ARTICLE IN PRESS

Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

European Journal of Operational Research xxx (2006) xxx–xxx

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/ejor

Production, Manufacturing and Logistics

An integrated production and inventory model to dampen upstream demand variability in the supply chain

Robert N. Boute [a,*], Stephen M. Disney [b], Marc R. Lambrecht [a], Benny Van Houdt [c]

[a] *Department of Applied Economics, K.U. Leuven, Naamsestraat 69, 3000 Leuven, Belgium*
[b] *Logistics Systems Dynamics Group, Cardiff Business School, Cardiff University, Aberconway Building, Colum Drive, Cardiff CF10 3EU, UK*
[c] *Department of Mathematics and Computer Science, University of Antwerp, Middelheimlaan 1, 2020 Antwerpen, Belgium*

Received 27 May 2005; accepted 18 January 2006

Abstract

We consider a two-echelon supply chain: a single retailer holds a finished goods inventory to meet an i.i.d. customer demand, and a single manufacturer produces the retailer's replenishment orders on a make-to-order basis. In this setting the retailer's order decision has a direct impact on the manufacturer's production. It is a well known phenomenon that inventory control policies at the retailer level often propagate customer demand variability towards the manufacturer, sometimes even in an amplified form (known as the bullwhip effect). The manufacturer, however, prefers to smooth production, and thus he prefers a smooth order pattern from the retailer. At first sight a decrease in order variability comes at the cost of an increased variance of the retailer's inventory levels, inflating the retailer's safety stock requirements. However, integrating the impact of the retailer's order decision on the manufacturer's production leads to new insights. A smooth order pattern generates shorter and less variable (production/replenishment) lead times, introducing a compensating effect on the retailer's safety stock. We show that by including the impact of the order decision on lead times, the order pattern can be smoothed to a considerable extent without increasing stock levels. This leads to a situation where both parties are better off.

Keywords: Inventory control; Queueing; Markov processes; Supply chain management

1. Introduction

Lee et al. (1997a,b) describe a problem frequently encountered in supply chains, called the *bullwhip effect*: demand variability increases as one moves up the supply chain. This distorted information throughout

* Corresponding author. Tel.: +32 16 32 69 66.
E-mail addresses: robert.boute@econ.kuleuven.ac.be (R.N. Boute), disneysm@cardiff.ac.uk (S.M. Disney), marc.lambrecht@econ.kuleuven.ac.be (M.R. Lambrecht), benny.vanhoudt@ua.ac.be (B. Van Houdt).

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2006.01.023

the supply chain can lead to inefficiencies: excessive inventory investment, poor customer service, lost revenues, misguided capacity plans, ineffective transportation and missed production schedules (Lee et al., 1997a).

Even when demand variability is not amplified but merely transmitted to the upstream echelons, this order variability can have large upstream cost repercussions. In a make-to-order supply chain, the upstream manufacturer – pursuing smooth production – prefers minimal variability in the replenishment orders from the (downstream) retailer. Balakrishnan et al. (2004) emphasize the opportunities to reduce supply chain costs by dampening upstream demand variability. This has led to the creation of new replenishment rules that are able to generate *smooth* order patterns, which we call ''smoothing replenishment rules''. Smoothing is a well-known method to reduce variability. A number of production level smoothing rules were developed in the 1950s and 1960s (Cf. Magee, 1956; Magee, 1958; Simon, 1952; Vassian, 1955; Deziel and Eilon, 1967). The more recent work on smoothing replenishment rules can be found in Dejonckheere et al. (2003), Balakrishnan et al. (2004) and Disney et al. (in press).

The production-smoothing model has also received a lot of attention in the macro-economic literature. Early theoretical investigations of optimal inventory and production behavior established that if production costs are convex, then it is optimal for a firm to only partially adjust output in response to a change in its inventory position. This resulted in the *production-smoothing hypothesis*, where we would expect to observe sales more variable than output. Among others, Blinder (1986), Blanchard (1983), West (1986), Miron and Zeldes (1988), Fair (1989), Krane and Braun (1991), and more recently Allen (1997) are all concerned with the question of whether production is smoothed relative to sales.

We have to be careful not to focus only on one side of the production smoothing ''coin''. The manufacturer does benefit from smooth production, but retailers, driven by the goal of reducing inventory (holding and shortage/backlog) costs, prefer to use replenishment policies that chase demand rather than dampen consumer demand variability. Dampening variability in orders may have a negative impact on the retailer's customer service due to inventory variance increases (Bertrand, 1986; Dejonckheere et al., 2002; Disney and Towill, 2003). Inventory acts as a buffer, absorbing increases or decreases in demand while production remains relatively steady (Buffa and Miller, 1979). This leads to a tension between the retailer's and manufacturer's preferred order variability.

However, we can model a two-stage make-to-order supply chain as a production-inventory system, where the retailer's inventory replenishment lead times are endogenously determined by the manufacturer's production facility. In this framework the choice of the retailer's replenishment policy (amplifying or dampening customer demand variability in the replenishment orders) determines the arrival process at the manufacturer's production queue and as such it affects the distribution of the production lead times. We expect that a smooth order pattern gives rise to shorter and less variable lead times due to the laws of factory physics (Hopp and Spearman, 2001). This may exercise a compensating effect on the retailer's safety stock.

In this paper we consider an inventory control policy that is able to dampen the upstream demand variability by generating a smooth order pattern. Moreover we integrate the impact of this order decision on the manufacturer's production system. We develop a procedure to estimate the lead time distribution given the explicit order pattern generated by our smoothing replenishment rule. We then focus on the resulting impact of order smoothing on the safety stock requirements to provide a given service level. Many papers on the control of production-inventory systems explicitly include the replenishment delay in their models, for example Axsäter (1976), Towill (1982), Riddalls and Bennett (2002), and Warburton (2004b), but the replenishment delay always remains fixed, independent of the replenishment policy. To the best of our knowledge, we think that an integrated production and inventory analysis of order smoothing has yet to be investigated.

The remainder of the paper is organized as follows. Our research model and its assumptions are presented in the next section. In Section 3 we describe the retailer's inventory control policy and we compare the standard base-stock replenishment policy with a smoothing replenishment rule. Section 4 analyses the manufacturer's production system and discusses the procedure to estimate the lead time distribution using matrix analytic methods. In Section 5 we combine both supply chain echelons by aggregating the inventory and production subsystems into a production/inventory model and we analyse the impact on customer service and safety stock. Numerical results are presented in Section 6, and Section 7 concludes.

2. Model description

We consider a two-echelon supply chain with a single retailer and a single manufacturer. Every period, the retailer observes customer demand. If there is enough on-hand inventory available, the demand is immediately satisfied. If not, the shortage is backlogged. To maintain an appropriate amount of inventory on hand, the retailer places a replenishment order with the manufacturer at the end of every period.

The manufacturer does not hold a finished goods inventory but produces the retailer's orders on a make-to-order basis. The manufacturer's production system is characterized by a single server queueing model that sequentially processes the orders which require stochastic unit processing times. Once the complete replenishment order is produced, it replenishes the retailer's inventory. The time from the period an order is placed to the period that it replenishes the retailer's inventory, is the replenishment lead time T_p. The queueing process at the manufacturer implies that the retailer's replenishment lead times are stochastic and correlated with the order quantity. A schematic of the model is shown in Fig. 1.

2.1. Assumptions

- The sequence of events in a period is as follows. The retailer first receives goods from the manufacturer, then he observes and satisfies customer demand and finally, he places a replenishment order with the manufacturer.
- Customer demand, D_t, is independently and identically distributed (i.i.d.) over time according to an arbitrary, finite, discrete distribution.
- The order quantity, O_t, is determined by the retailer's replenishment policy. The retailer's replenishment rule defines the variability in the orders placed on the manufacturer. In Section 3 we discuss how we can control this upstream demand variability.
- The replenishment orders are processed by a single server first-come-first-served. This excludes the possibility of order crossovers. When the server is busy, they join a queue of unprocessed orders.
- The service times of a single item are i.i.d. according to a phase type (PH) distribution. We provide an algorithm to match the first two moments of an arbitrary distribution to a 2-phase PH distribution in Appendix A. To ensure stability (of the queue), we assume that the utilization of the production facility (average batch production time divided by average batch interarrival time) is strictly smaller than one.
- The time from the moment the order arrives at the production queue to the point that the production of the entire batch is finished, is the *production* lead time or response time that we denote by T_r. Note that the production lead time is not necessarily an integer number of periods. Since in our inventory model events occur on a discrete time basis with a time unit equal to one period, the *replenishment* lead time, denoted by T_p, has to be expressed in terms of an integer number of periods. We therefore rely on the sequence of events. In our sequence of events, the retailer is always able to satisfy demand after the receipt of products from the manufacturer (see Fig. 2). For instance, suppose that the retailer places an order at the end of period t, and it turns out that the *production* lead time is 0.8 periods. This order quantity will be added to the inventory in the next period $t+1$, and can be used to satisfy demand in period $t+1$. Therefore

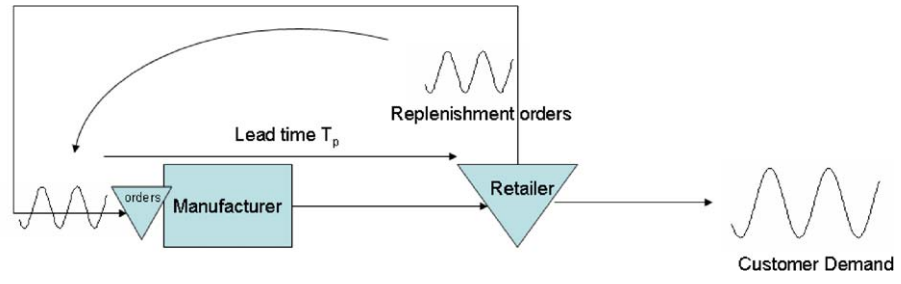


Fig. 1. A two-stage make-to-order supply chain.

