On the Benefits of Delayed Ordering

(European Journal of Operational Research)

A.A. Syntetos[1], R.H. Teunter[2], M.Z. Babai[3] and S. Transchel[4]

[1] Cardiff University, Wales (SyntetosA@cardiff.ac.uk)

[2] University of Groningen, The Netherlands (R.H.Teunter@rug.nl)

[3] Kedge Business School, France (Mohamed-Zied.Babai@kedgebs.com)

[4] Kühne Logistics University, Germany (sandra.transchel@the-klu.org)

Abstract

Practical experience and scientific research show that there is scope for improving the performance of inventory control systems by delaying a replenishment order that is otherwise triggered by generalised and all too often inappropriate assumptions. This paper presents the first analysis of the most commonly used continuous (*s*, *S*) policies with delayed ordering for inventory systems with compound demand. We analyse policies with a constant delay for all orders as well as more flexible policies where the delay depends on the order size. For both classes of policies and general demand processes, we derive optimality conditions for the corresponding delays. In a numerical study with Erlang distributed customer inter-arrival times, we compare the cost performance of the optimal policies with no delay, a constant delay and flexible delays. Sensitivity results provide insights into when the benefit of delaying orders is most pronounced, and when applying flexible delays is essential.

Keywords: Inventory control; Delayed ordering; Intermittent demand; (s, S) policy; Marginal cost analysis.

1. Introduction

1.1. Motivation and research relevance

It is well known that for periodic review inventory systems, the order level, order-up-to level (s, S) policy is optimal under quite general conditions (Scarf, 1959; Iglehart, 1963; Sahin, 1990). In particular, the optimality under concern, in the case of backordering of unfilled demand, is associated with: i) constant ordering cost; ii) linear stock-out and holding cost; iii) a fixed replenishment lead time.

The same is not true under continuous review, as is illustrated by the following simple example. Consider an item with a constant lead time L and a larger constant customer inter-arrival time I between unit-sized transactions. Then the optimal policy is obviously to have at most one unit on hand and always reorder $I - L$ time units after a transaction. In other words, compared to the (s=0, S=1) policy, each replenishment order should be delayed by $I - L$ time units. An alternative interpretation is that the order is being placed L time units before it is needed to satisfy the next demand, thereby avoiding any time in inventory.

More generally, delaying orders seems suitable whenever the customer inter-arrival times do not exhibit the memory-less property of the exponential distribution. There are various settings where this situation may naturally occur. One is that of a multi-echelon system, where lot-sizing is applied at lower levels. Another occurs in spare parts management, where parts used for corrective maintenance may wear. Empirical results by Porras and Dekker (2008) under a continuous (s, S) system confirm that assuming demand is driven by a Poisson process results in overstocking spare parts having 0/1 demands. Numerous papers in the area of spare parts

modelling assume a continuous review system; interested readers are referred to Kennedy *et al*. (2002) for an overview in this area.

The demand for spare parts is known to arrive sporadically/intermittently and to be driven by increasing failure rate (IFR) distributions. This is true not only for engineering spares but for service parts kept at the wholesaling/retailing level as well. The stock-bases in the military context, process industries, aerospace, automotive and IT sectors are also dominated by such items. Two very comprehensive benchmarking reports by the Aberdeen Group (2005) and Deloitte (2006) identify the increasing importance of after-sales service and parts business (please refer also to Inderfurth and Kleber, 2013). As stated in the latter report, the combined revenues of many of the world's largest manufacturing companies are more than US$1.5 trillion. Further, on average, service revenues account for more than 25% of the total business, so delaying orders for (expensive) spare parts can have a considerable effect on the bottom line. For example, Dickinson (2013) states that “the rotable pool of high value assets for the EuroFighter is managed through delayed ordering practices”. Similarly, in many organisations, Maintenance, Repair and Operations (MRO) inventory accounts for as much as 40% of the annual procurement budget (Donnelly, 2013). Thus, small improvements regarding the management of the relevant inventories may be translated to substantial cost savings; whereas it is also true to say that any research in this area has a direct relevance to a wide range of companies and industries.

In addition, it is also worthwhile noting that demand patterns in Business-to-Business environments (B2B) are all too often determined by the degree of heterogeneity of the client base (Bartezzaghi *et al*., 1999). Heterogeneous requests occur when the potential market consists of

customers with considerably different sizes, e.g., few large customers coexist with a number of small customers. (Similarly, in the MRO environment planned maintenance and breakdowns may also introduce differences in order inter-arrival time distributions.) The higher the heterogeneity of customers, the higher the demand lumpiness, since periods with high requests from a large customer alternate with periods with low or no requests at all from small customers. Alternatively, following a request from a large customer, it is unlikely that another demand will be received in the near future necessitating a delayed ordering mechanism on the part of the supplier. The potential correlation between customers' requests further induces lumpiness. Correlation may be due, amongst other reasons, to imitation and fashion, which induce similar behaviours in customers so that sudden peaks of demand may occur after periods of no requests.

Collective consumer behaviour may be modelled through what are often termed in the literature as 'censored Poisson' processes, whereby the p^{th} event of a Poisson process is only recorded, resulting in inter-event Erlang (of order p) distributions (e.g., Chatfield and Goodhardt, 1973). The discussion conducted in this section also illustrates the compound nature of the demand and the need to take this into consideration, if a realistic inventory model is to be developed.

1.2. Research background

Order delays in a continuous review setting have not received sufficient attention in the literature. To the best of our knowledge, Schultz (1987, 1989), Katircioglu (1996), Moinzadeh (2001), Moinzadeh and Zhou (2008) and Axsäter and Viswanathan (2012) are the only authors who discuss this issue. Schultz (1987) considers (S-1,S) policies and assumes for tractability that the probability of the sum of two demands being less than S is negligible, which is quite

restrictive. He shows that a constant delay in placing an order can result in significant holding-cost reductions with little additional risk or cost of stockouts.

Schultz (1989) discusses a different, but again very restrictive setting. He assumes unit-sized transactions and only considers the (s=0, S=1) policy. Furthermore, there is instantaneous emergency replenishment in case of shortages. Results are given for the optimal delay for customer inter-arrival distributions with increasing failure rates. Specific expressions for the optimal delay are given for several commonly used distributions, including the Erlang distribution.

Moinzadeh (2001) considers a somewhat more general setting, but still restricted to unit-sized transactions and (*S-1,S*) policies. Each order is delayed by a constant period of time, independent of demand activities during that period. For general customer inter-arrival times, Moinzadeh (op. cit.) develops an efficient heuristic for computing the policy parameters. He evaluates the performance of the heuristic via a numerical experiment for the cases with Erlang and Uniform customer inter-arrival times.

