Computers & Industrial Engineering 191 (2024) 110171

Contents lists available at ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie

Joint optimization of inventory replenishment and rationing policies for an omnichannel store with both in-store and online demands

Zengxu Guo, Haoxun Chen *

Logistics and Optimization of Industrial Systems, Laboratory of Computer Science and Digital Society, University of Technology of Troyes, Troyes, France

ARTICLE INFO

Keywords:
Inventory
Omnichannel retailing
Demand classes
Rationing policy
Optimization

ABSTRACT

Boosted by the COVID-19 pandemic, omnichannel retail has become a popular retail model and is expected to further grow. We study inventory replenishment and rationing of a product in an omnichannel store that fulfils both in-store and online demands. Both demands follow Poisson process. As each in-store demand must be satisfied immediately whereas each online order has a delivery time window, the two demand classes have different behaviors facing stockout: offline demand is lost while online order is backlogged with backorder cost charged only if it is delivered late. The inventory replenishment and rationing between the two demand classes in this omnichannel inventory system are controlled by a continuous review (Q, r) policy and a critical level policy, respectively. The intertwined nature of inventory replenishment and rationing in this system makes its analytical cost evaluation and its joint optimization of the two policies challenging. After deriving an analytical formula for evaluating the expected cost of the system, two heuristic algorithms are proposed to jointly optimize the two policies. Numerical experiments on randomly generated instances demonstrate that the multi-start local search algorithm and the scatter search algorithm can obtain a near-optimal solution quickly with percentage cost gap no larger than 0.81% and 0.18%, respectively, and the critical level policy can yield cost savings up to 10.91% compared to the first-come, first-served policy. Finally, a sensitivity analysis is conducted to evaluate the influences of system parameters on the optimal critical level.

1. Introduction

Omnichannel retailing, which integrates multichannel sales and marketing (Difrancesco et al., 2021; Hübner et al., 2022), aims to improve customer shopping experience by eliminating the barrier between online and offline channels. The integration of online platforms in conventional retailers makes using store inventory to meet the needs of online consumers possible (Xu & Cao, 2019). Smaller brick-and-mortar retailers are embracing this approach to serve online customers by converting their brick-and-mortar stores into distribution centers (Mou et al., 2018). One of omnichannel retailing modes that is gaining an increasing acceptance by retailers is known as buy-online-delivery-from-store (BODS) mode. This mode can benefit retailers. Firstly, it facilitates quick order fulfillment from nearby stores, reducing delivery times and costs, and improving customer satisfaction (Bayram & Cesaret, 2021; He et al., 2021, Goedhart et al., 2022). Secondly, it optimize resource allocation by leveraging existing store networks instead of maintaining separate warehouses dedicated solely to online orders (Yang & Zhang, 2020). Thirdly, it also aids in optimizing inventory turnover, particularly for perishable or time-sensitive products, mitigating the risk of product obsolescence (Bayram & Cesaret, 2021; Bendoly, 2004). It can provide flexible delivery options such as in-store pickup and same-day delivery (Goedhart et al., 2022).

In real business, many retailers have adopted the BODS strategy. One famous example is the retail giant Target, which has used this strategy to improve its ability to serve online demand. With the help of its widespread store network, Target can achieve faster online order fulfillment by using its store inventory, allowing customers to receive their orders on the same or next day. In the second quarter of 2023, Target successfully fulfilled 97.6% of its online orders through its stores (Stambor, 2023). Moreover, its last-mile delivery capability ensures that approximately 40% of orders are delivered on the next day.

Although using the inventory of a store to fulfill online orders provides advantages, it makes inventory management of the store challenging. Popular items may rapidly sell out in a physical store due to combined online and offline demand, disappointing offline shoppers. Therefore, effective inventory rationing between offline and online demand in a store is critical for improving customer shopping experiences

* Corresponding author.
E-mail addresses: zengxu.guo@utt.fr (Z. Guo), haoxun.chen@utt.fr (H. Chen).

https://doi.org/10.1016/j.cie.2024.110171
Received 21 November 2023; Received in revised form 11 April 2024; Accepted 20 April 2024
Available online 21 April 2024

while reducing inventory costs.

In omnichannel retail, if desired products are not available in a store, walk-in customers usually leave unsatisfied, resulting in lost sales. On the other hand, online orders can be backlogged since customers foresee their orders' delivery lead time and rarely cancel the orders (Enders et al., 2014; Vicil, 2021). Therefore, it is more realistic to assume lost sales for unsatisfied offline demands and backorders for unsatisfied online demands in case of stockout. In addition, a delivery time window is generally proposed for each online order, and an omnichannel retailer usually offers a compensation to online customers in case of late delivery of their orders, so backorder costs of an online order are only charged when it is delivered later than its promised delivery date. The backorder costs may also include the costs of loss of future online sales due to the late delivery.

Most previous studies consider homogeneous customer behaviors facing stockout (Hosseini-Motlagh et al., 2021; Wan & Cao, 2018; Y. Wang et al., 2020). They assume either all customers accept backorders or all demands are lost during stockout. Only a limited number of papers consider heterogeneous customer behaviors facing stockout by assuming both backorders and lost sales (Wang and Tang, 2014; Zhou and Zhao, 2010; Enders et al., 2014; Vicil, 2021; Huang et al., 2011). However, they assume a zero lead time for inventory replenishment, or do not consider on-hand inventory rationing between two demand classes, or consider a simpler inventory policy like one-to-one policy or base-stock policy for replenishment without considering ordering costs. Only one paper considers online order delivery time window (Huang et al., 2011), but it does not address rationing between two demand classes. To the best of our knowledge, no paper proposes algorithms other than enumeration for joint optimization of inventory replenishment and rationing policies for an omnichannel inventory system with positive lead time, (Q, r) policy for replenishment, a critical level policy for rationing, and online order delivery time window.

This paper is focused on analytical cost evaluation and optimization of a continuous-review retailing inventory system with online and offline demand and positive lead time. Both demands follow a Poisson process. The inventory replenishment of the system is controlled by a widely used (Q, r) policy. For rationing the on-hand inventory of the system between demands of the two classes when they occur, a critical level policy is adopted as in many papers (Wang et al., 2013; Wan & Cao, 2018; Escalona et al., 2019; Vicil, 2021). Specifically, when demands of the two classes occur, if the on-hand inventory is higher than a critical level K, both demands are satisfied on a first-come, first-served (FCFS) basis. Otherwise, only offline demand is fulfilled, while online demand is backlogged. For simplicity of exposition, the combination of the two policies is referred to as (Q, r, K) policy hereafter.

The theoretical novelty and practical significance of this paper can be encapsulated in three aspects. Firstly, our inventory model captures main features of inventory management in an omnichannel store. In case of stockout, we assume offline demand is lost whereas online demand is backlogged. This feature is strongly relevant to inventory management of an omnichannel store but is ignored in previous studies. In addition, we assume each online order is delivered to its customer within a specified time window, backorder costs of this order will not be incurred within the time window even if it cannot be immediately delivered, this is another new feature of our inventory model. This feature is particularly pertinent in e-commerce, where a delivery due date is often set for each online order. Notably, we provide an analytical method for the cost evaluation of this new inventory model. Secondly, we develop two effective heuristic algorithms for the optimization of the (Q, r, K) policy employed in the inventory system studied, which can avoid fully or partially enumerating the parameter values of the policy for its optimization as done in the previous literature. Thirdly, we demonstrate by extensive numerical experiments the superiority of the critical level policy over the FCFS policy for rationing on-hand inventory between two demand classes in an omnichannel store, and identify key system parameters affecting the performance of the first policy.

The contributions of this paper are summarized as follows:

1. We study for the first time inventory replenishment and rationing in a continuous review inventory system controlled by a (Q, r, K) policy with two demand classes of backorders and lost sales, positive lead time, and delivery time window for online orders.
2. An analytical expression for exactly evaluating the expected total cost per unit time of the system is derived. Its accuracy is validated through extensive numerical simulations.
3. We develop two efficient and effective heuristic algorithms, multi-start local search and scatter search, for joint optimization of the inventory replenishment and rationing policies.
4. Managerial insights are provided for inventory management of the studied system. Specifically, our numerical study reveals critical conditions under which the proposed rationing policy performs well.

This paper is organized as follows. Section 2 offers a comprehensive review of relevant literature. Section 3 describes the studied inventory system with two demand classes. The expected total cost of this system is analytically derived in Section 4. Section 5 presents two algorithms for parameter optimization of inventory replenishment and rationing policy for the system. The performances of the proposed algorithms are evaluated by numerical experiments in Section 6, with a sensitivity analysis to evaluate the impact of system parameters and derived management insights. Finally, this paper is concluded with a summary of its key findings and remarks for future research.

2. Literature review

This study falls within the research domain of managing inventory systems with multiple demand classes, with a specific focus on a critical level based rationing policy. The study of inventory rationing is dated to 1965, when Veinott proposed a critical level policy to cope with multiple demand classes in a periodic review inventory system (Veinott, 1965). Under this policy, when on-hand inventory drops below a given critical level, low-priority demand is backlogged while high-priority demand continues to be fulfilled. Since then, many researchers have examined inventory rationing in various model settings with different replenishment policies. A comprehensive review of previous studies in this domain is provided by Kleijn & Dekker (1999), Arslan et al. (2007), and Möllering & Thonemann (2008). In the following, we review the relevant literature in three streams involving inventory systems with multiple demand classes of backorder type only, lost sales type only, and mixed backorders and lost sales type, respectively. A demand class is called backorder (resp. lost sales) type if all unsatisfied demands of this class are backlogged (resp. lost) in case of stockout. A summary of the relevant literature and a comparison between our study and the existing ones are presented in Table 1.