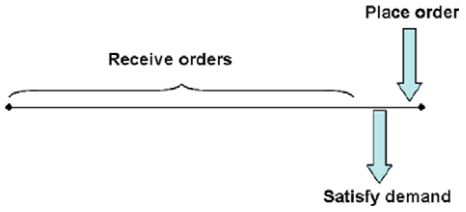


Fig. 2. Sequence of events in a period: (1) receive order, (2) satisfy customer demand, (3) place order.

the *replenishment* lead time is 0 periods. An order O_t with a production lead time of 1.4 periods is added to the inventory in period $t+2$ and can be used to satisfy demand D_{t+2}. Consequently we will treat the 1.4 period production lead time as an integer 1 period replenishment lead time. Hence, we round the response time T_r down to the nearest integer T_p (i.e., setting $T_p = \lfloor T_r \rfloor$) to obtain the (discrete) replenishment lead time.

3. Inventory control policy

There are many different types of replenishment policies, of which two are commonly used: the periodic review, order-up-to policy, and the continuous review, reorder point, order quantity model. Given the common practice in retailing to replenish inventories frequently (e.g., daily) and the tendency of manufacturers to produce to demand, we will focus our analysis on periodic review, base-stock, or order-up-to replenishment policies.

3.1. Standard base-stock replenishment policy

The standard periodic review base-stock replenishment policy is the (R,S) replenishment policy (Silver et al., 1998). At the end of every review period R, the retailer tracks his inventory position IP_t, which is the sum of the inventory on hand (that is, items immediately available to meet demand) and the inventory on order (that is, items ordered but not yet arrived due to the lead time), minus the backlog (that is, demand that could not be fulfilled and still has to be delivered). A replenishment order is then placed to raise the inventory position to an order-up-to or base-stock level S_t, which determines the order quantity O_t:

$$O_t = S_t - IP_t. \tag{1}$$

The base-stock level S_t is the inventory required to ensure a given customer service. Orders are placed every R periods, and after an order is placed, it takes T_p periods for the replenishment to arrive (with T_p being the stochastic replenishment lead time). Hence the risk period (the time between placing a replenishment order until receiving the subsequent replenishment order) is equal to the review period plus the replenishment lead time $R + T_p$. The base-stock level covers the forecasted demand during the risk period plus a buffer or safety stock SS to meet unexpected fluctuations in demand during this risk period. As customer demand is i.i.d., the best possible demand forecast is the average of all previous demands, $E(D)$, so that

$$S_t = \left[E\left(T_p\right) + R\right] \cdot E(D) + SS \tag{2}$$

remains constant over time, or $\forall t$: $S_t = S$, with S a constant. Note that, when forecasting the i.i.d. customer demand with e.g. the moving average or exponential smoothing forecasting technique, the order-up-to level S_t varies over time, or is *adaptive* (Kim et al., in press).

Placing a replenishment order every period t, the inventory position IP_t at the end of period t equals last period's inventory position (which is raised up to the base-stock level S_{t-1}) minus the observed customer demand D_t. Hence, similar to Chen et al. (2000a), we can rewrite (1) as

$$O_t = S_t - (S_{t-1} - D_t), \tag{3}$$

and substituting (2) into (3) we see that the order pattern is equal to the demand pattern:

$$O_t = D_t. \tag{4}$$

Consequently the standard base-stock policy generates orders whose variability is equal to the variability of customer demand. Thus, when customer demand is wildly fluctuating, this replenishment rule sends a highly variable order pattern to the manufacturer, which may impose high capacity and inventory costs on the manufacturer. The manufacturer not only prefers a leveled production schedule, the smoothed demand also allows him to minimize his raw materials inventory cost. Therefore, we discuss a smoothing replenishment policy that is able to reduce the variability of the orders transmitted upstream.

3.2. Smoothing replenishment policy

In this section we describe two ways to develop a replenishment rule that dampens the upstream demand variability. We show that they both come down to the same result under the assumptions of the model considered in this paper. The first approach stems from linear control theory and introduces a proportional controller into the standard base-stock policy. The second approach originates from Balakrishnan et al. (2004) who propose to set the order quantity equal to a convex combination of previous demand realisations in order to dampen the upstream demand variability.

We start with the linear control theory approach. Forrester (1961) and Magee (1958) propose not to recover the entire deficit between the base-stock level and the inventory position in one time period (contrary to what happens in (1)), but instead order only a fraction β of the inventory deficit:

$$O_t = \beta \cdot (S - IP_t). \tag{5}$$

Forrester (1961) refers to $1/\beta$ as the "adjustment time" and hence explicitly acknowledges that the deficit recovery should be spread out over time. This particular replenishment policy is recently used by, among others, Dejonckheere et al. (2003), Warburton (2004a,b) and Disney et al. (in press).

When customer demand is i.i.d., we forecast lead time demand with its average and consequently always order up to the same base stock level S, so that S is constant over time. This means that

$$O_t - O_{t-1} = \beta \cdot (S - IP_t) - \beta \cdot (S - IP_{t-1}) = \beta \cdot (IP_{t-1} - IP_t).$$

The inventory position IP_t is monitored after customer demand D_t is satisfied and before replenishment order O_t is placed. Hence

$$IP_t = IP_{t-1} + O_{t-1} - D_t,$$

so that we obtain

$$O_t - O_{t-1} = \beta \cdot (D_t - O_{t-1}),$$

or

$$O_t = (1 - \beta) \cdot O_{t-1} + \beta \cdot D_t. \tag{6}$$

The ordering quantity is a weighted combination of the previous order quantity and the last observed customer demand. Moreover, as in Eq. (4), we do not need the lead time distribution to make our order decision, although the base-stock level S in (5) does depend upon the lead time.

It is notable that the replenishment rule described by (6) is exactly the same as the *exponential smoothing policy* proposed by Balakrishnan et al. (2004) to decrease order variability. Balakrishnan et al. (2004) set the order quantity equal to a convex combination of previous demand realisations:

$$O_t = \sum_{k=0}^{\infty} \beta_k D_{t-k}. \tag{7}$$

Setting $\beta_k = \beta(1 - \beta)^k$ gives a similar result to Eq. (6). The authors of this paper and Balakrishnan et al. independently came to the same conclusion.

From Eq. (6) we can determine the variability in orders created by our smoothing rule:

$$\mathrm{Var}(O) = (1-\beta)^2\mathrm{Var}(O) + \beta^2\mathrm{Var}(D) + 2\beta(1-\beta)\mathrm{covar}(O_{t-1}, D_t) = (1-\beta)^2\mathrm{Var}(O) + \beta^2\mathrm{Var}(D),$$

and we obtain

$$\mathrm{Var}(O) = \frac{\beta}{2-\beta}\mathrm{Var}(D). \tag{8}$$

If we do not smooth, i.e. if $\beta = 1$, these expressions reduce to the standard base-stock policy, where $O_t = D_t$: we chase sales and thus the variability in orders equals the customer demand variability. For $1 < \beta < 2$ we amplify the demand variability in the replenishment orders (known as the bullwhip effect) and for $0 < \beta < 1$ we are able to dampen the demand variability and generate a smooth replenishment pattern.

With this *generalised* replenishment policy we can clearly reduce the variance transmitted upstream by decreasing β. Under a fixed lead time assumption, such a smoothing policy is justified when production (or ordering) and holding costs are convex or when there is a cost of changing the level of production (Veinott, 1966). When the production capacity is fixed and lead times result from a single server queueing system (as in the model described in this paper), this replenishment rule enables us to smooth the manufacturer's production, resulting in shorter order-to-delivery times and more balanced, peak shaving production schedules, which are beneficial for the manufacturer.

Besides the benefits realised through smoother planning, the manufacturer also realises cost savings on its own raw material and/or component inventories. Hosoda and Disney (2005) show that if one faces a first-order autoregressive demand pattern such as Eq. (6) and adopts the optimal base-stock policy (with minimum mean squared error forecasting), the inventory variance declines as β decreases, reducing the safety stock requirements. Balakrishnan et al. (2004) state that this replenishment rule serves to provide advanced order information to the manufacturer. The retailer's replenishment orders are not statistically independent, because from (6) we can derive that $corr(O_t, O_{t-x}) = (1-\beta)^x$, and the dependence between successive orders creates an opportunity for the manufacturer to use information embedded in past retailer orders.

Since order smoothing leads to a number of cost savings for the manufacturer, it seems to be a dominating operations strategy. We have to be careful not to focus only on one side of the production smoothing "coin" however. In developing a replenishment rule one has to consider the impact on the inventory variance as well. The manufacturer does benefit from smooth production, but dampening variability in orders may have a negative impact on the retailer's customer service due to inventory variance increases (Bertrand, 1986; Dejonckheere et al., 2002; Disney and Towill, 2003).

Disney et al. (in press) quantify the variance of the net stock and compute the required safety stock as a function of the smoothing intensity. Their main conclusion is that when customer demand is i.i.d., order smoothing comes at a price: in order to guarantee the same fill rate, more investment in safety stock is required. As a consequence, retailers, driven by the goal of reducing inventory costs (holding and shortage/backlog), prefer to use replenishment policies that chase demand rather than dampen consumer demand variability.

However, the manufacturer produces on a make-to-order base and (production/replenishment) lead times are determined by the manufacturer's queueing model. This implies that the retailer's order pattern determines the arrival process at the queue and as such it affects the distribution of the lead times. We expect that a smooth order pattern gives rise to shorter and less variable lead times due to the laws of factory physics (Hopp and Spearman, 2001). This in turn exercises a downward effect on the retailer's inventory level, which may compensate the increase in inventory variance. The impact of the smoothing decision on lead time reduction is the topic of the next section.

4. The impact of order variance dampening on lead times

4.1. Interaction between the retailer's inventory policy and the manufacturer's queueing system

Most inventory models proposed in the literature take the replenishment lead time, T_p, as a fixed constant or as an exogenous variable with a given probability distribution. However, the replenishment orders do in

fact load the production facilities. The nature of this loading process relative to the available capacity and the variability it creates are the primary determinants of lead times in the facility (Karmarkar, 1993). Therefore, the inventory control system should work with a lead time which is a good estimate of the real lead time, depending on the production load, the interarrival rate of orders, and the variability of the production system (Hopp and Spearman, 2001). Zipkin (2000, p. 246) states: "to understand the overall inventory system, we need to understand the supply system. For this purpose we can and do apply the results of queueing theory".