The studies by Katircioglu (1996) and Moinzadeh and Zhou (2008) are more general than those discussed so far in that they consider (a) unrestricted order levels ($s < S$) and (b) more sophisticated policies that end a delay when a new demand occurs. However, both models still assume unit sized demands. They obtain similar optimality conditions, albeit through different sorts of analysis. Both also provide numerical results that indicate significant potential savings from order delays. Katircioglu (1996) proves that the optimal policy is of this type. Moinzadeh

and Zhou (2008) extend their analysis and results to a two echelon setting with a single warehouse that delays orders and multiple retailers.

Axsäter and Viswanathan (2012) consider the case of a supplier who faces an Erlang demand process from a downstream customer with constant order sizes. They develop an algorithm to determine the optimal ordering time delay when the supplier controls its inventory according to a reorder point (R, nQ) installation stock policy and no information sharing takes place between the supplier and the customer. A numerical investigation shows substantial cost savings when the optimal time delay policy is used (instead of the installation stock policy without delay). These cost savings are also shown to be more substantial than those obtained when the installation stock policy without delay is used in conjunction with inventory information sharing between the customer and the supplier.

1.3. Contributions and organisation of the paper

In this paper, we provide the first analysis of (s, S) policies in a single echelon inventory system with order delays for compound demand processes. So, we drop the assumption that demands are unit-sized. As discussed before, this is an important generalisation since intermittent (spare parts) demand series, for which delaying orders is particularly suitable, are usually very lumpy (Boylan *et al*., 2008). Related to this more general setting, we also consider more flexible delay policies where the maximum delay depends on the order quantity. Like in the studies of Katircioglu (1996), Moinzadeh and Zhou (2008) and Axsäter and Viswanathan (2012), an order is only delayed for this long if no demand happens before then.

For general customer inter-arrival times, we derive conditions that can be used to determine the optimal maximum delay times for any order quantity. This is done using a marginal cost analysis. The exact form of the optimality conditions depends on the specific type of customer inter-arrival distribution. For the purpose of our (numerical) analysis, Erlang distributed customer inter-arrival times will be assumed. The case of Erlang distributed customer inter-arrival times is obtained if demand originates from the lot sizing by a single customer experiencing Poisson demand. It has been considered by many other authors, including Liu and Shi (1999), Schultz and Johansen (1999), Strijbosch *et al.* (2000) and those mentioned before. The Erlang demand process is also a building block in analysing multi-echelon systems (Deuermeyer and Schwarz, 1981; Moinzadeh and Lee, 1986; Lee and Moinzadeh, 1987; Svoronos and Zipkin, 1988; Andersson *et al*., 1998; Axsäter, 2000; Berling and Marklund, 2006, 2013).

We will also consider the much more restrictive policy with a constant delay time, independent of the order quantity. This policy was also studied by Katircioglu (1996) and Moinzadeh and Zhou (2008) for systems with unit-sized demands, and indeed shown to be optimal for those systems. This is clearly not the case for compound demand processes, but the optimal policy of this type will be easier to implement (in non-computerized systems) and can serve as a benchmark for the performance of the flexible delay policy.

The remainder of this paper is organised as follows. In Section 2, we introduce notations and present the inventory system and policy in detail. We derive the general optimality conditions for determining the maximum delays in the flexible delay policies and subsequently we do the same

for policies with a constant maximum delay time. The exact form of the optimality conditions for both types of policies is then provided assuming Erlang distributed customer inter-arrival times. In Section 3, we numerically study the effect of the order quantity on the maximum delay. Furthermore, we compare the costs of the optimal delay policies of the two types and also to the standard (*s*, *S*) policy without delays. We end with conclusions, discussion and directions for future research in Section 4.

2. Inventory system and policies

We first introduce some notations.

s: Order level

S: Order-up-to level

L: Lead time

$T(q)$: Maximum order delay for an order of quantity q

$O(\Delta)$: Term that is of large order (not to be confused with replenishment order) Δ

$o(\Delta)$: Term that is of small order (not to be confused with replenishment order) Δ

f_j: Probability of demand size j

f_j^k: Probability that k transactions give total demand j

h: Holding cost per item per time unit

b: Backorder cost per item per time unit

We consider a single item inventory system with a constant lead time. We include any demand process that satisfies the following two restrictions. First, both demand sizes and customer inter-arrival times are independent and identically distributed. Second, for any demand history, the

probability that one or more demands occur in the next Δ time units is of the order Δ (i.e. the probability is $O(\Delta)$). All well-known demand processes with unbounded customer inter-arrival time distributions, e.g. Exponential, Erlang, Gamma and Beta, satisfy this assumption. In order to avoid any confusion, in what remains we will refer to a customer demand for one or more items as a *transaction* and to the (total) number of items demanded in a transaction or time period as *demand*.

We consider two classes of policies, which we refer to as flexible delay policies and constant delay policies. The class of flexible delay policies that we consider are characterised by the order level s, the order-up-to level S and the maximum order delay $T(q)$ for an order of size $q, q = 1,2,\ldots$ As for the standard (*s*, *S*) policy, an order is triggered if the inventory position (inventory on hand + inventory on order - backorders) drops to or below s and the order quantity results from ordering up to S. However, an order is delayed by $T(q)$ time units or until the next demand occurs, whichever happens first. The class of constant delay policies are restricted by having the same maximum delay for all order quantities. In addition, we assume in our policies that order delays do not affect the sizes of the orders. This assumption, along with its implications, is further considered in the last section of our paper.

All demands that are not satisfied immediately are backordered. The objective is to minimize the average cost per time unit, including holding and backorder costs. We remark that ordering costs are not relevant for our study (and thus are not further considered), as the delay mechanism does not affect the number of orders. That is, the comparative performance of the policies considered here is not affected by the ordering costs.

2.1. Optimal policy with flexible delay

In this section, we derive optimality conditions for the optimal delays using marginal analysis. This is done by studying a marginal perturbation of a policy. The perturbed policy increases the maximum delay for a specific value of the order quantity q by a small amount $\Delta > 0$, but is otherwise the same. For ease of presentation, we will not introduce new notation for this specific value, but instead simply use q. For the same reason, we use the short notation π and $\pi(\Delta)$, for the original and the perturbed policy, respectively. We will also use T instead of $T(q)$.

Obviously, π and $\pi(\Delta)$ only lead to different delay decisions if the following event occurs: a demand occurs, say at time 0, that triggers an order of quantity q and no further demands occur for the next T time units. If this happens, then π places the order at time T while $\pi(\Delta)$ continues to delay that order for at most Δ time units.

So, let us consider this situation and compare the costs between π and $\pi(\Delta)$. Clearly, the inventory positions of policies π and $\pi(\Delta)$ can only differ in period $(T, T+\Delta)$. Hence, the inventory levels and costs can only differ in period $(L+T, L+T+\Delta)$. Let us denote the expected costs of π and $\pi(\Delta)$ in period $(L+T, L+T+\Delta)$ by C and $C(\Delta)$, respectively. The analysis that follows will show that $C - C(\Delta)$ is decreasing in T, which is intuitive since the marginal benefit of increasing a delay diminishes with the current length of the delay. This implies that the cost per time unit is at least quasi-convex in the maximum delay T. As a result, the optimal maximum delay is either zero, if $C - C(\Delta) < 0$ (even for $T = 0$), or the optimal value for T is that for which the following holds:

$$\lim_{\Delta \downarrow 0} \frac{C - C(\Delta)}{\Delta} = 0$$

or equivalently, $C - C(\Delta)$ is of small order Δ denoted by $o(\Delta)$ using the conventional "little or small o" notation, i.e.

$$C - C(\Delta) = o(\Delta).$$

What remains is to rewrite this optimality condition in terms of T and the other (policy) parameters, and show that $C - C(\Delta)$ is indeed decreasing in T.

