an arbitrary point t in time, it should also hold for the steady-state. Hence, in steady-state:

$$OH = IP - \widehat{D}(L) + B^c + B^n \tag{27}$$

To further analyze this formula, we need to consider the fact that even after replenishing the inventory, some backorders may not be fulfilled due to the threshold level. Despite the complexity of this analysis, we can simply determine the expected on-hand stock as following.

$$E[OH] = E[IP] - E[\widehat{D}(L)] + E[B^c] + E[B^n] \tag{28}$$

Since $IP \sim U[r + 1, r + Q]$, the expected inventory position is equal to the mean value of the interval $[r + 1, r + Q]$. Hence, the expected IP can be calculated exactly.

$$E[IP] = \frac{2r + Q + 1}{2} \tag{29}$$

The demand with net impact on the inventory level that is expected to arrive within a period of L time units can be calculated as following.

$$E[\widehat{D}(L)] = \lambda^c L + \lambda^n (L - H) \tag{30}$$

A critical order is backlogged if there is no on-hand stock at its due time. At a random time $t + L$, there are x critical backorders if the on-hand stock drops to the threshold level at $t_k \leq t + L$ ($OH(t_k) = K$), and there are $K + x$ critical orders with impact on the inventory level, placed in the interval $(t_k, T + L]$. Hence, the probability mass function of the expected time-weighted class-specific backorders at any point in time can be approximated as following:

$$P[B^c(t) = x] = P[t_k \leq t + L] \cdot P[\widehat{D}^c(t_k, t + L) = K + x] \tag{31}$$

Similar to the analysis of the service levels, we split the time interval $(t, t + L]$ to $(t, t - H + L]$ and $(t - H + L, t + L]$. For given inventory position and threshold level, we can approximate the probability that there are x critical backorders at a random time t.

$$\begin{aligned} P[B^c(IP(t), K) = x] = &\int_{}^{L-H} f_1(K, IP(t), y) \cdot e^{-\lambda^c (L-y)} \frac{[\lambda^c (L - y)]^{K+x}]}{(K + x)!} \, dy \\ &+ \int_{L-H}^{L} f_2(K, IP(t), y) \cdot e^{-\lambda^c (L-y)} \frac{[\lambda^c (L - y)]^{K+x}]}{(K + x)!} \, dy \end{aligned} \tag{32}$$

A non-critical order is backlogged at its due time when the on-hand stock is at or below the threshold level. At a random time $t + L$, there are x non-critical backorders if the on-hand stock reaches the threshold level at $t_k \leq t + L$ ($OH(t_k) = K$), and there are x non-critical orders with impact on the inventory level, placed in the interval $(t_k, T + L]$.

$$P[B^n(t) = x] = P[t_k \leq t + L] \cdot P[\widehat{D}^n(t_k, t + L) = x] \tag{33}$$

To calculate the above probabilities, consider that the non-critical orders are due H time units after their arrival. Thus, if $t_k > t - H + L$, no non-critical order arrived after t_k is backlogged before $t + L$. Hence, we use Equation 34 to approximate the corresponding probabilities.

$$P[B^n(IP(t),K)=x]=\int_0^{L-H} f_1(K,IP(t),y)\cdot e^{-\lambda^n(L-H-y)}\frac{[\lambda^n(L-H-y)]^x]}{x!}\,dy \tag{34}$$

Considering the assumptions that Q is large enough with respect to r, and r is larger than K, the expected class-specific backorders are given by the following equation.

$$E[B^i]=E[(IL^-)^i]=\frac{\sum_{x=1}^{\infty}\sum_{IP(t)=r+1}^{r+Q} x\cdot P\Big[B^i\big(IP(t),K\big)=x\Big]}{Q} \qquad \forall i\in\{c,n\} \tag{35}$$

As a result, we can compute the expected on-hand stock by using Equation 28, which is the objective of our optimization problem, as presented in Subsection 3.2.

Moreover, to calculate the expected number of non-critical orders that have been placed but not due yet ($E[Y^n]$), considering the Equation 5, we can use the following formula.

$$E[Y^n]=E\big[D^n(t-H,t)\big]=\lambda^n H \tag{36}$$

Therefore, the expected Y^n is the number of non-critical order arrivals in an interval period equal to H, regardless time t. In *Model 1*, critical demand is due immediately and thus, $E[Y^c]=0$.

3.6 Service Level Optimization Algorithm

As mentioned in Subsection 3.2, to optimize our inventory system, we need to minimize the holding cost, and thus the on-hand stock, while keeping sufficient class-specific service levels (see Equations 8 and 9). Assume that (Q^*,r^*,K^*) are the optimal policy parameters and S is the set that contains all possible combinations of the policy parameters. Then $(Q^*,r^*,K^*)\in S$. Though, to simplify our problem, we have made some assumptions (see Subsection 3.4). Thus, we introduce the subset $S^{'}\subseteq S$, which complies with the model assumptions and assume $(Q^*,r^*,K^*)\in S^{'}$.

For given minimum required critical and non-critical service levels, we find the optimal policy parameters based on a brute force approach, as shown in Algorithm 1.

In other words, the first step is to specify the subset $S^{'}\subseteq S$, which means to specify the limits of the inventory parameters (Q,r,K). This should be done according to the model assumptions. To be more precise, at first we set 1 as the lower limit of r. The reorder point cannot be smaller than 1, since $r>K\geq 0$. The upper limit of r is set empirically, to avoid checking too many possible combinations of policy parameters

Algorithm 1 Optimization algorithm for specified minimum required service levels

specify $S^{'}$
$OH^* \leftarrow large\ number$
for $(Q, r, K) \in S^{'}$ **do**
    `compute` $\beta^c(Q, r, K)$ `and` $\beta^n(Q, r, K)$`, using Equations 25`
    if $\beta^c(Q, r, K) \geq \bar{\beta^c}$ `and` $\beta^n(Q, r, K) \geq \bar{\beta^n}$ **then**
        `compute` OH`, using Equation 28`
        if $OH < OH^*$ **then**
            $OH^* \leftarrow OH$
            $(Q^*, r^*, K^*) \leftarrow (Q, r, K)$
        end if
    end if
end for
return (Q^*, r^*, K^*)
return $\beta^c(Q^*, r^*, K^*)$, $\beta^n(Q^*, r^*, K^*)$

and having high running times. The threshold level K can get the values $[1, r-1]$. According to our model, when rationing is applied, it holds that $\bar{\beta^c} \geq \bar{\beta^n}$, and the total demand fill rate (i.e. the percentage of total demand satisfied immediately from on-hand stock at their due dates) is $\beta \in [\bar{\beta^n}, \bar{\beta^c}]$. If we set the value of $\bar{\beta^n}$ as the value of the minimum required total demand fill rate ($\bar{\beta} = \bar{\beta^n}$), then we will be able to achieve the minimum required non-critical fill rate but we may fail to meet the critical fill rate requirement. However, we know that achieving a total demand fill rate below $\bar{\beta^n}$ will certainly lead to failure to meet the service level requirements. On the other hand, if we set the value of $\bar{\beta^c}$ as the value of the required total demand fill rate ($\bar{\beta} = \bar{\beta^c}$), then we expect that both required class-specific fill rates will be achieved. Moreover, we know that when all the other parameters are fixed, by increasing Q, the expected on-hand stock will be also increased. Since our goal is to minimize $E(OH)$, while meeting the service level requirements, we also aim to minimize Q, while meeting the service level requirements, if all the other parameters are fixed. Furthermore, since based on the model assumptions the order quantity of Q should be large enough with respect to the reorder point r, we assume that Q cannot be smaller than $2r$. Taking all the above into consideration, for given r, the lower limit of the quantity Q is set as the minimum required inventory that ensures the non-critical service level requirement is met, when Q meets the model assumptions i.e. $Q_{min}(r) = argmin\{Q \geq 2r : \bar{\beta}(Q, r) \geq \bar{\beta^n}\}$. We define $\bar{\beta}(Q, r)$ as the minimum required total demand fill rate when $K = 0$. If Q, r and the demand arrival rates are fixed, the highest non-critical service level can be achieved when $K = 0$. Thus, if all the other parameters are fixed, the minimum Q can be found when $K = 0$. Similarly, the upper bound of Q is set as the minimum required inventory that ensures the critical service level requirement is met, when Q meets the model assumptions, i.e. $Q_{max}(r) = argmin\{Q \geq 2r : \bar{\beta}(Q, r) \geq \bar{\beta^c}\}$. The minimum

required total demand fill rate for $K = 0$ is expressed as following.

$$\bar{\beta}(Q,r) = \frac{1}{Q}\sum_{j=1}^{r+Q}\sum_{k=max\{r+1,j\}}^{r+Q} P\big[\widehat{D}(t,t+L) = k-j\big] \tag{37}$$

Then, we set the initial expected on-hand stock as a large number. Each solution that meets the service level requirements leads to lower expected on-hand stock and thus, is a better solution than the initial. Rows 3-12 in Algorithm 1 represent a *for* loop, in which we evaluate the performance of our system for each combination (Q,r,K). If the new combination (Q,r,K) meets the service level requirements (row 5) and if the on-hand stock is smaller than the on-hand stock of the current best solution (row 7), then the combination (Q,r,K) becomes the current best solution and the on-hand stock which results from those policy parameters becomes the current minimum objective value. The best solution is the combination $(Q^*,r^*,K^*) \in S^{'}$ with the minimum objective value.