2.1. Inventory systems with multiple demand classes of backorder type only

Most studies on inventory systems with multiple demand classes assume unsatisfied demands are backordered (backlogged) for all classes. Nahmias and Demmy (1981) may be the first ones analyzing the performance of an inventory system with two demand classes of backorder type. Their analysis is based on an assumption that the system has at most one outstanding order at any time, this prevailing assumption is adopted in many studies. Kleijn and Dekker (1999) bring attention to the difficulty of analytical cost evaluation for an inventory system with two demand classes, particularly when it has a positive replenishment lead time. As a result, most studies adopt suboptimal but practical strategies, such as a threshold backorder clearing mechanism proposed by Deshpande et al. (2003). This mechanism has found widespread application in various studies, including Wan and Cao (2018), and Escalona et al. (2019). Wan and Cao (2018) study a periodic review inventory system

Table 1
Related literature.

Demand types	Literature	Inventory policy	Inventory review	Lead time	Rationing policy	Performance measure	Cost evaluation	Solution approaches
All backorders	Nahmias & Demmy (1981)	(Q, r)	continuous	positive	static	service level	approximation	none
	Deshpande et al., (2003)	(Q, r)	continuous	positive	static	cost	exact	enumeration
	Wan & Cao (2018)	(S, T)	periodic	positive	static	cost	exact	enumeration
	Escalona et al., (2019)	(Q, r)	continuous	positive	static	cost & service level	exact	enumeration
	D. Wang et al., (2013)	(Q, r)	continuous	positive	static	cost & service level	exact	heuristic
	Escalona et al., (2017)	(Q, r)	continuous	positive	static	cost & service level	exact	heuristic
	Möllering & Thonemann (2010)	base stock	periodic	positive	static	service level	exact	optimal & heuristic
	Y. Wang et al., (2013)	base stock	periodic	positive	static	service level	exact	enumeration
All lost sales	Melchiors et al., (2000)	(Q, r)	continuous	positive	static	cost	exact	enumeration
	Frank et al., (2003)	(s, S)	periodic	zero	dynamic & static	cost	exact	optimal & heuristic
Mixed backorder and lost sales	D. Wang & Tang, (2014)	(S, T)	periodic	zero	dynamic	cost	none	heuristic
	Zhou & Zhao (2010)	base stock	periodic	zero	none	cost	recursive (exact)	optimal
	Enders et al., (2014)	base stock	continuous	positive	static	cost	exact	enumeration
	Vicil (2021)	One-for-one	continuous	positive	static	cost	exact	enumeration
	Huang et al., (2011)	(Q, r)	continuous	positive	none	cost	exact	enumeration
	Guo & Chen, (2023)	(Q, r)	continuous	positive	static	cost	exact	none
	This study	(Q, r)	continuous	positive	static	cost	exact	heuristic

controlled by a (S, T) replenishment policy and a critical level rationing policy, and present a novel approach for accurately evaluating long-term average inventory costs of the system. Escalona et al. (2019) investigate the design of a critical level rationing policy to provide differentiated service levels for two demand classes of fast-moving items and propose global search algorithms to solve this problem. Möllering and Thonemann (2010) employ Markov chains to model and analyze a periodic review inventory system with two demand classes of backorder type. Wang et al. (2013) use a similar modeling and analysis approach, but they adopt an anticipated rationing policy for the system. This policy allocates on-hand inventory to higher-priority demands based on both a constant critical level and incoming replenishments.

2.2. Inventory systems with multiple demand classes of lost sales type only

The study of inventory systems with multiple demand classes of lost sales type only has received less attention than those with all backorder type demands. Analyzing an inventory model with lost sales is generally more intricate than the corresponding model with backorders. Melchiors et al. (2000) investigate a continuous review inventory system controlled by a (Q, r) replenishment policy and a critical-level rationing policy, they derive a formula for exactly calculating the average inventory cost of the system. Frank et al. (2003) study a periodic review inventory model with two demand classes of distinct priorities, one deterministic and the other stochastic, and characterize the structure of the optimal replenishment and rationing policy. They show that this policy is state dependent in general.

2.3. Inventory systems with multiple demand classes of mixed backorders and lost sales

Analyzing inventory models with both lost sales and backorders is notably more intricate compared to analyzing pure backorder models. For this reason, the study of inventory systems with multiple demand classes of mixed backorders and lost sales has received limited attention in the literature. Wang and Tang (2014) study an inventory system with a mix of backorder and lost sales demands, whose priorities change dynamically. By assuming zero replenishment lead time, they propose a dynamic rationing policy for the system. Zhou and Zhao (2010) also ignore the replenishment lead time in a periodic review inventory system with two demand classes of different priorities, where a demand of high-priority has a higher priority of using on-hand inventory to fulfill it than a demand of low-priority, without a particular rationing policy used to allocate on-hand inventory between the two demand classes. In case of stockout, the first demand is lost while the second demand is backordered. They formulate the inventory replenishment decision problem of this system as a dynamic programming problem and characterize the structure of the optimal replenishment policy. Enders et al. (2014) study a continuous review inventory system controlled by a base stock policy for replenishment and a critical level policy for rationing, they develop an efficient algorithm to determine the optimal critical level. Vicil (2021) investigates a continuous review inventory system employing a one-for-one replenishment policy a critical level rationing policy. They derive the steady-state probability distribution and the expected cost per unit time of the system and propose an efficient procedure to minimize the expected cost.

2.4. Research gap and objective

Our study differs from the relevant literature in several aspects, with the primary distinction on a common assumption made in most of the literature. That is, customers in all categories are assumed to have the same behavior when they face unfulfilled demands. More precisely, when immediate fulfillment is impossible, the existing literature usually treats the demands of all classes as backorders (Nahmias & Demmy, 1981; Deshpande et al., 2003; Möllering & Thonemann, 2010; D. Wang et al., 2013; Escalona et al., 2017, 2019; Wan & Cao, 2018) or alternatively as lost sales (Frank et al., 2003; Melchiors et al., 2000). However, in an omni-channel store serving both online and offline customers, unavailability of demanded products in the physical store usually leads to lost sales. On the other hand, online orders are usually not delivered immediately, but accumulate as backorders, which will be delivered/fulfilled within their delivery time window, with rare cancellations by customers. For the two reasons, we consider different behaviors for the two demand classes facing stockout, with unfulfilled online demand backordered and unsatisfied offline demand lost.

In comparison to the limited literature that examines inventory systems with multiple demand classes of both lost sales and backorders (Enders et al., 2014; Huang et al., 2011; Vicil, 2021; D. Wang & Tang, 2014; Zhou & Zhao, 2010), our study considers several distinguishing features. Firstly, our study differs from most of the aforementioned studies in terms of the replenishment policy employed in the inventory system considered. We study a continuous review inventory system controlled by a (Q, r) policy, whereas Wang & Tang (2014) and Zhou & Zhao (2010) consider a periodic review system controlled by a (S, T) policy (Wang & Tang, 2014), a base stock policy (Zhou & Zhao, 2010; Enders et al., 2014), or a one-for-one policy (Vicil, 2021). Secondly, our inventory model takes into account positive lead time for inventory replenishment, whereas Wang & Tang (2014) and Zhou & Zhao (2010) assume zero lead time.

The research conducted by Huang et al. (2011) is closely related to our study, as they explore a similar system controlled by a (Q, r) replenishment policy, with both in-store and online demands. They also treat unfulfilled online demands as backorders and unsatisfied offline demands as lost sales. However, their inventory model does not incorporate on-hand inventory rationing between the two demand classes. In contrast, we adopt a critical level policy for this rationing and our numerical experiments demonstrate that implementing the rationing policy can yield substantial cost reductions compared to applying a simple FCFS rule to fulfill both online and offline demand. In addition, our model takes into consideration an important feature of online shopping, that is, there is a time window for the delivery of each online order. Moreover, we propose two effective and efficient heuristics for optimizing the parameters of the inventory replenishment and rationing policies considered. The conference paper of Guo & Chen (2023) deals with the same problem as we consider in this paper. However, they do not propose any optimization method. In addition, they do not analyze the value of the rationing policy.

To recapitulate, our research makes a valuable contribution to the existing literature by investigating a continuously review inventory system characterized by having two distinct demand classes and a positive lead time, adopting a (Q, r, K) policy for replenishment and rationing, with both lost sales and backorders, and a delivery time window for each online order. We provide an analytical approach for evaluating the system's expected cost per unit time and develop two efficient and effective heuristics for parameter optimization of its inventory replenishment and rationing policies.

3. Problem description

We study inventory management of a general omnichannel store that fulfils both offline and online demands, under the condition that both demands are Poisson processes or can be approximated by Poisson processes. The Poisson demand model is widely adopted in the inventory literature (Turrini & Meissner, 2019), as it is more analytically tractable than other demand processes and provides a reasonable approximation to real demands especially for slow-moving products, which often account for 30% - 35% of the products sold by a retailer. The Poisson demand model is prevalently adopted in omnichannel inventory management research, as evidenced in Gabor et al. (2022), Goedhart et al. (2022) and Vicil (2021). It is worth noting that the analytical cost evaluation of the inventory system with Poisson demands to be presented in Section 4 can be extended to that with discrete compound Poisson demands, which can model the demands of most products in omnichannel retail. In addition, our optimization methods to be presented in Section 5 are independent of the demand model adopted.

Since the orders placed by online customers are usually delivered from their closest store, stores are relatively independent in terms of their online demands/orders. For this reason, in this paper, we focus on inventory replenishment and rationing of a product in a single omnichannel store. The results of this paper can be served as a basis for studying inventory management of an omnichannel retailer with multiple stores.

In the inventory system we study with the two classes of Poisson demand, the offline demand of class 1 with rate λ_1 has a higher priority, necessitating immediate fulfillment from the on-hand inventory of the store; otherwise, the demand is lost. The online demand of class 2 with rate λ_2 has a lower priority, it is backlogged if it cannot be immediately fulfilled.