By definition, the expected cost difference $C - C(\Delta)$ is equal to the sum of the cost difference of π and $\pi(\Delta)$ in period $(L+T, L+T+\Delta)$ over all possible demand scenarios multiplied by the corresponding probabilities that these scenarios occur. Recall that we only consider realistic demand processes for which the probability that two transactions occur in some interval of length Δ is $O(\Delta)$, i.e. that two transactions are unlikely to occur at almost the same time. Also, with all costs being proportional to time, the cost difference between any two scenarios in period $(L+T, L+T+\Delta)$ cannot be more than some constant (dependent on the cost parameter values) times the interval length Δ, i.e. that cost difference is of large order Δ denoted as $O(\Delta)$ using the conventional "large O" notation. This implies that the effect of any demand scenario where at least one transaction happens in period $(T, T+\Delta)$ and/or in period $(L+T, L+T+\Delta)$, on the expected cost difference $C - C(\Delta)$ is of $O(\Delta) \times O(\Delta) = o(\Delta)$. Next, we show that the same holds

for all other demand scenarios where no transactions occur in periods $(T, T+\Delta)$ and $(L+T, L+T+\Delta)$, which then implies that $C - C(\Delta) = o(\Delta)$.

So, let us consider the case where no transactions occur in periods $(T, T+\Delta)$ and $(L+T, L+T+\Delta)$. Since no transactions occur in periods $(T, T+\Delta)$, $\pi(\Delta)$ will apply the maximum delay and order at time $T+\Delta$. Let $D(T, L+T)$ denote the demand in period $(T, L+T)$. Then we can write the inventory level just before time $L+T$ for both π and $\pi(\Delta)$ as $S - q - D(T, L+T)$.

Let $N_T(k), k = 1,2,\ldots$, denote the probability that k transactions take place in period $(T, L+T)$, given that the last transaction before time T occurred at time 0. These probabilities depend on the distribution of customer inter-arrival times. In Section 2.3 we will derive expressions for the case of Erlang distributed customer inter-arrival times.

To obtain the distribution of the total demand in period $(T, L+T)$, we will combine the probabilities $N_T(k)$ of k transactions with the probabilities f_j^k that k transactions give total demand j. The latter probabilities can be calculated (Axsäter, 2006, p. 78-79) as

$$f_0^0 = 1, \qquad f_j^1 = f_j, \qquad f_j^k = \sum_{i=k-1}^{j-1} f_i^{k-1} f_{j-i}, \quad k = 2,3,\ldots, \qquad (1)$$

which can be solved recursively.

Let $P_T(j)$, $j=1,2,\ldots$, denote the probability that total demand in period $(T, L+T)$ is equal to j, given that the last transaction before time T occurred at time 0. By a slight misuse of notation, we will use $P_T(< j)$ for denoting the probability that the total demand in period $(T, L+T)$ is less than j for presentational ease. We have

$$P_T(0) = N_T(0) \qquad \text{and} \qquad P_T(j) = \sum_{k=1}^{j} N_T(k) f_j^k, \quad j=1,2,\ldots \qquad (2)$$

We will use these probabilities to derive the marginal cost difference. The incoming order is for q units and each unit is considered separately in determining the marginal cost difference. The *i*-th unit of the incoming order is needed to prevent a backorder in period $(L+T, L+T+\Delta)$ if demand in period $(T, L+T)$ is at least $(S-q)+i$; however, if the demand is less than $(S-q)+i$, the i-th unit generates an additional holding cost in period $(L+T, L+T+\Delta)$. So, the marginal cost difference over all q units can be expressed as

$$\begin{aligned} C - C(\Delta) &= \sum_{i=1}^{q} \left(hP_T(< S-q+i) - bP_T(\geq S-q+i) \right)\Delta \\ &= \sum_{i=1}^{q} \left((h+b)P_T(< S-q+i) - b \right)\Delta \qquad (3) \\ &= \left(-bq + (h+b)\sum_{i=1}^{q} \left(P_T(< S-q+i) \right) \right)\Delta \end{aligned}$$

Since the probability that no transaction occurs between time zero and T decreases as T increases, for any demand process, $P_T(< S-q+i)$ is obviously decreasing in T for all

$i = 1,...,q$. This, in turn, implies that the cost difference between π and $\pi(\Delta)$ is decreasing in T as well. Now, two scenarios can occur: First, the marginal cost difference at $T = 0$ is already negative and becomes more negative as T increases. Then, the optimal maximum delay is zero. If, however, $\sum_{i=1}^{q}\left(P_{T=0}(< S - q + i)\right) > \frac{bq}{h+b}$, then the cost difference is positive at $T = 0$ and decreases to $-bq < 0$ as T tends to infinity, implying that there exists a unique optimal delay T^*, which solves the following equation

$$(h+b)\sum_{i=1}^{q} P_T(< S - q + i) = bq\,. \qquad (4)$$

For the special case of unit-sized demand and $S = s + 1$, where each order is of quantity one, (4) simplifies to

$$P_T(< s+1) = \frac{b}{b+h} \qquad \text{or} \qquad P_T(< S) = \frac{b}{b+h}\,.$$

This is the same optimality condition as derived in Katircioglu (1996) and Moinzadeh and Zhou (2008), although formulated differently.

2.2. Optimal policy with a constant delay

The analysis of the previous section showed that the order quantity influences the cost effectiveness of a delay and thereby the optimality condition. So, in order to apply a similar

marginal analysis for the constant delay policy, we need to know the probabilities that an arbitrary order has a certain quantity q, which we will denote by p_q.