In this paper we explicitly model the two-stage supply chain described in Section 2 as a production-inventory system, where the retailer's inventory replenishment lead times are endogenously determined by the manufacturer's queueing model. It is essential to extend pure inventory systems with *exogenous* lead times to production-inventory systems with *endogenous* lead times. After all, inventory influences production by initiating orders, and production influences inventory by completing and delivering orders to inventory. In Fig. 3, the interaction between the retailer's inventory system and the manufacturer's production system is illustrated: the retailer's replenishment policy generates orders that constitute the arrival process at the manufacturer's production queue. The time until the order is produced (sojourn time in the production system), is the time to replenish the order. These production/replenishment lead times are load–dependent and affected by the current size of the order queue with the production system. The replenishment lead time in turn is a prime determinant in setting the safety stock requirements at the inventory system. In Boute et al. (2004), it is shown that ignoring endogenous lead times may lead to seriously underestimated customer service levels and/or excessive inventory holdings.

By analysing the characteristics of the replenishment orders, we implicitly analyse the characteristics of the production orders that arrive to the production queue (Van Nyen et al., 2005). In a periodic review base-stock policy, the arrival pattern consists of batch arrivals with a fixed interarrival time (equal to the review period, $R = 1$) and with variable batch sizes. The supply system is a *bulk queue* (Chaudry and Templeton, 1983), which tends to be difficult to analyse. Moreover as we can see from Eq. (6), the batch sizes generated by our smoothing rule are not i.i.d.; rather, they are autocorrelated. Queueing models with such arrival patterns can be solved with *matrix analytic methods* (MAM's). These methods are popular as modeling tools because they can be used to construct and analyse a wide class of stochastic models. They are applied in several areas, of which the performance analysis of telecommunication systems is one of the most notable (Latouche and Ramaswami, 1999).

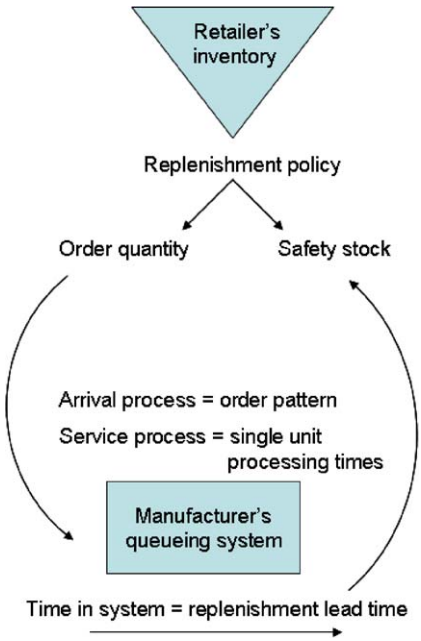


Fig. 3. Interaction between retailer's inventory system and manufacturer's production system.

4.2. Estimation of the manufacturer's production lead times

To estimate the lead time distribution we develop a discrete time queueing model. The arrival process consists of batch arrivals with a fixed interarrival time (1 period) and with autocorrelated batch sizes. The service times of a single item, denoted by M, are stochastic and i.i.d. according to a phase type (PH) distribution. The key idea behind PH distributions is to exploit the Markovian structure of the distribution to simplify the queueing analysis. Moreover, any general discrete distribution can be approximated in sufficient detail by means of a PH distribution (Horváth and Telek, 2002), since the class of discrete PH distributions is a versatile set that is dense within the set of all discrete distributions on the nonnegative integers (Neuts, 1989; Latouche and Ramaswami, 1999; Bobbio et al., 2003).

The computational complexity of our algorithm to compute the lead time distribution increases with the number of phases of the PH distributed service process. Therefore we want the service process to be PH-distributed with as few phases as possible. Since the lead time is expressed as an integer number of periods and the interarrival time is equal to one base period, we have the freedom to choose the time unit U of the queueing system in an appropriate manner (Bobbio et al., 2004b). When the time unit U is chosen as half of the mean service time of a single item, i.e., $U = E(M)/2$, Boute et al. (2004) are able to match the first two moments of the single unit service times, $E(M)$ and $\mathrm{Var}(M)$, by means of a PH distribution with only 2 phases. We provide this PH fitting procedure in Appendix A. The PH distribution is characterized by the pair (T, α), where T is a 2×2 substochastic matrix and α a 1×2 stochastic vector.

When we choose U to be the time unit of our queueing system, this implies that orders placed every period arrive at the queue at time epochs $0, d, 2d, \ldots$, where $d \times U = 1$. The order size at these time epochs evolves as a Markovian process with state space $\{x : 1 \leqslant x \leqslant m_D\}$, where m_D is the maximum demand, i.e., m_D is the smallest integer such that $\Pr[D > m_D] = 0$. Indeed, according to Eq. (6), the order size generated by our smoothing rule at time td is determined as

$$O_{td} = (1 - \beta) O_{(t-1)d} + \beta D, \tag{9}$$

where D is the customer demand random variable. Eq. (9) evolves as a Markovian process, since the probability of order quantity O_{td} can be determined given the value of the previous order quantity $O_{(t-1)d}$. Using induction on t we have $E(O_{td}) = E(D)$.

The order size that results from (9) can be a real number. However, because only an integer number of items can be produced, the actual batch size passed to the manufacturer at time t has size O^*_{td}:

$$O^*_{td} = \begin{cases} O_{td} & \text{if } O_{td} \in \mathbb{N}, \\ \lceil O_{td} \rceil & \text{with probability } O_{td} - \lfloor O_{td} \rfloor \text{ if } O_{td} \notin \mathbb{N}, \\ \lfloor O_{td} \rfloor & \text{with probability } \lceil O_{td} \rceil - O_{td} \text{ if } O_{td} \notin \mathbb{N}, \end{cases} \tag{10}$$

such that the batch size O^*_{td} is an integer number. Doing so, the expected value $E[O^*_{td}] = E[O_{td}] = E[D]$. Suppose for instance that the replenishment rule generates an order quantity of 5.8. Since 5.8 is not an integer, we round this to 5 units with a probability of 0.20 and to 6 units with a probability of 0.80. This (integer) number of units constitutes the batch size that has to be produced by the manufacturer.

In order to estimate the lead time distribution we start by defining the following additional random variables:

- t_n: the time of the nth observation point, which we define as the nth time epoch during which the server is busy,
- $a(n)$: the arrival time of the order in service at time t_n,
- B_n: the age of the order in service at time t_n, defined as the duration (expressed in the time unit of the queueing model, i.e., U) of the time interval $[a_n, t_n)$,
- C_n: the number of items part of the order in service that still need to start or complete service at time t_n,
- S_n: the service phase at time t_n.

All events, such as arrivals, transfers from the waiting line to the server and service completions are assumed to occur at instants immediately after the discrete time epochs. This implies that the age of an order in service at some time epoch t_n is at least 1.

Then, $(B_n, O_{a(n)}, C_n, S_n)$ forms a discrete time Markov process on the state space $\mathbb{N}_0 \times \{(x,c) : 1 \leqslant x \leqslant m_D, c \in \{1,2,\ldots,\lceil x \rceil\}\} \times \{1,2\}$, because B_n is a positive integer, $O_{a(n)}$ (the original order quantity of the order in service) is a real number between 1 and m_D, $C_n \leqslant \lceil O_{a(n)} \rceil$ and the PH service has two phases. We keep track of the original order quantity $O_{a(n)}$ instead of the rounded batch size $O^*_{a(n)}$, because it allows us to determine the order size of the next batch arrival precisely (see Eq. (9)). Since this original order quantity is a real number, the Markov process $(B_n, O_{a(n)}, C_n, S_n)$ has a continuous state space. Due to its continuous state space, it is very hard to find the steady state vector of this Markov process. Therefore, instead of keeping track of $O_{a(n)}$ in an exact manner, we will round it in a probabilistic way to the nearest multiple of $1/g$, where $g \geqslant 1$ is an integer termed the *granularity* of the system. Clearly, the larger g, the better the approximation. As a result, we obtain a Markov chain $(B_n, O^g_{a(n)}, C_n, S_n)$ on the discrete state space $\mathbb{N}_0 \times \{(x,c) : x \in \{1, 1+1/g, 1+2/g, \ldots, m_D\}, c \in \{1,2,\ldots,\lceil x \rceil\}\} \times \{1,2\}$. The quantity $O^g_{a(n)}$ evolves as follows. Let

$$O^c_{td} = (1-\beta) O^g_{(t-1)d} + \beta D.$$

Then the random variable O^g_{td} is determined as

$$O^g_{td} = \begin{cases} O^c_{td} & \text{if } O^c_{td} \in \mathbb{S}^g, \\ \lceil O^c_{td} \rceil_g & \text{with probability } \left(O^c_{td} - \lfloor O^c_{td} \rfloor_g\right) \cdot g \text{ if } O^c_{td} \notin \mathbb{S}^g, \\ \lfloor O^c_{td} \rfloor_g & \text{with probability } \left(\lceil O^c_{td} \rceil_g - O^c_{td}\right) \cdot g \text{ if } O^c_{td} \notin \mathbb{S}^g, \end{cases} \tag{11}$$

where $\mathbb{S}^g = \{1, 1+1/g, 1+2/g, \ldots, m_D\}$ and $\lceil x \rceil_g$ ($\lfloor x \rfloor_g$) is the smallest (largest) element in $\mathbb{S}^g$ that is larger (smaller) than x. Using induction on t, we have $E(O^g_{td}) = E(O_{td}) = E(D)$ for all g. As a result from Eq. (11), the conditional probabilities $\Pr[O^g_{td} = q' | O^g_{(t-1)d} = q]$ for $q, q' \in \mathbb{S}^g$, which we denote as $p_g(q,q')$, can be computed as

$$\begin{aligned} p_g(q,q') = \sum_{i \geqslant 1} \Pr[D=i] \Big\{ & 1_{\{q'-1/g < \bar\beta q + \beta i < q'\}} \left((\bar\beta q + \beta i) - \lfloor \bar\beta q + \beta i \rfloor_g \right) \cdot g + 1_{\{\bar\beta q + \beta i = q'\}} \\ & + 1_{\{q' < \bar\beta q + \beta i < q' + 1/g\}} \left(\lceil (\bar\beta q + \beta i) \rceil_g - (\bar\beta q + \beta i) \right) \cdot g \Big\}, \end{aligned} \tag{12}$$

where $\bar\beta = (1-\beta)$ and $1_{\{A\}}$ is 1 if the event A is true and 0 otherwise.