The replenishment of this inventory system is controlled by a continuous review (Q, r) policy with constant replenishment lead time denoted by L. The rationing between the two demand classes in this system is governed by a predefined critical level K. If the inventory on-hand exceeds K, all demands from both classes are fulfilled on a first-come, first-served basis. Otherwise, if the inventory on-hand falls to or below K, the system backlogs online orders, while it continues fulfilling offline orders until out of stock. Overall, the system is managed by a (Q, r, K) policy, which combines a (Q, r) replenishment policy and a critical level rationing with threshold K. It is assumed that the store has at most one outstanding order, common in inventory models with lost sales (Huang et al., 2011; Melchiors et al., 2000). This assumption is particularly reasonable in situations where the order cycle is longer than the replenishment lead time (Tempelmeier, 2006).

In online shopping, each order typically has a delivery time window, which is defined by its latest delivery time promised by the supplier. Based on this practice, we assume there is a maximum acceptable waiting time, represented as T_{cus}, for the delivery of each online order, T_{cus} is the width of the delivery time window. Differing from conventional backorder models where backorder costs are incurred immediately when an order cannot be fulfilled immediately, in our inventory model, no backorder cost is incurred for an online order if it can be delivered within its delivery time window, i.e., delivered before the maximum acceptable waiting time is reached. This feature is a unique characteristic of online shopping. Note that T_{cus} should be no longer than the replenishment lead time L. Otherwise, the fulfillment of an online order can be achieved without keeping any inventory dedicated to it. Consequently, this system can be reduced to a simplified one, where only offline demands need to be taken into account.

Before proceeding to analytical cost evaluation of the inventory system, we introduce the notations necessary for this analysis.

Decision variables	
r	Reorder point
K	Critical rationing level
Q	Ordering quantity
Parameters	
λ_i	Demand rate for channel i ($i = 1$ for the offline channel, $i = 2$ for the online channel)
π	Unit lost sales cost for offline demand
b	Unit backorder cost per unit time for online demand
h	Unit holding cost per unit time
F	Fixed ordering cost
L	Replenishment lead time
T_{cus}	Maximum acceptable waiting time or width of the delivery time window for each online order
Other notations	
t_B	Time when the inventory level hits the critical rationing level K
t_L	Time when the on-hand inventory is depleted
t_W	Starting time after which backlogged orders can be fulfilled within the maximum acceptable waiting time T_{cus}
t_W'	Time when the backorder cost starts to be incurred
B_T	Time-weighted backorder level in a cycle
EC	Expected cost per time unit
T	Length (duration) of a cycle
D_i	Demand from channel i ($i = 1$ for the offline channel, $i = 2$ for the online channel)
I	Time-weighted inventory held in a cycle
$I(t)$	Inventory level at time t
$D_i(t_1, t_2)$	Cumulative demand of class i from time t_1 to time t_2
LS	Lost sales quantity in a cycle

(continued on next page)

(*continued*)

u	Inventory level immediately prior to the arrival of a replenishment order
T_{LS}	Period of time during which lost sales occur
T_B	Period of time during which backorder costs are incurred
C	Total cost during a cycle

Fig. 1 illustrates the inventory dynamics of the system during an order cycle governed by the (Q, r, K) policy. The cycle starts at time t_0 when the inventory level of the system reaches the reorder point r, progresses through time t_0+L with the arrival of a new replenishment, and ends at time t_0+T when the inventory level of the system is raised to the reorder point r again. The assumption of at most one outstanding order in the system ensures that its inventory level $I(t)$ exhibits a regenerative property, whereby a cycle is regenerated each time a replenishment order is placed.

Specifically, at time t_0, the inventory level of the system reaches its reorder point r, triggering its placement of an order of quantity Q to replenish its inventory. When its on-hand inventory is higher than K, online and offline demands/orders are fulfilled in a first-come, first-served manner, making its inventory level decreasing at a rate denoted by λ ($\lambda = \lambda_1 + \lambda_2$). At time t_B, when the inventory level hits the critical threshold K, the critical level policy comes into effect, reserving the remaining inventory for offline demand. Consequently, only offline orders can be immediately satisfied, resulting in a decrease of on-hand inventory at a rate λ_1 (illustrated by the thin line in Fig. 1). Time t_L is reached upon the depletion of inventory. Passing this time, the offline demands that cannot be fulfilled are lost until a new replenishment arrives.

At time $t_0 + L$, the arrival of a replenishment raises the on-hand inventory, which enables the fulfilment of both online and offline demands again. This regenerative order cycle ends at time t_0+T with the placement of a new replenishment order. Because of the existence of a maximum acceptable waiting time T_{cus} for online order delivery, backorder costs are only incurred in the period from $t_W^{'}$ to $t_0 + L$. This period is denoted by T_B.

For simplicity of mathematical expression, for a Poisson demand with rate λt, we define its probability of j arrivals within a unit time, as follows.

$$p(j;\lambda t) = \frac{(\lambda t)^j e^{-\lambda t}}{j!} \tag{1}$$

The tail probability of this Poisson process is denoted by

$$q(j;\lambda t) = \sum_{i=j}^{\infty} p(i;\lambda t) \tag{2}$$

The evolution of the system in an infinite time horizon is composed

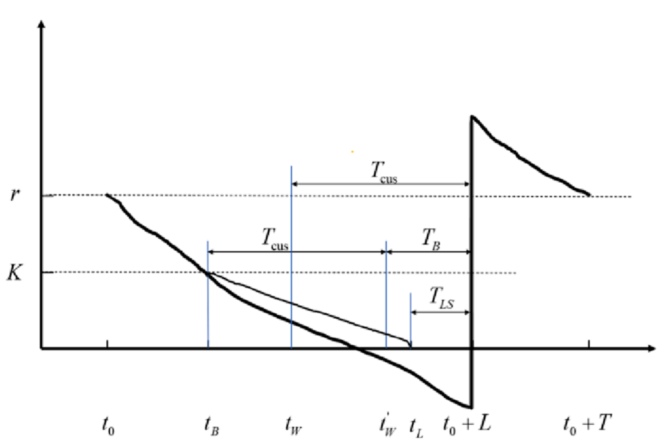

Fig. 1. The inventory process of an inventory system with both offline and online demands.

of renewal cycles. Each cycle ends with a regenerative epoch. Fig. 1 illustrates a renewal cycle of the system. The expected cost per unit of time of the system can be calculated by Eq. (3), where C represents the total cost incurred during a cycle, and T represents the duration of the cycle.

$$EC(Q,r,k) = \frac{E[C]}{E[T]} \tag{3}$$

The primary objective of this study is to minimize the expected cost of the inventory system per time unit by optimizing the parameter values of its (Q, r, K) policy. The analytical formula for calculating the expected cost EC defined in Eq. (3) needs to be derived. The costs of this system include fixed ordering, holding, lost sales, and backorder costs as indicated in Eq. (4). In the subsequent section, an analytical expression will be derived for each of the costs.

$$E[C] = F + hE[I] + \pi E[LS] + bE[B_T] \tag{4}$$

4. Cost evaluation

4.1. Cycle time

The expected order cycle time can be calculated as the sum of two expected times. The one is L, involving the period from t_0 to $t_0 + L$ in Fig. 1. The other involves the period from $t_0 + L$ to $t_0 + T$. During the second period, the inventory level of the system is decreased from $u + Q$ to r with the demand arrival rate λ, where u is a random variable representing the inventory level just before the arrival of a new replenishment. Note that $r - Q \leq u \leq r$ because of the presence of at most one outstanding order. Therefore, the expected order cycle time is given by:

$$\mathrm{E}[T] = L + \frac{Q + E[u] - r}{\lambda} \tag{5}$$

To calculate $\mathrm{E}[T]$, only $E[u]$ needs to be calculated in Eq. (5). We calculate it in two scenarios about r and Q. In the first scenario where r is greater than or equal to Q, all offline orders are fulfilled and no lost sales occur. In this scenario, the inventory level u can be calculated as the initial inventory level r subtracted by the total demand incurred during the period $(t_0, t_0 + L)$. Then,$E[u]$ can be calculated by the following formula:

$$\mathrm{E}[u] = r - \mathrm{E}[D(t_0, t_0 + L)] \tag{6}$$

In the second scenario where r is less than Q, lost sales may occur. In this scenario, it is imperative to directly analyze the probability distribution of u, which is given by:

$$p(u=i) = \begin{cases} p(D(t_0,t_0+L) = r-i) & \text{for } K < i \leq r \\ p(D(t_B,t_0+L) = K-i) & \text{for } 0 \leq i \leq K \\ p(\text{event 1}) \cup p(\text{event 2}) & \text{for } r-Q \leq i < 0 \end{cases} \tag{7}$$

where event 1 represents $D_1(t_B{,}t_0+L) \geq K{,}D_2(t_B{,}t_0+L) = -i$, and event 2 represents $D_1(t_B,t_0+L) < K{,}D_2(t_B,t_0+L) = K-D_1(t_B,t_0+L)-i$.

The probability of u in Eq. (7) is calculated in three cases. In the first case with $K < u \leq r$, the critical level policy remains inactive in the order cycle considered, u can then be calculated by deducting the demand during the period from t_0 to t_0+L from the initial inventory level r. In the second case with $0 \leq u \leq K$, the rationing policy is activated at time t_B and the on-hand inventory remains above zero without any lost sales. Subtracting the cumulative demand for duration (t_B, t_0+L) from the threshold K yields u. In the third case with $r-Q \leq u < 0$, two scenarios may happen during the period from t_B to t_0+L: with and without lost sales. In the presence of lost sales, the offline demand in this period exceeds the threshold K and equals the negative of the inventory level u. Otherwise, in the absence of lost sales, the offline demand in this period is less than K and equals K minus both the offline demand and inventory level u.