Clearly, $q \geq S - s$ and an order of size q can only happen, if the inventory position first reaches some level $j, j > s,$ and then a demand of size $q-(S-j) = q - S + j$ occurs so that the inventory position drops to $j-(q-S+j) = S-q$. So, if we let m_j denote the probability that the inventory position reaches level j, $j = s+1, s+2, \ldots, S$, before an order is triggered (that may be delayed), then we get

$$p_q = \sum_{j=s+1}^{S} m_j f_{q-S+j}, \quad S - s \leq q < \infty. \tag{5}$$

The probabilities m_j can be determined recursively (Axsäter 2006, p. 108) using

$$m_S = 1, \qquad m_j = \sum_{k=j+1}^{S} m_k f_{k-j}, \quad j = s+1, s+2, \ldots, S-1. \tag{6}$$

Similar to the derivation of (3), as shown in the appendix, we find the following expression for the difference in cost between policies that delay orders by T and $T+\Delta$ time units, respectively. The main difference to (3) is that the marginal cost difference is determined over all possible values of q multiplied with the probability of occurrence, i.e.

$$\left(-bE[q] + (h+b) \sum_{q=S-s}^{\infty} \sum_{i=1}^{q} p_q P_T(< S-q+i) \right) \Delta \qquad (7)$$

where $E[q]$ denotes the expected order size. As for (3), it is obvious that the cost difference (7) is strictly decreasing in T for customer inter-arrival distributions with an increasing failure rate. Thus, we get the following optimality condition for T.

$$(h+b) \sum_{q=S-s}^{\infty} \sum_{i=1}^{q} p_q P_T(< S-q+i) = bE[q] \qquad (8)$$

2.3. Erlang customer inter-arrival times

As a special case (that will also be used for numerical purposes in the next section), we consider customer inter-arrival times that follow an Erlang (λ, p) distribution. This is equivalent to assuming a 'censored' Poisson arrival process in which only every p^{th} event is recorded. This is a realistic representation of consumer purchasing behavior (e.g., Chatfield and Goodhardt, 1973). The mean customer inter-arrival time is $\frac{p}{\lambda}$ (or, equivalently, $\frac{\lambda}{p}$ is the number of demands per time unit) and p is the shape parameter that needs to be integer. For $p=1$ we obtain the exponential distribution. As p increases the distribution becomes less variable, and the variance goes to zero (i.e. constant times between demands) as p goes to infinity.

If the customer inter-arrival times are exponentially distributed, then it is well-known that base-stock policies are optimal under quite general conditions (e.g., Axsäter, 2006). So, there is no point in delaying orders in this case. However, if the customer inter-arrival times are less

variable, then delaying orders may be beneficial. Indeed, if the customer inter-arrival times are (almost) deterministic, then this is obviously the case. Because the Erlang distribution covers this entire 'range' from the exponential distribution ($p=1$) to the deterministic distribution ($p=\infty$), with a larger value for p implying less variation, this distribution is especially suitable for studying the benefit of delaying orders.

It is well-known that an Erlang (λ, p) distribution is the sum of p independent random 'phases' that are exponentially distributed with mean $1/\lambda$. So, the 'state' of the demand process is characterised by the number of phases $m, m = 0,1,\ldots, p-1,$ that have passed since the last demand occurred.

Clearly, the number of phases in an arbitrary period follows a Poisson distribution with rate λ . However, given that the last demand occurred T time units ago, the probability of being in phase $m, m = 0,1,\ldots, p-1,$ is (please refer to Cox, 1962)

$$\frac{e^{-\lambda T}(\lambda T)^m / m!}{\sum_{i=0}^{p-1} e^{-\lambda T}(\lambda T)^i / i!}.$$

If the demand process is in phase m at time T, then there are no demands in period $(T, L+T)$ if less then $p-m$ phase transitions occur in that period. Hence we get

$$N_T(0) = \sum_{m=0}^{p-1} \left(\frac{e^{-\lambda T}(\lambda T)^m / m!}{\sum_{i=0}^{p-1} e^{-\lambda T}(\lambda T)^i / i!} \sum_{n=0}^{p-m-1} e^{-\lambda L}(\lambda L)^n / n! \right). \qquad (9a)$$

If the demand process is in phase m at time T, then k demands occur in period $(T, L+T)$ if at least $kp-m$ and less than $(k+1)p-m$ phase transitions occur in that period. Hence we get

$$N_T(k) = \sum_{m=0}^{p-1} \left(\frac{e^{-\lambda T}(\lambda T)^m / m!}{\sum_{i=0}^{p-1} e^{-\lambda T}(\lambda T)^i / i!} \sum_{n=kp-m}^{(k+1)p-m-1} e^{-\lambda L}(\lambda L)^n / n! \right), \quad k = 1,2,\ldots. \qquad (9b)$$

Using (1), (2), (4), (9ab), the optimal delays for the flexible delay policy can be determined. Using (1), (2), (5), (6), (8) and (9ab), the optimal delay for the constant delay policy can be determined. We remark that, for numerical purposes, the infinite upper bound in the first summation of (9) can be replaced by $S-s+1$ plus the maximum demand size (that is likely to occur).

3. Numerical investigation and insights

To numerically analyse the performance of the constant and flexible delay policies, we have considered customer inter-arrival times that follow an Erlang (λ, p) distribution. As discussed in the previous section, for the memory-less case of $p=1$ base-stock policies are optimal and there is no cost benefit in delaying an order. Thus, this scenario is not considered further. Fixing $\lambda = 1$

and then varying the number of stages $p = 2,3....6$ allows us to progressively reduce the variability inherent in the process. The average customer inter-arrival time is then $p/\lambda = p$.

We have considered both constant and variable demand sizes (z). In the former case we have assumed $z = 1$, 2, 3. In that respect, we move beyond the assumption of unit-sized transactions which is the norm in the relevant literature. Further, and through the latter case, we also wish to study the effect of introducing the realistic assumption of demand sizes being variable. Demand sizes have a discrete uniform distribution $U(\alpha, \beta)$ with mean size $E(z) = (\alpha + \beta)/2$. Compared to other plausible candidates such as Geometric, Poisson and Log-series, the two parameters of the Uniform distribution facilitate fixing the mean and varying the standard deviation.

The (*s*, *S*) policies that have been extensively considered are the following: (*s*=0, *S*=1), (*s*=0, *S*=2), (*s*=0, *S*=3), (*s*=1, *S*=2), (*s*=1, *S*=3) and (*s*=2, *S*=3). Four lead time (L) values have been simulated: 1, 2, 3 and 4 periods. Finally, a wide range of b/h scenarios were considered by fixing $h = 1$ and varying $b = 1$, 2, 5, 10, 20.

For these settings, the order delays are optimised for both the constant and flexible delay policies along the lines presented in the previous section. In presenting the results, we will focus on the most interesting and insightful settings. The qualitative results for different constant demand sizes were very similar, and hence our focal point will be the case of compound demand as that is new to the literature. The sensitivity of the results with respect to key parameters such as *L* and *p* differ in size but not direction when the backorder cost (to holding cost ratio) varies, and we therefore report results for a fixed value of $b = 5$ only. However, our analysis also shows that the

higher the *b/h* ratio, the lower the benefits of allowing constant or flexible order delays. Hence, the sensitivity of the benefit to the *b/h* ratio is also presented, for a particular control parameters combination.

The entire simulation exercise was conducted within the *Rockwell ARENA* software. We remark that simulation is only needed to compare the costs of different policies, and that optimal delays are calculated exactly using the results in previous sections.

