Coordinating lead times and safety stocks under autocorrelated demand

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

(a) Technology & Operations Management Area, Vlerick Business School, Belgium
(b) Logistics Systems Dynamics Group, Cardiff Business School, United Kingdom
(c) Research Center for Operations Management, University KU Leuven, Belgium
(d) Department of Mathematics and Computer Science, University of Antwerp, Belgium

We consider a supply chain in which orders and lead times are linked endogenously, as opposed to assuming lead times are exogenous. This assumption is relevant when a retailer's orders are produced by a supplier with finite capacity and replenished when the order is completed. The retailer faces demands that are correlated over time – either positively or negatively – which may, for example, be induced by a pricing or promotion policy. The auto-correlation in demand affects the order stream placed by the retailer onto the supplier, and this in turn influences the resulting lead times seen by the retailer. Since these lead times also determine the retailer's orders and its safety stocks (which the retailer must set to cover lead time demand), there is a mutual dependency between orders and lead times. The inclusion of endogenous lead times and autocorrelated demand represents a better fit with real-life situations. However, it poses some additional methodological issues, compared to assuming exogenous lead times or stationary demand processes that are independent over time. By means of a Markov chain analysis and matrix analytic methods, we develop a procedure to determine the distribution of lead times and inventories, that takes into account the correlation between orders and lead times. Our analysis shows that negative autocorrelation in demand, although more erratic, improves both lead time and inventory performance relative to IID demand. Positive correlation makes matters worse than IID demand. Due to the endogeneity of lead times, these effects are much more pronounced and substantial error may be incurred if this endogeneity is ignored.

Keywords: production-inventory model, autocorrelated demand, operations/marketing interface

1. Introduction

In this paper we study the issue of coordinating the retailer's inventory decisions and the supplier's lead times. It is commonly known that supplier lead times have a direct impact on the retailer's

[*] Corresponding author. E-mail addresses of all authors: robert.boute@vlerick.com, disneysm@cardiff.ac.uk, marc.lambrecht@kuleuven.be, benny.vanhoudt@ua.ac.be

safety stocks: longer and more variable lead times require higher safety stocks. But in a make-to-order setting there is also an impact in the opposite direction: the lead times vary according to the order stream of the retailer and its variability. Complicating matters is the assumption that the retailer may be facing orders that are correlated over time. The degree of autocorrelation (and whether it is positive or negative) greatly impacts the level of fluctuations in the order stream, influencing in turn the lead time distribution. In addition we will show that, due to this autocorrelation, the order stream becomes dependent upon the lead time distribution. The objective of this paper is to study the interplay between the correlation in demand, the retailer's order policy (and its safety stocks) and the supplier's lead times. This interplay has, to the best of our knowledge, not been dealt with in the literature before. The resulting production/inventory system poses some challenging methodological issues.

Coordinating lead times and safety stocks is imperative in a supply chain where the supplier produces the retailer's orders on a make-to-order basis. Several reasons may motivate a make-to-order approach, ranging from a limited shelf life to frequent upgrades or customer's specific packaging requirements. In such an environment the supplier may opt to not hold inventory, but the retailer does hold safety stocks to satisfy immediate consumer demand. We have encountered several examples where a make-to-order policy is employed for customized products, and where the insights obtained in this paper can be applied. For instance, an industrial bakery, producing authentic specialities in the biscuit and cake world, employs a make-to-order policy for a major retailer due to specific packaging requirements with the retailer's label on the product, sometimes combined with a specific, temporary promotion. As the products have a limited shelf life and the retailer's orders fluctuate every period, a make-to-stock policy is not suitable for these products. Another example is a supplier of feminine-care and baby-care products (diapers, baby wipes, tampons, etc.) who manufactures retailer brands. In their quest to compete with A-brands, they rely heavy on promotions. Due to the high fluctuations in demand, in combination with the retailer-specific requirements, the supplier does not keep any stock, instead he produces to order. The retailer however holds the product in inventory to ensure immediate availability to the consumer.

The only abstraction we make from these practical settings is the fact that we apply our methodology to a single item; however our insights can be generalized to a multi-item setting, where a safety stock is held per item. Our model is also capable of representing a firm that replenishes its finished goods inventories from its own production facilities. This firm must plan releases into the production system in such a manner as to maintain safety stocks at its inventory points facing customers.

In such a make-to-order setting the nature of the order stream (variability in inter-arrival time and order sizes) affects the sojourn times within the supplier's queue, and thus the lead times observed by the retailer. By modeling a two-echelon (retailer-manufacturer) supply chain as a production/inventory system, we treat lead times as *endogenous* variables; this means that we do not merely assume the replenishment lead time to be a fixed or random *exogenous* variable. Instead we include the impact of a replenishment decision on the production lead times and use these lead times in our inventory model. We use an iterative procedure to cope with this interaction effect.

The inclusion of autocorrelation (or time-correlation) in demand, as opposed to assuming IID demand, is valid in many high-tech and consumer goods industries (see e.g. Dong and Lee, 2003). In these industries, consumers are typically highly sensitive to marketing actions. We analysed a large number of consumer demand patterns (weekly POS data) for consumer packaged non-food products, both branded products and private label products. For the regular 'turn' business, positively autocorrelated demand patterns seem to dominate. This is confirmed by Erkip et al. (1990) and Disney et al. (2006) who also find that positively correlated consumer demand was most commonly observed. However, in the presence of recurring weekly promotions, a retailer may observe negative autocorrelation as well; this is due to consumers stockpiling during the promotion period and cannibalising demand before (and after) the promotion. In the marketing literature, this is referred to as pre- and post-promotion dips (Macé and Neslin, 2004). This promotion strategy may create negative period-to-period correlation in demand.

We show that correlation in demand has an important impact on the performance of the supply chain in terms of safety stocks and lead times. The inclusion of autocorrelation in demand illustrates that price control mechanisms can be used to manage supply chains, reinforcing once more the importance of coordinating marketing and operations decisions along the chain. Note that in this paper we focus on the impact of autocorrelation, rather than on the overall variability in demand, which can also be influenced by price promotions. We refer to Raju (1992) who relates the promotional activity in a product category to its variability in sales and Boute et al. (2007) who study the operational impact of demand variability on lead times and safety stocks.

The remainder of this paper is organized as follows. The next section presents an overview of the related literature. Section 3 describes our research model and derives expressions for the orders generated by the retailer. Section 4 develops an iterative procedure to determine the endogenous supply lead times and Section 5 is devoted to the analysis of the safety stocks in the combined production/inventory system. Section 6 provides a numerical experiment and Section 7 concludes.

2. Literature review

This paper studies the interplay between autocorrelation in demand, the retailer's inventory policy (facing the autocorrelated demand), and the supplier's lead times (producing the retailer's orders). In the literature inventory models are discussed with either autocorrelated demand, but assuming exogenous lead times; or production/inventory models with endogenous lead times, assuming IID demand. In the following we review the literature on both streams of research. Finally we briefly discuss the literature on our methodology used.

Several papers discuss supply chains with autocorrelated demand and constant (exogenous) lead times. Fotopoulos et al. (1988) provide an upper bound for the safety stock when daily demands are autocorrelated and lead times follow an arbitrary distribution. Erkip et al. (1990) derive optimal stocking levels as a function of the autocorrelation coefficient. Dong and Lee (2003) develop a lower bound for the optimal stocking levels in serial multi-echelon systems under time-correlated demand. Kahn (1987) and Lee et al. (1997) demonstrate the existence of variance amplification upstream in the chain (aka the bullwhip effect) when the retailer follows a base-stock policy and demand is positively correlated. Zhang (2004b) studies the role of forecasting for AR(1) demands and concludes that the minimum Mean Squared Error (MSE) forecasting method minimises the variance of the forecasting error among all linear forecasting methods, and therefore leads to the lowest inventory costs. Alwan et al. (2003) employ this optimal MSE forecasting scheme and determine the underlying time-series model of the resulting order process. They show that when consumer demand is negatively correlated (with AR demand), the variability in orders is dampened with respect to the observed demand. This result is of great importance for our paper.

Negative correlation may occur, for instance, due to consumers stockpiling during the promotion period and deceleration before and after the promotion. Stockpiling is the propensity of consumers to increase their inventories above normal levels either by purchasing the category earlier, or by purchasing greater-than-normal quantities (Neslin et al., 1985). Deceleration is the willingness of consumers to deplete their inventories below normal levels by 'holding out' for an anticipated promotion (Mela et al., 1998). The impact of promotions on consumer demand behavior is extensively discussed in the marketing literature as they may influence profitability (Blattberg and Neslin, 1993; Hendel and Nevo, 2006). Macé and Neslin (2004) empirically studied the relationships between pre- and post-promotion dips in weekly store data, and find that these dips are stronger for high-priced, frequently promoted, mature, high-market-share products.

The interaction between inventory policies and lead times is generally studied in production/inventory systems with endogenous lead times. Graves (1988) provides an excellent review and critique of the research literature on safety stocks for manufacturing systems, and proposes a model to include consideration of the flexibility of the production stage in planning safety stocks. Base-stock controlled production/inventory systems in continuous time with exponential (single unit) demand processes have been studied widely, among many others, by Gavish and Graves (1980, 1981), Song and Zipkin (1996), Sox et al. (1997) and Jemaï and Karaesmen (2005). Ettl et al. (2000) and Liu et al. (2004) model a supply network with multiple storage locations by means of an inventory-queue model assuming Poisson demand. Boute et al. (2007) propose a solution method for production/inventory systems in discrete time with a random IID integer consumer demand.