Let us have a look at the evolution of the Markov chain $(B_n, O^g_{a(n)}, C_n, S_n)$. At each transition step, there are three possibilities. First, the current item in production stays in service and the phase of the service process may change. Second, the current item in service finishes production, and a new item of the same batch enters production. Third, the current item in service finishes production and when this is the last item of the batch, the complete batch is produced. The order quantity of the new batch that is taken in production is given by $p_g(q,q')$ according to Eq. (12), and the batch size is equal to $\lceil q' \rceil$, $\lfloor q' \rfloor$ or q' according to Eq. (10), such that the batch size is an integer number of units.

Let $(P_g)_{(a,q,r,s),(a',q',r',s')}$ be the transition probabilities of the Markov chain $(B_n, O^g_{a(n)}, C_n, S_n)$. These probabilities are then given by

$$(P_g)_{(a,q,r,s),(a',q',r',s')} = \begin{cases} T_{s,s'} & a' = a+1,\ q' = q,\ r' = r, \\ t_s \alpha_{s'} & a' = a+1,\ q' = q,\ r' = r-1 \geqslant 1, \\ t_s \alpha_{s'} p_g(q,q') (\lceil q' \rceil - q') & a' = \max(a-d+1, 1),\ r' = \lfloor q' \rfloor,\ q' \notin \mathbb{N},\ r = 1, \\ t_s \alpha_{s'} p_g(q,q') (q' - \lfloor q' \rfloor) & a' = \max(a-d+1, 1),\ r' = \lceil q' \rceil,\ q' \notin \mathbb{N},\ r = 1, \\ t_s \alpha_{s'} p_g(q,q') & a' = \max(a-d+1, 1),\ r' = q' \in \mathbb{N},\ r = 1, \\ 0 & \text{else}, \end{cases} \tag{13}$$

with $t = (e - Te)$ denoting the probability that the current unit in service finishes production. As a consequence, we have the following form for the transition matrix P_g of $(B_n, O^g_{a(n)}, C_n, S_n)$:

$$P_g = \begin{bmatrix} A_d & A_0 & 0 & \dots & 0 & 0 & \dots \\ A_d & 0 & A_0 & \dots & 0 & 0 & \dots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \ddots \\ A_d & 0 & 0 & \dots & A_0 & 0 & \dots \\ 0 & A_d & 0 & \dots & 0 & A_0 & \ddots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \ddots & \ddots \end{bmatrix}, \tag{14}$$

where A_0 and A_d are square matrices of dimension m_{tot}, with m_{tot} the number of elements in the set $\{(x,c) : x \in \{1, 1+1/g, 1+2/g, \dots, m_D\},\ c \in \{1,2,\dots,\lceil x \rceil\}\} \times \{1,2\}$, i.e., $m_{\text{tot}} = g(m_D+1)m_D - 2(g-1)$. The matrices A_0 represent the probabilities that the service of the batch continues and are given by the first two equations of (13), while the matrices A_d represent the probabilities that the service of the batch finishes and are given by the 3rd, 4th and 5th equation of (13).

The MC characterized by Eq. (14) is of the GI/M/1 type (Neuts, 1981). From an operational point of view it is clear that the proposed queueing system is stable if and only if its utilization ρ is strictly smaller than one, or equivalently if the average production time of a batch order is strictly smaller than the average inter-arrival time of a batch order. Since we have chosen the time unit of our queueing model such that the average production time of a single unit is equal to 2, and the average batch order size is equal to the average demand $E(D)$, the average production time of a batch order is $2E(D)$. The inter-arrival time of an order is one (review) period, or, when we express it in the time unit of our queueing model, equal to d time units. Hence the stability condition can be rephrased as $2E(D) < d$. This condition is not restrictive, as a system with a load $\rho > 1$ leads to infinite lead times as the demand is greater than the production capacity.

For an ergodic MC of the GI/M/1 type, one computes the steady state vector π of P_g, that is, $\pi P_g = \pi$ and $\pi e = 1$, as follows:

$$\pi_1 = \pi_1 (I - R^d)(I - R)^{-1} A_d, \tag{15}$$

$$\pi_i = \pi_1 R^{i-1}, \tag{16}$$

where $\pi = (\pi_1, \pi_2, \dots)$ and π_i is a $1 \times m_{\text{tot}}$ vector, for all $i > 0$. The vector π_1 is normalized as $\pi_1 (I - R)^{-1} e = 1$ and the $m_{\text{tot}} \times m_{\text{tot}}$ rate matrix R is the smallest nonnegative solution to the matrix equation $R = A_0 + R^d A_d$ and can be numerically solved with a variety of algorithms, e.g., Neuts (1981), Ramaswami (1988), Alfa et al. (2002).

Having obtained the steady state vector $\pi = (\pi_1, \pi_2, \dots)$, we can obtain the response time using the following observation: The probability that an order has a response time of a time units can be calculated as the expected number of orders with an age of a time units that complete their service at an arbitrary time instant, divided by the expected number of orders that get completed during an arbitrary time instant (that is, $1/d$ for a queue with $\rho < 1$). As such, denoting T_{r} as the response time (expressed in the time unit U) we have

$$\Pr[T_{\text{r}} = a] = d\rho \sum_{q,s} (\pi_a)_{(q,1,s)} (t)_s, \tag{17}$$

where $(\pi_a)_{(q,r,s)}$ represents the steady state probability of being in state (a,q,r,s). Notice, to make sure that an order completes its service, the number of remaining customers requiring service cannot be more than one.

We chose the time unit U of our queueing system as half of the mean production time of a single item (i.e., $E(M)/2$). Thus, if we want to express the lead time in terms of the number of periods needed to deliver the order to the retailer, we still need to make the following conversion:

$$\Pr[T_{\text{p}} = i] = \sum_j \Pr[T_{\text{r}} = j] \cdot 1_{\{\lceil j/d \rceil = i\}}, \tag{18}$$

where $1_{\{A\}}$ is 1 if the event A is true and 0 otherwise. Note that this conversion at the same time rounds the (possibly fractional) response time T_r to the discrete replenishment lead time, T_p, expressed in an integer number of periods.

5. Impact of the smoothing replenishment rule on safety stock

When demand is probabilistic, there is a definite chance of not being able to satisfy some of the demand directly out of stock. Therefore, a buffer or safety stock is required to meet unexpected fluctuations in demand. The goal is to reduce inventory without diminishing the level of service provided to customers. When the retailer faces (and satisfies) a variable customer demand, but replenishes through a smooth order pattern, this comes at the cost of an increase in its inventory variability. As a consequence, in order to provide the same service level a larger safety stock will be needed than the traditional standard base-stock replenishment policy, where orders have the same variability as customer demand (Dejonckheere et al., 2002; Disney and Towill, 2003; Disney et al., in press).

However, since we include the impact of the order decision on production, a smooth order pattern generates shorter and less variable lead times. This introduces a compensating effect on the required safety stock. The aim is to find values for the smoothing parameter $0 < \beta < 1$ where the decrease in lead times compensates the increase in inventory variance. In that case we can smooth production without having to increase inventory levels to provide the same customer service.

Similar to Graves (1999) and Disney et al. (in press), we characterize the inventory random variable and use it to find the safety stock requirements for the system. In Graves (1999) and Disney et al. (in press), the inventory control system operates with fixed lead times. Hence it is known exactly when a replenishment order is received in inventory. In this environment the retailer's inventory is replenished every period with the order that was placed $t_p + 1$ periods ago (with t_p the fixed lead time).

In our model however, the inventory is controlled by stochastic lead times. As a consequence the inventory is not necessarily replenished every period and we do not know when exactly a replenishment occurs. Moreover, the queueing analysis implies that it takes a longer time to produce (and consequently replenish) a larger order quantity. Hence the order quantity and its replenishment lead time are undoubtedly correlated, affecting the inventory distribution. Therefore the analysis is more involved. We refer to Song and Zipkin (1996), Song et al. (1999), Song and Yao (2002) and Lu et al. (2003) where the same problem is encountered.

In this section we first describe how to determine the optimal base-stock level S that is required to meet a target service level and then we explain how to find the corresponding safety stock SS.

5.1. Determination of optimal base-stock level

We study the fill rate, which is a popular metric of customer service. It measures the proportion of the demand that can immediately be delivered from the inventory on hand (Zipkin, 2000):

$$\text{Fill rate} = 1 - \frac{\text{expected number of backorders}}{\text{expected demand}}.$$

To calculate the fill rate, we monitor the inventory on hand after customer demand is observed and we retain the number of shortages when a stockout occurs. To do so, we observe the system at the end of every period t, after customer demand D_t is satisfied and after replenishment order O_t has been placed with the manufacturer (and before a possible order delivery occurs at the retailer in period $t + 1$), and we characterize the inventory random variable. At that time there may be $k \geqslant 0$ orders waiting in the production queue and there is always 1 order in service (since the observation moment is immediately after an order placement) which is placed k periods ago (O_{t-k}). Although k is a function of t, we write k as opposed to $k(t)$ to simplify the notation.

The inventory on hand or net stock NS_t is equal to the initial inventory on hand plus all replenishment orders received in inventory minus total observed customer demand. At the time instant we observe the system, the order O_{t-k} is currently in service. Hence, the orders placed more than k periods ago, i.e., O_{t-i}, $i \geqslant k + 1$, are already delivered in inventory, while customer demand is satisfied up to the current period t.

For convenience, we define the demand D_t, and hence the order quantity O_t, to be zero for $t \leqslant 0$, in which case the initial inventory level is equal to the base-stock level S. Hence we can write:

$$NS_t = S + \sum_{i=k+1}^{t} O_{t-i} - \sum_{i=0}^{t} D_{t-i}. \tag{19}$$

After backward substitution of Eq. (6) or directly using Eq. (7) we find that the order quantity is a convex combination of previous demand realisations (Balakrishnan et al., 2004)

$$O_{t-i} = \sum_{j=0}^{t-i} \beta(1-\beta)^j D_{t-i-j}. \tag{20}$$

Substituting (20) into (19) results in an expression for the net stock in function of customer demand only

$$\begin{aligned} NS_t &= S + \sum_{i=k+1}^{t} \sum_{j=0}^{t-i} \beta(1-\beta)^j D_{t-i-j} - \sum_{i=0}^{t} D_{t-i} = S - \sum_{i=0}^{k} D_{t-i} + \sum_{i=k+1}^{t} \left(\sum_{j=0}^{t-i} \beta(1-\beta)^j D_{t-i-j} - D_{t-i} \right) \\ &= S - \sum_{i=0}^{k} D_{t-i} - \sum_{i=k+1}^{t} \left(1 - \sum_{j=0}^{i-(k+1)} \beta(1-\beta)^j \right) D_{t-i} = S - \sum_{i=0}^{k} D_{t-i} - \sum_{i=k+1}^{t} (1-\beta)^{i-k} D_{t-i}. \end{aligned} \tag{21}$$

We need to determine the steady state distribution NS of the net stock random variable NS_t characterized by (21). As S is a constant (for a given β), we focus on the steady state distribution Z of $Z_t = S - NS_t$:

$$Z_t = \sum_{i=0}^{k} D_{t-i} + \sum_{i=k+1}^{t} (1-\beta)^{i-k} D_{t-i}. \tag{22}$$

Some care must be taken when evaluating (22) as the value of D_{t-k} influences the age k of the order in service: the larger the demand size, the larger the order size and consequently the longer it takes to produce the order. Moreover, since the order quantity is also affected by previously realised customer demand (see (20)), the demand terms D_{t-i}, $i = k+1, \ldots, t$ also influence the order's age k. Since k determines the number of demand terms in the summation terms in Eq. (22), there is correlation between the different terms that make up Z_t.