In Section 3.1, we will discuss the benefits of constant and flexible delay policies over no delay policies for specific (*s*,*S*) policies, when the values of *s* and *S* are fixed and therefore may not be optimal (for all considered parameter combinations). Note that in this case, where *s* and *S* are given as policy parameters, the delay can be used to correct for a suboptimal *s* value. In Section 3.2, we study the relevant benefits when *s* and *S* are optimised.

3.1. The benefit of delayed ordering: fixed *s* and *S*

We first elaborate on the (*s*=0, *S*=2) case, for a constant demand size $z = 2$. Note that every demand (of size 2) will trigger an order of size 2 and hence using a flexible delay policy is not relevant for this case. The benefits of the (constant) delay policy versus the no delay policy are presented in Figure 1 below for the various values of *L* and *p*, ($b/h = 5$).

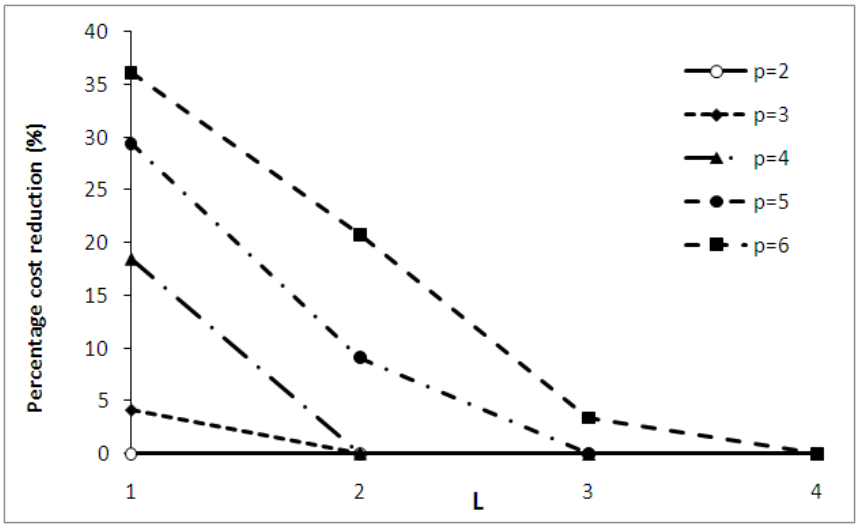


Figure 1: Cost reduction (%) of employing the optimal (constant) delay policy over the no delay policy for fixed (s=0, S=2), $z = 2, b/h = 5$.

The results demonstrate that the effects of non-unit sized transactions carry over from what is known for the unit sized ones (Moinzadeh, 2001) as far as the number of Erlang stages (p) and lead times (L) are concerned. When the lead-time increases (and consequently the lead time demand increases as well), the optimal delay decreases and so does the corresponding benefit. The delay policies exploit the information that less demand is expected over the next next L time units if a transaction has just occured (that triggered the order which can be delayed). That information is more valuable if the average inter-arrival time is larger compared to the lead time. This explains why delaying orders is more beneficial for smaller lead times.

It is also apparent from Figure 1 that the benefit decreases as the number of Erlang phases decreases, which is expected as the process then increasingly resembles the memory-less case. Conversely, as p increases the process tends towards the less variable or deterministic case and

the optimal delay increases implying more benefits. To summarize, the cost benefit offered by the constant delay policy increases in p and decreases in L.

We continue with the same (s=0, S=2) policy, but stochastic $U(1,3)$ demand sizes. Note that for this combination, there may already be a backorder when an order is placed, for instance when the demand size is 2 and the inventory position just before the demand is 1. It is therefore not surprising that the constant delay policy never delays orders, i.e. that the optimal constant delay is $T^*=0$ for all considered parameter combinations. This renders this scenario a particularly suitable one for studying the benefits of flexible over constant delay. The percentage cost reductions resulting from the employment of the flexible delay policy over the constant delay policy for (s=0, S=2) and $L=1$ are presented in Figure 2 below.

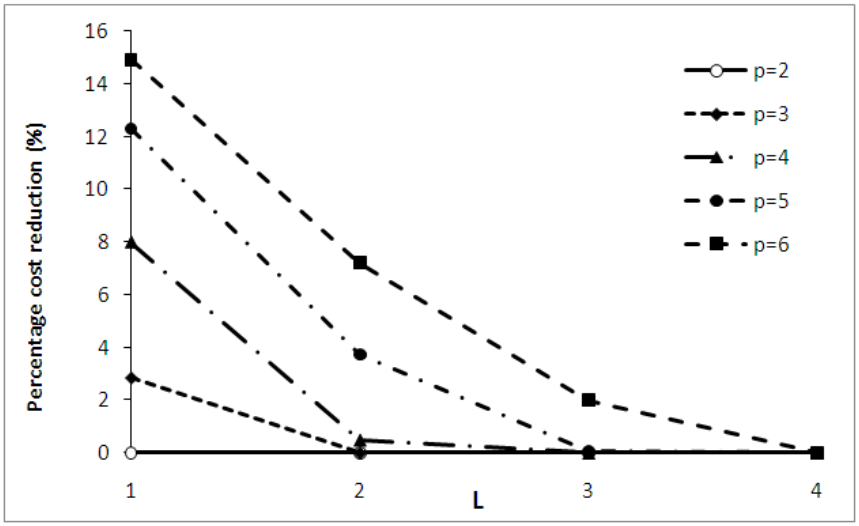


Figure 2: Cost reduction (%) of employing the optimal flexible delay policy over the optimal constant/no delay policy for fixed (s=0, S=2) and $z: U(1,3)$

Figure 2 shows that the benefit of allowing delays can be considerable, leading to a cost reduction of up to 15%.

Next, we show that the added flexibility still pays of if constant delays already provide a considerable benefit over no delays. To that end, we consider the results for the (s=1, S=3) policy and $U(1,3)$ demand sizes. Figures 3 and 4 show the benefits of constant versus no delay and flexible versus constant delay, respectively. It appears that although constant delays can be very beneficial with cost reductions of up to 20% compared to no delays, flexible delays still bring a considerable additional cost reduction of up to 7%. Experimentation under other (s, S) policies with the variance increasing from a constant size equal to 3, and then $U(2,4)$ and $U(1,5)$ leads to similar insights.