However, none of these papers consider autocorrelation in demand. The interaction between order release models and lead times is also related to this problem. Pahl et al. (2005) provide an overview of the literature on production planning models with load dependent lead times (see also Orcun et al. (2009)). They consider lead times to be dependent on the current load in the system, and make use of clearing functions to incorporate this dependency. Selcuk et al. (2009) discuss a *lead time syndrome*, which may arise in this setting: the cyclic interaction between planned lead times and order sizes may result in uncontrolled order release patterns.

Our methodology is based on Markov chains of the GI/M/1 type (Neuts, 1981), phase type (PH) distributions (see e.g. O'Cinneide (1990)) and matrix analytic methods (Latouche and Ramaswami, 1999). The domain of matrix analytic techniques was advocated by Neuts (1981, 1989). 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. We refer to Bini et al. (2005) for an overview of recent algorithmic developments. Software tools both in Fortran and MATLAB were made available by Bini et al. (2006). The use of matrix analytic techniques in the production/inventory models is fairly scarce. Riaño (2002) uses matrix analytic methods and PH distributions to accomplish analysis similar to our paper. Here, they model the transient behavior of multi-class queueing networks with load-dependent lead times.

This paper contributes to the existing literature in three ways. First, we set up a four-dimensional Markov chain that is able to compute the lead time distribution in a setting where production orders are generated by a periodic review base-stock policy with a correlated AR(1) demand process and MSE forecasting with known parameters. Second, we use this lead time distribution to generate orders, thereby tackling the mutual dependency that arises in this context (orders are dependent on the lead time distribution and vice versa). Third, we determine the inventory distribution and the safety stock requirements of the corresponding production/inventory system, taking into account the correlation between orders and lead times. These contributions are made via the application of matrix analytic methods. We generate insights into the effects of the endogeneity of lead times and the correlation in demand on supply chain performance.

3. Model description

We consider a two-echelon supply chain consisting of one retailer and one supplier. Consumer demand, D_t, is observed at the beginning of a time period t, but need not be fulfilled until the end of the period. Unfilled demand is backordered. Retailer's inventory levels are reviewed after demand is satisfied, and an order O_t is placed with the supplier. The supplier does not hold a finished goods inventory, but produces on a make-to-order basis. The supplier's capacity is finite and operates like a discrete time queuing system. A single server sequentially processes items one at the time on a first-come-first-served basis. When the server is busy, the order joins the queue of unprocessed orders. Orders are only shipped when completed. Let T_p denote the discrete distribution function of the replenishment lead time variable (i.e. the time from the period an order is placed to the period it replenishes the inventory). The supply process implies that lead times are endogenous, and thus correlated with the current load of the queue and the actual order size (larger orders increase the batch production time). Randomness in the suppliers production process, combined with batch ordering and delivery, lead to difficulty in characterising the production and replenishment lead times.

In the following, we describe in more detail the consumer demand process, its forecast and we derive an expression for the order process placed with the supplier.

3.1 Consumer demand process

There are a number of potential stochastic processes that can be assumed to model consumer demand, ranging from simple IID to non-stationary processes. One industrially relevant, flexible, correlative demand process that has often been studied in the supply chain literature is the first-order autoregressive or AR(1) model. Traditionally, an AR(1) demand is given by

$$D_t = \mu + \phi D_{t-1} + \varepsilon_t, \qquad |\phi| < 1, \tag{1}$$

where D_t is the demand observed in period t, ϕ the first-order autocorrelation coefficient, μ a constant, and ε_t an IID random error with mean 0 and variance σ^2. The assumption of $|\phi| < 1$ assures that the demand process is covariance stationary. Sometimes Eq. (1) is re-written as a mean-centered demand pattern, $D_t = E(D) + \phi\,(D_{t-1} - E\,(D)) + \varepsilon_t$, which omits the parameter μ.

For the purpose of this paper, we use a slightly different notation, but there is no fundamental difference to (1). We assume consumer demand follows the correlated process

$$D_t = \phi\ D_{t-1} + (1-\phi)\,G_t, \tag{2}$$

with $G_t = (\mu + \varepsilon_t)/(1-\phi)$ a random IID term with mean $E(G) = \mu/(1-\phi) > 0$ and variance $Var(G) = \sigma^2/(1-\phi)^2$. In this notation the error term is given by $(1-\phi)\,G_t$. With the initial condition $D_0 = G_0$, the average demand under this notation equals $E(D) = E(G)$ for $t \geq 0$, and its long run variance $Var(D) = \frac{1-\phi}{1+\phi}\ Var(G)$ for $t \to \infty$. Observe as well that under this notation the demand variance decreases as ϕ increases towards 1.

As we will discuss later in this paper, this notation yields some elegant formulations and remarkable similarities between the demand pattern and the order pattern when demand is forecasted using the MSE forecasting technique. This reduces the complexity of the queueing analysis, which is used to compute the lead time distribution in our model.

When $0 \leq \phi < 1$, the minimum and maximum demand are given by the minimum and maximum values of G, or $d_{min} = g_{min}$ and $d_{max} = g_{max}$. When $-1 < \phi \leq 0$, the minimum and maximum demand are given by $d_{min} = (g_{min} + \phi\ g_{max})/(1 + \phi)$ and $d_{max} = (g_{max} + \phi\ g_{min})/(1 + \phi)$. We use these relations to set conditions on g_{min}, g_{max} and ϕ to avoid negative demand.

For $-1 < \phi < 0$, the demand process is negatively correlated and will exhibit period-to-period oscillatory behavior. For $0 < \phi < 1$, the demand process is positively correlated, characterised by a wandering or meandering sequence of observations. One can view ϕ as a marketing parameter related to the impact of promotion on demand. A negative value for ϕ could mean that the consumer's buying behavior is highly influenced by a promotion in the sense that consumers increase their purchases in the promotion week, and decelerate before and after the promotion. A positive ϕ value denotes a less aggressive reaction to the promotion: product demand is related to previous period's demand, rather than influenced by a price promotion. Note that an AR(1) process means that the demand autocorrelation is one period apart. This is different from, for example, correlations lasting for several periods. That is, a period of high demand due to forward buying by consumers who buy several periods worth of product, leading to low demand for several following periods. In that case we may need an AR(p) process.

Several techniques are available to forecast lead time demand. The moving average (MA) and exponential smoothing (ES) forecast methods are widely employed because of their simplicity and ease of implementation. However, knowing that demand follows an AR(1) process, the minimum Mean Squared Error (MSE) forecasting method is the preferred forecasting scheme as it minimises the forecast error (Zhang, 2004b). It explicitly takes the correlated demand structure into account, which is not the case with the non-optimal ES and MA forecasts. Any other forecast method will lead to an increased forecast error and hence higher inventory costs. Therefore we will proceed using the MSE forecasting procedure. This forecasting technique assumes that the underlying parameters of the demand model are constant and known or that a suitable amount of demand data is available to estimate these parameters accurately. The MSE forecast is the conditional expectation of future demand, given current and previous demand observations (Box et al., 1994). Hence, for our assumed demand process in (2), the i-period ahead demand forecast is given by

$$\widehat{D}_{t+i} = \phi^i D_t + \left(1 - \phi^i\right) \cdot E(G). \tag{3}$$

3.2 Retailer's order stream

We adopt the standard periodic review base-stock policy. This policy is optimal for the retailer in absence of a fixed ordering cost and when holding and shortage costs are convex and proportional to the volume of on-hand inventory or shortage (Nahmias, 1997; Zipkin, 2000). Let S_t be the base-stock level, which equals the inventory position after placing the order in period t. The base-stock level is also the sum of the forecasted lead time demand and the safety stock. Lead time demand is here defined as the demand during the *risk period* l, with the risk period $l = 1 + t_p$ (review period plus replenishment lead time), or

$$D_t^l = \sum_{i=1}^{l} D_{t+i}. \tag{4}$$

Let $\widehat{D}_t^l = \sum_{i=1}^{l} \widehat{D}_{t+i}$ denote the lead time demand forecast, and I_s the safety stock required to achieve a desired service level. Then, $S_t = \widehat{D}_t^l + I_s$.

The base-stock level S_t is *adaptive* over time in the sense that we update the demand forecast when a new demand realisation occurs, since the current demand holds information about the future demands during the replenishment lead time. One could also adapt the value of l every period. Indeed, the current workload in the queue contains information on the time it takes to replenish the placed order. However, this excessive adaptivity might lead to instability: when for instance the queue is highly congested, the replenishment time will be long, which inflates the order size; this inflated order increases the workload in the queue even more (and so does its time to replenish), inflating the order even further, and so on. This instability is known as the lead time syndrome and is to be avoided (Selcuk et al., 2009). Therefore, instead of updating the lead time every period, we use its steady state variable $L = 1 + T_p$ in our decision rule to generate orders:

$$S_t = \widehat{D}_t^L + I_s. \tag{5}$$

Observe that even under this assumption we obtain a mutual dependency: the base-stock level (5) assumes a lead time distribution T_p. But T_p is determined by the order stream, loading the queue, which is generated by the base-stock level (5). We cope with this mutual dependency by assuming an initial lead time distribution T_p^n to generate orders, and we compute the lead time distribution T_p^{n+1} according to this order stream. This new lead time distribution is then used to update the base-stock level, and we continue this procedure until the lead time distribution converges. We discuss this iterative procedure in section 4.