With $p(u=i)$ in Eq. (7), since $r - Q \leq u \leq r$, the expectation of u is given by:

$$\mathrm{E}[u] = \sum_{i=r-Q}^{r} i \times p(u=i) \tag{8}$$

4.2. Lost sales cost

Consider the cycle in Fig. 1, time t_B at which the inventory level hits the critical level K is a random variable. Denote the probability density function of t_B by $f_{t_B}(t)$. From time t_0 to t_B, $r-K$ units of demand arrive at a constant rate of λ. Consequently, the distribution of t_B follows an Erlang distribution with shape $r-K$ and rate λ. Given time t_B, the lost sales occurred in the cycle, denoted by LS, can be calculated as the part of the offline demand occurred during the time period (t_B, t_0+L) exceeding the reserved on-hand inventory K, that is, $LS = [D_1(t_B, t_0+L) - K]^+$. Consequently, we can derive the following distribution of LS:

$$p(LS=i) = \begin{cases} p(D_1(t_B, t_0+L) = i+K) & \text{for } i>0 \\ p(D_1(t_B, t_0+L) \le K) & \text{for } i=0 \end{cases} \tag{9}$$

For simplicity of exposition, for any function φ with two variables, demand and duration, we use $\mathbb{E}_{(D_i(t_B,t_0+L),t_B)}[\varphi(D_i(t_B, t_0+L), t_B)]$ to denote the mathematical expectation of $\varphi(D_i(t_B, t_0+L), t_B)$ with respect to $D_i(t_B, t_0+L)$ and t_B. Consequently, the expectation of the lost sales per cycle is given by Eq. (10).

$$\mathbb{E}_{(D_1(t_B,t_0+L),t_B)}[LS] = \int_0^{\infty} \sum_{i=1}^{\infty} ip(D_1(t_B, t_0+L) = i+K) f_{t_B}(t)dt \tag{10}$$

where $f_{t_B}(t) = \lambda p(r-k-1; \lambda t)$ is the probability density function of Erlang random variable t_B with shape and rate parameters $r-K$ and λ, respectively.

4.3. Backorder cost

Starting from time t_B in the cycle, online orders are backlogged until the arrival of a new replenishment at time t_0+L. The backorders occurred in the time period from t_W to t_0+L do not incur any backorder cost since these orders are delivered within their maximum acceptable waiting time T_{cus}. Consequently, the backorder level that needs to be counted is $B(t_W^{'}, t) = D_2(t_B, t-T_{\text{cus}})$ for all time $t \in (t_W^{'}, t_0+L)$. The backorder level weighted by time within the cycle, denoted by B_T, is derived by Eq. (11).

$$B_T = \int_{t_W^{'}}^{t_0+L} B\left(t_W^{'}, t\right)dt \tag{11}$$

Therefore, the expectation of B_T is given by Eq. (12).

$$\begin{aligned}\mathbb{E}_{(D_2(t_B,t_0+L),t_B)}[B_T] &= \mathbb{E}_{(D_2(t_B,t_0+L),t_B)}\left[\int_{t_W^{'}}^{L} D_2(t_B, t-T_{\text{cus}})dt\right] \\ &= \int_0^{t_W}\left(\left(\int_s^{t_W} \lambda_2(t-s)dt\right) f_{t_B}(s)ds\right)\end{aligned} \tag{12}$$

4.4. Holding cost

The expected holding cost in the cycle can be derived based on the findings of Tijms (1994). According to the author, the holding costs can be calculated by aggregating the accrued holding costs for each encountered inventory level. More precisely, the holding cost at each level i results from the multiplication of three values: the level i, the expected time spent at that level, and the holding cost h per unit of product per unit of time. If n units of demand occur within a time interval of duration t, the expected time at which each inventory level spends can be computed in two scenarios: if the time interval concludes with a demand unit, the expected time is t/n; if the time interval concludes without a demand unit, the expected time is $t/(n+1)$.

The expected inventory held weighted by time of the system for the cycle can be obtained from $E[I_1]$ and $E[I_2]$, which correspond to the time-weighted inventory in the periods before and after the arrival of a replenishment, respectively.

4.4.1. Holding cost before the arrival of a replenishment

The computation of the holding cost prior to the replenishment arrival in the cycle can be divided into two scenarios based on the presence or absence of backorders. Let the subscripts $1a$ and $1b$ denote these two scenarios respectively. In the case where the cumulative demand in the time interval (t_0, t_0+L) does not exceed $r-K$, the inventory level is distributed in the range from $r-D(t_0, t_0+L)$ to r, with the expected duration of $L/(D(t_0, t_0+L)+1)$ at each inventory level i with $r-D(t_0, t_0+L) \le i \le r$. Consequently, Eq. (13) gives the expected inventory held weighted by time for the period prior to the replenishment arrival.

$$\begin{aligned}\mathbb{E}_{D(t_0,t_0+L)}[I_{1a}] &= \mathbb{E}_{D(t_0,t_0+L)}\left[\sum_{i=r-D(t_0,t_0+L)}^{r} i\frac{L}{D(t_0, t_0+L)+1}\right] \\ &= \sum_{j=0}^{r-K}\sum_{i=r-j}^{r} i\frac{L}{j+1}p(j; \lambda L)\end{aligned} \tag{13}$$

In the second scenario where $D(t_0, t_0+L)$ exceeds $r-K$, the first backorder occurs at time t_B in the time interval between t_0 and t_0+L, this time interval can be partitioned into two subintervals: before and after time t_B, represented by subscripts $1b1$ and $1b2$, respectively. Prior to time t_B, the distribution of the inventory level spans from K to r, with the expected time of $(t_B-t_0)/(r-K)$ at each inventory level. Hence, the expected inventory weighted by time in the period (t_0, t_B) can be formulated as:

$$\mathbb{E}_{t_0<t_B<t_0+L}[I_{1b1}] = \mathbb{E}_{t_0<t_B<t_0+L}\left[\sum_{i=K}^{r} i\frac{t_B-t_0}{r-K}\right] = \int_0^L\left(\sum_{i=K}^{r} i\frac{t_B-t_0}{r-K}\right)f_{t_B}(t)dt \tag{14}$$

Since time t_0 can be taken as zero, the above formula can be rewritten as

$$\mathbb{E}_{t_0<t_B<t_0+L}[I_{1b1}] = \sum_{i=K}^{r}\frac{i}{r-K}\int_0^L tf_{t_B}(t)dt \tag{15}$$

Next, we calculate the expected inventory weighted by time for the time subinterval $(t_B, t_0+L]$. At any given time t in this subinterval, the inventory held can be calculated as $[K - D_1(t_B, t)]^+$. Therefore, the expected inventory weighted by time in this subinterval can be formulated as:

$$\begin{aligned}\mathbb{E}_{(D_1(t_B,t_0+L),t_0<t_B<t_0+L)}[I_{1b2}] &= \mathbb{E}_{(D_1(t_B,t_0+L),t_0<t_B<t_0+L)}\left[\int_{t_B}^{L}[K-D_1(t_B, t)]^+dt\right] \\ &= \sum_{i=0}^{K-1}\frac{(K-i)}{\lambda_1}\int_0^L q(i+1; \lambda_1(L-s))f_{t_B}(s)ds\end{aligned} \tag{16}$$

4.4.2. Holding cost after the arrival of the replenishment

The inventory level just after the arrival of the replenishment at time t_0+L is $u+Q$. At time t_0+T, the inventory level is decreased to r. In the time interval from t_0+L to t_0+T, both online and offline orders are fulfilled with a total demand rate λ, leading to the decrease of the inventory level. The expected time to be spent at each inventory level equals $1/\lambda$. The cumulative holding costs in this time interval can be calculated from the holding costs incurred at each inventory level. Considering u as a random variable, the expected inventory weighted by time in this interval can be calculated as:

$$\mathbb{E}_u[I_2] = \mathbb{E}_u\left[\sum_{i=r+1}^{Q+u} i\frac{1}{\lambda}\right] = \frac{1}{\lambda}\sum_{j=r-Q+1}^{r}\sum_{i=r+1}^{Q+j} ip(u=j) \tag{17}$$

The expected holding cost in this cycle can thus be computed by adding up the expected holding costs calculated above as given by Eq. (18).

$$hE[I] = h(\mathbb{E}[I_{1a}] + \mathbb{E}[I_{1b1}] + \mathbb{E}[I_{1b2}] + \mathbb{E}[I_2]) \tag{18}$$

5. Optimization

The last section has presented an analytical expression for cost evaluation of the inventory system considered. In this section, we try to develop efficient methods for optimizing the parameters of the inventory replenishment and rationing policies of the system. Several studies adopting a critical level policy for an inventory system with two demand classes, including Melchiors et al. (2000) and Wan & Cao (2018), employ enumeration and bounding methods as optimization tools. However, these methods are quite time-consuming when applied to instances of larger size. In light of this, we devise two heuristics that can avoid the enumeration: one based on local search and the other based on scatter search. Both heuristics work in a finite domain containing the optimal parameter values of the (Q, r, K) policy employed.

5.1. Bounds for the decision variables

To define the finite domain of the decision variables Q, r, and K, it is necessary to determine their upper and lower bounds. For order quantity Q, denote its optimal value by Q^*. According to Escalona et al. (2017), Q^* usually exceeds the economic order quantity Q_{EOQ} obtained by the EOQ model assuming constant demand. Our resolution of a large number of representative instances using the enumeration method indicates that the ratio Q^*/Q_{EOQ} typically falls in the range of 1.05 to 1.11. Without compromising the generality of the findings, the range of Q is set to be 0.8 to 1.2 times Q_{EOQ}. Furthermore, we consider K with $0 \le K \le r$.

To determine a upper bound for the reorder point r, we adopt a conjecture of Melchiors et al. (2000). Denote by $r^*(Q)$ the optimal reorder point of the inventory model considered for its given order quantity Q. Melchiors et al. (2000) conjecture that $r^*(Q)$ would not exceed the optimal reordering point $\bar{r}(Q)$ of the model without implementing a critical level policy (equivalent to $K = 0$). Their conjecture is supported by numerical experiments.