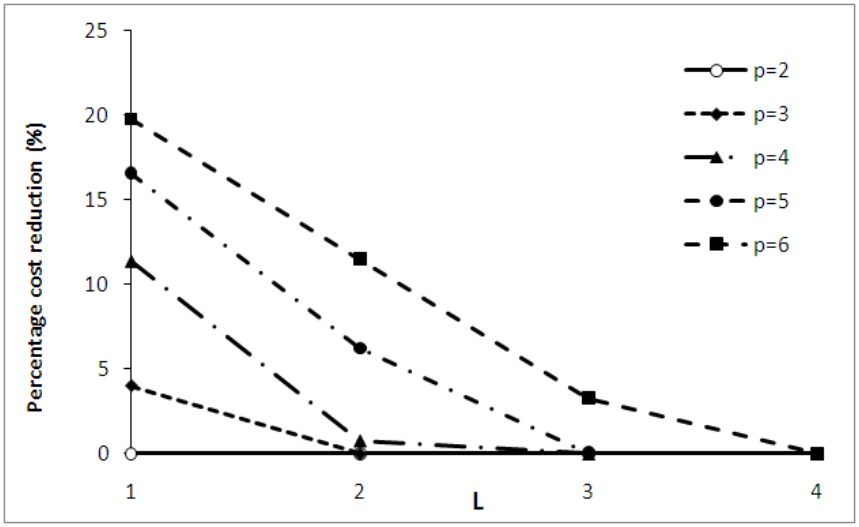


Figure 3: Cost reduction (%) of employing the optimal constant delay policy over the no delay policy for fixed (s=1, S=3) and $z : U(1,3)$

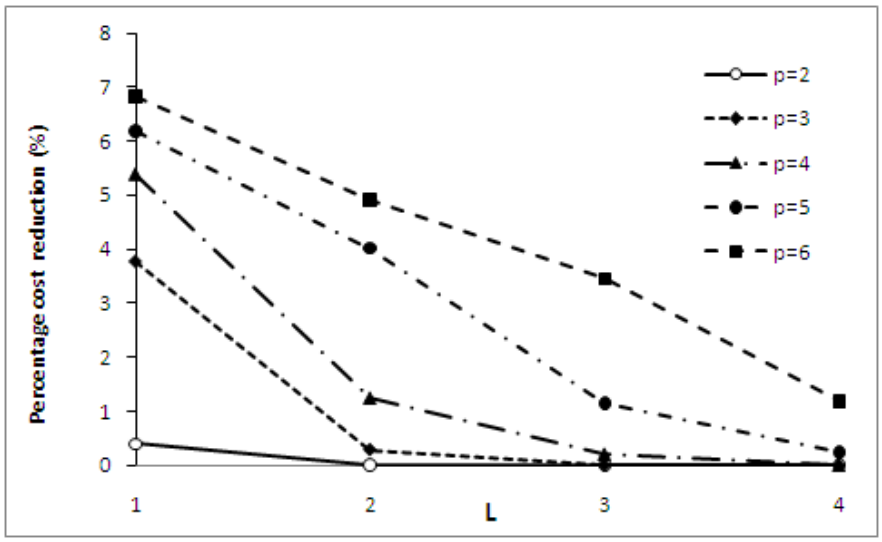
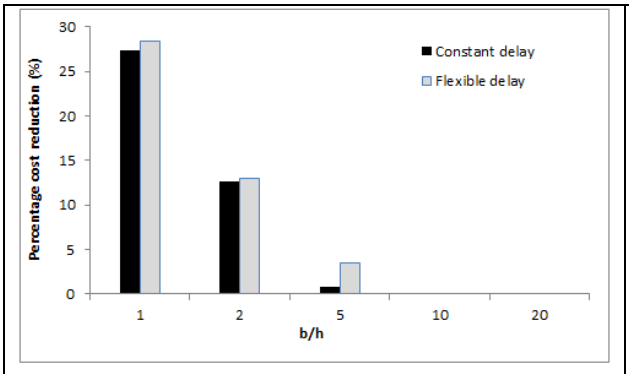
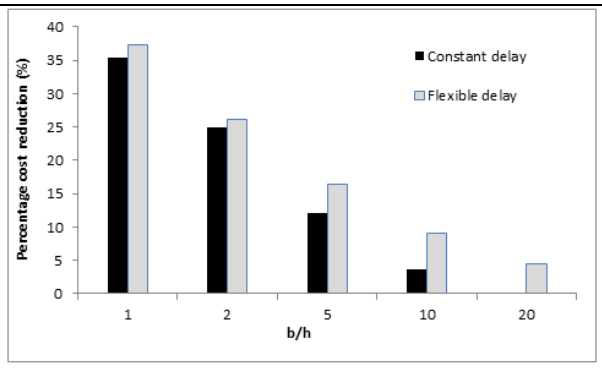


Figure 4: Cost reduction (%) of employing the optimal flexible delay policy over the optimal constant delay policy for fixed (s=1, S=3) and $z : U(1,3)$

We conclude this subsection by showing in Figure 5 the sensitivity of the benefit derived from (constant or flexible) delayed ordering to the b/h ratio for a fixed value L = 2 and two sub-cases of p = 4, 6.



p = 4



p = 6

Figure 5: Cost reduction (%) of employing the constant and flexible delay policies over the no delay policy for fixed (s=1, S=3), $z : U(1,3)$

Figure 5 shows that in all control parameter combinations the benefit obtained by using the flexible delay policy is higher than that obtained by the constant delay one, which is expected since the latter is a special a case of the former. It is also apparent that the benefit derived from delayed ordering is decreasing with the ratio *b/h* which can be attributed to the decrease in optimal delays. Finally, as *b/h* increases flexibility becomes more essential to reap the benefits of delaying orders.

3.2. The benefit of delayed ordering: optimal *s* and *S*

In this section we are contrasting the best flexible delay policy against the best constant delay one. Demand sizes are uniformly distributed between 1 and 3. We only consider policies for which $S - s = 2$. Varying $S - s$ would imply varying the order frequency and thereby order cost in practice. Rather than including order cost in our study and numerical investigation, we choose to fix $S - s$ to 2. The optimal values of *s* and *S* (under this restriction) turned out to be identical for the three types of policies (no-delay, constant and flexible delay) for all considered parameter combinations. These optimal values are reported in Table 1. Recall, that the cost was shown to be quasi-convex in the maximum delay and so the optimal maximum delay is easily determined for any given values of *s* (and *S*), allowing us also to consider a wide range of values for *s* (and *S*) in order to determine the best one(s). We remark that the costs appeared to be unimodal in *s* (and *S*) for all considered cases. If this could be shown to hold in general (which is not straightforward and beyond the scope of this research), even more efficient search procedures could be applied. Especially for larger values of *s* and *S* (than considered here), this could be an important advantage.

L		1	2	3	4
p=2	**(s,S)***	(1,3)	(2,4)	(2,4)	(5,7)
p=3	**(s,S)***	(0,2)	(1,3)	(2,4)	(3,5)
p=4	**(s,S)***	(0,2)	(1,3)	(1,3)	(2,4)
p=5	**(s,S)***	(0,2)	(1,3)	(1,3)	(1,3)
p=6	**(s,S)***	(0,2)	(1,3)	(1,3)	(1,3)

Table 1: Optimal (*s*, *S*) values under the restriction that $S - s = 2,\ z : U(1,3)$)

The benefits of the best constant delay policy over the best policy with no delay and those of the best flexible delay over the best constant delay policy are given in Figures 6 and 7 respectively.