We also assume the safety stock I_s to be stationary over time. We do take into account that a larger order involves a longer supply time and as such we include the correlation between orders and lead time in determining the inventory distribution (see section 5). However, we do not periodically adapt the safety stock depending on the most recent demand or lead time observation. This is in line with inventory literature, where typically the base-stock level is periodically adjusted as the demand forecast changes, but the safety stock is assumed to be stationary over time (see e.g. Erkip et al., 1990; Fotopoulos et al., 1988; Graves, 1999; Zhang, 2004a,b). Moreover, we believe the problem becomes mathematically intractable with non-stationary safety stocks.

The timing of events (first receive goods from supplier, then satisfy demand and finally place the order) and the conservation of flow implies that

$$\begin{aligned} O_t &= S_t - S_{t-1} + D_t \\ &= D_t + \left(\widehat{D}_t^L - \widehat{D}_{t-1}^L \right). \end{aligned} \tag{6}$$

Observe that L is a random variable, so $\widehat{D}_t^L = \sum_{l=1}^{\infty} \Pr(L = l)\widehat{D}_t^l$. Then, using (3-4) we find

$$\begin{aligned} \widehat{D}_t^L &= \sum_{l=1}^{\infty} \Pr(L = l) \left(\sum_{i=1}^{l} \phi^i D_t + \sum_{i=1}^{l} (1 - \phi^i) E(G) \right) \\ &= \sum_{l=1}^{\infty} \Pr(L = l) \left(\frac{\phi(1 - \phi^l)}{1 - \phi} D_t + \left(l - \frac{\phi(1 - \phi^l)}{1 - \phi} \right) E(G) \right) \\ &= \frac{\phi(1 - E(\phi^L))}{1 - \phi} D_t + \left(E(L) - \frac{\phi(1 - E(\phi^L))}{1 - \phi} \right) E(G), \end{aligned} \tag{7}$$

with $E(\phi^L) = \sum_{l=1}^{\infty} \Pr(L = l)\phi^l$. Substituting (7) into (6) returns the retailer's order process:

$$O_t = \frac{1 - E\left(\phi^{L+1}\right)}{1 - \phi} D_t - \frac{\phi\left(1 - E\left(\phi^L\right)\right)}{1 - \phi} D_{t-1}. \tag{8}$$

The retailer's order quantity is a linear combination of the observed demand in the current period and the previous period. Substituting (2) into (8) provides

$$O_t = E\left(\phi^{L+1}\right) \cdot D_{t-1} + \left(1 - E\left(\phi^{L+1}\right)\right) \cdot G_t, \tag{9}$$

which is very similar to the expression of the demand process (Eq. (2)). This order process actually has an ARMA(1,1) structure, similar but different to the AR(1) process (Zhang, 2004a). Observe that this order stream is dependent on the lead time distribution L.

As these orders are sent to the supplier's production queue, it is worthwhile analysing some characteristics of this process. First, the order size has the same bounds as the demand size. Amongst others, this implies that if we provide a condition on G and ϕ to avoid negative demand, this automatically precludes negative order sizes. Next, from (9) we find that the variance in the order stream is given by

$$\begin{aligned} Var(O) &= \left(E\left(\phi^{L+1}\right)\right)^2 Var(D) + \frac{(1 + \phi)\left(1 - E\left(\phi^{L+1}\right)\right)^2}{1 - \phi} Var(D) \\ &= \left[1 + \frac{2\phi\left(1 - E\left(\phi^L\right)\right)\left(1 - E\left(\phi^{L+1}\right)\right)}{1 - \phi} \right] Var(D). \end{aligned} \tag{10}$$

Using (10), we derive that the order variance is amplified with respect to the demand variance when there is positive correlation in demand. This phenomenon is referred to as the bullwhip effect:

$$\begin{aligned} Var(O) > Var(D) \quad &\Leftrightarrow \quad 1 + \frac{2\phi\left(1 - E\left(\phi^{L}\right)\right)\left(1 - E\left(\phi^{L+1}\right)\right)}{1-\phi} > 1 \\ &\Leftrightarrow \quad 2\phi\left(1 - E\left(\phi^{L}\right)\right)\left(1 - E\left(\phi^{L+1}\right)\right) > 0 \\ &\Leftrightarrow \quad \phi > 0. \end{aligned} \tag{11}$$

Analogous to (11), we find that when the autocorrelation coefficient is negative, there is an anti-bullwhip, or *de-whip* effect, which means that the orders are smoothed compared to the demand:

$$Var(O) < Var(D) \quad \Leftrightarrow \quad \phi < 0. \tag{12}$$

This result contrasts with the traditional, non-optimal, MA and ES forecasting techniques, which always produce bullwhip, independent of the assumed demand (Dejonckheere et al., 2003). A similar conclusion was obtained by Alwan et al. (2003). This is important for our analysis. The sign of the correlation coefficient determines whether orders are amplified in variability towards the supplier, or not. In case of no autocorrelation, consumer demand is IID and we obtain orders equal to demand, i.e., no amplification nor dampening. Since the supplier produces on a make-to-order basis, this will impact the lead time distribution. Positively correlated demand amplifies variability in orders, with increasing average supply lead times as a consequence. Negative period-to-period correlation in demand dampens the order variability, resulting in shorter lead times on average. We establish and analyse the supplier's queueing process, and therefore, the lead time distribution seen by the retailer, in the following section.

4. Computation of the lead time distribution

The supplier's operation acts as a discrete time queueing system. The replenishment orders described by (9) load the production queue. This means that the arrival process at the queue consists of batch arrivals (equal to the size of the replenishment orders) and deterministic inter-arrival times (equal to one review period). A single server sequentially processes single items with stochastic service times. An order is shipped only when the production of the order is completed. Thus, larger orders increase the batch production time. In addition, the lead time is also a function of the current load (work in process) of the queue. Hence, to compute the lead time distribution, we need to set up a queueing model taking these considerations into account.

4.1 Lead time dependency

The nature of the order stream, loading the queue, determines the distribution of the lead times. At the same time, from (9) we know that the order stream itself is dependent on the lead time distribution. In other words, we have a mutual dependency between the order process and the lead time distribution.

To cope with this mutual dependency, we develop an iterative procedure. We start with an initial guess for the lead time distribution T_p^0 (typically, we select T_p^0 deterministically, equal to 0 periods). Next, for $n > 0$, we make use of T_p^{n-1} to determine the order process in (9). Given this order stream, we determine the new lead time distribution T_p^n and repeat this procedure. We have carried out extensive numerical experiments, and we find that the lead time distribution converges towards the actual lead time distribution when $|\phi| < 1$. This assumption is not restrictive as $|\phi| < 1$ also assures that the demand process is covariance stationary[1].

Since the Markov chain analysis used to find the lead time distribution is based on a numerical procedure, we do not have a formal proof that an equilibrium distribution T_p exists. However, assuming it does exist, we note that the coupling between the order process and the lead time depends only on the scalar $E\left(\phi^{T_p+1}\right)$. If $0 \leq \phi < 1$, $E\left(\phi^{T_p+1}\right)$ takes values in $[0, 1]$ for any distribution T_p (on the positive integers). If we define $x = E\left(\phi^{T_p+1}\right)$, with T_p an arbitrary distribution, and compute the lead time distribution T_p^* based on the corresponding order stream, we can define $f(x) = E\left(\phi^{T_p^*+1}\right)$. In other words, we have a mapping f from $[0, 1]$ to $[0, 1]$. Thus, if we can show that f is continuous, there must be a fixed point for f in $[0, 1]$ (due to Brouwer's fixed-point theorem) and thus an equilibrium distribution. Although the continuity of f seems intuitively obvious, it is hard to prove formally as f relies on a numerical procedure. For $-1 < \phi < 0$ a similar argument can be given except that f takes values in $[-1, 1]$.

It is noteworthy to re-emphasize that, if we would use the transient value of t_p, instead of its steady state distribution T_p to generate orders, convergence would not be guaranteed due to the lead time syndrome discussed earlier.

4.2 Assumptions of the queueing model

To estimate the lead time distribution at iteration n, we develop the following discrete time queueing model. The retailer's base-stock policy, assuming AR(1) demand and MSE forecasting, generates batch arrivals with a fixed inter-arrival time (equal to the review period, i.e. 1 period) and with variable batch sizes, which are correlated (see (9)). The service times of a single item, denoted by M, are stochastic and IID according to a phase type (PH) distribution[2].

[1] The convergence was numerically studied for several hundred randomly generated systems. For each of these systems we used five different randomly chosen starting values for $E(\phi^{T_p+1})$ and found that convergence occurred to the same fixed point in each case.

[2] 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 (O'Cinneide, 1990), 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 (Bobbio et al., 2003; Latouche and Ramaswami, 1999; Neuts, 1989).

The computational complexity of our queueing algorithm increases with the number of phases of the PH distributed service process. We use the moment matching procedure described by Boute et al. (2007) to match the first two moments of the single unit service times to a discrete PH distribution with a minimal number of phases (including more moments leads to a higher number of phases). Since the lead time is expressed as an integer number of periods and the inter-arrival time of orders is equal to one base period, we have the freedom to choose the time unit U of the queueing system as desired (Bobbio et al., 2004). When the time unit U is chosen as half the mean service time of a single item, i.e., $U = E(M)/2$, it is possible to match the first two moments of the single unit service times by means of a PH distribution with only 2 phases, characterised by the pair (T, α):

$$\alpha = (\beta, 1-\beta) \ , \quad T = \begin{bmatrix} 1-\beta & \beta \\ 0 & 0 \end{bmatrix}, \text{ with } \quad \beta = \frac{1}{1+2cv^2(M)}, \tag{13}$$

and $cv^2(M)$ the squared coefficient of variation of the single unit service times (Boute et al., 2007).