At first sight the correlation between the different terms of Z_t seems to necessitate some kind of approximation. However, the Markov chain $(B_n, O^g_{a(n)}, C_n, S_n)$ used to determine the lead time distribution, retains the age k of the order in service (here denoted by B_n) and the order quantity O_{t-k} (rounded to the nearest multiple of $1/g$ with g the granularity, here denoted by $O^g_{a(n)}$). According to (20) the order quantity O_{t-k} is equal to

$$O_{t-k} = \sum_{i=k}^{t} \beta(1-\beta)^{i-k} D_{t-i}. \tag{23}$$

Dividing (23) by β and substituting into (22) gives

$$Z_t = \sum_{i=0}^{k-1} D_{t-i} + \frac{O_{t-k}}{\beta}, \tag{24}$$

where both the order quantity O_{t-k} and the age k of the order in service[1] are measures that can be obtained from the Markov chain $(B_n, O^g_{a(n)}, C_n, S_n)$. Denote by $B^{(b)}$ and $O^{(b)}$ the steady state random variables of the age of the order in service and the original size of the order in service, provided that the server is busy, respectively. The joint distribution $(B^{(b)}, O^{(b)})$ is then available via the steady state vector π by summing the appropriate terms.

We are now able to compute the steady state probabilities Z of Z_t. We distinguish between two cases: At the moment of observation, the replenishment order O_t finds $k > 0$ orders pending at the manufacturer's queue, or it finds the queue empty, $k = 0$. Let the random variable F_t equal 0 if the order O_t finds the (manufacturer's)

[1] Recall that the time unit U of our queueing analysis is equal to $1/d$ periods. Hence k periods corresponds to kd time units.

queue empty and there was no service completion at the order placement time, and 1 otherwise. First, consider the case where a new order finds $k > 0$ orders pending at the manufacturer:

$$\Pr[Z = s, F = 1] = \lim_{t\to\infty} \Pr[Z_t = s, F_t = 1] = \sum_{k>0} \Pr[B^{(b)} = dk, O^{(b)} = q]\rho d \cdot \Pr[D^{k*} = s - q/\beta]. \tag{25}$$

Indeed, $\Pr[B^{(b)} = dk, O^{(b)} = q]\rho d$ is the joint probability that a new order finds k orders at the manufacturer with the original size of the one in service equaling q (the factor ρ drops the busy condition, while dividing by $1/d$ conditions the probability on an arrival event). Now, $\Pr[D^{k*} = s - q/\beta]$ gives the probability that the total demand that was taken from the inventory during the last k periods equals $s - q/\beta$ due to Eq. (24), where D^{k*} denotes the k-fold convolution of the demand D.

Second, we focus on the case where a new order finds the queue idle ($F_t = 0$):

$$\Pr[Z = s, F = 0] = \lim_{t\to\infty} \Pr[Z_t = s, F_t = 0] = \sum_{a=1}^{d-1} \sum_{q\in\mathbb{S}^g} \Pr[T_r = a, O^g = q] \cdot p_g(q, s\beta). \tag{26}$$

Recall, $\Pr[T_r = a]$ is the probability that an arbitrary order has a response time of a time units. An arbitrary tagged order will find the queue empty upon arrival if the previous order (which is just as arbitrary) has a response time of less than d time units, since d is the interarrival time of the orders. The joint probability $\Pr[T_r = a, O^g = q] = \Pr[O^g = q | T_r = a] \cdot \Pr[T_r = a]$ gives the probability that the order preceding the tagged one has a size q and its lead time equals a time units. Multiplying this with $p_g(q, s\beta)$ thus gives us the probability that the tagged order finds the queue empty upon arrival and has size $s\beta$. Recall that $p_g(q, s\beta)$ represents the conditional probability that the new order quantity equals $s\beta$, given the previous order quantity q (see Eq. (12)). The probabilities $\Pr[T_r = a, O^g = q]$ can be retrieved from the steady state vector π as

$$\Pr[T_r = a, O^g = q] = \sum_s (\pi_a)_{(q,1,s)}(t_s). \tag{27}$$

The steady state probabilities of the net stock $\Pr[NS = k] = \lim_{t\to\infty} \Pr[NS_t = k]$ can then be computed from Eqs. (25) and (26):

$$\Pr[NS = k] = \Pr[Z = S - k, F = 0] + \Pr[Z = S - k, F = 1], \tag{28}$$

for $k \leqslant S - 1$ (as the minimum size of an order is at least one). The probability of a stock-out is given by

$$\Pr[NS < 0] = \Pr[Z > S], \tag{29}$$

and the average number of shortages when a stock-out occurs is given by

$$E(NS^-) = E\big([Z - S]^+\big), \tag{30}$$

where $x^+ := \max\{0, x\}$. Finally, the fill rate can then be calculated as

$$\text{Fill rate} = 1 - \frac{E\big([Z - S]^+\big)}{E(D)}. \tag{31}$$

In practice, decision makers often have to find the minimal base-stock level that is required to achieve a given fill rate. From (31) we can compute the minimal base-stock level S that is required such that an imposed fill rate is met. In the next subsection we will describe how to find the corresponding safety stock SS given this base-stock level S.

5.2. Determination of corresponding safety stock

To determine the base-stock level, we observed the system after customer demand is satisfied and after a replenishment order is placed. However, in order to find the safety stock requirements, we characterize the inventory random variable at time instants just before a replenishment occurs. Indeed, by definition the safety stock is the average level of the net stock just before a replenishment arrives (Silver et al., 1998).

Suppose that the n'th replenishment order (i.e., the order placed at the end of period n) is to be delivered in period $n + t_p + 1$. We monitor the inventory on hand just before this replenishment occurs and denote this

time instant by $n*$. At the time instant just before replenishment n occurs, the orders placed before period n are delivered in inventory and due to the sequence of events (satisfy demand *after* receipt of the orders) customer demand is satisfied up to period $n + t_p$. Then, the net stock or inventory on hand NS_{n*} is equal to

$$NS_{n*} = S + \sum_{k=1}^{n-1} O_k - \sum_{k=1}^{n+t_p} D_k. \tag{32}$$

Similar to (20) we can write the order quantity as a convex combination of previous demand realisations:

$$O_k = \sum_{j=1}^{k} \beta(1-\beta)^{k-j} D_j, \tag{33}$$

and substituting (33) into (32) results in a result similar to Eq. (21):

$$NS_{n*} = S - \sum_{k=n}^{n+t_p} D_k - \sum_{k=1}^{n-1} (1-\beta)^{n-k} D_k. \tag{34}$$

Note that Eqs. (21) and (34) are different, however. In Eq. (34) the age of the order in service (t_p periods) does represent the replenishment lead time, since the order will be delivered immediately after the observed time instant. The value of t_p is consequently a realisation of the lead time variable T_p. In Eq. (21) the age of the order in service (k periods) does not necessarily represent the lead time, since it is possible that the order in service does not finish production within the next period, and hence the order in service will not replenish the inventory in the next period.

Eq. (34) monitors the net stock at time instants just before a replenishment occurs. By definition the expected value of this expression represents the safety stock:

$$SS \equiv E(NS_{n*}) = S - E\left(\sum_{k=n}^{n+t_p} D_k\right) - E\left(\sum_{k=1}^{n-1} (1-\beta)^{n-k} D_k\right). \tag{35}$$

Since t_p is a realisation of the lead time distribution T_p, we obtain an elegant result for the safety stock that can be used to determine the safety stock SS, given a base-stock level S:

$$SS = S - \left[E(T_p) + 1\right] \cdot E(D) - \frac{(1-\beta)}{\beta} \cdot E(D). \tag{36}$$

6. Numerical experiments

To illustrate our findings, we set up a numerical experiment where the retailer daily observes a random discrete customer demand between 1 and 20 units with an average $E(D) = 10.5$. We assume three different demand patterns, each with a different variability: a symmetric bell shaped demand distribution with variability $\mathrm{Var}(D) = 11.99$ (case 1), a uniform demand pattern with variability $\mathrm{Var}(D) = 33.25$ (case 2), and finally a symmetric U-shaped demand pattern with variability $\mathrm{Var}(D) = 56.59$ (case 3). We acknowledge that this last demand pattern is somewhat artificial and rarely observed in reality, but it provides a good illustration of a wildly fluctuating customer demand. The (discrete) probability distributions of the three demand patterns are shown in Fig. 4.

The retailer satisfies this daily customer demand from his inventory on hand and replenishes according to the smoothing rule discussed in Section 3.2. We assume that the manufacturer's production operates 10 hours per day and the service time of a single unit is PH distributed with an average $E(M) = 48$ minutes and a coefficient of variation $\mathrm{cv}(M) = 1$. Hence the average production load is $(10.5 \times 48)/(60 \times 10) = 0.84$.

The retailer has to determine the parameter β to control his inventory. When he sets $\beta = 1$, the retailer places orders equal to demand and hence the customer demand variability is fully transmitted to the manufacturer. With the procedure described in Section 4.2, we can compute the lead times that result from the manufacturer's production system corresponding to this order decision. We provide the first two moments of this lead time distribution in Table 1.