3.7 Cost Optimization Algorithm

To optimize our inventory system, in terms of its expected cost, we need to minimize the total inventory cost, as shown in Equations 10 and 11. To achieve that, we follow the same logic and similar steps with the service level optimization algorithm. To be more precise, assume that (Q^*,r^*,K^*) are the optimal parameters which belong to the subset $S^{'}$, as described in the previous subsection. The parameters (Q^*,r^*,K^*) can be found, using Algorithm 2, which is based on a brute force approach.

Algorithm 2 Cost optimization algorithm

1: specify $S^{'}$
2: $C^* \leftarrow large\ number$
3: **for** $(Q,r,K) \in S^{'}$ **do**
4: `compute` $C(Q,r,K)$`, using Equation 10`
5: **if** $C(Q,r,K) < C^*$ **then**
6: $C^* \leftarrow C(Q,r,K)$
7: $(Q^*,r^*,K^*) \leftarrow (Q,r,K)$
8: **end if**
9: **end for**
10: **return** (Q^*,r^*,K^*)
11: **return** $C(Q^*,r^*,K^*)$

Again, the first step is to specify the subset $S^{'} \subseteq S$. Then we set the initial minimum total cost as a large number. Each combination (Q,r,K) leads to better solution than the initial one. The combination from the set $S^{'}$ that results in the minimum total cost consists the optimal inventory parameters and the objective value is the total expected cost when applying these parameters.

4 Numerical Study

Since there is no existing algorithm in the literature that computes the steady-state probabilities of our problem, to evaluate the performance of our heuristic, we compare the results against a simulation.

At first, in Subsection 4.1, we compute the class-specific fill rates, for given parameters. We compare the results of our approximation with those of a simulation and identify how the system performs when its parameters are changed. In Subsection 4.2, we use Algorithm 1 to find the optimal inventory control policy parameters, that minimize the expected inventory holding cost, while the system meets some specified service level requirements. Similarly, in Subsection 4.3, we use Algorithm 2 to find the optimal parameters that minimize the expected total inventory cost. The results are analyzed and they lead to significant conclusions. Finally, in Subsection 4.4, we evaluate the performance of our model compared to models in which either inventory rationing or demand lead time is not applied, and we indicate the benefits of incorporating them into our model.

4.1 Fill Rates Calculation

First, we compare the results of our heuristic against those of the simulation. The simulation is performed for $1,000,000$ order arrivals, regardless of their priority. For given policy parameters, arrival rates, and lead times, we simulate the inventory system to find the corresponding class-specific fill rates and performance measures. Next, for the same parameters we make the corresponding estimations according to the formulas presented in Section 3.

Tables 2 and 3 contain results for both *Model 1* and *Model 2*. The acronyms *sim* and *approx* refer to simulation and approximation, respectively, whereas *AE* stands for the absolute error, which is the difference between their service levels. For all the instances, we set the replenishment lead time L to 0.5 and the demand lead time H to 0.1. We randomly choose the rest parameters for some instances. Then, for each instance, we fix all of those parameters except for one. Table 2 presents the instances which lead to critical service level of at least 99%, whereas Table 3 the instances which lead to critical service level between 90% and 99%, according to the simulation. For example, the first row of Table 2 represents the instance in which $\lambda^c = 1, \lambda^n = 4, r = 3, Q = 7$ and $K = 2$. Then, we fix all those parameters except for λ^c and we evaluate the system for all $\lambda^c \in [1, 7]$. The inventory system leads to service level of at least 99% only for $\lambda^c = 1$ and thus the rest instances with service level above 90% are presented in Table 3. The same methodology is followed for all parameters.

We observe that for critical service levels above 99%, the approximations are very close to the simulation results. The maximum absolute error of the 18 instances in Table

λ^c	λ^n	r	Q	K	DLT: non-critical (Model 1)				DLT: critical (Model 2)			
					$\beta^n_{exact}(\%)$	$\beta^c_{sim}(\%)$	$\beta^c_{approx}(\%)$	$AE^c(\%)$	$\beta^n_{exact}(\%)$	$\beta^c_{sim}(\%)$	$\beta^c_{approx}(\%)$	$AE^c(\%)$
1	4	3	7	2	82.54	99.73	99.52	**0.21**	78.72	99.77	99.77	**0.00**
7	10	10	20	5	86.42	99.91	99.75	**0.15**	85.10	99.86	99.95	**0.09**
8	10	10	20	5	84.20	99.78	99.53	**0.25**	83.30	99.73	99.90	**0.17**
10	10	10	20	5	79.58	99.33	98.71	**0.62**	79.58	99.30	99.68	**0.38**
10	7	10	20	5	85.10	99.54	99.22	**0.32**	86.42	99.65	99.82	**0.17**
10	8	10	20	5	83.30	99.50	99.07	**0.44**	84.20	99.56	99.78	**0.22**
10	10	10	20	5	79.58	99.33	98.71	**0.61**	79.58	99.29	99.68	**0.39**
10	12	10	20	5	75.75	99.18	98.30	**0.88**	74.79	99.02	99.56	**0.54**
7	10	7	20	5	72.47	99.34	98.73	**0.61**	70.98	99.34	99.60	**0.26**
7	10	9	20	5	82.08	99.81	99.55	**0.25**	80.66	99.81	99.89	**0.09**
7	10	12	20	5	93.32	99.98	99.93	**0.04**	92.36	99.98	99.99	**0.01**
10	7	9	20	5	80.66	99.21	98.66	**0.55**	82.08	99.35	99.64	**0.29**
10	7	10	20	5	85.10	99.55	99.22	**0.33**	86.42	99.65	99.82	**0.17**
10	7	12	20	5	92.36	99.88	99.78	**0.11**	93.32	99.91	99.96	**0.05**
10	7	10	20	3	92.36	99.17	98.71	**0.45**	93.32	99.37	99.66	**0.29**
10	7	10	20	5	85.10	99.53	99.22	**0.31**	86.42	99.65	99.82	**0.17**
10	7	10	20	6	80.66	99.65	99.41	**0.24**	82.08	99.74	99.87	**0.13**
10	7	10	20	8	70.98	99.80	99.66	**0.14**	72.47	99.85	99.93	**0.08**

Table 2: Performance of approximation for random parameters and critical service level at least 99 ($L = 0.5, H = 0.1$)

λ^c	λ^n	r	Q	K	DLT: non-critical (Model 1)				DLT: critical (Model 2)			
					$\beta^n_{exact}(\%)$	$\beta^c_{sim}(\%)$	$\beta^c_{approx}(\%)$	$AE^c(\%)$	$\beta^n_{exact}(\%)$	$\beta^c_{sim}(\%)$	$\beta^c_{approx}(\%)$	$AE^c(\%)$
1	4	3	7	2	76.11	98.64	97.87	**0.76**	73.46	98.26	98.95	**0.70**
7	10	10	20	5	69.45	96.50	94.99	**1.51**	68.11	96.14	97.42	**1.28**
8	10	10	20	5	62.72	93.26	90.96	**2.30**	62.72	93.29	95.14	**1.85**
10	10	10	20	5	78.72	96.01	95.55	**0.46**	82.54	97.28	97.71	**0.43**
10	7	10	20	5	73.46	95.09	94.10	**0.99**	76.11	96.09	96.91	**0.81**
10	8	10	20	5	62.72	93.29	90.96	**2.32**	62.72	93.32	95.14	**1.82**
10	10	10	20	5	52.10	91.44	87.65	**3.79**	49.52	90.45	93.22	**2.77**
10	12	10	20	5	77.20	98.88	98.07	**0.82**	77.68	98.94	99.49	**0.55**
7	10	7	20	5	74.79	98.36	97.23	**1.13**	75.75	98.56	99.23	**0.67**
7	10	9	20	5	67.50	98.86	98.00	**0.86**	66.00	98.86	99.28	**0.42**
7	10	12	20	5	66.00	96.24	94.80	**1.44**	67.50	96.79	97.84	**1.06**
10	7	9	20	5	70.98	97.59	96.53	**1.06**	72.47	98.02	98.74	**0.72**
10	7	10	20	5	75.90	98.56	97.79	**0.77**	77.37	98.80	99.30	**0.51**
10	7	12	20	5	65.86	97.22	95.92	**1.30**	67.62	97.69	98.52	**0.83**
10	7	10	20	3	69.45	97.47	96.35	**1.12**	71.03	97.91	98.67	**0.77**
10	7	10	20	5	74.77	97.93	96.98	**0.94**	76.06	98.23	98.90	**0.67**
10	7	10	20	6	78.50	98.24	97.43	**0.80**	79.61	98.52	99.07	**0.54**

Table 3: Performance of approximation for random parameters and critical service level between 90 - 99 ($L = 0.5, H = 0.1$)

2 is less than 1%. To be more precise, the average absolute errors of the critical service levels for *Model 1* and *Model 2* are 0.36% and 0.20%, respectively. The non-critical service levels are found exactly by Equation 13 and verified through the simulation. Moreover, the absolute errors for *Model 1* are higher than those for *Model 2*. This is due to the fact that when the critical demand has demand lead time, we take into account a smaller number of critical order arrivals, as we ignore the ones which arrive in the examined time interval but have due time outside of this interval. Thus, there is less room for errors.

Similar observations can be made according to Table 3. The maximum absolute error of the 17 instances is less than 4%. The average absolute errors of the critical service levels for *Model 1* and *Model 2* are 1.32% and 0.96%, respectively. Thus, our heuristic seems to perform better for higher critical service levels.

The parameters of the instances in Table 4 are chosen randomly. Table 4 presents the performance of the inventory system for different combinations of policy parameters, arrival rates, replenishment times, and demand lead times.