However, it is difficult to directly calculate $\bar{r}(Q)$ even if K is 0, because our model with both backorders and lost sales is more complex than theirs. To overcome this difficulty, we divide the single stock of our model into two stocks, one stock serving offline demands and the other serving online orders. The inventory models associated with these two stocks can be treated as a pure lost sales model with demand rate λ_1 and a pure backorder model with demand rate λ_2 respectively. For the two models, we can derive upper bounds of their reorder points, denoted by $\widetilde{r}_1(Q_1)$ and $\widetilde{r}_2(Q_2)$, respectively, where Q_1 and Q_2 are the order quantities of the two independent inventory models, respectively. Considering the inventory pooling effect, the sum of $\widetilde{r}_1(Q_1)$ and $\widetilde{r}_2(Q_2)$ is taken as an upper bound for $r^*(Q)$.

To obtain the upper bound $\widetilde{r}_1(Q_1)$ of the reorder point of the lost sales model, we employ the Hadley-Whitin heuristic (Hadley & Whitin, 1963). That is, for any given value of Q_1, $\widetilde{r}_1(Q_1)$ is given by:

$$\widetilde{r}_1(Q_1) = \min\left\{r \ge 0 : P(D_1(t_0, t_0 + L) \le r) \ge (\pi\lambda_1/Q_1)\big/(h + \pi\lambda_1/Q_1)\right\} \tag{19}$$

The upper bound $\widetilde{r}_1(Q_1)$ is a decreasing function of Q_1. According to Hadley & Whitin (1963), the optimal order quantity Q_1^* of the lost sales model is no less than economic order quantity $Q_{1\mathrm{EOQ}}$. Therefore, we have $\widetilde{r}_1(Q_{1\mathrm{EOQ}}) \ge \widetilde{r}_1(Q_1^*)$, then we can take $\widetilde{r}_1(Q_{1\mathrm{EOQ}})$ as an upper bound of the reorder point of the pure lost sales model. Similarly, an upper bound $\widetilde{r}_2(Q_2)$ of the reorder point of the pure backorder model can also be derived by applying the Hadley-Whitin heuristic as given below.

$$\widetilde{r}_2(Q_2) = \min\{r \ge 0 : P(D_2(t_0, t_0 + L) \le r) \ge b/(h + b)\} \tag{20}$$

This upper bound does not depend on order quantity Q_2 as Eq. (20) implies.

5.2. Multi-start local search

In this section, we introduce our first heuristic, referred to as Multi-Start Local Search (MLS for short) method, for optimizing the (Q, r, K) policy employed in the inventory system studied. MLS is based on local search. Its procedure, referred to as procedure MLS, is given in Table 2. This method performs independent multiple local searches starting from different initial solutions. The number of initial solutions generated is determined according to the solution space size of an instance of the optimization problem to be solved, with the application of specific rules described in the numerical study of the method in the next section.

The key intuition behind this heuristic is that by generating a large number of initial solutions relatively diversified in the solution space and performing local search starting from each initial solution, it would be very likely to find a high quality near-optimal solution or even the optimal solution. To achieve the diversification of initial solutions, the value range of each decision variable is partitioned into four equidistant intervals. Subsequently, a two-step approach is adopted to generate each initial solution. Firstly, the probability of selecting a subinterval in which the value of a decision variable is located is set inversely proportional to its frequency of being selected. Secondly, the value of each decision variable in an initial solution is randomly generated in its subinterval selected. It is necessary to calculate the selection frequency of each subinterval during the generation of the initial solutions.

The local search starting from each initial solution is realized by the while loop in Step 3 of this procedure. In this main loop, in each iteration, the neighbor solutions of the current solution are generated in the following way: each variable can take three possible values, that is, its value in the current solution, its current value plus 1, and its current value minus 1. In this way, totally 26 ($3 \times 3 \times 3 - 1$) neighbor solutions are generated in each iteration for this optimization problem with three variables. For each neighbor solution, its expected cost is calculated using the analytical cost evaluation method presented in Section 4 and the costs of all neighbor solutions are compared, the best neighbor solution with the smallest cost can then be obtained. If the cost reduction in percentage of the best neighbor solution with respect to the current solution exceeds a given percentage *imp* which is 0.5% in our numerical

Table 2
Procedure MLS.

	Inputs: $\lambda_1, \lambda_2, L, h, \pi, b, F, T_{\mathrm{cus}}$
Step 1	Calculate the bounds of the variable Q, r and K
Step 2	Generate initial seed solutions
Step 3	Local search is applied to each initial solution to obtain an improved solution **while** there are still initial solutions that have not been applied local search, **do**: Set the current solution to an initial solution Generate the neighbors of the current solution Get the best neighbor **if** the best neighbor is better than the current solution with the percentage cost reduction no less than *imp*: Set the current solution to the best neighbor Update the current cost by the cost of the best neighbor solution **end if** Record the final solution obtained by local search and its cost
Step 4	Sort all solutions obtained by multi-start local search according to their costs, and output the best solution with the smallest cost

study, the current solution is replaced by the best neighbor solution, this local search iteration will be repeated until no neighbor solution that reduces the cost of the current solution by the preset percentage can be found. Finally, all solutions found by local search are sorted according to their costs and the best solution with the smallest cost is output.

5.3. Scatter search

Our second heuristic for the optimization of the (Q, r, K) policy, referred to as Scatter Search (SS) method, is based on scatter search, credited to Glover (1977). This population-based iterative method combines elements of the existing solutions, using diversified and intensified search strategies. Martí et al. (2006) outlined the implementation of a scatter search algorithm in five key steps: diversification generation, solution improvement, subset generation, solution combination, and reference set update. The procedure of our SS method is given in Table 3 below.

The diversification generation step generates diversified initial solutions, akin to that in the MLS method. The solution improvement step applies local search to improve each initial solution. The local search method is terminated after three iterations for each initial solution. This yields a population of improved initial solutions.

The reference set in SS, denoted by Refset, contains two distinct subsets: Refset$_1$ and Refset$_2$. Refset$_1$ and Refset$_2$ contain respectively the n_1 best solutions and the n_2 most diverse solutions from the population of improved initial solutions, with n_1 and n_2 set to 3 and 2 respectively in our numerical study. The best solutions are selected based on cost, whereas the diverse solutions are selected based on the Euclidean distance. To obtain Refset$_2$, the minimum Euclidean distance between each candidate solution x in the initial solution set and all solutions in Refset$_1$, denoted by $d_{\min}(x)$, is calculated. The solution that maximizes the minimum distance is selected and added to Refset$_2$. This process is repeated until n_2 solutions are added to Refset$_2$. Given n solutions in Refset ($n = n_1 + n_2$), the subset generation step creates $n(n\text{-}1)/2$ subsets of two solutions from Refset, categorized into three groups: Refset$_1$- Refset$_1$, Refset$_1$- Refset$_2$, and Refset$_2$- Refset$_2$. A subset of two solutions is categorized into group Refset$_i$- Refset$_j$ if one solution from Refset$_i$ and the other from Refset$_j$, $i = 1, 2, j = 1, 2$.

In the solution combination step, two reference solutions x' and x'' in each subset are combined to generate new solutions using three types of linear combinations: Com$_1$, $x = x' - d$; Com$_2$, $x = x' + d$; and Com$_3$, $x = x'' + d$, where $d = e(x'' - x')/2$, with e randomly generated in the interval (0, 1). If both x' and x'' belong to Refset$_1$, four new solutions are generated: one from Com$_1$, one from Com$_3$, and two from Com$_2$. If only one of x' and x'' is in Refset$_1$, three new solutions are generated: one from each of Com$_1$, Com$_2$, and Com$_3$. If neither x' nor x'' is in Refset$_1$, only two new solutions are generated: one from Com$_2$ and the other from randomly selected Com$_1$ or Com$_3$. Each new solution is then improved by applying the local search method and replace the worst solution in Refset$_1$ if it is better or a solution in Refset$_2$ if it is more diverse. The process continues until a predefined iteration number *Iter* (100 in our numerical study) is reached, outputting the best solution in Refset$_1$ and its cost.

Table 3
Procedure SS.

	Inputs: $\lambda_1, \lambda_2, L, h, \pi, b, F, T_{\text{cus}}$
Step 1	Calculate the bounds of the variable Q, r and K
Step 2	Generate a population (set) of initial solutions with the diversification generation method
Step 3	Apply the local search method to improve each initial solution in the population
Step 4	Bulid reference set Refset containing two subsets Refset$_1$ and Refset$_2$
Step 5	Reference set update
	Set *FindNewSolution* = *TRUE* and *Iteration* = 0
	while *FindNewSolution* = *TRUE* and *Iteration* $\leq$ *Iter*
	Construct all possible subsets of Refset that contain two solutions with one new solution at least. Let *SubsetNum* denote the number of the possible subsets
	Set *FindNewSolution* = *FALSE*
	for (*SubsetCount* = 1 to *SubsetNum*)
	Generate the next subset s from RefSet with the Subset Generation Method
	Apply the Solution Combination Method to s to generate two, three, or four new solutions x_s
	Apply the local search method to improve x_s to obtain an improved solution x_s^*
	if $x_s^* \notin$ *Refset* and the cost of x_s^* is larger than that of the worst solution in Refset$_1$, **then**
	Replace the worse solution by x_s^* in Refset$_1$.
	Set *FindNewSolution* = *TRUE* and *Iteration* = *Iteration* + 1
	else
	if $x_s^* \notin$ *Refset*$_2$ and $d_{\min}(x_s^*) > d_{\min}(x)$ for a solution x in Refset$_2$ **then**
	Replace the solution x by x_s^* in Refset$_2$.
	Set *FindNewSolution* = *TRUE* and *Iteration* = *Iteration* + 1
	end if
	end if
	end for
	end while
Step 6	Sort all solutions in Refset according to their costs, and output the best solution with the smallest cost

6. Numerical study

In this section, numerical experiments are conducted to evaluate the efficiency and effectiveness of the optimization methods MLS and SS. In addition, we want to answer the following questions: What is the quantifiable benefit associated with the adoption of the rationing policy considered? How do various system parameters influence the performance of the rationing policy?.