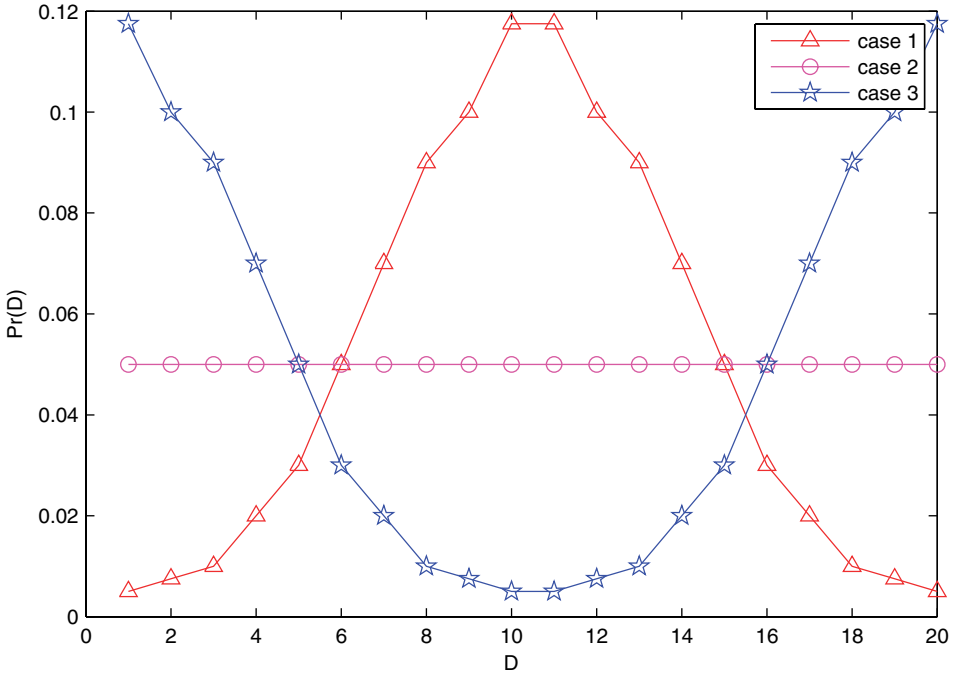


Fig. 4. Three customer demand patterns with different demand variability.

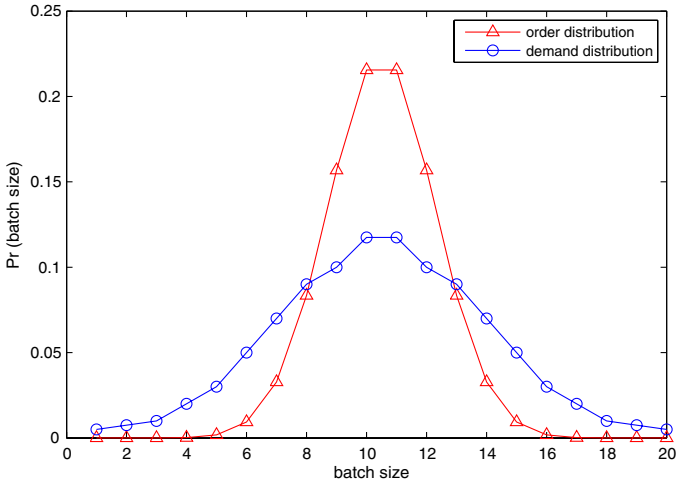
Table 1
Comparison of no order smoothing and order smoothing with exogenous and endogenous lead times

	β	Var(O)	$E(T_p)$	Var(T_p)	SS Exogenous LT	SS Endogenous LT
Case 1	1	11.99	0.6365	0.4823	7.7195	19.9971
	0.4	2.9975	0.6365	0.4823	8.4890	
	0.4	2.9975	0.5336	0.4163		19.7157
Case 2	1	33.25	1.0233	1.1255	17.2655	40.5134
	0.4	8.3125	1.0233	1.1255	18.5879	
	0.4	8.3125	0.7814	0.9044		40.0613
Case 3	1	56.69	1.4626	2.2351	25.7532	65.6799
	0.4	14.1475	1.4626	2.2351	28.1991	
	0.4	14.1475	1.0886	1.8114		64.6523

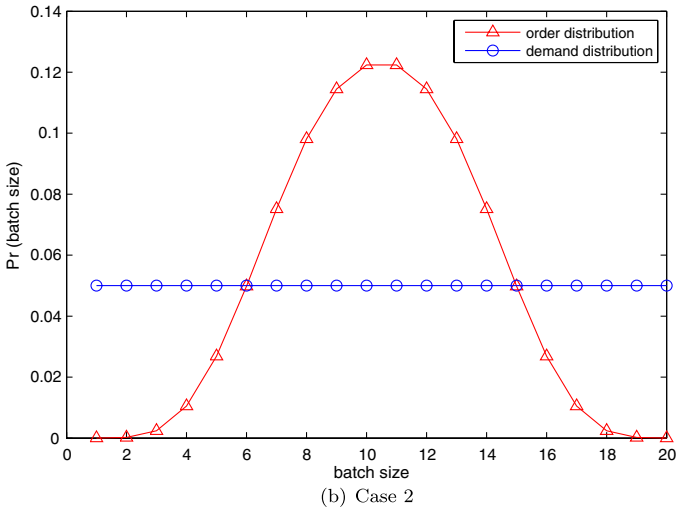
Suppose first that we treat the lead time to be an *exogenous* variable with the probability distribution that results from the queueing analysis corresponding to this order pattern, i.e., we assume lead times to be stochastic, but we arbitrarily assign a lead time to an order quantity. This means that we do not take the correlation between the order quantity and its production (replenishment) lead time into account, or equivalently, we ignore the correlation between O_{t-k} and k in Eq. (24). The safety stock that is required to maintain a 98% fill rate can then be calculated directly via Eq. (21) and is provided in Table 1 (second last column).

When the retailer chooses to smooth his orders with a parameter $\beta = 0.4$, the upstream demand variability is considerably dampened. In Fig. 5(a)–(c) we plot the order pattern resulting from this smoothing decision together with the observed customer demand pattern. Recall that when $\beta \neq 1$, the order pattern is correlated, while customer demand is i.i.d.

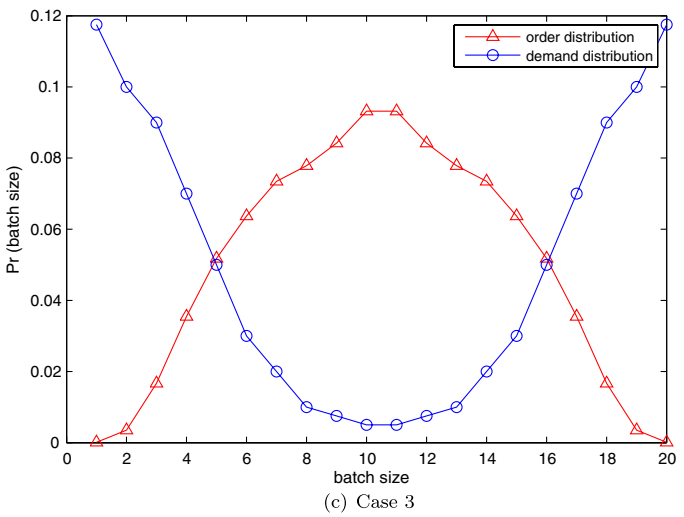
This smoothing decision leads to an increase in inventory variance, since inventory absorbs the variability in demand while the replenishments are relatively steady. When we would *not* consider the impact of this dampened order variability on lead time reduction, a higher safety stock has to be kept in order to maintain the same fill rate. As an example, we take the same *exogenous* lead time distribution as in the case where $\beta = 1$. In this



(a) Case 1



(b) Case 2



(c) Case 3

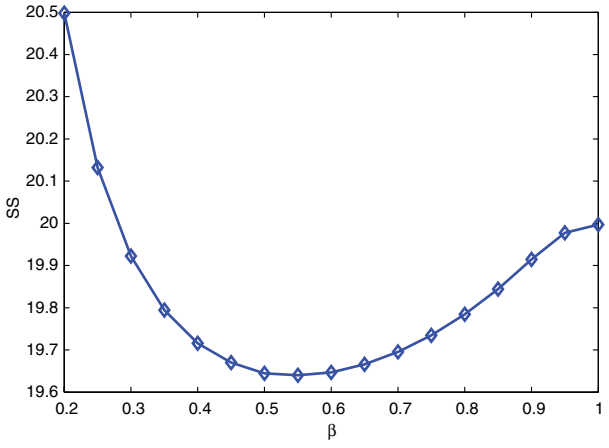
Fig. 5. Customer demand patterns and corresponding order patterns when $\beta = 0.4$.

case, smoothing with $\beta = 0.4$ indeed leads to an increased safety stock (see Table 1, second last column): the manufacturer can smooth his production, but at the expense of an increase in the retailer's inventory.

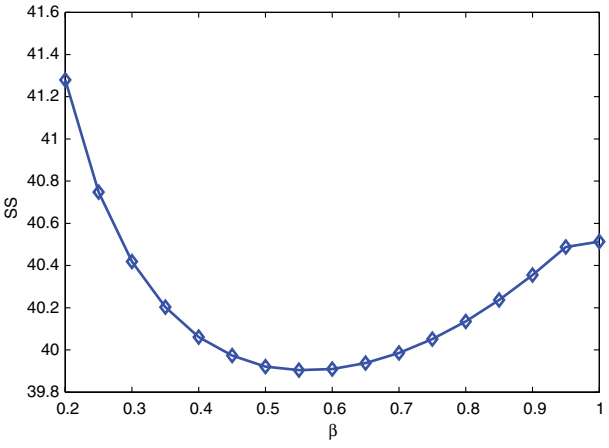
Working with exogenous lead times is, however, incomplete. First of all, we may not simply ignore the correlation between the order quantity and its lead time. It undoubtedly takes a longer time to produce an order with a large batch size. Including the correlation between O_{t-k} and k in Eq. (24) is indispensable for a correct representation of the model. When we compute the safety stock required to meet a 98% fill rate with *endogenous* lead times (with the analysis described throughout Section 5) for $\beta = 1$, we obtain different results (see Table 1, last column). Using exogenous lead times seriously underestimates the required safety stock; consequently, customer service will dramatically degrade.