When U is the time unit of our queueing system, orders placed every period arrive at the queue at times $0, e, 2e, \ldots$, where $e \times U = 1$ period. The order sizes are driven by an underlying Markov process with state space $\{d_{min}, d_{min}+1, ..., d_{max}\}$, where d_{min} and d_{max} are respectively the min and max demand size as defined in section 3.1. Indeed, according to (9), the order placed in period t, or equivalently, at time te if expressed in the time unit U, is determined by

$$O_{te} = E\left(\phi^{L+1}\right) D_{(t-1)e} + \left(1 - E\left(\phi^{L+1}\right)\right) G_{te}, \tag{14}$$

where the demand process itself evolves as

$$D_{te} = \phi D_{(t-1)e} + (1-\phi) G_{te}, \tag{15}$$

which has a Markovian nature. Using induction on t we find that $E(O) = E(D) = E(G)$. Hence, if we know the value of $D_{(t-1)e}$, we can define the transition to both D_{te} and O_{te} (and their respective probabilities) based on G_{te} (and its probability function). This reduces the complexity of the Markov analysis considerably as we only need to keep track of the demand $D_{(t-1)e}$ to determine the transition probabilities to both the demand D_{te} and order size O_{te} in the subsequent period.

The demand and order size resulting from (14) and (15) can be a real number. As it is more natural to have demands of integer size, the actual demand (determining the order size) is stochastically rounded[3] to have size D^*_{te}:

$$D^*_{te} = \begin{cases} D_{te} & \text{if } D_{te} \in \mathbb{N}, \\ \lceil D_{te} \rceil & \text{with probability } D_{te} - \lfloor D_{te} \rfloor \text{ if } D_{te} \notin \mathbb{N}, \\ \lfloor D_{te} \rfloor & \text{with probability } \lceil D_{te} \rceil - D_{te} \text{ if } D_{te} \notin \mathbb{N}. \end{cases} \tag{16}$$

Analogously, because only an integer number of items can be produced, the batch size passed to the manufacturer at time t is also stochastically rounded to size O^*_{te}:

$$O^*_{te} = \begin{cases} O^+_{te} & \text{if } O^+_{te} \in \mathbb{N}, \\ \lceil O^+_{te} \rceil & \text{with probability } O^+_{te} - \lfloor O^+_{te} \rfloor \text{ if } O^+_{te} \notin \mathbb{N}, \\ \lfloor O^+_{te} \rfloor & \text{with probability } \lceil O^+_{te} \rceil - O^+_{te} \text{ if } O^+_{te} \notin \mathbb{N}, \end{cases} \tag{17}$$

where O^+_{te} is found by (14) when replacing $D_{(t-1)e}$ by $D^*_{(t-1)e}$. In order to simplify the notation, however, we will use respectively D_{te} and O_{te} instead of D^*_{te} and O^*_{te}, and assume in the remainder of this section that D_{te} and O_{te} are rounded according to (16) and (17) respectively. Discretising the range of the demand and order sizes on the integer values is not only more natural, but also helps in computing the lead time distribution in an efficient manner. That is, it allows us to construct a Markov chain that has a considerably smaller state space, leading to less demanding time and memory requirements for the numerical procedure involved.

Let $p^{(g)}(k, k')$ denote the transition probabilities characterising the Markovian demand process, defined by (15), so that $p^{(g)}(k, k') = \Pr(G_{te} = g, D_{te} = k' | D_{(t-1)e} = k)$ for k, k' in $\{d_{min}, d_{min} + 1, ..., d_{max}\}$ and g in $\{1, \ldots, g_{max}\}$. Then, due to the stochastic rounding to integer demand values (Eq. (16)), these conditional probabilities are given by:

$$p^{(g)}(k, k') = \Pr(G = g) \cdot \Big\{ 1_{\{k'-1 < \phi k + \bar{\phi} g < k'\}} \left((\phi k + \bar{\phi} g) - \lfloor \phi k + \bar{\phi} g \rfloor \right) + \\ 1_{\{\phi k + \bar{\phi} g = k'\}} + 1_{\{k' < \phi k + \bar{\phi} g < k'+1\}} \left(\lceil (\phi k + \bar{\phi} g) \rceil - (\phi k + \bar{\phi} g) \right) \Big\}, \tag{18}$$

where we denote $\bar{\phi} = (1 - \phi)$ and the indicator function $1_{\{A\}}$ is 1 if the event A is true and 0 otherwise. Similarly, we can derive the transition probabilities characterising the order process, defined by (14). Let $p_{[q]}(k, k')$ denote the conditional probabilities $\Pr(O_{te} = q, D_{te} = k' | D_{(t-1)e} = k)$. Then, (18) combined with (17) leads to

$$p_{[q]}(k, k') = \sum_{g=1}^{g_{max}} p^{(g)}(k, k') \cdot \left\{ 1_{\{q-1 < \gamma k + \bar{\gamma} g < q\}} \left((\gamma k + \bar{\gamma} g) - \lfloor \gamma k + \bar{\gamma} g \rfloor \right) + \right. \\ \left. 1_{\{\gamma k + \bar{\gamma} g = q\}} + 1_{\{q < \gamma k + \bar{\gamma} g < q+1\}} \left(\lceil (\gamma k + \bar{\gamma} g) \rceil - (\gamma k + \bar{\gamma} g) \right) \right\}, \tag{19}$$

[3] For instance suppose that the demand process generates a demand of size 5.8. We round this to 5 units with a probability of 0.20 and to 6 units with a probability of 0.80. This (integer) demand size is used to determine the batch order that is sent to the queue. This rounding does not affect the expected value, $E(D^*_{te}) = E(D_{te}) = E(D)$.

where we denote $\gamma = E(\phi^{L+1})$ and $\bar{\gamma} = (1 - \gamma)$.

4.3 Markov chain analysis

To create a Markov chain that is able to find the lead time distribution, we define the following random variables:

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

We assume that all events (e.g., order arrival, service start and service completion) occur immediately after the discrete time epochs of the Markov chain. This implies that the age of any order in service at an arbitrary epoch t_n is at least 1 time unit. Then, $(V_n, D_{a(n)}, C_n, S_n)$ forms a Markov chain on the state space $\mathbb{N}_0 \times \{x : x = d_{min}, d_{min}+1, ..., d_{max}\} \times \{c \in \{1, 2, \dots, d_{max}\}\} \times \{1, 2\}$, as: V_n is a positive integer; $D_{a(n)}$ (the demand size in the period when the order in service was placed) is an integer between d_{min} and d_{max}; C_n an integer between 1 and d_{max}; and the PH service has two phases.

The Markov chain $(V_n, D_{a(n)}, C_n, S_n)$ evolves as follows. At each transition step, there are three possibilities. First, the current serviced item remains in service and the phase of the service process may change. Second, the current serviced item completes its service, and a new item of the same batch starts service. Third, the current serviced item completes its service and when this is the last item of the batch, it means that the complete batch is produced and a new order starts service with batch size given by $p_{[q]}(k, k')$ according to (19). Let $(P)_{(a,k,r,s),(a',k',r',s')}$ denote the transition probabilities of the Markov chain $(V_n, D_{a(n)}, C_n, S_n)$. Based on the evolution of this Markov chain, these probabilities are then given by

$$(P)_{(a,k,r,s),(a',k',r',s')} = \begin{cases} T_{s,s'} & a' = a+1, k' = k, r' = r, \\ T_s^* \alpha_{s'} & a' = a+1, k' = k, r' = r-1 \geq 1, \\ T_s^* \alpha_{s'} p_{[q]}(k, k') & a' = \max(a - e + 1, 1), r' = q, r = 1, \\ 0 & \text{otherwise,} \end{cases} \tag{20}$$

with $T^* = (e - Te)$ denoting the probability that the current serviced item completes its service. We obtain the following form for the transition matrix P of $(V_n, D_{a(n)}, C_n, S_n)$:

$$P = \begin{bmatrix} A_e & A_0 & 0 & \dots & 0 & 0 & \dots \\ A_e & 0 & A_0 & \dots & 0 & 0 & \dots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \ddots \\ A_e & 0 & 0 & \dots & A_0 & 0 & \dots \\ 0 & A_e & 0 & \dots & 0 & A_0 & \ddots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \ddots & \ddots \end{bmatrix}, \tag{21}$$

where A_0 and A_e are square matrices of dimension $m_{tot} = 2(d_{max} - d_{min} + 1)d_{max}$. The matrix A_0 represents the probabilities that the service of the batch continues and is given by the first two equations in (20), while the matrix A_e represents the probabilities that the service of the batch is completed and is given by the 3rd equation in (20).

The MC characterised by (21) is of the GI/M/1 type (Neuts, 1981). The queueing system is stable if and only if its utilization ρ is strictly smaller than one (a system with load $\rho > 1$ leads to infinite lead times). This means that the average service time of a batch order should be 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 service time of a single item equals 2, and since the average batch order size equals the average demand $E(D)$, the average service time of a batch order is $2E(D)$. The time between two order arrivals is one (review) period, or, when we express it in the time unit of our queueing model, equal to e time units. Hence the stability condition can be rephrased as $2E(D) < e$.