From Tables 2, 3 and 4 we come to the following conclusions:

- When the arrival rates get higher, more demand orders need to be satisfied, and if the inventory policy remains the same, both critical and non-critical fill rates decrease.

- If r increases, while all the other parameters are fixed, the on-hand stock is replenished more often, and the maximum on-hand stock increases. Thus, both fill rates increase. Actually, the non-critical fill rate witnesses a great increase. The larger r is compared to K, the more chances exist for a replenishment to arrive before the on-hand stock drops to the threshold level and no further non-critical orders are satisfied until the next replenishment arrival.

- While Q rises and the other parameters remain constant, since the replenishment orders are larger, we achieve higher service levels.

- If K increases, we store more inventory units exclusively for the high-priority customers. Therefore, the critical service level increases. On the other hand, we stop satisfying our low-priority customers sooner, and thus the non-critical service level decreases.

- For fixed (Q, r, K) policy parameters, higher replenishment lead time L means higher backorder levels and thus, lower service levels. The non-critical service level is more dependent on the replenishment lead time, as there is less inventory that can satisfy low-priority customers.

λ^c	λ^n	r	Q	K	L	H	DLT: non-critical (Model 1)				DLT: critical (Model 2)			
							$\beta^n_{exact}(\%)$	$\beta^c_{sim}(\%)$	$\beta^c_{approx}(\%)$	$AE^c(\%)$	$\beta^n_{exact}(\%)$	$\beta^c_{sim}(\%)$	$\beta^c_{approx}(\%)$	$AE^c(\%)$
10	10	10	20	3	0.5	0.1	87.97	98.57	97.60	**0.97**	87.97	98.59	99.31	**0.72**
10	10	10	20	3	1	0.1	40.39	68.81	64.19	**4.63**	40.39	68.79	73.08	**4.29**
10	10	10	20	3	1	0.5	59.96	85.04	74.70	**10.34**	59.96	85.07	97.51	**12.45**
15	10	10	20	3	0.5	0.1	76.99	93.57	91.13	**2.44**	79.31	94.81	96.68	**1.88**
15	10	10	20	3	1	0.1	19.19	43.89	36.45	**7.43**	20.99	46.10	47.73	**1.63**
15	10	10	20	3	1	0.5	35.70	60.24	50.18	**10.05**	47.63	71.64	90.85	**19.21**
8	8	10	20	4	0.2	0.1	99.92	100.00	100.00	**0.00**	99.92	100.00	100.00	**0.00**
8	8	10	20	4	0.4	0.1	96.22	99.96	99.90	**0.06**	96.22	99.96	99.99	**0.03**
8	8	10	20	4	0.6	0.1	84.88	99.13	98.58	**0.55**	84.88	99.17	99.53	**0.36**
8	8	10	20	4	0.8	0.1	69.85	95.34	93.75	**1.59**	69.85	95.33	96.71	**1.38**
8	8	10	20	4	1	0.1	54.02	87.16	84.50	**2.66**	54.02	87.17	89.64	**2.47**
8	8	10	20	4	0.5	0.1	91.32	99.79	99.53	**0.25**	91.32	99.76	99.90	**0.14**
8	8	10	20	4	0.5	0.2	94.01	99.90	99.61	**0.30**	94.01	99.90	99.99	**0.09**
8	8	10	20	4	0.5	0.3	96.22	99.97	99.78	**0.19**	96.22	99.96	100.00	**0.04**
8	8	10	20	4	0.5	0.4	97.90	99.98	99.92	**0.07**	97.90	99.99	100.00	**0.01**
8	8	10	20	4	0.5	0.5	99.02	99.99	99.98	**0.01**	99.02	100.00	100.00	**0.00**

Table 4: Performance of approximation with varying system parameters

- As H increases, both service levels increase. High demand lead time allows the inventory system to be better prepared for the demands. The larger H is, the more we can focus on the class with no DLT. For example, if $L = 0.5$ and $H = 0.4$, the purpose of the current stock is mainly to satisfy the class with no DLT, whereas the class with DLT will be satisfied mainly by the outstanding orders.

The expected performance measures of all the instances in those 3 tables are presented in Appendix C.

4.2 Service Level Optimization Study

In this subsection, we perform an optimization study, according to Algorithm 1. We compare the results of a simulation and our heuristic method. The set $S^{'}$ of the feasible policy parameters is defined as presented in Subsection 3.6. As $S^{'}$ is too large, to reduce the running time of the simulation, we reduce the number of demand arrivals to $10,000$. Deshpande et al. (2003), who also study a (Q, r, K) model, state that approximately $10,000$ demand arrivals comprise sufficiently large sample that ensures stability of the estimations. After finding the optimal policy parameters, we perform simulations of $1,000,000$ demand arrivals to calculate the expected on-hand stock in each case.

Table 5 presents the optimal policy parameters of the simulation and our approximation, for 10 instances. In all instances, the replenishment lead time is 0.5, the demand lead time 0.1, the minimum required critical fill rate 99% and the minimum required non-critical fill rate 80%, while the combination of the class-specific demand arrival rates differ in each case. Five columns correspond to each model. The first 2 columns present the optimal policy parameters, while the next 2 columns, the expected on-hand stock according to the simulation and the approximation, respectively. The last column presents the percentage differences of the approximation from the simulation results.

For both models, our heuristic method finds a solution close enough to the optimal solution of the simulation. In 4 out of 20 instances, the approximated expected on-hand stock is more than 10% higher than the approximated. Though, in 6 out of 20 instances, the optimal policies are the same. Specifically, for *Model 1*, the highest difference is around 17.5%, and the average is approximately 7.7%, while for *Model 2*, the highest around 6% and the average around 1.2%. Note that the expected on-hand stocks for the optimal policy parameters, as found from both the simulation and the heuristic, were given by a simulation of $1,000,000$ demand arrivals. To be more specific, the optimal policy parameters found by the heuristic approach are used as input to the simulation. Through this process, we obtain the $E[OH]$ ($E[OH]_{approx}$ in Table 5) when the policy parameters of the heuristic are used in the model. Then

λ^c	λ^n	DLT: non-critical (Model 1)					DLT: critical (Model 2)				
		$(Q,r,K)_{sim}$	$(Q,r,K)_{approx}$	$E[OH]_{sim}$	$E[OH]_{approx}$	$\Delta\big[E(OH)\big]^*$	$(Q,r,K)_{sim}$	$(Q,r,K)_{approx}$	$E[OH]_{sim}$	$E[OH]_{approx}$	$\Delta\big[E(OH)\big]^*$
6	1	(12, 6, 0)	(12, 6, 0)	9.097	9.097	**0.00%**	(10, 5, 1)	(10, 5, 3)	7.590	7.603	**0.18%**
6	2	(12, 6, 1)	(12, 6, 3)	8.708	8.727	**0.21%**	(12, 5, 3)	(13, 5, 3)	8.152	8.640	**5.99%**
6	3	(12, 6, 3)	(13, 6, 4)	8.360	8.885	**6.28%**	(12, 6, 1)	(12, 6, 1)	8.621	8.621	**0.00%**
6	4	(13, 6, 3)	(14, 7, 2)	8.478	9.925	**17.07%**	(13, 6, 3)	(13, 6, 3)	9.144	9.144	**0.00%**
6	5	(15, 6, 3)	(14, 7, 3)	9.115	9.565	**4.94%**	(14, 6, 3)	(14, 6, 3)	8.707	8.707	**0.00%**
6	6	(14, 7, 3)	(14, 7, 4)	9.178	9.271	**1.01%**	(14, 7, 3)	(14, 7, 3)	9.188	9.188	**0.00%**
7	6	(15, 7, 4)	(16, 8, 4)	9.265	10.698	**15.47%**	(15, 7, 4)	(15, 7, 4)	9.367	9.367	**0.00%**
8	6	(16, 8, 4)	(18, 8, 5)	10.233	11.269	**10.12%**	(16, 8, 2)	(16, 8, 3)	10.335	10.374	**0.38%**
9	6	(19, 8, 5)	(18, 9, 5)	11.282	11.741	**4.07%**	(16, 8, 4)	(17, 8, 4)	10.043	10.528	**4.84%**
10	6	(18, 9, 4)	(20, 10, 4)	11.194	13.151	**17.48%**	(18, 9, 3)	(18, 9, 4)	11.547	11.606	**0.51%**

Table 5: Comparison between optimal policy parameters of simulation and approximation, based on the service level optimization model ($L = 0.5, H = 0.1, \bar{\beta^c} = 99\%, \bar{\beta^n} = 80\%$)

* $\Delta\big[E(OH)\big]$ is calculated as the fraction $\frac{E[OH]_{approx} - E[OH]_{sim}}{E[OH]_{sim}}$

the results of these simulations are checked for feasibility. We found that in all the instances, after simulating the optimal policy parameters of the heuristic, the service level requirements were not violated.

It is important to recall that the set $S^{'}$ is restricted to the feasible solutions based on the model assumptions. Thus, the inventory control policy parameters in Table 5 are optimal when those assumptions are applied. In different circumstances, there may be inventory control policy parameters that lead to lower expected on-hand stock.

4.3 Cost Optimization Study

Cost optimization is performed according to Algorithm 2. We compare the expected costs that result from simulations of $1,000,000$ demand arrivals when applying the optimal policy parameters according to the heuristic method and a simulation.