Second, when the retailer smoothes his orders, he sends a less variable arrival pattern to the manufacturer's queue (see Fig. 5(a)–(c)), which inevitably results in different lead times. We, therefore, have to include the impact of this decreased order variability on production (queueing). Indeed, when we estimate the lead time distribution if the retailer sends a smooth order pattern with $\beta = 0.4$ to the manufacturer's production (we used a granularity equal to 8), we observe that order smoothing leads to lower and less variable lead times (see Table 1). This introduces a compensating effect on the safety stock SS. In the last column of Table 1 we observe that the safety stock for $\beta = 0.4$ (computed with *endogenous* lead times) is even slightly lower than when we do not smooth the orders ($\beta = 1$).

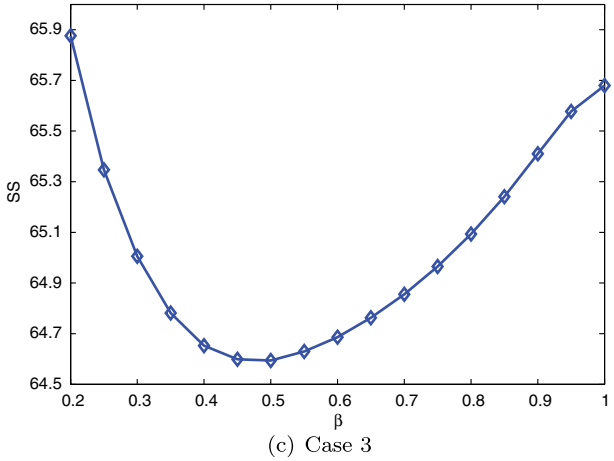
In Fig. 6 we show the effect of order smoothing on the retailer's safety stock for a smoothing parameter $\beta = 0.2$ to $\beta = 1$. When we include the effect of order smoothing on production, the safety stock is a U-shaped



(a) Case 1



(b) Case 2



(c) Case 3

Fig. 6. Safety stocks required to ensure a 98% fill rate for three demand patterns with different variability.

function of the smoothing intensity. We can smooth the replenishment orders to some extent while decreasing the safety stock. However, as of a certain point (around $\beta = 0.5$) the safety stock increases exponentially. When β approaches zero, the lead time reduction cannot compensate the increase in inventory variability anymore and the safety stock exceeds the safety stock that is required when the orders are not smoothed ($\beta = 1$).

These results imply that the retailer can dampen the upstream demand variability without having to increase his safety stock to maintain customer service at the same target level. Moreover, the retailer can even decrease his safety stock when he smoothes his orders. This is clearly a better situation for both the retailer and the manufacturer. The manufacturer receives a less variable order pattern and the retailer can decrease his safety stock while maintaining the same fill rate, so that the cooperative surplus is realised. This Pareto-improving policy may require contractual arrangements between the supply chain partners so that the lead time reduction is effectively implemented (Tsay et al., 1999).

7. Conclusions

Disney and Towill (2003) question "to what extent can production rates be smoothed in order to minimise production adaptation costs without adversely increasing inventory costs". This is an important trade-off because, if a perfectly level production rate is used, then large inventory deviations are found and hence large inventory costs are incurred. Conversely, if inventory deviations are minimised (by "passing on orders"), highly variable production schedules are generated and hence production adaptation costs are incurred. We have shown that by including the impact of the order decision on production, we can turn this conflicting situation into a situation where both parties are better off. A smooth order pattern gives rise to shorter and less variable (production/replenishment) lead times. This introduces a compensating effect on the retailer's inventory level. In this paper we showed that we can smooth the order pattern to a considerable extent without increasing stock levels. This may motivate the retailer to generate a smooth ordering pattern, resulting in a better situation for both supply chain echelons.

Acknowledgements

The authors thank the reviewers for their insightful comments and suggestions that have led to improvements of the paper. This research contribution is supported by contract grant G.0051.03 from the Research Programme of the Fund for Scientific Research – Flanders (Belgium) (F.W.O.-Vlaanderen). Benny Van Houdt is a postdoctoral Fellow of F.W.O.-Vlaanderen.

Appendix A. Matching the first two moments of an arbitrary distribution to a 2 phase discrete PH distribution

In this section we provide the procedure to fit the first two moments of the single unit service time, $E(M)$ and $\mathrm{Var}(M)$, to a 2 phase PH distribution, as developed by Boute et al. (2004). We assume that $E(M) \geqslant 2$ is an integer; this condition is not necessary, but allows some simplification in the fitting procedure. By definition, a discrete PH distribution is the distribution of the number of steps prior to final absorption in an absorbing Markov chain (Nelson, 1995). A PH distribution X is characterized by the triple (n, T, α), where $n > 0$ is an integer, referred to as the number of phases of the distribution or the number of transient states in an absorbing Markov chain, T is an $n \times n$ substochastic matrix, delineating the transition probabilities between the transient states and α is a stochastic $1 \times n$ vector, which defines the probabilities α_i that the process is started in the transient state i. The transition probabilities between the transient states and the absorbing state are given by t, which is an $n \times 1$ substochastic vector equal to $(e - Te)$, where e is a $n \times 1$ column vector with all its entries equal to one. Hence the probability that k steps are taken prior to absorption is given by

$$\Pr[X = k] = \alpha T^{k-1} t, \tag{37}$$

where $k > 0$. Its mean and variance obey the following equations:

$$E(X) = \alpha(I - T)^{-1}e, \tag{38}$$

$$\mathrm{Var}(X) = \alpha(I - T)^{-1}(2(I - T)^{-1} + (1 - E(X))I)e, \tag{39}$$

with I an $n \times n$ identity matrix. In order to match the mean single unit service time $E(M)$ and its variance $\mathrm{Var}(M)$, we need to find a PH distribution characterized by a triple (n, T, α) such that $E(M) = E(X)$ and $\mathrm{Var}(M) = \mathrm{Var}(X)$. Moreover, since the algorithm used to compute the lead time distribution speeds up with a smaller n, we want a representation (n, T, α) that fits the two first moments with n as small as possible (including higher moments will lead to a higher number of phases).

Denote $\mathrm{cv}^2(M)$ as the squared coefficient of variation, that is, $\mathrm{cv}^2(M) = \mathrm{Var}(M)/[E(M)]^2$. By applying a theorem by Telek (2000, Theorem 1), and the fact that $E(M)$ is an integer, we find that the minimum number of phases needed to match $E(M)$ and $\mathrm{Var}(M)$ equals

$$n = \max\left(2, \left\lceil \frac{E(M)}{E(M) \cdot \mathrm{cv}^2(M) + 1} \right\rceil\right). \tag{40}$$

Since the lead time is expressed as an integer number of periods and the interarrival time is equal to one base period, we have the freedom to choose the time unit U of the queueing system in an appropriate manner (Bobbio et al., 2004b). When we choose the time unit of our queueing system U equal to half of the mean single unit production time, i.e., $U = E(M)/2$, and denote $E(M_U)$ and $\mathrm{Var}(M_U)$ as the mean and variance of the production time expressed as multiples of U, then by definition we find $E(M_U) = 2$ and $\mathrm{Var}(M_U) = 4\mathrm{Var}(M)/[E(M)]^2$, implying that $\mathrm{cv}^2(M_U) = \mathrm{cv}^2(M)$. Consequently, we only need $n = 2$ phases, because

$$n = \max\left(2, \left\lceil \frac{2}{1 + 2\mathrm{cv}^2(M)} \right\rceil\right) = 2. \tag{41}$$

Meaning, we can always match the two first moments of the service process of a single item using a 2 state PH distribution.

Next, we choose the 1×2 vector α and the 2×2 matrix T as follows. These choices are motivated by a variety of results when matching continuous time PH distributions (Telek and Heindl, 2002; Bobbio et al., 2004a):

$$\alpha = (\delta, 1 - \delta), \tag{42}$$

$$T = \begin{bmatrix} 1 - p_1 & p_1 \\ 0 & 1 - p_2 \end{bmatrix}. \tag{43}$$

This leaves us with 3 parameters: δ, p_1 and p_2, and two equations: $E(M_U) = E(X)$ and $\mathrm{Var}(M_U) = \mathrm{Var}(X)$. Therefore, we add an additional constraint which demands that the stationary vector of the matrix $(T + t\alpha)$ is the uniform vector (1/2, 1/2). This constraint implies that the probability that the PH Markov process is in phase i is equal for each phase $i = 1, 2$, and hence equal to 1/2. As a consequence the average sojourn time in phase i equals $E(M_U)/2$.

Due to (42) and (43) the PH Markov process starts in phase 1 with probability δ and every time slot it may depart from phase 1 with probability p_1. Hence the time in phase 1 is geometrically distributed with an average of δ/p_1. The PH process subsequently passes to phase 2 and departs from phase 2 with probability p_2. Hence the average time in phase $i = 2$ equals $1/p_2$. Setting $E(M_U) = E(X)$ poses the following conditions on δ, p_1 and p_2:

$$\begin{aligned} p_1 &= 2\delta/E(M_U), \\ p_2 &= 2/E(M_U). \end{aligned} \tag{44}$$

Since $E(M_U) = 2$, we find that $p_2 = 1$ and $0 \leqslant p_1 = \delta$.

Thus, it remains to determine δ, with $0 \leqslant \delta \leqslant 1$, based on the remaining condition $\mathrm{Var}(M_U) = \mathrm{Var}(X)$. Therefore, we define the generating function $G(z) = \sum_k P[X = k]z^k$ of the PH distribution X characterized by (42) and (43). Recall that a discrete PH distribution denotes the distribution of the number of steps prior to final absorption in an absorbing Markov chain. Due to the special structure of (42) and (43), the process

always passes through phase 2 and runs through phase 1 with probability δ. The time that the process spends in phase $i = 1, 2$ is geometrically distributed and independent of the time that the process spends in phase j, $j \neq i$. Since the generating function of a geometric distribution is a well known result and equal to

$$G_{\mathrm{GEO}}(z) = \frac{pz}{1-(1-p)z},$$

we find that the generating function $G(z) = \sum_k P[X = k]z^k$ of the PH distribution X is then given by

$$G_{(z)} = \left\{\delta\left(\frac{p_1 z}{1-(1-p_1)z}\right) + (1-\delta)\right\}\left(\frac{p_2 z}{1-(1-p_2)z}\right). \tag{45}$$

The condition $\mathrm{Var}(M_U) = \mathrm{Var}(X)$ can then be rephrased as

$$\mathrm{Var}(M_U) = \left.\frac{\mathrm{d}^2 G(z)}{\mathrm{d}z^2}\right|_{z=1} + E(M_U)(1 - E(M_U)). \tag{46}$$

Some careful calculations show that this equation is solved by setting δ equal to

$$0 \leqslant \delta = \frac{2E(M_U)}{2E(M_U) + 2(2 - E(M_U) + 2\mathrm{cv}^2(M_U)E(M_U))} \leqslant 1, \tag{47}$$

or

$$0 \leqslant \delta = \frac{1}{1 + 2\mathrm{cv}_M^2} \leqslant 1, \tag{48}$$

where the first and last inequality in Eqs. (47) and (48) is due to Eq. (40).