For an ergodic MC of the GI/M/1 type, the steady state vector of P, denoted by π, i.e., $\pi P = \pi$ and $\pi \mathbf{1} = 1$, is computed as follows:

$$\pi_1 = \pi_1 (I - R^e)(I - R)^{-1} A_e, \tag{22}$$

$$\pi_i = \pi_1 R^{i-1}, \tag{23}$$

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

Once the steady state vector $\pi = (\pi_1, \pi_2, \dots)$ is obtained, we can find the response (or sojourn) time in our queueing system by making the following observation: the probability that an order has a response time of a time units is equivalent to the expected number of orders of age a that complete service at an arbitrary time instant, divided by the expected number of orders that complete service

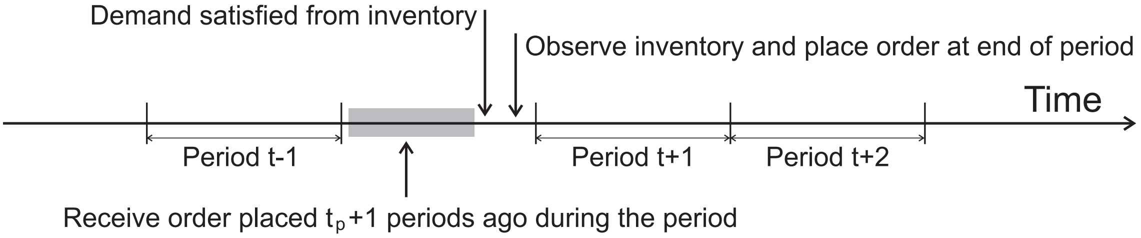
Demand satisfied from inventory
Observe inventory and place order at end of period
Time
Period t-1
Period t+1
Period t+2
Receive order placed t_p+1 periods ago during the period

Figure 1: Sequence of events within a period: 1. receive produced orders, 2. satisfy demand, 3. place order

during an arbitrary time instant, irrespective of their age (which is $1/e$ for a queue with $\rho < 1$). This number is obtained by observing the state of the Markov chain only at the service completion times. Let T_r denote the response time variable and let $\pi_a(k, r, s)$ represent the steady state probability of being in state (a, k, r, s). Then,

$$\Pr(T_r = a) = e\rho \sum_{k,s} \pi_a(k, 1, s)\ (T^*)_s. \tag{24}$$

Eq. (24) is derived as follows: a batch order can only complete its service if the current serviced item is the last one in the batch that needs to be serviced (i.e., $r = 1$). The term $(T^*)_s$ originates from the PH distributed service times, and denotes the probability that the current serviced item completes its service when in phase s. Finally, we need to multiply with the average load ρ since the Markov chain is only defined when the server is busy.

Note that the response time t_r resulting from the queueing analysis, is expressed in the time unit U (defined as $U = E(M)/2$) and 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 t_p is expressed in an integer number of periods. We derive the replenishment lead time, t_p, from the response time, t_r, by relying on the sequence of events in a period: the demand need not be fulfilled until the end of the period, i.e., after the receipt of produced items in inventory, and an order is placed only after demand is satisfied (see Fig. 1). Therefore, rounding the response time down to the nearest integer (i.e., setting $t_p = \lfloor t_r \rfloor$) leads to the (discrete) replenishment lead time. For instance, an order placed at the end of period t with response time $t_r = 0.7$ periods is added to inventory in the next period and can be used to satisfy demand in this period; thus the 0.7 period response time corresponds to a replenishment lead time $t_p = 0$.

$$\Pr(T_p = i) = \sum_j \Pr(T_r = j) \cdot 1_{\{\lceil j/e \rceil = i\}}. \tag{25}$$

This lead time distribution, T_p, is then used to start a new iteration.

5. Characterisation of inventory distribution and safety stocks

The stationarity of the lead time distribution T_p does not necessarily imply that there is no correlation between lead times. On the contrary, the endogeneity of lead times implies that it takes a longer time to produce (and consequently replenish) a larger order, which means that the order quantity and its replenishment lead time are correlated. The lead time for an order is also affected by the current load of the queue, which is dependent on the previously placed orders. Hence, since orders are correlated over time, it is fair to assume that lead times are also autocorrelated. Therefore, if we want to characterise the inventory distribution and determine the safety stock levels in an appropriate way, we need to take this endogeneity into account. In this section we first characterise the transient evolution of the net inventory and then derive its steady state distribution. Based on this inventory distribution the safety stock is then defined to ensure a desired fill rate.

5.1 Transient evolution of the net inventory

Let I_t denote the retailer's net inventory at the end of period t, after demand D_t is satisfied. If we monitor the system after the replenishment order O_t has been placed, there may be $l \geq 0$ orders in the queue and the order O_{t-l} is in service (since the observation moment is immediately after an order placement). Note that l is a function of t, but we write l as opposed to $l(t)$ to simplify notation. Let the initial inventory level I_0 be the control variable (which includes the safety stock I_s), then

$$I_t = I_0 + \sum_{i=l+1}^{t} O_{t-i} - \sum_{i=0}^{t} D_{t-i}.$$

According to (8), the order process $O_t = \tau_1 D_t - \tau_2 D_{t-1}$ with $\tau_1 = \frac{1-E\left(\phi^{L+1}\right)}{1-\phi}$ and $\tau_2 = \frac{\phi\left(1-E\left(\phi^{L}\right)\right)}{1-\phi}$. Then,

$$\begin{aligned} I_t &= I_0 + \sum_{i=l+1}^{t} \left(\tau_1 D_{t-i} - \tau_2 D_{t-i-1}\right) - \sum_{i=0}^{t} D_{t-i} \\ &= I_0 + \sum_{i=l+2}^{t} \left(\tau_1 - \tau_2 - 1\right) D_{t-i} + \tau_1 D_{t-l-1} - \sum_{i=0}^{l+1} D_{t-i} \\ &= I_0 + \tau_2 D_{t-l-1} - \sum_{i=0}^{l} D_{t-i}, \end{aligned} \tag{26}$$

as $\tau_1 - \tau_2 = 1$. Note that $I_0 = I_s + \left(E(T_p) + 1 - \tau_2\right) E(D)$ to satisfy $E(I) = I_s$.

5.2 Steady state distribution of the net inventory

We focus our analysis on the evolution of $Z_t = -\tau_2 D_{t-l-1} + \sum_{i=0}^{l} D_{t-i}$, as it determines the evolution of I_t (as $I_t = I_0 - Z_t$). Due to the autoregressive nature of D_t, the terms D_{t-i} in $\sum_{i=0}^{l} D_{t-i}$ are correlated. To avoid this correlation, we recursively substitute $D_{t-i} = (\phi D_{t-i-1} + (1-\phi)G_{t-i})$ to obtain

$$\sum_{i=0}^{l} D_{t-i} = \phi \frac{1-\phi^{l+1}}{1-\phi} D_{t-l-1} + \sum_{i=0}^{l} (1-\phi^{i+1}) G_{t-i}, \tag{27}$$

which is a function of D_{t-l-1}, the demand that was observed in the period *before* the order which is currently in service, was placed, and a sum of independent error terms G_{t-i}. This gives

$$Z_t = \frac{\phi}{1-\phi} \left(E\left(\phi^L\right) - \phi^{l+1} \right) D_{t-(l+1)} + \sum_{i=0}^{l} \left(1-\phi^{i+1}\right) G_{t-i}, \tag{28}$$

where G_{t-i} are IID random variables. Let Z be the steady state distribution of Z_t. Some care must be taken when evaluating Z, since there is still correlation between the terms that make up Z_t. From (9) we know that the terms G_{t-l} and $D_{t-(l+1)}$ determine the order size O_{t-l}. This means that these terms also affect the time that this order spends in production, and thus, the number of batches that have joined the queue after this order. Since a new order joins the queue every period, there must be l orders in the queue when order O_{t-l} is in service. We can take this correlation into account by tracking the joint probability of having an order in service with age l at the end-of-period, while $G_{t-l} = g$ and $D_{t-(l+1)} = k$. We denote these probabilities as $\Pr\left(\hat{B} = l, \hat{G} = g, \hat{D} = k\right)$, with $\hat{B}$ the limiting distribution of $l(t)$ as t goes to infinity.

In order to find these joint probabilities, we could extend the 4-dimensional Markov chain $(V_n, D_{a(n)}, C_n, S_n)$, created to find our lead time distribution, to a 6-dimensional Markov chain $(V_n, D_{a(n)-e}, D_{a(n)}, G_{a(n)}, C_n, S_n)$, which additionally tracks the error term $G_{a(n)}$ and the demand $D_{a(n)-e}$ (remember that our Markov analysis works with a time unit U, where $e \times U = 1$). However, doing so will increase the dimensions of the block matrices of the transition matrix (21) with a factor $g_{max}(d_{max} - d_{min} + 1)$. This increases the time and memory requirements of the numerical procedure to find the steady state probabilities of the corresponding Markov chain.

Instead, we derive these joint probabilities from the (known) steady state vector π of the previously used Markov chain $(V_n, D_{a(n)}, C_n, S_n)$ in a number of steps. We first determine the system state probabilities at the start of service of the $n+1$'th order, by observing the Markov chain just before the service completion of the preceding order. In the transition to the start of service of order $n+1$, we keep track of the error term $G(n+1)$, the order quantity $O(n+1)$ and the value of $D(n)$. This defines Lemma 1. Then, we observe the system at an arbitrary busy moment and derive its steady state vector. This is nearly identical to the steady state vector π, but additionally contains the values of $G(n+1)$ and $D(n)$. This results in Lemma 2. In the last step, we restrict our observation moment to arrival instants only, which corresponds to the end of a period. This allows us to determine the end-of-period probabilities $\Pr\left(\hat{B}=l, \hat{G}=g, \hat{D}=k\right)$, enabling us to find the distribution of the net inventory (Theorem 1). We refer to the Appendix for the derivation of Lemma 1 and Lemma 2.