Table 6 presents the optimal policy parameters based on the heuristic and the simulations, their corresponding expected total inventory cost, and the percentage difference between these costs. Specifically, it presents 11 instances. All of them have fixed critical demand arrival rate $\lambda^c = 6$, replenishment lead time $L = 0.5$, demand lead time $H = 0.1$, inventory holding cost $h = 1$, and critical backorder cost $b_c = 6,000$. The chosen cost parameters are the same with those of Deshpande et al. (2003) with the intention our scenarios to be closer to real-life cases. For the first 3 instances, we fix the non-critical fill rate $\lambda^n = 6$, and the non-critical backorder cost $b_n = 300$, but we change the ordering/setup cost to reflect to three industries (Deshpande et al. (2003)): high-tech industries ($A = 200$), computers and telecommunication industries ($A = 100$), and commodity and package good industries ($A = 0$). Then, we fix the ordering/setup cost, and we increase the cost of non-critical backorders until it equals the critical backorder cost. Finally, while keeping the other parameters stable, we change the non-critical demand arrival rate. Hence, we can observe the performance of our model for different parameter combinations. An important remark is that the expected cost of the approximation method ($E[C]_{approx}$ in Table 6) is found after simulating the inventory system when the optimal policy parameters of the heuristic are applied.

Our approximation method finds results close enough to those of the simulation. To be more accurate, for *Model 1*, the maximum cost deviation is 19.44%, and the average is 7.20%, while for *Model 2*, the maximum is 22.35%, and the average 11.71%. It is important to mention that the expected inventory cost for the optimal policy parameters, as found from both the simulation and the heuristic, were given by a simulation of $1,000,000$ demand arrivals. From Table 6, we also come to the following conclusions:

- A change to the ordering/setup cost has little effect on the optimal policy, based on the chosen cost parameters.

A	b_n	λ^n	DLT: non-critical (Model 1)					DLT: critical (Model 2)				
			$(Q,r,K)_{sim}$	$(Q,r,K)_{approx}$	$E[C]_{sim}$	$E[C]_{approx}$	$\Delta\big[E[C]\big]^*$	$(Q,r,K)_{sim}$	$(Q,r,K)_{approx}$	$E[C]_{sim}$	$E[C]_{approx}$	$\Delta\big[E[C]\big]^*$
200	300	6	(8, 4, 4)	(8, 4, 3)	1658.13	1665.13	**0.42%**	(8, 4, 3)	(8, 3, 2)	1547.26	1772.13	**14.53%**
100	300	6	(8, 4, 4)	(8, 4, 3)	1595.40	1596.86	**0.09%**	(8, 4, 3)	(8, 3, 2)	1484.77	1697.70	**14.34%**
0	300	6	(8, 4, 4)	(8, 4, 3)	1518.27	1525.90	**0.50%**	(8, 4, 3)	(8, 3, 2)	1412.88	1634.92	**15.72%**
200	600	6	(10, 5, 3)	(8, 4, 3)	1783.29	1865.40	**4.60%**	(8, 4, 2)	(8, 3, 2)	1710.67	1968.53	**15.07%**
200	1200	6	(10, 5, 2)	(9, 4, 3)	1921.62	2195.65	**14.26%**	(9, 4, 1)	(9, 3, 2)	1940.60	2287.85	**17.89%**
200	1500	6	(10, 5, 2)	(9, 4, 3)	1984.88	2370.67	**19.44%**	(9, 4, 1)	(9, 3, 2)	2013.85	2463.86	**22.35%**
200	3000	6	(10, 5, 1)	(10, 5, 0)	2194.30	2263.63	**3.16%**	(10, 5, 0)	(9, 4, 0)	2025.54	2310.39	**14.06%**
200	6000	6	(12 ,6, 0)	(10, 5, 0)	2268.41	2445.40	**7.80%**	(12, 6, 0)	(10, 5, 0)	2225.06	2227.78	**0.12%**
200	1500	1.5	(8, 4, 2)	(8, 4, 0)	1709.08	1787.90	**4.61%**	(6, 3, 1)	(6, 3, 2)	1484.58	1506.87	**1.50%**
200	1500	3	(10, 5, 2)	(8, 4, 0)	1832.23	2025.17	**10.53%**	(8, 4, 1)	(7, 3, 2)	1615.86	1783.50	**10.37%**
200	1500	4.5	(10, 5, 2)	(8, 4, 3)	1881.92	2140.46	**13.74%**	(8, 4, 1)	(8, 4, 2)	1770.32	1819.87	**2.80%**

Table 6: Comparison between optimal policy parameters of simulation and approximation, based on the cost optimization model ($\lambda^c = 6, L = 0.5, H = 0.1, h = 250, b_c = 6000$)

* $\Delta\big[E[C]\big]$ is calculated as the fraction $\frac{E[C]_{approx} - E[C]_{sim}}{E[C]_{sim}}$

- Increasing the non-critical backorder cost, while keeping the other parameters the same, means increasing the importance of non-critical demand over the importance of critical demand. Therefore, rationing becomes less useful and parameter K reduces. Deshpande et al. (2003) have stated that if $b_c = b_n$, then the optimal threshold level $K^* = 0$. Since the penalties for both critical and non-critical backorders are the same, there is no reason to differentiate the demand.

4.4 Importance of Rationing and Demand Lead time

Now, we study the benefits of inventory rationing in a (Q, r, K) system. Table 7 provides the results of 10 instances, with different arrival rates, but fixed replenishment and demand lead times and minimum required service levels. For each model, the first two columns show the optimal policy parameters and the expected on-hand stock, when rationing is not applied, while the next two columns the cost savings in terms of on-hand stock according to the simulation and heuristic results, as presented in Table 5, where rationing is allowed.

λ^c	λ^n	DLT: non-critical (Model 1)				DLT: critical (Model 2)			
		(Q, r)	$E[OH]$	$S_{sim}(\%)$	$S_{approx}(\%)$	(Q, r)	$E[OH]$	$S_{sim}(\%)$	$S_{approx}(\%)$
6	1	(12, 6)	9.097	**0.00**	**0.00**	(12, 5)	8.604	**11.79**	**11.63**
6	2	(14, 7)	10.708	**18.68**	**18.51**	(12, 6)	9.100	**10.42**	**5.05**
6	3	(14, 7)	10.305	**18.88**	**13.78**	(14, 7)	10.600	**18.68**	**18.68**
6	4	(19, 7)	12.388	**31.57**	**19.88**	(14, 7)	10.108	**9.54**	**9.54**
6	5	(16, 8)	11.507	**20.79**	**16.87**	(16, 8)	11.607	**24.98**	**24.98**
6	6	(19, 8)	12.611	**27.22**	**26.48**	(18, 9)	13.116	**29.95**	**29.95**
7	6	(18, 9)	12.599	**26.47**	**15.09**	(18, 9)	12.703	**26.26**	**26.26**
8	6	(20, 10)	14.106	**27.45**	**20.11**	(19, 9)	12.821	**19.39**	**19.09**
9	6	(20, 10)	13.610	**17.10**	**13.73**	(20, 10)	13.911	**27.81**	**24.32**
10	6	(22, 11)	15.083	**25.78**	**12.81**	(21, 10)	14.011	**17.59**	**17.17**

Table 7: Comparison between optimal policy parameters, with and without rationing $(L = 0.5, H = 0.1, \bar{\beta^c} = 99, \bar{\beta^n} = 80)$

Table 7 shows that inventory rationing can lead to significant savings, up to around 30%. Similar conclusions were also made in a $(S - 1, S)$ review policy by Koçağa and Şen (2007). On average, for the 10 instances, inventory rationing saves 20.5% of the on-hand stock based on the simulation results and 17.2% based on the heuristic. To be more accurate, for *Model 1* the simulation results in 21.4% savings and the heuristic in 15.7%. For *Model 2*, the savings amount to 19.6% and 18.7% for the simulation and the heuristic, respectively.

To further investigate the importance of inventory rationing in a (Q, r) model, we compare the result of the cost optimization model, if rationing is allowed or not (Table

