# Lot sizing and lead time decisions in production/inventory systems

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

[a]*Research Center for Operations Management, University KU Leuven, Belgium*
[b]*Technology & Operations Management Area, Vlerick Business School, Belgium*
[c]*Department of Mathematics and Computer Science, University of Antwerp, Belgium*

**Abstract**

Traditionally, lot sizing decisions in inventory management trade-off the cost of placing orders against the cost of holding inventory. However, when these lot sizes are to be produced in a finite capacity production/inventory system, the lot size has an important impact on the lead times, which in turn determine inventory levels (and costs). In this paper we study the lot sizing decision in a production/inventory setting, where lead times are determined by a queueing model that is linked endogenously to the orders placed by the inventory model. Assuming a continuous review $(s, S)$ inventory policy, we develop a procedure to obtain the distribution of lead times and the distribution of inventory levels, when lead times are endogenously determined by the inventory model. We numerically show that ignoring the endogeneity of lead times may lead to inappropriate lot sizing decisions and significantly higher costs. This cost discrepancy is very outspoken if the lot size based on the economic order quantity deviates significantly from desirable production lot sizes. In these cases, the endogenous treatment of lead times is of particular importance.

*Keywords:* Production/inventory system, Lot sizing, Markov chain analysis

## 1. Introduction

A century ago, Ford Whitman Harris presented the Economic Order Quantity (EOQ) model as a simple, yet powerful model to determine how many parts to make (or order) at once, so as to balance the fixed costs per lot against the carrying costs (Harris 1913). Although various assumptions are underlying the model, the EOQ model proves to be a robust solution to many lot sizing decisions in practice. To apply the EOQ model, it is common practice to additionally define a reorder point based on the distribution of demand

*Corresponding author

*Email addresses:* `ann.noblesse@kuleuven.be` (Ann M. Noblesse), `robert.boute@kuleuven.be` (Robert N. Boute), `marc.lambrecht@kuleuven.be` (Marc R. Lambrecht), `benny.vanhoudt@ua.ac.be` (Benny Van Houdt)

during lead time, so that a fixed order quantity $Q$ (equal to the EOQ) is ordered as soon as the inventory position reaches the order point $r$.

Arrow et al. (1951) introduced a slightly modified version of this model, i.e. the $(s, S)$ inventory policy, in which an order point $s$ and an order-up-to level $S$ are established: no order is placed until inventories fall to $s$ or below, whereupon an order is placed to restore the inventory position to the level $S$. In other words, orders are placed with a lot size which is always larger than or equal to the value of $S - s$ (in many heuristics, the batching parameter $S - s$ is set equal to the EOQ), but in this case the order size is stochastic: the more the inventory level falls below $s$ (which happens, for instance, in case of a large demand size), the more the order quantity will exceed $S - s$; we call this the random *overshoot*. Several authors showed, in different settings, that an $(s, S)$ policy is optimal when a fixed order cost is present (Scarf 1960, Iglehart 1963, Veinott 1966, Porteus 1971). Today, the $(s, S)$ inventory policy is still of main importance to inventory theory and ordering policies and is incorporated in business software of many companies all over the world (Caplin and Leahy 2010).

The traditional $(s, S)$ inventory literature treats lead times exogenously with respect to the inventory policy. This means that the lot sizing decision is made in a local inventory environment, where production lead times are assumed to be exogenous and independent with respect to the lot size. Treating lead times as exogenous to the inventory model is justified when both production and inventory are decoupled through a large inventory at the production; if the owner of the production system guarantees a fixed delivery date; or if transportation lead times are much longer than production lead times (Benjaafar et al. 2005). In these environments, the inventory policy does not have a (significant) impact on lead times. For some recent examples of inventory systems with exogenous lead times, we refer to Glock (2012a) and Hoque (2013).

However, these assumptions do not hold in integrated production/inventory systems. In a production environment, there is a relationship between lot sizes and the lead times. Karmarkar (1987) shows that very small lot sizes may cause an increase in traffic intensity at the production if there is a setup time per batch, resulting in lengthy queues and long waiting times. At the other extreme, if lot sizes are very large, lead times approach an increasing function of the lot size. Therefore, in such a system, we need to take the dependency between lot sizes and lead times into account to determine the optimal lot size and reorder point parameters.

In this paper, we examine the lot sizing decision in a production/inventory environment, in which the order quantities generated by the inventory model determine the production lot sizes, and thus the (production) lead times. These lead times in turn effect the parameters of the inventory model. We show that the inclusion of endogenous lead times (as opposed to assuming that lead times are exogenous) leads to different lot sizing decisions. Ignoring the endogeneity in lead times may lead to incorrect lot sizing decisions and, as a result, to higher costs.