Lemma 1 defines the system state probabilities at the start of service. Let $\bar{\pi}_{a'}(g,k,r)$ denote the probability that immediately after we start serving an order (say at time t), we observe an order with age a', an order size equal to r, while $G_{t-ea'} = g$ and $D_{t-(a'+1)e} = k$. Then,

Lemma 1 $\bar{\pi}_{a'}(g,k,r) = e\rho \sum_{a,s} \pi_a(k,1,s)\ (T^*)_s\ 1_{\{a'=[a-e]^+\}}\ p_{[r]}^{(g)}(k)$, *where* $[x]^+ = \max(0,x)$.

Given the system state probabilities at the start of service, Lemma 2 establishes an expression for the probability vector of the system at an arbitrary busy moment. Denote $\tilde{\pi}_a(g,k,r',s)$ as the probability of having an order in service with age a, with r' items of the order still remaining to be served, and with service phase s, provided that the system is busy (say at time t), while $G_{t-ea} = g$ and $D_{t-(a+1)e} = k$. Observe that $\tilde{\pi}_a(g,k,r',s)$ and $\pi_a(k',r',s)$ have a nearly identical interpretation, except that k' is the demand in the period the order in service was placed (D_{t-ea}), while k is the demand in the preceding period ($D_{t-(a+1)e}$), and g reflects the realisation of G_{t-ea}.

Lemma 2 $\tilde{\pi}_{a'}(g,k,r',s) = \frac{1}{2E(G)} \sum_{u,a,r} \bar{\pi}_a(g,k,r)\ p_{\langle s\rangle}(u,r,r')\ 1_{\{a'=a+u\}}$.

Given Lemma 1 and 2, we are in a position to compute the probabilities at arrival instants by observing that all time epochs, where the age of the customer is a multiple of e, correspond to an arrival instant. This results in the following Theorem.

Theorem 1

$$\begin{aligned} \textit{For } l>0,\ \Pr\left(\hat{B}=l,\hat{G}=g,\hat{D}=k\right) &= \rho e \sum_{r,s} \tilde{\pi}_{el}(g,k,r,s), \\ \textit{and} \qquad \Pr\left(\hat{B}=0,\hat{G}=g,\hat{D}=k\right) &= \rho e \left(\sum_{s} \sum_{a=1}^{e-1} \pi_a(k,1,s)(T^*)_s \right) \Pr(G=g). \end{aligned}$$

Observe that when $\hat{B}=0$, we use the steady state vector π of our original Markov chain instead of $\tilde{\pi}$. When an order arrives at an empty queue, the demand corresponding to the previous order is in fact the demand corresponding to the order that just finished service. This demand value can be derived from the steady state vector π.

Using Theorem 1 and Eqs. (26-28) we can find the steady state distribution of Z and the end-of-period net inventory distribution I.

Corollary 1

$$\begin{aligned}
\Pr(Z = z) &= \lim_{t\to\infty} \Pr(Z_t = z) \\
&= \sum_{b=0}^{\infty} \sum_{g_b,k} \Pr\left(\hat{B} = b, \hat{G} = g_b, \hat{D} = k\right) \cdot \sum_{g_0,g_1,\ldots,g_{b-1}} \left(\prod_{j=0}^{b-1} \Pr(G = g_j) \right) \\
&\quad \cdot 1_{\{\sum_{i=0}^{b}(1-\phi^{i+1})g_i + \phi k (E(\phi^L)-\phi^b)/(1-\phi) = z\}}.
\end{aligned}$$

Corollary 2 $\Pr(I = i) = \Pr(Z = I_0 - i)$, *with* $I_0 = I_s + (E(T_p) + 1 - \tau_2)\, E(D)$.

5.3 Determination of safety stocks

To measure customer service, we use the P2 fill rate measure, which measures the proportion of demand that can be immediately fulfilled from the inventory on hand (Zipkin, 2000). Although this is only an approximation, it is rather accurate near 100 percent fill rate (Sobel, 2004).

$$\text{Fill rate} = \frac{E\,(I)^+}{E(D)}. \tag{29}$$

The safety stock level I_s that provides a target fill rate can be found using Corollary 2.

6. Numerical experiment

In this section we use our procedure to numerically investigate the impact of autocorrelation in demand on lead times and its resulting effect on safety stocks. In the first two experiments we demonstrate how including or not including the impact of the order stream on lead times yields different results. In a third experiment we contrast the use of exogenous lead times with the use of endogenous lead times, i.e., with exogenous lead times, a lead time is arbitrarily assigned to an order; with endogenous lead times, an order's supply lead time is explicitly related with its order size and the current load of the system.

We consider a daily autoregressive demand, given by $D_t^{AR} = \phi D_{t-1} + (1-\phi)G_t$, with G uniformly distributed between 6 and 15, so that $\Pr(G = g) = 0.1$ for $g \in \{6, 7, ..., 15\}$ and $\Pr(G = g) = 0$, else. The production load at the supplier is 84%, i.e., it is available 10 hours per day and it takes on average 48 minutes to produce a single unit, with a coefficient of variation equal to 1. Orders arrive at the queue every day, or, setting the time unit of the queuing model to $U = 24$ mins, this is equivalent to an arrival every $\frac{10\cdot 60}{24} = 25$ time units. Single unit service times, expressed in time unit U, are then on average 2 time units (with a standard deviation of 2). The PH distribution matching these first two moments is represented by $\alpha = (1/3, 2/3)$ and $T = \begin{bmatrix} 2/3 & 1/3 \\ 0 & 0 \end{bmatrix}$, see

Eq. (13).

In order to study the effect of the autocorrelation, we compare the AR demand with its corresponding IID demand, i.e., the stationary distribution of D_t^{AR}. Previous studies have shown that inventory stocking levels are increasing with more positive autocorrelation in demand, and decreasing when there is more negative correlation in demand, given a random lead time independent of the order stream. In our first experiment we numerically confirm these studies. We take a random lead time distribution, e.g. the lead time distribution corresponding to a demand process G_t (which is equivalent to setting $\phi = 0$ in the above demand processes), and we treat this lead time distribution exogenously, independent of the order stream. This is, we use standard inventory theory and determine stocking levels based on the convolution of demand during the random lead time. Table 1 reports the safety stock requirements to provide a 98% fill rate, for the AR demand (I_S^{AR}) and the equivalent IID demand (I_S^{IID}) for different values of ϕ, together with the difference between both (Δ). Indeed, more negative correlation leads to lower safety stocks compared to the corresponding uncorrelated (IID) demand. Positive correlation requires higher safety stocks compared to the corresponding IID demand. We do observe, however, that in this example the difference in safety stocks is decreasing again as ϕ approaches one. Note that for both demand processes, safety stocks go down as the value of ϕ increases, which is due to the decrease in demand variance as ϕ goes to one.

ϕ	-0.3	-0.15	0	0.15	0.30	0.45	0.60	0.75
I_S^{AR}	14.82	14.67	14.52	14.35	14.13	13.86	13.53	13.19
I_S^{IID}	16.02	15.13	14.52	14.14	13.79	13.50	13.29	13.12
Δ	-1.20	-0.46	0	0.21	0.34	0.36	0.24	0.07

Table 1: Impact of serial demand correlation on safety stocks (excluding the lead time impact)

We discussed earlier in this paper (section 3.2) that in a make-to-order setting, positive correlation in demand amplifies the variability in the order stream, resulting in more variability at the production queue and hence we expect longer lead times on average. Negative correlation, on the other hand, dampens the order variability, leading to shorter lead times on the average. In contrast, the order stream under an uncorrelated, IID demand, is neither amplified, nor dampened, in variability; its orders equal the demand stream. That is why this policy is sometimes called a chase sales policy. In Table 2 we report the average lead time, $E(T_p)$, which corresponds to both the AR and IID demand process for different values of ϕ, together with their difference (Δ). We only display the average lead times, but the entire distribution is found using the procedure described in section 4 (Eqs. 22-25). The results confirm our expectations: lead times are on average shorter for AR demand compared to IID when there is negative correlation due to the dampening effect in its orders. The inverse is true for more positive correlation. In that case, the amplification in the order stream due to the autocorrelation increases average lead times compared to IID.

ϕ	-0.3	-0.15	0	0.15	0.30	0.45	0.60	0.75
$E(T_p)^{AR}$	0.5702	0.5719	0.5727	0.5714	0.5656	0.5514	0.5273	0.4949
$E(T_p)^{IID}$	0.7038	0.6291	0.5727	0.5352	0.5037	0.4785	0.4583	0.4426
Δ	-0.1336	-0.0572	0	0.0326	0.0619	0.0729	0.0688	0.0523

Table 2: Impact of serial demand correlation on average lead times

In our second experiment we take this lead time distribution, corresponding to respectively the AR and IID demand for each value of ϕ, and use it to find the safety stock requirements. Similar to our first experiment, we use the lead times exogenously, i.e., stocking levels are based on the convoluted demand during the random lead time, but here we use the lead time that results from the effective order stream that is sent to the queue. We observe the same trends as before, i.e. more positive correlation leads to higher inventories compared to its IID equivalent, and vice versa, more negative correlation now leads to much lower safety stocks than its IID equivalent due to its dampening effect on lead times. However, the difference is now much more significant, due to the reinforcing effect of the lead times. Table 3 summarizes the safety stock results when we include the lead time impact. Hence, although the same conclusions as before are still valid, we clearly observe that the lead time effect is strong and important in determining the safety stock requirements in the presence of correlation. This emphasizes that when the impact on lead times is included, ignoring the autocorrelation in demand can seriously underestimate safety stocks in the presence of positive correlation, and overestimate inventories in the presence of negative correlation.