A	b_n	λ^c	H	DLT: non-critical (Model 1)					DLT: critical (Model 2)				
				(Q,r)	(Q,r,K)	$E[C(Q,r)]$	$E[C(Q,r,K)]$	Savings (%)	(Q,r)	(Q, r, K)	$E[C(Q,r)]$	$E[C(Q,r,K)]$	Savings (%)
200	300	6	0.1	(4, 8)	(5, 6, 3)	1678.38	1434.25	**14.55**	(4, 7)	(4, 6, 2)	1553.25	1392.27	**10.36**
100	300	6	0.1	(3, 8)	(3, 7, 2)	1530.68	1288.36	**15.83**	(3, 7)	(3, 6, 2)	1392.84	1235.72	**11.28**
0	300	6	0.1	(2, 8)	(2, 7, 2)	1306.75	1050.25	**19.63**	(2, 7)	(2, 6, 2)	1157.80	1022.86	**11.65**
200	600	6	0.1	(4, 8)	(4, 7, 2)	1682.67	1503.97	**10.62**	(4, 7)	(4, 6, 1)	1571.28	1483.99	**5.56**
200	1200	6	0.1	(4, 8)	(5, 7, 2)	1692.68	1608.35	**4.98**	(4, 7)	(4, 7, 1)	1591.06	1569.81	**1.34**
200	1500	6	0.1	(4, 8)	(5, 7, 1)	1700.58	1630.10	**4.14**	(4, 7)	(4, 7, 1)	1607.83	1595.34	**0.78**
200	3000	6	0.1	(4, 8)	(4, 8, 1)	1732.65	1712.70	**1.15**	(4, 7)	(4, 7, 0)	1670.11	1670.11	**0.00**
200	6000	6	0.1	(4, 8)	(4, 8, 0)	1773.77	1773.77	**0.00**	(4, 8)	(4, 8, 0)	1773.77	1773.77	**0.00**
200	1500	1.5	0.1	(4, 4)	(4, 4, 1)	1233.43	1219.47	**1.13**	(4, 4)	(4, 4, 0)	1208.02	1208.02	**0.00**
200	1500	3	0.1	(4, 5)	(4, 5, 1)	1428.56	1355.08	**5.14**	(4, 5)	(4, 5, 0)	1364.85	1364.85	**0.00**
200	1500	4.5	0.1	(4, 7)	(4, 6, 1)	1578.03	1498.06	**5.07**	(4, 6)	(4, 6, 1)	1492.77	1483.23	**0.64**
200	1500	6	0	(5, 8)	(5, 8, 1)	1788.31	1709.91	**4.38**	(4, 8)	(5, 7, 1)	1699.56	1677.34	**1.31**
200	1500	6	0.2	(4, 7)	(4, 7, 1)	1609.51	1541.96	**4.20**	(4, 6)	(4, 6, 1)	1517.61	1501.87	**1.04**
200	1500	6	0.3	(4, 6)	(4, 6, 1)	1520.15	1447.11	**4.80**	(4, 5)	(4, 5, 1)	1434.01	1417.63	**1.14**
200	1500	6	0.4	(3, 6)	(4, 5, 1)	1434.01	1372.30	**4.30**	(3, 5)	(3, 5, 0)	1320.89	1320.89	**0.00**
200	1500	6	0.5	(3, 5)	(3, 5, 1)	1328.73	1298.38	**2.28**	(3, 4)	(3, 4, 0)	1129.52	1129.52	**0.00**
200	300	6	0	(5, 8)	(4, 7, 3)	1750.29	1492.04	**14.75**	(5, 7)	(5, 6, 2)	1636.32	1467.68	**10.31**
200	300	6	0.2	(4, 7)	(4, 6, 2)	1586.83	1371.77	**13.55**	(4, 6)	(4, 5, 2)	1460.18	1314.10	**10.00**
200	300	6	0.3	(4, 6)	(4, 5, 2)	1487.48	1299.64	**12.63**	(4, 5)	(4, 4, 2)	1371.04	1221.31	**10.92**
200	300	6	0.4	(3, 6)	(3, 5, 2)	1417.63	1250.78	**11.77**	(4, 4)	(4, 3, 2)	1257.58	1098.16	**12.68**
200	300	6	0.5	(3, 5)	(3, 4, 2)	1330.81	1237.18	**7.04**	(2, 3)	(3, 2, 1)	979.21	860.18	**12.16**

Table 8: Comparison between optimal policy parameters, with and without rationing ($L = 0.5, \bar{\beta}^c = 99, \bar{\beta}^n = 80, b_c = 6000$)

8). We only base our results on simulations, in order to avoid any restrictions due to model assumptions. Furthermore, to be more certain about the outcome, we apply the cost optimization algorithm (Algorithm 2) multiple times. First, we apply it for a wide range of possible parameter combinations, and $10,000$ demand arrivals. Next, we select the 10 combinations with the lowest cost. We perform the algorithm once again for those 10 combinations, for $100,000$ demand arrivals. The top 4 combinations with the lowest cost are selected and used as input for the algorithm with $1,000,000$ demand arrivals. The policy parameters with the lowest cost among those 4 combinations are the optimal ones.

Table 8 presents 21 instances, in which λ_n, b_c, h, L are fixed, while λ_c, A, b_n, H vary. The first two columns of each model include the optimal policy parameters in case of no rationing and rationing. The next two columns show the corresponding expected costs while the last one the percentage of cost savings when rationing is allowed.

We observe that according to those 21 instances, we can achieve cost savings up to approximately 20%. For *Model 1*, the average savings are 7.7%, whereas for *Model 2*, 4.8%.

When the backorder costs differ per class, the models have different optimal values even in the case that rationing is not applied. On the contrary, when $b_c = b_n$, the outcome is the same with or without rationing, regardless of the examined model.

The last 10 instances highlight the influence of demand lead time on inventory cost. Different DLT means different optimal policy, which means different expected cost. However, for any value of DLT, inventory rationing leads to significant savings, most times.

To further analyze the importance of incorporating demand lead time into a continuous (Q, r) model, we simulate the first 11 instances of Table 8, without considering their DLT. In other words, we keep the same parameters except for H, which becomes negligible. The results are represented in Table 9.

From Table 9, it is obvious that incorporating DLT into our inventory system leads to notable improvements, in terms of cost minimization. According to 11 instances, for *Model 1*, DLT results in 5.16% cost savings on average and 7.66% at most (see column $Savings_1$), whereas for *Model 2*, it results in 7.01% cost savings on average and 9.10% at most (see column $Savings_2$).

To conclude, incorporating DLT to a continuous (Q, r) review policy can decrease the expected total inventory cost significantly (up to approximately 10%). Incorporating inventory rationing can lead to even further improvements, which according to the examined instances can be even 20% of cost savings.

A	b_n	λ^c	(Q, r, K)	$E[C]$	$Savings_1(\%)$	$Savings_2(\%)$
200	300	6	(4, 7, 3)	1492.04	**3.87**	**6.80**
100	300	6	(4, 7, 3)	1348.80	**4.48**	**8.65**
0	300	6	(2, 8, 2)	1096.64	**4.23**	**7.11**
200	600	6	(5, 7, 2)	1589.38	**5.37**	**6.76**
200	1200	6	(4, 8, 2)	1687.24	**4.68**	**7.15**
200	1500	6	(5, 8, 1)	1709.91	**4.67**	**6.83**
200	3000	6	(5, 8, 1)	1805.41	**5.13**	**7.69**
200	6000	6	(4, 9, 0)	1877.53	**5.53**	**5.53**
200	1500	1.5	(4, 5, 1)	1320.57	**7.66**	**9.10**
200	1500	3	(4, 6, 1)	1444.51	**6.19**	**5.58**
200	1500	4.5	(4, 7, 1)	1575.41	**4.91**	**5.88**

Table 9: Comparison between optimal policy parameters with and without demand lead time ($L = 0.5, H = 0.1, \bar{\beta}^c = 99, \bar{\beta}^n = 80, b_c = 6000$)

5 Conclusions and Extensions

In this thesis, we consider an inventory system of a single item that supports two demand classes, differentiating in their priority and ADI. To be more precise, we study two models. In *Model 1*, the high-priority demand is due immediately, at its arrival time, whereas the low-priority demand is due after a deterministic demand lead time. In *Model 2*, the low-priority demand is due immediately, whereas the high-priority demand is due after a deterministic demand lead time. Our study is based on a static threshold rationing policy, which is incorporated into the original continuous (Q, r) review framework. It is the first one in the existing literature that combines inventory rationing and ADI/DLT in a continuous (Q, r) model.

Despite the high complexity of the examined model, we derive an exact expression for the non-critical demand service level and a sufficient approximate expression for the critical demand service level. To achieve that, some important assumptions are made, which simplify our analysis. Specifically, the order quantity is assumed to be large enough with respect to the reorder point, and the reorder point is assumed to be greater than the threshold level. Hence, we are able to approximately define the steady-state probabilities of the on-hand stock level, and thus the critical service level.

Comparing the approximations with simulation results, we claim that our method performs sufficiently well, especially for high service levels. Based on the examined instances, for critical service levels of at least 99%, the approximations deviate by 0.28% from the simulations, on average. For critical service levels of $90\% - 99\%$, the approximations deviate by 1.14%, on average. Conclusions on how the service levels are affected by differentiating the inventory policy parameters, the demand arrival rates

and the replenishment and demand lead times are made.

To further evaluate our approximations, we develop two algorithms (i.e. the service level optimization and the cost optimization algorithms). In the service level optimization algorithm, we use a brute force approach to find the optimal inventory control parameters, which minimize the expected inventory holding cost subject to class-specific service level requirements. According to the examined instances, our heuristic method finds holding costs that differ by 4.46% from the minimum, on average. Moreover, in 6 out of 20 instances, our heuristic method finds the same optimal parameters as the simulation. In the cost optimization algorithm, we use a similar approach to find the optimal inventory control policy parameters that minimize the expected total inventory cost. The average deviation between the approximations and the simulation results is 9.46%. A variety of different cost parameter combinations is examined, which enables us to identify how the total inventory cost fluctuates under different circumstances.

Furthermore, through simulations, the importance of both inventory rationing and ADI/DLT is highlighted. Based on the examined instances, inventory rationing can lead to total inventory cost savings of up to 20%, and ADI/DLT up to 10%. Thus, they are highly recommended.

In conclusion, incorporating inventory rationing and ADI/DLT into a continuous (Q, r) review policy has never been studied in the existing literature. However, through our study, we find that both can lead to independently significant cost savings and we hope that this study will encourage more researchers to further analyze this topic. This study can be extended in multifarious ways. Some of them are: extending our model to be applicable to several demand classes, deriving further expressions to overcome the limitations due to our model assumptions, considering stochastic DLT and/or stochastic replenishment lead time, allowing flexible deliveries, examining different demand distributions.

References

Arslan, H., Graves, S. C., and Roemer, T. A. (2007). A single-product inventory model for multiple demand classes. *Management Science*, 53(9):1486–1500.

Benjaafar, S., Cooper, W. L., and Mardan, S. (2011). Production-inventory systems with imperfect advance demand information and updating. *Naval Research Logistics (NRL)*, 58(2):88–106.

Dekker, R., Kleijn, M. J., and De Rooij, P. (1998). A spare parts stocking policy based on equipment criticality. *International Journal of production economics*, 56:69–77.