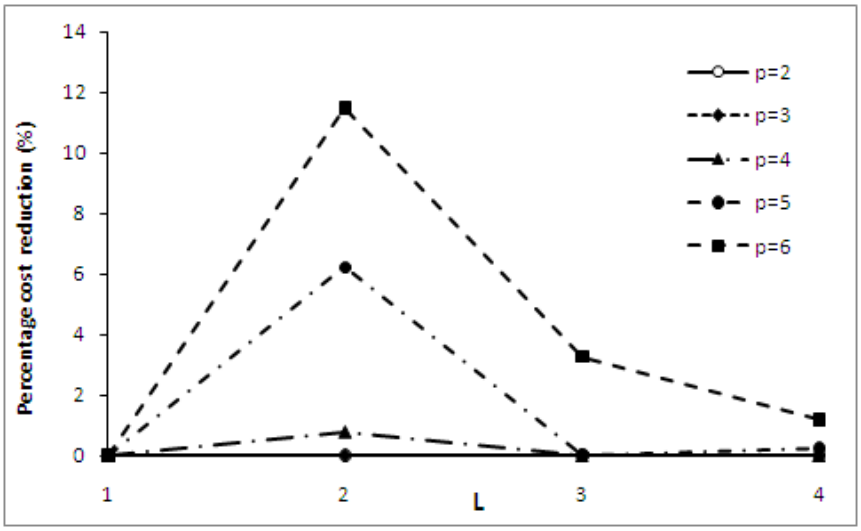


Figure 6: Cost reduction (%) of employing the best constant delay policy over the best policy with no delay for $z : U(1,3)$

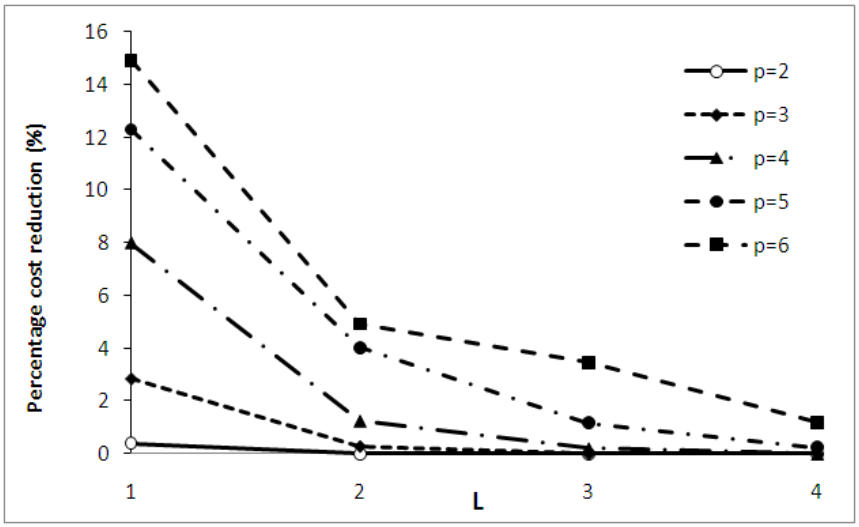


Figure 7: Cost reduction (%) of employing the best flexible delay policy over the best constant delay policy for $z : U(1,3)$

The results in Figures 6 and 7 demonstrate the considerable benefits resulting from the introduction of flexible delays. Cost reductions of up to 15% are achieved compared to the constant delay policies, which in turn outperform the base case of no delays by as much as 12%. Overall the benefits of flexible delays are substantial enough to out-weigh any implementation related advantages (in non-computerized systems) associated with the constant delay policies.

4. Discussion and conclusions

We have studied for the first time the benefits of delayed ordering in a continuous (*s*, *S*) inventory control setting facing compound demand. In addition to constant delays, flexible delay policies have also been considered that take into account the order quantity (and corresponding undershoot). The modelling features of our study constitute collectively a realistic representation of the problem in hand and form an important extension to previously published work in this area. General optimality conditions have been derived and the exact form of these conditions has

been provided for the case of Erlang distributed customer inter-arrival times. Subsequently, the performance of these solutions has been assessed through an extensive numerical investigation.

The numerical analysis illustrates the considerable benefits of delaying ordering, which increase as the variability of the demand process declines and decrease with the lead time length. Constant delay policies have been found to considerably outperform classical no-delay inventory control by offering cost reductions of up to 12%. Moreover, especially in situations with lumpy demand where constant delays are not very effective, the dynamic nature of flexible delays provides (additional) savings of up to 15%. The trade-off between cost reduction and implementation related requirements is an important one but the results justify the introduction of flexible delays in any real world system. This leads to an interesting direction for further research, namely to consider the integration of the flexible policies with real time business planning automated solutions.

Ordering costs have not been considered in our work and such an inclusion would facilitate studying the effect of the delay structure on the optimal order quantity. Furthermore, in our work, as commonly considered in the literature, we have assumed for our proposed policies that order delays do not affect the sizes of orders. This has allowed us to analyse the cost benefit of purely delaying orders and not the potential additional benefit of adjusting order sizes and thereby the number of orders as well. This explains why ordering costs are not affected by the delays for our policies and therefore not included in our work. Note also that the reported cost benefits are conservative in this sense, since additional benefit might be achieved by updating replenishment orders if a new demand arrives that cuts a delay short. Obviously, an alternative would be to update order sizes if a ‘delay is stopped’ because a new demand arrives, which would be an

interesting avenue for further research. Of course, ordering costs should be included in such an analysis.

Consideration of variable lead times would also add further realism to the policies discussed here. Also, given the importance of delayed ordering in many inventory situations, experimentation with real-world data and research on the empirical performance of the policies discussed in this paper is merited. Finally, a comparison with periodic policies would potentially allow interesting insights to emerge. Periodic formulations may hide the effects of delayed ordering, since delays shorter than the review period may not be effectively realised. It is important to note that periodic control situations where the lead time may be shorter than the average inter-demand interval have already been explored in the literature (Syntetos *et al*., 2009).

Before we close this paper we discuss the potential role of delayed ordering in achieving more effective inter-departmental co-operation between Operations (or whatever department the inventory task is performed under) and Finance. The latter function generally views inventory as a liability, or a depreciating asset; i.e. as far as Finance is concerned the arrival of these units could have been delayed without harm. Dickinson (2013) confirms that "ordering practices may have a considerable impact on cash flow" and thereby the considerable potential benefit of delayed ordering mechanisms. In addition, the effect of delaying orders may be similar to that related to deferred payment terms from the supplier. Such a practice can be considered a blanket approach to inventory cost reduction. While this approach does not directly reduce the amount of inventory on hand, it does delay the amount of cash tied up in carrying inventory. Delayed ordering in conjunction with extended supplier payment terms can become a profit center for an organisation.

Appendix: Cost difference derivation for a constant delay

The constant delay policy applies the same maximum delay for all order quantities. So, different from a flexible delay policy where the maximum delay $T(q)$ depends on the order size q, the constant delay policy is associated with a single delay parameter T. So, the expected cost difference between either applying T or the marginally larger $T+\Delta$ should consider all possible order sizes and their respective probabilities. In the remainder of this appendix, we will therefore take this into account when deriving the cost difference (7) between policies that delay orders by a maximum of T or $T+\Delta$ time units, respectively. The derivation is otherwise similar to that for the cost difference (3) under a flexible delay, where we recall that for ease of presentation T was also used in that derivation rather than $T(q)$.