6.1. Generation of instances

A total of 54 representative instances were generated for the inventory system we study. Table 4 presents the parameter settings of these instances along with their respective classifications. We take hour as time unit. The demand rate parameter λ was varied among three values, namely small, medium, and large, corresponding to 0.3, 0.6, and 1.5, respectively. Three different market shares of online and offline demands are considered: the offline demand rate is equal to, twice of, and half of the online demand rate.

The unit holding cost (h) was set to 0.5. The ratio of unit backorder cost (b) to unit holding cost (h) was set to 10:0.5, representing a service level of slightly higher than 95%. The fixed ordering cost F is set based on the average cycle time of the inventory system estimated based on EOQ model. Let Q_{EOQ} denote the economic order quantity of the system assuming constant demand, $Q_{\text{EOQ}} = \sqrt{2F\lambda/h}$, the average cycle time can then be estimated as $Q_{\text{EOQ}}/\lambda = \sqrt{2F/\lambda h}$. For each instance, the inventory replenishment lead time of the system was set to 24 h (one day), the maximum acceptable waiting time for each customer order was set to 6 h. Two average order cycle times were considered for each instance, namely 48 h (2 days) and 120 h (5 days) respectively, with the corresponding $F = 48^2\lambda h/2$and $120^2\lambda h/2$, respectively.

We set unit lost sales cost π according to unit backorder cost b and the average waiting time of a backorder before its fulfilment, that is, the lost sales cost of a unit demand is set equal to the total backorder cost of this demand incurred during its backordered period (its waiting time before fulfilment) in the corresponding backorder model. The average waiting time of each unit demand backordered before its fulfilment in the backorder model is estimated as half of the average cycle time, i.e., $Q_{\text{EOQ}}/2\lambda$. However, the real waiting time of a backorder varies from one to another. With this analysis, for each instance, we consider π under

Table 4
Parameter setting of the instances.

Group	Instance	Demand rate	Demand dominance	Order cycle	Unit lost sale cost
Group1	instance1	small	equal	short	$\pi = Q_{EOQ}b/2\lambda$
	instance2		offline demand		
	instance3		online demand		
Group2	instance4	small	equal	long	
	instance5		offline demand		
	instance6		online demand		
Group3	instance7	medium	equal	short	
	instance8		offline demand		
	instance9		online demand		
Group4	instance10	medium	equal	long	
	instance11		offline demand		
	instance12		online demand		
Group5	instance13	high	equal	short	
	instance14		offline demand		
	instance15		online demand		
Group6	instance16	high	equal	long	
	instance17		offline demand		
	instance18		online demand		
Group7-12	instance19-36	the parameter values of each instance are the same as those of the corresponding instance in instances 1–18 except π.			$\pi = Q_{EOQ}b/\lambda$
Group13-18	instance37-54	the parameter values of each instance are the same as those of the corresponding instance in instances 1–18 except π.			$\pi = Q_{EOQ}b/4\lambda$

three scenarios:$\pi = Q_{EOQ}b/2\lambda$,$\pi = Q_{EOQ}b/\lambda$, and $\pi = Q_{EOQ}b/4\lambda$.

Instances 1–18 are taken as benchmark instances with $\pi = Q_{EOQ}b/2\lambda$. Instances 19–36 and 37–54 have $\pi = Q_{EOQ}b/\lambda$ and $\pi = Q_{EOQ}b/4\lambda$, respectively. The 54 instances are divided into 18 groups. Each group has three instances which have the same parameter values except for their demand rates. The numerical experiments were performed on a PC with a CPU of Intel Core i7-10610 and a RAM of 32 GB. All algorithms were coded using Matlab.

As stated in section 5.2, the number of initial solutions in the MLS method is set based on the solution space size of an instance. The 18 benchmark instances can be classified into four categories. The first category includes instances 1 to 9, whose number of initial solutions is set to 10. The second category includes instances 10 to 12, whose number of initial solutions is set to 20. The subsequent three instances form the third category with 30 initial solutions set for MLS. The final category consists of the last three instances, with 40 initial solutions for MLS.

For a scatter search algorithm, it is recommended to set its population size at least 10 times of the size of its reference set, as mentioned in Martí et al. (2006). For our specific SS algorithm, the reference set contains 5 solutions, suggesting that a minimum of 50 initial solutions should be generated. However, through our preliminary numerical experiments on some representative instances, we found our SS algorithm with 40 initial solutions yielded more favorable results. Hence, it generates 40 initial solutions for each instance.

6.2. The effectiveness of the proposed heuristics

For instances 1–12, 19–30 and 37–48 (totally 36 instances), the enumeration method obtains an optimal solution within a reasonable time. Hence, we compare the solution qualities of the two heuristics for the remaining 18 instances with larger solution spaces. The percentage cost deviation (gap) between the solutions obtained the two heuristics and the optimal solution is illustrated in Fig. 2. The maximum relative gap observed does not exceed 0.81%, which occurs for the MLS method applied to instance 16. This indicates that both heuristics consistently generate high quality solutions. The SS method consistently performs better, with a maximum relative gap of 0.18%, and can find the optimal solution for one third of the instances.

In terms of computation time, for the 36 instances with relatively small solution spaces, the enumeration method quickly finds an optimal solution, making its comparison with the two heuristics irrelevant. Table 5 presents the computation times required to solve the remaining 18 larger instances using the enumeration, MLS, and SS method, respectively. The MLS and SS methods reduce computation time by at least 71.61% and 77% respectively compared to the enumeration method. The SS method is proved more efficient for instances with high demand rate, whereas the enumeration method becomes time-consuming, surpassing 12 h for some of the instances. In contrast, the MLS and SS methods never exceed two and one hour, respectively. Therefore, we recommend using direct enumeration for small-size instances, enumeration and MLS for medium-size instances, and the SS method for large-size instances.

6.3. The value of rationing

To evaluate the value of rationing, numerical experiments are conducted to compare the critical level based rationing policy considered and its non-rationing counterpart or the FCFS rule previously adopted in Huang et al. (2011). The combination of the (Q, r) policy and the FCFS rule is referred to as (Q, r, FCFS) policy hereafter. To ensure a comprehensive comparison, the 54 instances we tested are divided into 18 groups. Each group contains 3 instances. For all instances in each group, their parameters are the same except for their two demand rates. Specifically, the three instances in each group have the ratio of λ_2 to λ as 1:3, 1:2, and 2:3, respectively. The percentage cost reduction of the critical level policy with respect to the FCFS policy for each instance is presented in Table 6.

Firstly, the adoption of the critical level rationing policy leads to cost

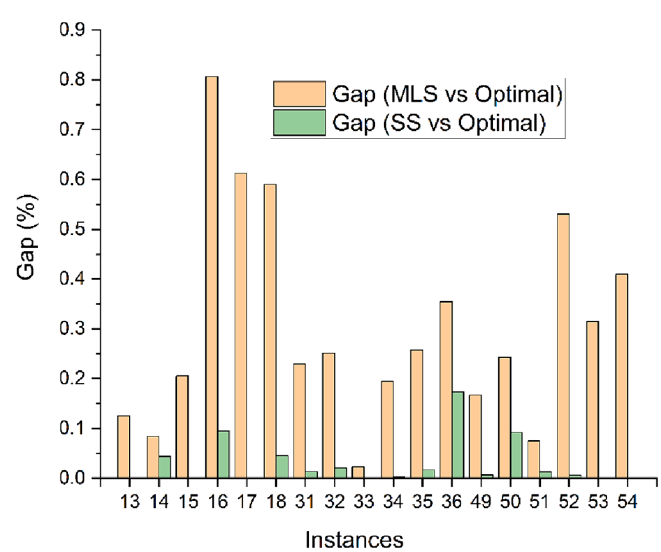

Fig. 2. Comparison between the optimal solution and the solutions obtained by MLS and SS.

Table 5
CPU time.

Instances	Computation time (s)			Time saving (%)	
	Enumeration	MLS	SS	MLS vs Enu	SS vs Enu
13	8960.50	1630.36	1690.03	81.81	81.14
14	8860.98	1486.84	1979.30	83.22	77.66
15	8865.62	1710.76	1682.68	80.70	81.02
16	40851.97	3219.31	3337.50	92.12	91.83
17	40815.76	2993.68	2430.44	92.67	94.05
18	40805.12	5405.31	3352.22	86.75	91.78
31	9983.73	1888.49	1363.63	81.08	86.34
32	9871.14	1982.94	1746.15	79.91	82.31
33	9400.80	1812.15	1669.52	80.72	82.24
34	45517.68	3221.51	2992.97	92.92	93.42
35	45353.37	4071.12	2407.90	91.02	94.69
36	42944.03	3822.51	3514.23	91.10	91.82
49	8007.71	2273.32	1836.48	71.61	77.07
50	8000.98	1847.43	1617.77	76.91	79.78
51	8021.72	2012.19	1614.83	74.92	79.87
52	36592.84	3506.84	3330.51	90.42	90.90
53	36617.41	3396.00	2179.48	90.73	94.05
54	36735.97	4411.82	2571.50	87.99	93.00

savings with respect to FCFS. Secondly, the cost reduction becomes more significant when the ratio of the online demand rate to the total demand rate increases. This is because when a larger proportion of the total demand is online orders, it becomes more imperative to allocate the on-hand inventory to offline orders as priority rather than to make this allocation simply according to the FCFS rule. Thirdly, a larger ratio of lost sales cost to backorder cost leads to a more significant cost reduction, with Group (11) realizing the largest cost reduction of 10.91%. This implies that the higher the priority assigned to offline demand, the more the critical level rationing policy should be employed to reserve inventory for it. Table 7 provides the optimal parameter values for the (Q, r, K) policy and the (Q, r, FCFS) policy for the 18 benchmark instances. We observe that the presence of rationing based on critical level reduces the reorder point r and order quantity Q for most instances.