Deshpande, V., Cohen, M. A., and Donohue, K. (2003). A threshold inventory rationing policy for service-differentiated demand classes. *Management science*, 49(6):683–703.

Evans, R. V. (1968). Sales and restocking policies in a single item inventory system. *Management Science*, 14(7):463–472.

Gallego, G. and Özer, Ö. (2001). Integrating replenishment decisions with advance demand information. *Management science*, 47(10):1344–1360.

Gayon, J.-P., Benjaafar, S., and De Véricourt, F. (2009). Using imperfect advance demand information in production-inventory systems with multiple customer classes. *Manufacturing & Service Operations Management*, 11(1):128–143.

Ha, A. Y. (1997a). Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Science*, 43(8):1093–1103.

Ha, A. Y. (1997b). Stock-rationing policy for a make-to-stock production system with two priority classes and backordering. *Naval Research Logistics (NRL)*, 44(5):457–472.

Ha, A. Y. (2000). Stock rationing in an m/ek/1 make-to-stock queue. *Management Science*, 46(1):77–87.

Hariharan, R. and Zipkin, P. (1995). Customer-order information, leadtimes, and inventories. *Management Science*, 41(10):1599–1607.

Kaplan, A. (1969). Stock rationing. *Management Science*, 15(5):260–267.

Karaesmen, F., Buzacott, J. A., and Dallery, Y. (2002). Integrating advance order information in make-to-stock production systems. *IIE transactions*, 34(8):649–662.

Karaesmen, F., Liberopoulos, G., and Dallery, Y. (2004). The value of advance demand information in production/inventory systems. *Annals of Operations Research*, 126(1):135–157.

Kleijn, M. J. and Dekker, R. (1999). An overview of inventory systems with several demand classes. *New trends in distribution logistics*, pages 253–265.

Koçağa, Y. L. and Şen, A. (2007). Spare parts inventory management with demand lead times and rationing. *IIE Transactions*, 39(9):879–898.

Nahmias, S. and Demmy, W. S. (1981). Operating characteristics of an inventory

system with rationing. *Management Science*, 27(11):1236–1245.

Özer, Ö. and Wei, W. (2004). Inventory control with limited capacity and advance demand information. *Operations Research*, 52(6):988–1000.

Simpson Jr, K. F. (1958). In-process inventories. *Operations Research*, 6(6):863–873.

Tan, T., Güllü, R., and Erkip, N. (2007). Modelling imperfect advance demand information and analysis of optimal inventory policies. *European Journal of Operational Research*, 177(2):897–923.

Tan, T., Güllü, R., and Erkip, N. (2009). Using imperfect advance demand information in ordering and rationing decisions. *International Journal of Production Economics*, 121(2):665–677.

Topan, E., Tan, T., van Houtum, G.-J., and Dekker, R. (2018). Using imperfect advance demand information in lost-sales inventory systems with the option of returning inventory. *IISE Transactions*, 50(3):246–264.

Topkis, D. M. (1968). Optimal ordering and rationing policies in a nonstationary dynamic inventory model with n demand classes. *Management Science*, 15(3):160–176.

Veinott Jr, A. F. (1965). Optimal policy in a dynamic, single product, nonstationary inventory model with several demand classes. *Operations Research*, 13(5):761–778.

Vicil, O. (2021a). Inventory rationing on a one-for-one inventory model for two priority customer classes with backorders and lost sales. *IISE Transactions*, 53(4):472–495.

Vicil, O. (2021b). Optimizing stock levels for service-differentiated demand classes with inventory rationing and demand lead times. *Flexible Services and Manufacturing Journal*, 33(2):381–424.

Vicil, O. (2022). Cost optimization in the $(s-1, s)$ backorder inventory model with two demand classes and rationing. *Flexible Services and Manufacturing Journal*, 34(1):101–124.

Vicil, O. and Jackson, P. (2016). Computationally efficient optimization of stock pooling and allocation levels for two-demand-classes under general lead time distributions. *IIE Transactions*, 48(10):955–974.

Vicil, O. and Jackson, P. (2018). Stock optimization for service differentiated demands with fill rate and waiting time requirements. *Operations Research Letters*, 46(3):367–372.

Wang, D. and Tang, O. (2014). Dynamic inventory rationing with mixed backorders and lost sales. *International Journal of Production Economics*, 149:56–67.

Wang, T. and Toktay, B. L. (2008). Inventory management with advance demand information and flexible delivery. *Management Science*, 54(4):716–732.

A Model 2: Critical Orders with Demand Lead Time

In section 3, we derive all the necessary equations for the case in which non-critical demand is due after a demand lead time. With small modifications to some of these equations, we can derive the necessary equations for *Model 2*, where critical demand is due after a fixed demand lead time H, and non-critical demand is due immediately. The modified equations are presented in this appendix.

First instead of Equations 4 and 5, we use Equations 38 and 39 to calculate the demand with net impact on the inventory and the number of critical orders that have been placed but not yet due. Note that non-critical orders are due immediately and thus, $Y^n = 0$.

$$\widehat{D}(t, t+L) = D^c(t, t-H+L) + D^n(t, t+L) \tag{38}$$

$$Y^c(t) = D^c(t-H, t) \tag{39}$$

Then, instead of using Equations 13 and 15, we can calculate the non-critical and critical stock-out probabilities as following.

$$\beta^n(K, IP(t)) = \sum_{i=0}^{IP(t)-K-1} e^{-[\lambda^n L + \lambda^c(L-H)]} \frac{\left[\lambda^n L + \lambda^c(L-H)\right]^i}{i!} \tag{40}$$

An easy way to calculate the critical demand service level is by determining it as the probability of no having a stock-out at a random time $t + L$.

$$\beta^c(K, IP(t)) = 1 - P_{out}(K, IP(t)) \tag{41}$$

where $P_{out}(K, IP(t))$ is the probability that the inventory system will be in a stock-out situation at a random time $t + L$, when the threshold level is K and the inventory position at time t is $IP(t)$. The corresponding probability can be computed by the following formula.

$$\begin{aligned} P_{out}(K, IP(t)) = \int_0^{L-H} & f_1(K, IP(t), y) \\ & \cdot \left[1 - \sum_{i=0}^{K-1} e^{-\lambda^c(L-H-y)} \frac{[\lambda^c(L-H-y)]^i]}{i!}\right] dy \end{aligned} \tag{42}$$

It is important to mention that in the examined model, we have a stock-out situation if and only if on-hand reaches the threshold level before $L-H$ and the critical demand

that arrives between t_k and $L - H$ is at least K. If T (i.e. the time from t until t_k, $T = t_k - t$) is within the interval $(t + L - H, t + L]$, then our system cannot be in a stock-out situation. As the critical demand placed in $(t+L-H, t+L]$ is due later than $t + L$ and do not have net impact on the inventory until that time, then the critical demand should be over before $t + L - H$, in order to have a stock-out. Thus, we only consider the interval $(t, t + L - H]$.

Since non-critical demand is due immediately, it holds that $E[Y^n] = 0$, while for the critical orders that are placed and not yet due, it holds the following.

$$E[Y^c] = E\big[D^c(t - H, t)\big] = \lambda^c H \tag{43}$$

Similarly to *Model 1*, for given inventory position, the probabilities that the class-specific backorders will equal a specific number x can be found as following:

$$P[B^c(IP(t)) = x] = \int_O^{L-H} f_1(K, IP(t), y) \cdot e^{-\lambda^c(L-H-y)} \frac{[\lambda^c(L - H - y)]^{K+x}]}{(K + x)!}\, dy \tag{44}$$

$$\begin{aligned} P[B^n(IP(t)) = x] = &\int_O^{L-H} f_1(K, IP(t), y) \cdot e^{-\lambda^n(L-y)} \frac{[\lambda^n(L - y)]^x]}{x!}\, dy \\ &+ \int_{L-H}^{L} f_2(K, IP(t), y) \cdot e^{-\lambda^n(L-y)} \frac{[\lambda^n(L - y)]^x]}{x!} \end{aligned} \tag{45}$$

where f_2 denotes the probability density function of T for the interval $(t+L-H, t+L]$, which should be calculated according to Equations 46 and 47.

$$F_2(K, IP(t), y) = P[T \leq y] = P[D^c(t, t + L - T) + D^n(t, t + y) \geq IP(t) - K] \tag{46}$$

$$f_2(K, IP(t), y) = \lambda^n e^{-(\lambda^n y + \lambda^c(L-H))} \frac{\big[\lambda^n y + \lambda^c(L - H)\big]^{IP(t)-K-1}}{(IP(t) - K - 1)!} \tag{47}$$

B Proofs

Proof of Equation 19.