ϕ	-0.3	-0.15	0	0.15	0.30	0.45	0.60	0.75
I_S^{AR}	14.86	14.68	14.52	14.35	14.11	13.67	13.03	12.23
I_S^{IID}	20.61	17.09	14.52	12.84	11.45	10.28	9.25	8.34
Δ	-5.75	-2.41	0	1.51	2.66	3.39	3.78	3.89

Table 3: Impact of serial demand correlation on safety stocks, including its effect on lead times

In both these experiments we used the lead time as a random, exogenous, distribution in our safety stock calculations, meaning that a given lead time realisation is arbitrarily assigned to an order and we can use the convolution of demand during the random lead time. This is actually incomplete in a true make-to-order setting, where lead times are to be treated as endogenous variables. This means that an order's supply time depends on its size and on the current load in the system at the moment it is placed. Also, in the presence of correlation in demand, it is fair to assume that its lead times also have correlation within it. In a third experiment we calculate safety stocks when we take this *endogeneity* of the lead times into account, using the extensive procedure described in section 5 to calculate the inventory distribution and corresponding safety stocks (I_S^{endo}), and we compare these results with the safety stocks when lead times are assumed to be exogenous, as we did in our previous experiment (I_S^{exo}). Table 4 reports the results for the AR demand (similar results are obtained for the IID demand). Clearly, the relaxation of the endogenous lead time assumption consistently underestimates safety stocks and consequently degrades fill rates and customer service (Δ denotes the difference between the use of exogenous and endogenous lead times), and a substantial error is incurred when this endogeneity is ignored.

ϕ	-0.3	-0.15	0	0.15	0.30	0.45	0.60	0.75
I_S^{exo}	14.86	14.68	14.52	14.35	14.11	13.67	13.03	12.23
I_S^{endo}	16.46	16.28	16.12	15.94	15.68	15.22	14.50	13.57
Δ	-1.6	-1.6	-1.6	-1.59	-1.57	-1.55	-1.47	-1.34

Table 4: Safety stock comparison for AR demand with exogenous vs. endogenous lead times

7. Concluding remarks

Much of the management science literature separates the questions of production and inventory control. However, inventory influences production by initiating orders, and production influences inventory by completing and delivering orders to inventory. Modeling a two-echelon supply chain (retailer-manufacturer) as a production/inventory system complies with this research question and explicitly analyses the interaction between the retailer's inventory and the manufacturer's production management. This results in new insights. For instance, Boute et al. (2007) have shown that an increased demand variability has a double impact on supply chain performance: it not only increases inventory variability (thereby inflating safety stocks), lead times go up as well due to the increased order variability, which reinforces the increase in safety stocks. Therefore, decoupling the inventory and production systems, thereby treating lead times as (exogenous) IID variables, underestimates the required safety stocks and consequently results in lower fill rates.

In this paper we studied the autocorrelation in demand, rather than its variability. The inclusion of autocorrelation in demand poses some additional methodological issues, compared to assuming IID demand processes. The order stream becomes dependent on the lead time distribution. Since the lead time distribution itself depends on the order stream (in a make-to-order environment), we encounter a mutual dependency problem, which we tackle through an iterative procedure. The lead time distribution at each iteration is found via a four-dimensional Markov chain, which we solve using matrix analytic methods. To determine optimal stocking levels, we explicitly take the correlation between orders and lead times into account, making use of the same Markov chain analysis. This is the methodological contribution of this paper.

Empirically, time-correlated demands are commonly observed (e.g., see Disney et al. (2006); Erkip et al. (1990)). It is a better match with real-life demand patterns in many high-tech and consumer goods industries and it is indeed used in many recent supply chain management research studies (Dong and Lee, 2003). Autocorrelated demand behavior can for instance be impacted by marketing promotions. For example, negative autocorrelation can be caused when consumers increase their purchases in a promotion period, and strongly decrease their demand in the periods preceding and subsequent to a promotion period, resulting in erratic sales. Positive autocorrelation, on the other hand, denotes a wandering, meandering sales pattern.

This paper illustrates that price control mechanisms can be used to manage inventories. Our analysis shows that when we consider the demand variance to be the same, the erratic, negatively correlated demand results in an improved supply chain performance compared to stationary independent demand, both in terms of lead times and safety stocks, whereas meandering, positively correlated sales makes performance even worse than IID. When there is positive autocorrelation in demand, the order variance is amplified compared to consumer demand, which implies that in a make-to-order environment, this increased order variance will on average result in longer supply lead times. This in turn inflates the safety stock requirements downstream in the chain. The inverse is true when demand is negatively autocorrelated. In that case there is a natural smoothing in the replenishment orders when the optimal MSE forecasting scheme is employed, with on average shorter lead times as a consequence, decreasing the safety stock requirements compared to IID. In other words, the endogeneity of lead times reinforces the impact of autocorrelation in demand on stocking levels. Ignoring this endogeneity may result in substantial errors.

This sheds new light on Sales & Operations Planning (S&OP) meetings, where sales and marketing managers decide, amongst others, on pricing their products, and link it with required inventories and production lead times, which is the responsibility of operations managers. Typically, operations managers tend to constrain the pricing flexibility for sales managers since they may create vexing ripple effects in operations. However, as we show in this paper, we need to consider both the variability and the autocorrelation in demand caused by promotions, since they both have an impact on the operational performance of the supply chain. Given the same variability, a price promotion policy leading to negatively autocorrelated demand provides better performance. It is important to note that the demand correlation as defined in this paper, is one period apart and recurring. In other words, an AR(1) demand process. The story is different when the correlation lasts for several periods, which is represented by an AR(p) process. In that case, the long periods of low demand following the promotions contribute to increased variability in addition to the autocorrelation, violating the assumptions of the paper.

Retailers can influence the level of autocorrelation in its demand stream through pricing and promotion incentives. Chen et al. (2010), e.g., discuss inventory-based dynamic pricing strategies and their impact on the demand properties. Care should be taken, however, that the pricing policy is designed in such a way that it primarily influences the level of serial correlation, rather than increasing the overall variability in demand. It happens all too often that price promotions increase the level of variability in demand, so that the overall effect on the supply chain is negative. Upasani and Uzsoy (2008) provide an overview on integrative production/marketing models and discuss the value of information sharing between marketing and production, a specific example of which is demand management through price promotions to attain smooth production plans.

In terms of further work it may be interesting to look at other methods for smoothing the orders placed on the manufacturing. For example, deliberate demand process mis-specification (Hosoda and Disney, 2009) have shown to be an effective smoothing mechanism. However this rather unorthodox approach changes the structure of the demand process placed on the manufacturer from ARMA(1,1) to ARMA(1,2). The proportional feedback controller approach as exemplified by Hosoda and Disney (2006) is also an interesting smoothing mechanism, however here the AR(1) demand is transformed into an ARMA(1,∞) process. It is hard to predict what the precise consequences of these structural changes are as the correlation in the order process has a strong effect of the behavior of the manufacturer's queue. Another smoothing mechanism that is interesting to investigate is constraints at the retailer to perhaps reflect transport capacity (Schoenmeyr and Graves, 2009). These capacity constraints would effectively produce a smoothed response, although the difference equation approach we use here to characterise the retailer's order stream would present serious difficulties – a Markovian approach may be much more insightful. Finally, Baganha and Cohen (1998) show that reduction of demand variability is possible at the plant level when it is passed through a distribution center compared to when it is received directly from replenishment orders issued by retailers. In this paper we didn't introduce a distribution center – we require retailers to smooth demand variance through the replenishment and forecasting tools we propose – but it is an interesting thought for further research.

Appendix: Derivation of Lemma 1 and Lemma 2

Lemma 1 $\bar{\pi}_{a'}(g,k,r) = e\rho \sum_{a,s} \pi_a(k,1,s)\ (T^*)_s\ 1_{\{a'=[a-e]^+\}}\ p_{[r]}^{(g)}(k)$, where $[x]^+ = \max(0,x)$.

Lemma 1 can be explained as follows. $\sum_{a,s} \pi_a(k,1,s)\ (T^*)_s$ provides the expected number of orders that complete service, with a demand k. Dividing this by the expected number of orders that start (or complete) service during an arbitrary time instant – that is, $1/e$ for a queue with $\rho < 1$ – returns the probability that the previous order has a demand k.

After service completion, the subsequent order starting service has age a'. This order has age $a' = 0$ when the previous order completes service before the next order arrival at the queue, or, equivalently, when a is smaller than the inter-arrival time e. This is because the Markov chain is only defined at time slots when the server is busy. When $a > e$, the next order was in the queue for $a - e$ time instants before starting service, and consequently $a' = a - e$.

The term $p_{[r]}^{(g)}(k)$ defines the probability that the new order in service has size r and the error term equals g, given previous demand k, or $p_{[r]}^{(g)}(k) = \Pr\left(O_{te} = r, G_{te} = g | D_{(t-1)e} = k\right)$. These probabilities look similar to (18), but in this case we are not interested in the next period's demand size k', but in the next period's order quantity r. These probabilities can be found from

$$\begin{aligned} p_{[r]}^{(g)}(k) = \Pr(G = g) \cdot \big\{ & 1_{\{r-1<\gamma k+\bar{\gamma} g<r\}} \left((\gamma k + \bar{\gamma} g) - \lfloor \gamma k + \bar{\gamma} g \rfloor\right) + \\ & 1_{\{\gamma k+\bar{\gamma} g=r\}} + 1_{\{r<\gamma k+\bar{\gamma} g<r+1\}} \left(\lceil (\gamma k + \bar{\gamma} g) \rceil - (\gamma k + \bar{\gamma} g)\right) \big\} . \end{aligned}$$

Finally, multiplying these probabilities by the average load ρ, shifts from busy time slots to all time slots.