We also analyzed the impact of total demand rate on the effectiveness of the critical level based rationing. This analysis was conducted for two scenarios with different order cycle times. In case with a short order cycle for the groups numbered in odd in Table 6, the total demand rate has a significant impact on the rationing effectiveness, with an enhanced rationing effect when the total demand rate increases. However, in case with a long order cycle for the groups numbered in even in Table 6, the impact of the total demand rate becomes less significant. This phenomenon may be explained as follows: when the order cycle is short, for a Poisson total demand, a higher arrival rate implies a high variation of the total demand during the cycle, which amplifies the holding, backorder, and lost sales costs of the inventory system studied, thereby increases the benefits associated with the application of a critical level rationing policy.

6.4. Parameter sensitivity analysis for critical level K

The influence of various system parameters on the optimal critical level K is examined, with results presented in Table 8. This table shows how K is affected by the total demand rate λ, the order cycle time, the unit lost sale cost, and the ratio of the online demand rate to the total demand rate. An increase in the total demand rate results in a higher optimal K, which can be intuitively interpreted by that a greater demand rate necessitates a higher inventory for the fulfillment of all demands and a higher inventory reserved for the fulfilment of offline demands. Similarly, with the same demand rate, a longer order cycle results in a higher order quantity, and consequently a higher inventory reserved for offline demands. Notably, a high K signifies a high level of

Table 8
Optimal critical level K.

Demand rate	Order cycle	π	$\lambda_2/\lambda = 1/3$	$\lambda_2/\lambda = 1/2$	$\lambda_2/\lambda = 2/3$
$\lambda = 0.3$	short	$Q_{EOQ}b/2\lambda$	2	2	2
		$Q_{EOQ}b/\lambda$	4	3	2
		$Q_{EOQ}b/4\lambda$	0	1	1
	long	$Q_{EOQ}b/2\lambda$	0	3	2
		$Q_{EOQ}b/\lambda$	5	4	3
		$Q_{EOQ}b/4\lambda$	0	0	0
$\lambda = 0.6$	short	$Q_{EOQ}b/2\lambda$	6	5	3
		$Q_{EOQ}b/\lambda$	7	6	4
		$Q_{EOQ}b/4\lambda$	0	3	2
	long	$Q_{EOQ}b/2\lambda$	0	6	5
		$Q_{EOQ}b/\lambda$	9	7	5
		$Q_{EOQ}b/4\lambda$	0	0	3
$\lambda = 1.5$	short	$Q_{EOQ}b/2\lambda$	13	10	7
		$Q_{EOQ}b/\lambda$	15	12	8
		$Q_{EOQ}b/4\lambda$	11	8	6
	long	$Q_{EOQ}b/2\lambda$	0	13	9
		$Q_{EOQ}b/\lambda$	19	15	10
		$Q_{EOQ}b/4\lambda$	0	11	8

Table 6
The impact of the demand rates on the rationing effect.

Cost reduction (%)											
$\pi = Q_{EOQ}b/2\lambda$				$\pi = Q_{EOQ}b/\lambda$				$\pi = Q_{EOQ}b/4\lambda$			
λ_2/λ	1/3	1/2	2/3	λ_2/λ	1/3	1/2	2/3	λ_2/λ	1/3	1/2	2/3
Group1	1.08	2.71	3.67	Group7	2.16	5.02	6.92	Group13	0.00	0.08	0.43
Group2	0.00	0.98	2.34	Group8	1.56	3.15	4.94	Group14	0.00	0.00	0.00
Group3	1.92	4.38	6.17	Group9	3.47	6.44	9.02	Group15	0.00	1.06	2.33
Group4	0.00	1.55	3.26	Group10	1.62	3.54	5.23	Group16	0.00	0.00	1.04
Group5	3.29	6.30	8.78	Group11	4.72	7.90	10.91	Group17	0.71	3.86	5.72
Group6	0.00	2.25	4.10	Group12	1.62	3.61	5.36	Group18	0.00	0.60	2.65

Table 7
Comparison of optimal parameter values for (Q, r, K) and (Q, r, FCFS) policies.

Instances	Q, r, K	Q, r, FCFS	Instances	Q, r, K	Q, r, FCFS	Instances	Q, r, K	Q, r, FCFS
1	[17,8,2]	[17,9]	7	[32,15,5]	[33,17]	13	[76,35,10]	[78,40]
2	[17,9,2]	[17,9]	8	[32,16,6]	[34,17]	14	[76,38,13]	[79,41]
3	[17,7,2]	[17,8]	9	[32,14,3]	[32,16]	15	[76,32,7]	[78,38]
4	[38,8,3]	[41,8]	10	[75,14,6]	[79,16]	16	[185,32,13]	[190,39]
5	[43,8,0]	[43,8]	11	[82,16,0]	[82,16]	17	[199,38,0]	[199,38]
6	[38,7,2]	[39,8]	12	[75,12,5]	[78,15]	18	[185,28,9]	[187,38]

differentiation between two demand classes, whereas a low K indicates a low level of differentiation. In case $\pi = Q_{\mathrm{EOQ}}b/4\lambda$ where the unit sales lost cost is relatively small, it is more likely that K equals 0. This is intuitive because the purpose of rationing is to reduce lost sales costs induced by unsatisfied offline demands. Furthermore, when the ratio λ_2/λ is high, with a low proportion of high-priority offline demand, it intuitively leads to a lower optimal critical level K, except for the case with $K = 0$.

6.5. A case study

To validate the applicability and effectiveness of our proposed model and optimization methods, we conduct a case study utilizing data from an e-commerce company in China, our industrial collaborator. An omnichannel retail store of this company serves offline customers and also sell products on its own app. For this case study, we focus on a specific product – a case of milk with 16 bottles of 250 ml, with the relevant data introduced in the next paragraph. Currently, the store applies a continuous-review (Q, r) policy to replenish the inventory of this product. It fulfills online and offline demands of the product on a FCFS basis and does not reserve inventory for offline demand.

The unit selling pricing for the milk is 85 CNY for both online and offline channels. Historical sales data over the past year indicates that the demand of this product in offline and online channels approximately follows Poisson distributions with the arrival rate per hour of 0.66 and 1.25, respectively. The retailer estimates the annual holding cost of this product at 20% of the selling price, resulting in its holding cost per hour h of 0.00194 CNY ($= (85 \times 20\%)/(365 \times 24)$). The lost sales cost for offline demand of the product is estimated to be its gross profit margin, which is 20% of the selling price, resulting in per unit lost sales π of 17 CNY ($=85 \times 20\%$). The retailer aims for a service level of 99% for online channel. The backorder cost per unit per hour b is estimated by utilizing a commonly employed formula: service level $= b/(h + b)$, resulting in $b = 0.192$ CNY. The product is replenished every three days on average with the average order quantity of 138 units. Each replenishment order placed by the store to a FDC (Front Distribution Center) of the company can be received by this store within 24 h with a replenishment lead time of one day. The distance between the store and the FDC is about 25 km. The transportation cost for this product is estimated at 0.5 CNY per cubic meter per kilometer. Given this product's packaging size of 0.006 cubic meter (13 cm $\times$ 21 cm $\times$ 22 cm), the ordering cost for each order is estimated as its transportation cost, which is about 11 CNY. The store serves online customers with a promised delivery time of no more than 6 h (half a day), hence the maximum acceptable waiting time T_{cus} is set to 6 h.

Based on the aforementioned real business data, we use our analytical cost evaluation method to estimate all inventory-related costs of the product in this store that adopts (Q, r, K) policy or (Q, r, FCFS) policy and use our two algorithms to optimize the parameters of the two policies, with the results given in Table 9, where "cost" stands for the total cost per hour.

From Table 9, we can see, compared to the store's current (Q, r, FCFS) policy without rationing, the adoption of the (Q, r, K) policy in this omnichannel store can help it reduce 6.61% and 6.29% of its total cost if the parameters of the policy are set by the SS and the MLS algorithm respectively. The computation times of the SS and MLS algorithm are 3714.00 and 11674.80 s respectively. It takes 82408.21 s to obtain the optimal parameter values of the policy through enumeration, which can reduce the total cost by 6.75% with respect to the (Q, r, FCFS) policy. For this instance, both the SS and MLS algorithm obtain a solution very close to the optimal one, but the SS algorithm outperforms the MLS algorithm in terms of solution quality and computation time.

7. Conclusion

This paper studies an inventory system appeared in an omnichannel retailer with both Poisson online and offline demands and controlled by a continuous review (Q, r, K) policy. We propose an inventory model that captures main features of inventory management in an omnichannel store, where any online order that cannot be delivered immediately is treated as a backorder but its backorder cost is incurred only after passing its maximum delivery waiting time, whereas all unsatisfied offline demands are lost. For this model, we develop an analytical method for its cost evaluation and two heuristic algorithms for joint optimization of its inventory replenishment and rationing policies. The effectiveness of the algorithms is demonstrated by extensive numerical experiments and a case study. The two algorithms are quite general and applicable to inventory systems with other stochastic demands if the expected cost of the underlying inventory system can be evaluated analytically or by simulation. The modeling, analysis, and optimization methodologies proposed in this study provide a foundation for inventory management of multiple omnichannel stores.