Since we examine the case in which $y \in (t, t + L - H]$, demands arrive according to Poisson distribution with rate $\lambda^c + \lambda^n$. The probability that exactly i number of orders arrive in the time interval $(t, t + y]$ is:

$$P[D(t, t + y) = i] = e^{-(\lambda^c+\lambda^n)y} \frac{[(\lambda^c + \lambda^n)y]^i}{i!} \tag{48}$$

Thus, for $F_1(t, y)$ it holds:

$$\begin{aligned} F_1(t,y) &= P[D(t,t+y) \geq IP(t) - K] \\ &= 1 - P[D(t,t+y) < IP(t) - K] \\ &= 1 - P[D(t,t+y) \leq IP(t) - K - 1] \\ &= 1 - \sum_{i=0}^{IP(t)-K-1} e^{-(\lambda^c+\lambda^n)y} \frac{[(\lambda^c+\lambda^n)y]^i}{i!}, \quad y \geq 0 \end{aligned} \tag{49}$$

The cumulative distribution function is given by the formula:

$$F_1(t,y) = \begin{cases} 1 - \sum_{i=0}^{IP(t)-K-1} e^{-(\lambda^c+\lambda^n)y} \frac{[(\lambda^c+\lambda^n)y]^i}{i!} & y \geq 0 \\ 0 & y < 0 \end{cases} \tag{50}$$

The derivative of the above formula with respect to y, for $y \geq 0$, is:

$$\begin{aligned} f_1(t,y) &= \frac{dF_1(t,y)}{dy} \\ &= (\lambda^c+\lambda^n) \sum_{i=0}^{IP(t)-K-1} e^{-(\lambda^c+\lambda^n)y} \frac{[(\lambda^c+\lambda^n)y]^i}{i!} \\ &\quad - \sum_{i=0}^{IP(t)-K-1} e^{-(\lambda^c+\lambda^n)y} \frac{(\lambda^c+\lambda^n) i [(\lambda^c+\lambda^n)y]^{i-1}}{i!} \\ &= (\lambda^c+\lambda^n) \sum_{i=0}^{IP(t)-K-1} e^{-(\lambda^c+\lambda^n)y} \frac{[(\lambda^c+\lambda^n)y]^i}{i!} \\ &\quad - \sum_{i=0}^{IP(t)-K-1} e^{-(\lambda^c+\lambda^n)y} \frac{(\lambda^c+\lambda^n) [(\lambda^c+\lambda^n)y]^{i-1}}{(i-1)!} \\ &= (\lambda^c+\lambda^n)^{IP(t)-K} e^{-(\lambda^c+\lambda^n)y} \frac{y^{IP(t)-K-1}}{(IP(t)-K-1)!}, \quad y \geq 0 \end{aligned} \tag{51}$$

Thus, the probability density function of the variable T is given by:

$$f_1(t,y) = \begin{cases} (\lambda^c+\lambda^n)^{IP(t)-K} e^{-(\lambda^c+\lambda^n)y} \frac{y^{IP(t)-K-1}}{(IP(t)-K-1)!} & y \geq 0 \\ 0 & y < 0 \end{cases} \tag{52}$$

Proof of Equation 21.

Since, we examine the case in which $y \in (t + L - H, t + L]$, demands with net impact on the inventory level arrive according to Poisson distribution with rate $\lambda^c + \lambda^n$,

until time $t + L - H$, and then with rate λ^c. Thus, the total arrivals until y are $(\lambda^c + \lambda^n)(L - H) + \lambda^c\big[y - (L - H)\big] = \lambda^c y + \lambda^n(L - H)$. The probability that exactly i number of orders with net impact on the inventory level arrive in the time interval $(t, t + y]$ is:

$$P[\widehat{D}(t, t + y) = i] = e^{-[\lambda^c y + \lambda^n (L-H)]} \frac{[\lambda^c y + \lambda^n (L - H)]^i}{i!} \tag{53}$$

Similarly to proof of Equation 19, we can prove that Equation 21 holds.

λ^c	λ^n	r	Q	K	DLT: non-critical (Model 1)								
					$E[OH]_{sim}$	$E[OH]_{approx}$	AE	$E[B_c]_{sim}$	$E[B_c]_{approx}$	AE	$E[B_n]_{sim}$	$E[B_n]_{approx}$	AE
1	4	3	7	2	5.009	5.046	**0.037**	0.000	0.001	**0.001**	0.112	0.140	**0.027**
7	10	10	20	5	12.806	13.081	**0.276**	0.000	0.001	**0.001**	0.154	0.080	**0.074**
8	10	10	20	5	12.414	12.591	**0.177**	0.001	0.002	**0.001**	0.188	0.088	**0.101**
10	10	10	20	5	11.630	11.611	**0.018**	0.004	0.006	**0.003**	0.259	0.102	**0.157**
10	7	10	20	5	12.832	12.766	**0.066**	0.003	0.005	**0.002**	0.126	0.061	**0.065**
10	8	10	20	5	12.480	12.380	**0.100**	0.003	0.005	**0.002**	0.166	0.074	**0.091**
10	10	10	20	5	11.771	11.611	**0.159**	0.004	0.006	**0.003**	0.262	0.102	**0.160**
10	12	10	20	5	11.100	10.849	**0.251**	0.005	0.008	**0.003**	0.380	0.132	**0.248**
7	10	7	20	5	10.487	10.167	**0.319**	0.003	0.005	**0.002**	0.472	0.162	**0.310**
7	10	9	20	5	12.263	12.108	**0.155**	0.001	0.002	**0.001**	0.233	0.105	**0.127**
7	10	12	20	5	15.071	15.040	**0.031**	0.000	0.000	**0.000**	0.059	0.039	**0.020**
10	7	9	20	5	11.905	11.788	**0.117**	0.005	0.008	**0.003**	0.186	0.080	**0.106**
10	7	10	20	5	12.846	12.766	**0.080**	0.003	0.005	**0.002**	0.125	0.061	**0.063**
10	7	12	20	5	14.747	14.733	**0.014**	0.001	0.001	**0.001**	0.050	0.031	**0.019**
10	7	10	20	3	12.746	12.739	**0.006**	0.005	0.008	**0.003**	0.051	0.031	**0.019**
10	7	10	20	5	12.808	12.766	**0.042**	0.003	0.005	**0.002**	0.128	0.061	**0.066**
10	7	10	20	6	12.897	12.783	**0.114**	0.002	0.003	**0.001**	0.186	0.080	**0.106**
10	7	10	20	8	13.071	12.822	**0.249**	0.001	0.002	**0.001**	0.365	0.119	**0.246**

Table 10: Performance of approximation for random parameters and critical service level at least 99% ($L = 0.5, H = 0.1$)

λ^c	λ^n	r	Q	K	DLT: critical (Model 2)								
					$E[OH]_{sim}$	$E[OH]_{approx}$	AE	$E[B_c]_{sim}$	$E[B_c]_{approx}$	AE	$E[B_n]_{sim}$	$E[B_n]_{approx}$	AE
1	4	3	7	2	4.760	4.776	**0.016**	0.000	0.000	**0.000**	0.170	0.158	**0.012**
7	10	10	20	5	12.884	12.788	**0.095**	0.001	0.000	**0.000**	0.088	0.180	**0.092**
8	10	10	20	5	12.521	12.394	**0.127**	0.001	0.000	**0.001**	0.093	0.206	**0.113**
10	10	10	20	5	11.760	11.606	**0.154**	0.004	0.002	**0.002**	0.102	0.261	**0.159**
10	7	10	20	5	13.097	13.057	**0.040**	0.002	0.001	**0.001**	0.056	0.109	**0.054**
10	8	10	20	5	12.674	12.572	**0.102**	0.002	0.001	**0.001**	0.070	0.147	**0.077**
10	10	10	20	5	11.759	11.606	**0.152**	0.004	0.002	**0.002**	0.102	0.258	**0.155**
10	12	10	20	5	10.926	10.648	**0.278**	0.006	0.002	**0.004**	0.138	0.406	**0.269**
7	10	7	20	5	10.487	9.873	**0.614**	0.003	0.002	**0.001**	0.171	0.472	**0.302**
7	10	9	20	5	12.263	11.815	**0.449**	0.001	0.000	**0.000**	0.114	0.233	**0.119**
7	10	12	20	5	15.071	14.745	**0.326**	0.000	0.000	**0.000**	0.045	0.059	**0.014**
10	7	9	20	5	12.170	12.076	**0.094**	0.004	0.002	**0.002**	0.074	0.166	**0.092**
10	7	10	20	5	13.132	13.057	**0.075**	0.002	0.001	**0.001**	0.056	0.108	**0.052**
10	7	12	20	5	15.057	15.028	**0.029**	0.000	0.000	**0.000**	0.028	0.041	**0.014**
10	7	10	20	3	13.033	13.030	**0.003**	0.004	0.002	**0.002**	0.028	0.041	**0.014**
10	7	10	20	5	13.109	13.057	**0.052**	0.002	0.001	**0.001**	0.056	0.110	**0.054**
10	7	10	20	6	13.176	13.075	**0.102**	0.001	0.001	**0.001**	0.074	0.163	**0.090**
10	7	10	20	8	13.322	13.114	**0.208**	0.001	0.000	**0.000**	0.113	0.333	**0.220**

Table 11: Performance of approximation for random parameters and critical service level at least 99% ($L = 0.5, H = 0.1$)