Lemma 2 $\tilde{\pi}_{a'}(g, k, r', s) = \frac{1}{2E(G)} \sum_{u,a,r} \bar{\pi}_a(g, k, r)\ p_{\langle s \rangle}(u, r, r')\ 1_{\{a'=a+u\}}$.

$\bar{\pi}_a(g, k, r)$ defines the system state probabilities at the start of service of an order with size r. Then, if we observe the system at an arbitrary busy moment t_b, the probability that t_b falls within the service of an order of size r, is given by

$$\frac{\sum_v \Pr(O = r) \Pr(M^{r*} = v)\, v}{E(O)\, E(M)} = \frac{\Pr(O = r)\, r}{E(G)},$$

where $\sum_v \Pr(M^{r*} = v)\, v$ defines the expected service time of a batch of size r, which is equal to $E(M) \cdot r$, and $E(O) = E(G)$.

The probability that t_b is located in the u-th time epoch of a length v interval, is $1/v$. Thus, the probability of observing the system during the u-th time slot after starting service of an order of size r, means that the service has to last for at least u time slots, which implies that this probability is given by

$$\frac{\Pr(O = r)\, r}{E(G)} \left(\sum_{v \geq u} \frac{\Pr(M^{r*} = v)\, v}{E(M^{r*})} (1/v) \right) = \frac{\Pr(O = r) \Pr(M^{r*} \geq u)}{2E(G)}.$$

The term $p_{\langle s \rangle}(u, r, r')$ denotes the probability that an order of size r requires at least u time slots to complete, r' equals the number of remaining items that require service completion and s is the service phase after u time units. These probabilities are computed from the matrix T and α.

References

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

L. C. Alwan, J. J. Liu, and D. G. Yao. Stochastic characterization of upstream demand processes in a supply chain. *IIE Transactions*, 35, pp 207–219, 2003.

M. Baganha and M. Cohen. The stabilizing effect of inventory in supply chains. *Operations Research*, 46(3), pp S72–S83, 1998.

D. Bini, G. Latouche, and B. Meini. *Numerical methods for structured Markov chains.* Oxford University Press, Oxford and New York, 2005.

D. Bini, B. Meini, B. Van Houdt, and S. Steffe. Structured markov chains solver: software tools. In *Proceedings of SMCtools'06*, Pisa (Italy), 2006. ACM Press.

R. C. Blattberg and S. A. Neslin. Sales promotion models. In E. J. and L. G. L., editors, *Handbooks in Operations Research and Management Science: Marketing*, pp 553–609. North Holland, Amsterdam, 1993.

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

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

R. N. Boute, M. R. Lambrecht, and B. Van Houdt. Performance evaluation of a production/inventory system with periodic review and endogenous lead times. *Naval Research Logistics*, 54(4), pp 462–473, 2007.

G. E. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control.* Prentice Hall, Englewood Cliffs, NJ, 3rd edition, 1994.

H. Chen, O. Wu, and D. Yao. On the benefit of inventory-based dynamic pricing strategies. *Production and Operations Management*, 19(3), pp 249–260, 2010.

J. Dejonckheere, S. M. Disney, M. R. Lambrecht, and D. R. Towill. Measuring and avoiding the bullwhip effect: A control theoretic approach. *European Journal of Operational Research*, 147(3), pp 567–590, 2003.

S. M. Disney, I. Farasyn, M. R. Lambrecht, D. R. Towill, and W. Van de Velde. Taming bullwhip whilst watching customer service in a single supply chain echelon. *European Journal of Operational Research*, 173(1), pp 151–172, 2006.

L. Dong and H. L. Lee. Optimal policies and approximations for a serial multiechelon inventory system with time-correlated demand. *Operations Research*, 51(6), pp 969–980, 2003.

N. Erkip, W. H. Hausman, and S. Nahmias. Optimal centralized ordering policies in multi-echelon inventory systems with correlated demands. *Management Science*, 36(3), pp 381–392, 1990.

M. Ettl, G. Feigin, G. Lin, and D. Yao. A supply network model with base-stock control and service requirements. *Operations Research*, 48(2), pp 216–232, 2000.

S. Fotopoulos, M.-C. Wang, and S. Rao. Safety stock determination with correlated demands and arbitrary lead times. *European Journal of Operational Research*, 35, pp 172–181, 1988.

B. Gavish and S. C. Graves. A one-product production/inventory problem under continuous review policy. *Operations Research*, 28(5), pp 1228–1236, 1980.

B. Gavish and S. C. Graves. Production/inventory systems with a stochastic production rate under a continuous review policy. *Computers & Operations Research*, 8(3), pp 169–183, 1981.

S. C. Graves. Safety stocks in manufacturing systems. *Manufacturing and Service Operations Management*, 1(1), pp 67–101, 1988.

S. C. Graves. A single-item inventory model for a nonstationary demand process. *Manufacturing & Service Operations Management*, 1(1), pp 50–61, 1999.

I. Hendel and A. Nevo. Measuring the implications of sales and consumer inventory behavior. *Econometrica*, 74(6), pp 1637–1673, 2006.

T. Hosoda and S. M. Disney. The governing dynamics of supply chains: The impact of altruistic behavior. *Automatica*, 42, pp 1301–1309, 2006.

T. Hosoda and S. M. Disney. Impact of market demand mis-specification on a two-level supply chain. *International Journal of Production Economics*, 121(2), pp 739–751, 2009.

Z. Jemaï and F. Karaesmen. The influence of demand variability on the performance of a make-to-stock queue. *European Journal of Operational Research*, 164(1), pp 195–205, 2005.

J. A. Kahn. Inventories and the volatility of production. *The American Economic Review*, 77(4), pp 667–679, 1987.

G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods and Stochastic Modeling.* SIAM, Philadelphia, 1999.

H. L. Lee, V. Padmanabhan, and S. Whang. Information distortion in a supply chain: The bullwhip effect. *Management Science*, 43(4), pp 546–558, 1997.

L. Liu, X. Liu, and D. Yao. Analysis and optimization of a multistage inventory-queue system. *Management Science*, 50(3), pp 365–380, 2004.

S. Macé and S. A. Neslin. The determinants of pre- and postpromotion dips in sales of frequently purchased goods. *Journal of Marketing Research*, 16, pp 339–350, 2004.

C. F. Mela, K. Jedidi, and D. Bowman. The long-term impact of promotions on consumer stockpiling behavior. *Journal of Marketing Research*, 35, pp 250–262, 1998.

S. Nahmias. *Production and Operation Analysis.* McGraw-Hill, 3rd edition, 1997.

S. A. Neslin, C. Henderson, and J. Quelch. Consumer promotions and the acceleration of product purchases. *Marketing Science*, 4(2), pp 147–165, 1985.

M. Neuts. *Matrix-Geometric Solutions in Stochastic Models, An Algorithmic Approach.* John Hopkins University Press, 1981.

M. Neuts. *Structured Stochastic Matrices of M/G/1 type and their applications.* Marcel Dekker, Inc, New York and Basel, 1989.

C. O'Cinneide. Characterization of phase-type distributions. *Communications in Statistics: Stochastic Models*, 6(1), pp 1–57, 1990.

S. Orcun, R. Uzsoy, and K. Kempf. An integrated production planning model with load-dependent lead-times and safety stocks. *Computers and Chemical Engineering*, 33, pp 2159–2163, 2009.

J. Pahl, S. Voss, and D. L. Woodruff. Production planning with load dependent lead times. *4OR, A quarterly journal of operations research*, 3(4), pp 257–302, 2005.

J. S. Raju. The effect of price promotions on variability in product category sales. *Marketing Science*, 11(3), pp 207–220, 1992.

V. Ramaswami. Nonlinear matrix equations in applied probability - solution techniques and open problems. *SIAM review*, 30(2), pp 256–263, June 1988.

G. Riaño. *Transient behavior of stochastic networks: application to production planning with load-dependent lead times.* PhD thesis, Georgia Institute of Technology, 2002.

T. Schoenmeyr and S. Graves. Strategic safety stocks in supply chains with capacity constraints. Working Paper, Sloan School of management, M.I.T., 2009.

B. Selcuk, I. J. Adan, A. G. De Kok, and J. C. Fransoo. An explicit analysis of the lead time syndrome: stability condition and performance evaluation. *International Journal of Production Research*, 47(9), pp 2507–2529, 2009.

M. Sobel. Fill rates of single stage and multistage supply systems. *Manufacturing and Service Operations Management*, 6(1), pp 41–52, 2004.

J.-S. Song and P. Zipkin. The joint effect on leadtime variance and lot size in a parallel processing environment. *Management Science*, 42(9), pp 1352–1363, 1996.

C. R. Sox, L. J. Thomas, and J. O. McClain. Coordinating production and inventory to improve service. *Management Science*, 43(9), pp 1189–1197, 1997.

A. Upasani and R. Uzsoy. Incorporating manufacturing lead times in joint production-marketing models: A review and some future directions. *Annals of OR*, 161, pp 171–188, 2008.

X. Zhang. Evolution of arma demand in supply chains. *Manufacturing & Service Operations Management*, 6, pp 195–198, 2004a.

X. Zhang. The impact of forecasting methods on the bullwhip effect. *International Journal of Production Economics*, 88, pp 15–27, 2004b.

P. H. Zipkin. *Foundations of Inventory Management.* McGraw-Hill, New York, 2000.