Our numerical study provides some managerial insights for inventory management of an omnichannel retailer. Firstly, the (Q, r, K) policy is more effective than the (Q, r, FCFS) policy in most cases and can significantly reduce the total cost and reorder point. In our numerical instances, it has the potential to achieve cost reductions of up to 10.91%. This value is 6.61% in a case study based on real industry data. Secondly, our study reveals the conditions under which the (Q, r, K) policy can achieve a better performance. Specifically, a retailer with a major portion of online demand should prioritize the implementation of the (Q, r, K) policy than that with a minor portion of online demand. Furthermore, a retailer with a shorter replenishment order cycle should pay attention to the interplay between its total demand rate and rationing efficiency. As the total demand rate increases, the potential cost reduction of its rationing amplifies significantly. Moreover, the increased divergence in priority levels among demands positively influences the heightened efficiency of inventory rationing. Lastly, our study reveals the interdependence between the parameters of the underlying inventory system and its critical level K. Managers are recommended to set a higher critical level K as the total demand rate, order cycle time, and the ratio of the offline demand to the total demand increase.

Our study has some limitations especially in the demand model we consider. We assume the demand of each class follows a Poisson process. Although Poisson demand is widely adopted in the inventory literature and analytically more tractable, the demand of a product may follow another stochastic process/distribution, such as compound Poisson distribution and normal distribution.

Future research may follow the following directions. Firstly, alternative demand models that describe more real-world demand scenarios are worthy to explore with the corresponding study of analytical cost evaluation of the underlying inventory system. Secondly, a more general inventory system without the restriction of existence of at most one

Table 9
Comparison of (Q, r, K) policy and (Q, r, FCFS) policy for a real instance.

Policy	Optimization method	Q	r	K	Cost	Cost reduction (%)	Computation time (s)
(Q, r, FCFS)	Enumeration	151	59	0	0.3182		852.52
(Q, r, K)	Enumeration	151	47	11	0.2968	6.75	82408.21
	SS	153	47	12	0.2972	6.61	3714.00
	MLS	149	48	13	0.2982	6.29	11674.80

outstanding order is worthy to investigate. Thirdly, consideration may be given to the integration of other features of retail operations in the inventory model studied. For instance, our inventory model may be extended to include possible order transfers (transshipments) among multiple omnichannel stores. This extension should incorporate order allocation decisions into the model in addition to its inventory replenishment and rationing decisions. However, the high complexity of the extended model makes its cost evaluation and joint optimization of inventory replenishment, rationing, and order transfer decisions very challenging.

CRediT authorship contribution statement

Zengxu Guo: Writing – original draft, Methodology. **Haoxun Chen:** Writing – review & editing, Supervision, Methodology.

Data availability

Data will be made available on request.

Acknowledgement

The work of this paper was partially supported by China Scholarship Council with grant number 202006280023.

References

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

Bayram, A., & Cesaret, B. (2021). Order fulfillment policies for ship-from-store implementation in omni-channel retailing. *European Journal of Operational Research, 294*(3), 987–1002. https://doi.org/10.1016/j.ejor.2020.01.011

Bendoly, E. (2004). Integrated inventory pooling for firms servicing both on-line and store demand. *Computers & Operations Research, 31*(9), 1465–1480. https://doi.org/10.1016/S0305-0548(03)00102-3

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

Difrancesco, R. M., van Schilt, I. M., & Winkenbach, M. (2021). Optimal in-store fulfillment policies for online orders in an omni-channel retail environment. *European Journal of Operational Research, 293*(3), 1058–1076. https://doi.org/10.1016/j.ejor.2021.01.007

Enders, P., Adan, I., Scheller-Wolf, A., & van Houtum, G.-J. (2014). Inventory rationing for a system with heterogeneous customer classes. *Flexible Services and Manufacturing Journal, 26*(3), 344–386. https://doi.org/10.1007/s10696-012-9148-1

Escalona, P., Angulo, A., Weston, J., Stegmaier, R., & Kauak, I. (2019). On the effect of two popular service-level measures on the design of a critical level policy for fast-moving items. *Computers & Operations Research, 107*, 107–126.

Escalona, P., Ordóñez, F., & Kauak, I. (2017). Critical level rationing in inventory systems with continuously distributed demand. *OR Spectrum, 39*, 273–301.

Frank, K. C., Zhang, R. Q., & Duenyas, I. (2003). Optimal Policies for Inventory Systems with Priority Demand Classes. *Operations Research, 51*(6), 993–1002. https://doi.org/10.1287/opre.51.6.993.24923

Gabor, A. F., van Ommeren, J. K., & Sleptchenko, A. (2022). An inventory model with discounts for omnichannel retailers of slow moving items. *European Journal of Operational Research, 300*(1), 58–72.

Glover, F. (1977). Heuristics for Integer Programming Using Surrogate Constraints. *Decision Sciences, 8*(1), 156–166. https://doi.org/10.1111/j.1540-5915.1977.tb01074.x

Goedhart, J., Haijema, R., & Akkerman, R. (2022). Inventory rationing and replenishment for an omni-channel retailer. *Computers & Operations Research, 140*, Article 105647. https://doi.org/10.1016/j.cor.2021.105647

Guo Z, Chen H. (2023). Cost Evaluation of a (Q, r, K) Inventory Model with Two Demand Classes of Lost Sales and Backorders. *Proceedings of IFIP International Conference on Advances in Production Management Systems (APMS 2023),* Trondheim, Norway, 17-21 September, 2023, pp. 257-270. Springer.

Hadley, G., & Whitin, T. M. (1963). Analysis of inventory systems (No. 658.787 H3).

He, Y., Xu, Q., & Shao, Z. (2021). "Ship-from-store" strategy in platform retailing. *Transportation Research Part E: Logistics and Transportation Review, 145*, Article 102153. https://doi.org/10.1016/j.tre.2020.102153

Hosseini-Motlagh, S.-M., Choi, T.-M., Johari, M., & Nouri-Harzvili, M. (2021). A profit surplus distribution mechanism for supply chain coordination: An evolutionary game-theoretic analysis. *European Journal of Operational Research.* https://doi.org/10.1016/j.ejor.2021.10.059

Huang, S., Liao, B., & Robb, D. J. (2011). Inventory management in A (Q, r) inventory model with two demand classes and flexible delivery. *International Journal of Innovation Computing, Information and Control, 7*(3), 1243–1254.

Hübner, A., Hense, J., & Dethlefs, C. (2022). The revival of retail stores via omnichannel operations: A literature review and research framework. *European Journal of Operational Research, 302*(3), 799–818. https://doi.org/10.1016/j.ejor.2021.12.021

Kleijn, M. J., & Dekker, R. (1999). *An overview of inventory systems with several demand classes.* Springer.

Martí, R., Laguna, M., & Glover, F. (2006). Principles of scatter search. *European Journal of Operational Research, 169*(2), 359–372. https://doi.org/10.1016/j.ejor.2004.08.004

Melchiors, P., Dekker, R., & Kleijn, M. J. (2000). Inventory rationing in an (s, Q) inventory model with lost sales and two demand classes. *Journal of the Operational Research Society, 51*, 111–122.

Möllering, K. T., & Thonemann, U. W. (2008). An optimal critical level policy for inventory systems with two demand classes. *Naval Research Logistics (NRL), 55*(7), 632–642.

Möllering, K. T., & Thonemann, U. W. (2010). An optimal constant level rationing policy under service level constraints. *OR Spectrum, 32*(2), 319–341. https://doi.org/10.1007/s00291-009-0167-6

Mou, S., Robb, D. J., & DeHoratius, N. (2018). Retail store operations: Literature review and research directions. *European Journal of Operational Research, 265*(2), 399–422. https://doi.org/10.1016/j.ejor.2017.07.003

Nahmias, S., & Demmy, W. S. (1981). Operating Characteristics of an Inventory System with Rationing. *Management Science, 27*(11), 1236–1245. https://doi.org/10.1287/mnsc.27.11.1236

Stambor, Z. (2023). Target pivots to a regional distribution network to speed up deliveries. Available at https://www.insiderintelligence.com/content/target-pivots-regional-distribution-network-speed-up-deliveries. (accessed August 24, 2023).

Tempelmeier, H. (2006). Supply chain inventory optimization with two customer classes in discrete time. *European Journal of Operational Research, 174*(1), 600–621.

Tijms, H. C. (1994). *Stochastic models: An algorithmic approach* (Vol. 303). John Wiley & Sons Incorporated.

Turrini, L., & Meissner, J. (2019). Spare parts inventory management: New evidence from distribution fitting. *European Journal of Operational Research, 273*(1), 118–130.

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

Vicil, O. (2021). 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.

Wan, G., & Cao, Y. (2018). A continuous cost evaluation approach for periodic review inventory systems with threshold rationing policy. *Computers & Industrial Engineering, 126*, 75–87.

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

Wang, D., Tang, O., & Huo, J. (2013). A heuristic for rationing inventory in two demand classes with backlog costs and a service constraint. *Computers & Operations Research, 40*(12), 2826–2835.

Wang, Y., Zhang, S. H., & Sun, L. (2013). Anticipated rationing policy for two demand classes under service level constraints. *Computers & Industrial Engineering, 65*(2), 331–340. https://doi.org/10.1016/j.cie.2013.02.005

Wang, Y., Zhang, S. H., Zhou, S. X., & Zhang, Y. (2020). Anticipated rationing policy for inventory systems with two demand classes and backlogging costs. *International Journal of Production Research, 58*(20), 6300–6314.

Xu, J., & Cao, L. (2019). Optimal in-store inventory policy for omnichannel retailers in franchising networks. *International Journal of Retail & Distribution Management, 47*(12), 1251–1265. https://doi.org/10.1108/IJRDM-09-2018-0199

Yang, D., & Zhang, X. (2020). Omnichannel operations with ship-from-store. *Operations Research Letters, 48*(3), 257–261. https://doi.org/10.1016/j.orl.2020.03.009

Zhou, Y., & Zhao, X. (2010). A two-demand-class inventory system with lost-sales and backorders. *Operations Research Letters, 38*(4), 261–266.