λ^c	λ^n	r	Q	K	DLT: non-critical (Model 1)								
					$E[OH]_{sim}$	$E[OH]_{approx}$	AE	$E[B_c]_{sim}$	$E[B_c]_{approx}$	AE	$E[B_n]_{sim}$	$E[B_n]_{approx}$	AE
2	4	3	7	2	4.562	4.584	**0.022**	0.003	0.007	**0.004**	0.163	0.159	**0.003**
3	4	3	7	2	4.129	4.137	**0.008**	0.011	0.021	**0.010**	0.218	0.175	**0.043**
4	4	3	7	2	3.707	3.716	**0.009**	0.030	0.045	**0.015**	0.277	0.186	**0.090**
4	1	3	7	2	4.659	4.684	**0.025**	0.019	0.036	**0.017**	0.039	0.043	**0.003**
4	2	3	7	2	4.319	4.345	**0.026**	0.022	0.039	**0.017**	0.098	0.088	**0.010**
4	4	3	7	2	3.706	3.716	**0.010**	0.030	0.045	**0.015**	0.278	0.186	**0.091**
4	6	3	7	2	3.182	3.192	**0.010**	0.037	0.049	**0.012**	0.542	0.287	**0.255**
11	10	10	20	5	11.239	11.124	**0.115**	0.008	0.010	**0.003**	0.300	0.109	**0.192**
12	10	10	20	5	10.849	10.638	**0.211**	0.013	0.015	**0.002**	0.336	0.115	**0.222**
7	10	6	20	5	9.638	9.200	**0.438**	0.006	0.008	**0.002**	0.636	0.191	**0.445**
10	7	6	20	5	9.230	8.871	**0.359**	0.033	0.031	**0.003**	0.482	0.140	**0.342**
10	7	7	20	5	10.090	9.840	**0.250**	0.019	0.020	**0.001**	0.368	0.119	**0.248**
10	7	8	20	5	10.973	10.813	**0.160**	0.010	0.013	**0.003**	0.266	0.099	**0.167**
10	7	7	17	5	8.669	8.370	**0.300**	0.022	0.024	**0.002**	0.431	0.141	**0.291**
10	7	7	19	5	9.613	9.348	**0.265**	0.020	0.021	**0.002**	0.383	0.126	**0.258**
10	7	7	23	5	11.533	11.322	**0.211**	0.016	0.018	**0.002**	0.320	0.104	**0.216**
10	7	7	27	5	13.473	13.304	**0.169**	0.014	0.015	**0.001**	0.274	0.089	**0.185**

Table 12: Performance of approximation for random parameters and critical service level between 90% - 99% ($L = 0.5, H = 0.1$)

λ^c	λ^n	r	Q	K	DLT: critical (Model 2)								
					$E[OH]_{sim}$	$E[OH]_{approx}$	AE	$E[B_c]_{sim}$	$E[B_c]_{approx}$	AE	$E[B_n]_{sim}$	$E[B_n]_{approx}$	AE
2	4	3	7	2	4.399	4.398	**0.001**	0.004	0.003	**0.000**	0.197	0.177	**0.020**
3	4	3	7	2	4.050	4.035	**0.015**	0.012	0.011	**0.001**	0.236	0.182	**0.054**
4	4	3	7	2	3.712	3.695	**0.017**	0.030	0.024	**0.006**	0.276	0.186	**0.090**
4	1	3	7	2	4.933	4.959	**0.026**	0.011	0.018	**0.007**	0.028	0.035	**0.007**
4	2	3	7	2	4.497	4.517	**0.021**	0.016	0.021	**0.004**	0.081	0.080	**0.002**
4	4	3	7	2	3.710	3.695	**0.015**	0.030	0.024	**0.005**	0.278	0.186	**0.092**
4	6	3	7	2	3.045	2.985	**0.059**	0.044	0.027	**0.017**	0.610	0.303	**0.307**
11	10	10	20	5	11.383	11.214	**0.169**	0.007	0.003	**0.004**	0.290	0.106	**0.184**
12	10	10	20	5	11.025	10.823	**0.202**	0.011	0.004	**0.006**	0.318	0.110	**0.207**
7	10	6	20	5	9.638	8.903	**0.735**	0.006	0.003	**0.003**	0.636	0.200	**0.436**
10	7	6	20	5	9.476	9.147	**0.329**	0.026	0.013	**0.014**	0.446	0.134	**0.312**
10	7	7	20	5	10.340	10.121	**0.219**	0.015	0.007	**0.007**	0.333	0.113	**0.220**
10	7	8	20	5	11.232	11.098	**0.135**	0.008	0.004	**0.004**	0.240	0.093	**0.146**
10	7	7	17	5	8.911	8.647	**0.264**	0.017	0.009	**0.008**	0.392	0.133	**0.259**
10	7	7	19	5	9.870	9.628	**0.242**	0.016	0.008	**0.008**	0.349	0.119	**0.229**
10	7	7	23	5	11.786	11.605	**0.181**	0.013	0.006	**0.007**	0.291	0.099	**0.192**
10	7	7	27	5	13.764	13.589	**0.175**	0.011	0.005	**0.006**	0.245	0.084	**0.161**

Table 13: Performance of approximation for random parameters and critical service level between 90% - 99% ($L = 0.5, H = 0.1$)

λ^c	λ^n	r	Q	K	L	H	DLT: non-critical (Model 1)								
							$E[OH]_{sim}$	$E[OH]_{approx}$	AE	$E[B_c]_{sim}$	$E[B_c]_{approx}$	AE	$E[B_n]_{sim}$	$E[B_n]_{approx}$	AE
10	10	10	20	3	0.5	0.1	11.643	11.575	**0.069**	0.009	0.012	**0.003**	0.118	0.060	**0.058**
10	10	10	20	3	1	0.1	4.042	3.754	**0.288**	0.660	0.179	**0.481**	1.912	0.298	**1.614**
10	10	10	20	3	1	0.5	6.591	7.603	**1.013**	0.215	0.126	**0.089**	0.886	0.200	**0.686**
15	10	10	20	3	0.5	0.1	9.340	9.182	**0.158**	0.073	0.053	**0.020**	0.271	0.092	**0.179**
15	10	10	20	3	1	0.1	1.862	2.575	**0.713**	2.107	0.381	**1.726**	3.295	0.323	**2.971**
15	10	10	20	3	1	0.5	3.449	6.427	**2.978**	1.154	0.299	**0.855**	1.798	0.257	**1.541**
8	8	10	20	4	0.2	0.1	18.115	18.100	**0.014**	0.000	0.000	**0.000**	0.000	0.000	**0.000**
8	8	10	20	4	0.4	0.1	14.927	14.919	**0.008**	0.000	0.000	**0.000**	0.021	0.019	**0.002**
8	8	10	20	4	0.6	0.1	11.862	11.784	**0.078**	0.005	0.007	**0.002**	0.165	0.076	**0.089**
8	8	10	20	4	0.8	0.1	9.081	8.728	**0.353**	0.044	0.031	**0.012**	0.523	0.151	**0.372**
8	8	10	20	4	1	0.1	6.592	5.975	**0.617**	0.179	0.078	**0.101**	1.141	0.230	**0.911**
8	8	10	20	4	0.5	0.1	13.372	13.346	**0.026**	0.001	0.002	**0.001**	0.070	0.043	**0.027**
8	8	10	20	4	0.5	0.2	14.136	14.132	**0.004**	0.000	0.002	**0.002**	0.041	0.030	**0.011**
8	8	10	20	4	0.5	0.3	14.933	14.920	**0.013**	0.000	0.001	**0.001**	0.020	0.019	**0.001**
8	8	10	20	4	0.5	0.4	15.681	15.711	**0.030**	0.000	0.000	**0.000**	0.009	0.010	**0.002**
8	8	10	20	4	0.5	0.5	16.494	16.505	**0.011**	0.000	0.000	**0.000**	0.002	0.005	**0.003**

Table 14: Performance of approximation with varying system parameters

λ^c	λ^n	r	Q	K	L	H	DLT: critical (Model 2)								
							$E[OH]_{sim}$	$E[OH]_{approx}$	AE	$E[B_c]_{sim}$	$E[B_c]_{approx}$	AE	$E[B_n]_{sim}$	$E[B_n]_{approx}$	AE
10	10	10	20	3	0.5	0.1	11.612	11.566	**0.046**	0.009	0.003	**0.005**	0.119	0.060	**0.059**
10	10	10	20	3	1	0.1	4.044	3.709	**0.335**	0.660	0.135	**0.526**	1.902	0.298	**1.604**
10	10	10	20	3	1	0.5	6.591	7.489	**0.899**	0.212	0.012	**0.199**	0.890	0.200	**0.689**
15	10	10	20	3	0.5	0.1	9.805	9.639	**0.166**	0.055	0.020	**0.035**	0.225	0.083	**0.142**
15	10	10	20	3	1	0.1	2.055	3.000	**0.945**	1.967	0.314	**1.653**	3.041	0.316	**2.725**
15	10	10	20	3	1	0.5	4.863	8.635	**3.772**	0.653	0.055	**0.598**	1.173	0.209	**0.963**
8	8	10	20	4	0.2	0.1	18.095	18.100	**0.006**	0.000	0.000	**0.000**	0.000	0.000	**0.000**
8	8	10	20	4	0.4	0.1	14.905	14.919	**0.014**	0.000	0.000	**0.000**	0.022	0.019	**0.003**
8	8	10	20	4	0.6	0.1	11.873	11.780	**0.093**	0.005	0.002	**0.003**	0.164	0.076	**0.089**
8	8	10	20	4	0.8	0.1	9.061	8.714	**0.348**	0.043	0.016	**0.026**	0.525	0.151	**0.374**
8	8	10	20	4	1	0.1	6.606	5.949	**0.657**	0.176	0.052	**0.124**	1.133	0.230	**0.903**
8	8	10	20	4	0.5	0.1	13.367	13.344	**0.023**	0.001	0.001	**0.001**	0.071	0.043	**0.028**
8	8	10	20	4	0.5	0.2	14.139	14.130	**0.009**	0.000	0.000	**0.000**	0.041	0.030	**0.011**
8	8	10	20	4	0.5	0.3	14.919	14.919	**0.000**	0.000	0.000	**0.000**	0.021	0.019	**0.002**
8	8	10	20	4	0.5	0.4	15.699	15.711	**0.011**	0.000	0.000	**0.000**	0.010	0.010	**0.000**
8	8	10	20	4	0.5	0.5	16.497	16.505	**0.008**	0.000	0.000	**0.000**	0.005	0.005	**0.000**

Table 15: Performance of approximation with varying system parameters