Increasing the maximum delay from T (policy π) or $T+\Delta$ (policy $\pi(\Delta)$) only matters if the following event occurs: a demand occurs, say at time 0, that triggers an order and no further demands occur for the next T time units. If this happens, then π places the order at time T while $\pi(\Delta)$ continues to delay that order for at most Δ time units, and so the costs can only differ in period $(L+T, L+T+\Delta)$. Given the order size q for the considered event, the cost difference is exactly that as given in (3) under a flexible delay. By summing over all possible values of q and multiplying with the corresponding probability p_q that an arbitrary order is indeed of size q, we get that the expected cost difference under constant delay is

$$\sum_{q=S-s}^{\infty} p_q \left(-bq + (h+b) \sum_{i=1}^{q} \left(P_T(< S-q+i) \right) \right) \Delta$$
$$= \left(-bE(q) + (h+b) \sum_{q=S-s}^{\infty} \sum_{i=1}^{q} p_q \left(P_T(< S-q+i) \right) \right) \Delta$$

as is given in (7).´

References

Aberdeen Group (2005). *The service parts management solution selection report*. SPM Strategy and Technology Selection Handbook. Boston, MA: Aberdeen Group.

Andersson, J., Axsäter, S., and Marklund, J. (1998). Decentralized Multi-Echelon Inventory Control. *Production and Operations Management*, *7*(4), 370-386

Axsäter, S. (2000). Exact Analysis of Continuous Review (R, Q) Policies in Two-Echelon Inventory Systems with Compound Poisson Demand. *Operations Research*, *48*(5), 686-696.

Axsäter, S., and Viswanathan, S. (2012). On the value of customer information for an independent supplier in a continuous review inventory system. *European Journal of Operational Research*, *221*(2), 340-347.

Axsäter, S. (2006). *Inventory control*. 2nd ed. New York, NY: Springer-Verlag.

Bartezzaghi, E., Verganti, R., and Zotteri, G. (1999). A simulation framework for forecasting uncertain lumpy demand. *International Journal of Production Economics*, *59*(1-3), 499-510.

Berling, P. and Marklund, J. (2006). Heuristic Coordination of Decentralized Inventory Systems Using Induced Backorder Costs, *Production and Operations Management*, *15*(2), 294–310.

Berling, P., and Marklund, J. (2013). A model for heuristic coordination of real life distribution inventory systems with lumpy demand. *European Journal of Operational Research*, *230*(3). 515-526.

Boylan, J.E., Syntetos, A.A., and Karakostas, G.C. (2008). Classification for forecasting and stock control: a case study. *Journal of the Operational Research Society*, *59*(4), 473-481.

Chatfield, C., and Goodhardt, G.J. (1973). A consumer purchasing model with Erlang inter-purchase times. *Journal of the American Statistical Association*, *68*(4), 828-835.

Cox, D.R. (1962). *Renewal theory*. London: Methuen.

Deloitte (2006). *The service revolution in global manufacturing industries*. New York, NY: Deloitte Research.

Deuermeyer, B.L., and Schwarz, L.B. (1981). A model for the analysis of system service level in ware- house/retailer distribution systems: the identical retailer case. In L.B. Schwartz (Ed.), *Multi-level production/inventory control systems: Theory and practice*. Amsterdam: Elsevier Science Ltd, 163-193.

Dickinson, P. (2003). Managing Consultant, Tata Consultancy Services (http://www.tcs.com). Private communication to the authors.

Donnelly, J.M. (2013). The case for managing MRO inventory. *Supply Chain Management Review*, *17*(2), 18-25.

Inderfurth, K., and Kleber, R. (2013). An advanced heuristic for multiple-option spare parts procurement after end-of-production. *Production and Operations Management*, *22*(1), 54-70.

Iglehart, D. (1963). Optimality of (s,S) policies in the infinite horizon dynamic inventory problem. *Management Science*, *9*(2), 259-267.

Katircioglu, K. (1996). *Essays in Inventory control*. Doctoral dissertation, The University of British Columbia, Vancouver, British Columbia, Canada.

Kennedy, W.J., Patterson, J.W., and Fredendall, L.D. (2002). An overview of recent literature on spare parts inventories. *International Journal of Production Economics*, *76*(2), 201-215.

Lee, H.L., and Moinzadeh, K. (1987). Operating characteristics of a two-echelon inventory system for repairable and consumable items under batch ordering and shipment policy. *Naval Research Logistics Quarterly*, *34*(3), 365-380.

Liu, L., and Shi, D.H. (1999). An (s, S) model for inventory with exponential lifetimes and renewal demands. *Naval Research Logistics*, *46*(1), 39-56.

Moinzadeh, K. (2001). An improved ordering policy for continuous review inventory systems with arbitrary inter-demand time. *IIE Transactions*, *33*(2), 111-118.

Moinzadeh, K., and Lee, H.L. (1986). Batch size and stocking levels in multi-echelon repairable systems. *Management Science*, *32*(12), 1567-1581.

Moinzadeh, K., and Zhou, Y.P. (2008). Incorporating a delay mechanism in ordering policies into multi-echelon distribution systems. *IIE Transactions*, *40*(4), 445-458.

Porras, E., and Dekker, R. (2008). An inventory control system for spare parts at a refinery: An empirical comparison of different re-order point methods. *European Journal of Operational Research*, *184*(1), 101-132.

Sahin, I. (1990). *Regenerative inventory systems*. New York, NY: Springer-Verlag.

Scarf, H.E. (1959). The optimality of (s,S) policies in the dynamic inventory problem. *Mathematical Methods in the Social Sciences*. *Proceedings of the First Stanford Symposium*. Stanford, CA: Stanford University Press, 196-202.

Schultz, C.R. (1987). Forecasting and inventory control for sporadic demand under periodic review. *Journal of the Operational Research Society*, *38*(5), 453-458.

Schultz, C.R. (1989). Replenishment delays for expensive slow-moving items. *Management Science*, *35*(12), 1454-1462.

Schultz, H., and Johansen, S.G. (1999). Can-order policies for coordinated inventory replenishment with Erlang distributed times between ordering. *European Journal of Operational Research***,** *113*(1), 30-41

Strijbosch, L.W.G., Heuts, R.M.J., and van der Schoot, E.H.M. (2000). A combined forecast-inventory control procedure for spare parts. *Journal of the Operational Research Society*, *51*(10), 1184-1192.

Svoronos, A., and Zipkin, P. (1988). Estimating the performance of multi-level inventory systems. *Operations Research*, *36*(1), 57-72.

Syntetos, A.A., Babai, M.Z., Dallery, Y., and Teunter, R.H. (2009). Periodic control of intermittent demand items: theory and empirical analysis. *Journal of the Operational Research Society*, *60*(5), 611-618.