In conclusion, the PH distribution fitted to the mean service time of a single item $E(M)$ and its variance $\mathrm{Var}(M)$ is characterized by

$$n = 2, \tag{49}$$

$$\alpha = \left(\frac{1}{1 + 2\mathrm{cv}_M^2}, 1 - \frac{1}{1 + 2\mathrm{cv}_M^2}\right), \tag{50}$$

$$T = \begin{bmatrix} 1 - \frac{1}{1+2\mathrm{cv}_M^2} & \frac{1}{1+2\mathrm{cv}_M^2} \\ 0 & 0 \end{bmatrix}. \tag{51}$$

References

Alfa, A., Sengupta, B., Takine, T., Xue, J., 2002. A new algorithm for computing the rate matrix of GI/M/1 type Markov chains. In: Proceedings of the 4th International Conference on Matrix Analytic Methods, Adelaide, Australia, pp. 1–16.

Allen, D.S., 1997. Do inventories moderate fluctuations in output? Review, Federal Reserve Bank of St. Louis, July/August, pp. 39–50.

Axsäter, S., 1976. Coordinating control of production-inventory systems. International Journal of Production Research 14 (6), 669–688.

Balakrishnan, A., Geunes, J., Pangburn, M., 2004. Coordinating supply chains by controlling upstream variability propagation. Manufacturing & Service Operations Management 6 (2), 163–183.

Bertrand, J.W.M., 1986. Balancing production level variations and inventory variations in complex production systems. International Journal of Production Research 24 (5), 1059–1074.

Blanchard, O., 1983. The production and inventory behavior of the American automobile industry. Journal of Political Economy 91, 365–400.

Blinder, A., 1986. Can the production smoothing model of inventory behavior be saved?. Quarterly Journal of Economics 101 431–453.

Bobbio, A., Horváth, A., Scarpa, M., Telek, M., 2003. Acyclic discrete phase type distributions: Properties and a parameter estimation algorithm. Performance Evaluation 54 (1), 1–32.

Bobbio, A., Horváth, A., Telek, M., 2004a. 3-Moments matching with minimal, positive acyclic phase-type distributions. Research Report. Department of Telecommunication, Technical University of Budapest.

Bobbio, A., Horváth, A., Telek, M., 2004b. The scale factor: A new degree of freedom in phase type approximation. Performance Evaluation 56 (1–4), 121–144.

Boute, R.N., Lambrecht, M.R., Van Houdt, B., 2004. Periodic review base-stock replenishment policy with endogenous lead times. Research Report 0448. Department of Applied Economics, Katholieke Universiteit Leuven, Belgium.

Buffa, E.S., Miller, J.G., 1979. Production-Inventory Systems: Planning and Control, third ed. Irwin, Homewood Illinois.

Chaudry, M., Templeton, J., 1983. A first course in bulk queues. Wiley, New York.
Chen, F., Drezner, Z., Ryan, J., Simchi-Levi, D., 2000a. Quantifying the bullwhip effect in a simple supply chain: The impact of forecasting, lead times, and information. Management Science 46 (3), 436–443.
Dejonckheere, J., Disney, S.M., Farasyn, I., Janssen, F., Lambrecht, M.R., Towill, D.R., Van de Velde, W., 2002. Production and inventory control: The variability trade-off. In: Proceedings of the 9th EUROMA Conference, Copenhagen, Denmark.
Dejonckheere, J., Disney, S.M., Lambrecht, M.R., Towill, D.R., 2003. Measuring and avoiding the bullwhip effect: A control theoretic approach. European Journal of Operational Research 147, 567–590.
Deziel, D., Eilon, S., 1967. A linear production-inventory control rule. The Production Engineer 43, 93–104.
Disney, S.M., Towill, D.R., 2003. On the bullwhip and inventory variance produced by an ordering policy. Omega 31, 157–167.
Disney, S.M., Farasyn, I., Lambrecht, M.R., Towill, D.R., Van de Velde, W., in press. Taming bullwhip whilst watching customer service in a single supply chain echelon. European Journal of Operational Research, doi:10.1016/j.ejor.2005.01.026.
Fair, R.C., 1989. The production smoothing model is alive and well. Cowles Foundation Discussion Papers 896. Cowles Foundation, Yale University.
Forrester, J., 1961. Industrial Dynamics. MIT Press, Cambridge, MA.
Graves, S.C., 1999. A single-item inventory model for a nonstationary demand process. Manufacturing & Service Operations Management 1 (1), 50–61.
Hopp, W.J., Spearman, M.L., 2001. Factory Physics, second ed. Irwin, McGraw-Hill.
Horváth, A., Telek, M., 2002. PhFit: A general phase-type fitting tool. In: Proceedings of Performance TOOLS 2002, London, UK.
Hosoda, T., Disney, S.M., 2005. On variance amplification in a three-echelon supply chain with minimum mean square error forecasting. Omega 34 (4), 344–358.
Karmarkar, U.S., 1993. Manufacturing lead times, order release and capacity loading. In: Graves, S.C., Rinnooy Kan, A.H.G., Zipkin, P.H. (Eds.), Logistics of Production and Inventory, Hand-books in Operations Research and Management Science, vol. 4. Elsevier Science Publishers BV, pp. 287–329.
Kim, J.G., Chatfield, D.C., Harrison, T.P., Hayya, J.C., in press. Quantifying the bullwhip in a supply chain with stochastic lead time. European Journal of operational research, doi:10.1016/j.ejor.2005.01.043.
Krane, S., Braun, S.N., 1991. Production smoothing evidence from physical-product date. Journal of Political Economy 99, 558–581.
Latouche, G., Ramaswami, V., 1999. Introduction to Matrix Analytic Methods and Stochastic Modeling. SIAM, Philadelphia.
Lee, H.L., Padmanabhan, V., Whang, S., 1997a. The bullwhip effect in supply chains. Sloan Management Review 38 (3), 93–102, Spring.
Lee, H.L., Padmanabhan, V., Whang, S., 1997b. Information distortion in a supply chain: The bullwhip effect. Management Science 43 (4), 546–558.
Lu, Y., Song, J.-S., Yao, D., 2003. Order fill rate, leadtime variability, and advance demand information in an assemble-to-order system. Operations Research 51 (2), 292–308.
Magee, J.F., 1956. Guides to inventory control, part II. Harvard Business Review, 103–116.
Magee, J.F., 1958. Production Planning and Inventory Control. McGraw-Hill, New York.
Miron, J.A., Zeldes, S.P., 1988. Seasonality, cost shocks, and the production smoothing model of inventories. Econometrica 56, 877–908.
Nelson, R., 1995. Probability, Stochastic Processes, and Queueing Theory. Springer-Verlag.
Neuts, M., 1981. Matrix-Geometric Solutions in Stochastic Models, An Algorithmic Approach. John Hopkins University Press.
Neuts, M., 1989. Structured Stochastic Matrices of M/G/1 Type and their Applications. Marcel Dekker, Inc., New York and Basel.
Ramaswami, V., 1988. Nonlinear matrix equations in applied probability—solution techniques and open problems. SIAM Review 30 (2), 256–263.
Riddalls, C.E., Bennett, S., 2002. Production-inventory system controller design and supply chain dynamics. International Journal of Systems Science 33 (3), 181–195.
Silver, E.A., Pyke, D.F., Peterson, R., 1998. Inventory Management and Production Planning and Scheduling, third ed. John Wiley & Sons, New York.
Simon, H., 1952. On the application of servomechanism theory in the study of production control. Econometrica 20, 247–268.
Song, J.S., Yao, D.D., 2002. Performance analysis and optimization of assemble-to-order systems with random lead times. Operations Research 50 (5), 889–903.
Song, J.-S., Zipkin, P., 1996. The joint effect on leadtime variance and lot size in a parallel processing environment. Management Science 42 (9), 1352–1363.
Song, J.S., Xu, S.H., Liu, B., 1999. Order-fulfillment performance measures in an assemble-to-order system with stochastic lead times. Operations Research 47 (1), 131–149.
Telek, M., 2000. Minimal coefficient of variation of discrete phase type distributions. In: 3rd International Conference on Matrix Analytic Methods in Stochastic Models. Notable Publications Inc., Leuven, Belgium, pp. 391–400.
Telek, M., Heindl, A., 2002. Matching moments for acyclic discrete and continuous phase-type distributions of second order. International Journal of Simulation Systems, Science & Technology, Special issue on Analytical & Stochastic Modelling Techniques 3 (3–4).
Towill, D.R., 1982. Dynamic analysis of an inventory and order based production control system. International Journal of Production Research 20 (6), 671–687.
Tsay, A.A., Nahmias, S., Agrawal, N., 1999. Modeling supply chain contracts: A review. In: Tayur, S., Ganeshan, R., Magazine, M. (Eds.), Quantitative Models for Supply Chain Management. Kluwer Academic Publishers (Chapter 10).
Van Nyen, P.L.M., Bertrand, J.W.M., Van Ooijen, H.P.G., Vandaele, N.J., 2005. A heuristic to control integrated multi-product multi-machine production-inventory systems with job shop routings and stochastic arrival, set-up and processing times. OR Spektrum 27, 399–434.

Vassian, H.J., 1955. Application of discrete variable servo theory to inventory control. Operations Research 3, 272–282.
Veinott, A.F., 1966. The status of mathematical inventory theory. Management Science 12 (11), 745–777.
Warburton, R.D.H., 2004a. An analytical investigation of the bullwhip effect. Production and Operations Management 13 (2), 150–160.
Warburton, R.D.H., 2004b. An exact analytical solution to the production inventory problem. International Journal of Production Economics 92, 91–96.
West, K.D., 1986. A variance bounds test of the linear quadratic inventory model. Journal of Political Economy 94, 374–401.
Zipkin, P.H., 2000. Foundations of Inventory Management. McGraw-Hill, New York.