## 2. Model assumptions and notations

We consider a continuous time, single item production/inventory system. We assume customer demand arrives according to a compound Poisson process with a general and finite distribution of discrete demand sizes per customer. Inventory is managed using a continuous review $(s, S)$ policy to exploit the economies of scale when ordering. A finite-capacity production system produces these orders on a make-to-order basis. Under a continuous review $(s, S)$ policy, the order arrival process at the production queue consists of a combination of batch order quantities, which are stochastic (due to the stochastic overshoot) and also the time between orders is stochastic, where the order quantities and time between orders can be correlated.

One processor sequentially produces individual units on a first-come-first-served basis. We assume that each order requires a phase-type distributed setup time and the time to produce each single unit of the order is also random and phase-type distributed. We make use of the phase-type distribution, since its Markovian nature allows for an exact queueing analysis and the class of phase-type distributions is dense in the set of all positive-valued distributions, meaning any positive-valued distribution can be approximated arbitrarily close by a phase-type distribution (Latouche and Ramaswami 1987).

When the entire order is produced, it replenishes the inventory (there is no delivery until the ordered batch is completed). The time between the moment an order is placed (by the $(s, S)$ policy), and the moment it is received in inventory (after setup and production time), is the replenishment lead time. This replenishment lead time thus consists of a waiting time in queue (if the system is busy), a setup time and a production time. In other words, this lead time is stochastic and depends on the way orders are placed and its production process.

The goal is to derive the parameters of the $(s, S)$ inventory policy which minimize the expected total cost. We assume a fixed cost per order placed and a holding (resp. shortage) cost per unit in inventory (resp. short) per unit of time.

To determine the $(s, S)$ parameters that minimize the total cost in this production/inventory setting, we take into account the impact of the lot size $(S - s)$ on the lead times, as this will influence the inventory levels and thus the corresponding inventory costs. To do so, we derive the lead time distribution which results from a given set of $(s, S)$ parameters. This is the topic of Section 4. The analysis of the steady state distribution of inventory levels (at a random point in time), given a set of $(s, S)$ parameters and assuming endogenous lead times is discussed in Section 5. This analysis allows us to evaluate the expected total (inventory and ordering) cost of our $(s, S)$ controlled production/inventory model and to determine the optimal $(s, S)$ parameters that minimize this total cost function. In Section 6, we numerically illustrate the performance of our integrated approach, and compare it with the traditional local inventory approach, where lead times are assumed to be exogenous.

Throughout this paper we will adopt the following notations:

- The compound Poisson demand has arrival rate $\lambda$, and demand sizes are independent and identically distributed and follow a general discrete distribution with maximum demand size $m$. We use $d_i$ to denote the probability of a demand of size $i$, with $d_i = 0$ for $i > m$.

- Inventory is controlled by a continuous review $(s, S)$ inventory policy (consequently orders can be placed at any time); in case of a stockout, unmet demand is backlogged. Order quantities vary between $S - s$ and $S - s + m - 1$, depending on the observed customer demand prior to the moment the order was placed (which determines the overshoot).

- The probability distribution of inventory levels is defined by the probability of having $S - i$ units on hand, which we denote as $\phi_i$ with $i \in \{0, 1, \ldots\}$.

- The time needed to produce a single unit has an order $n_p$ phase-type representation with parameters $(\gamma_p, U_p)$, and the setup time has an order $n_s$ phase-type representation $(\gamma_s, U_s)$. Hence, the density function of the production and setup time is given by $\gamma_p \exp(U_p x)(-U_p e_{n_p})$ and $\gamma_s \exp(U_s x)(-U_s e_{n_s})$, respectively, where $e_n$ is a column vector of size $n$ with all its entries equal to one.

- The workload of production (without setup times) equals

$$\rho_{work} = \lambda \left(\gamma_p(-U_p)^{-1} e_{n_p}\right) \sum_{i=1}^{m} i d_i, \tag{1}$$

  with $\lambda$ the arrival rate of customers, $\sum_{i=1}^{m} i d_i$ the expected demand size per customer, and $\left(\gamma_p(-U_p)^{-1} e_{n_p}\right)$ the expected time to produce one unit. Based on $\rho_{work}$, we define the overall load/utilization as

$$\rho = \rho_{work} + (\gamma_s(-U_s)^{-1} e_{n_s})/\mu_{ot}, \tag{2}$$

with $(\gamma_s(-U_s)^{-1}e_{n_s})$ the average setup time and $\mu_{ot}$ the average time between two orders.

- We define $q_{k,n}$ as the joint probability that the current order in production is of size $k$ and $n$ demand arrivals (with random demand size) have occurred since the order in production (of size $k$) was placed. This joint probability is needed to calculate the inventory levels.

- A fixed ordering cost $K$, a penalty cost $p$ per unit backlog per time unit and a holding cost $h$ per unit in inventory per time unit are taken into account. The variable procurement cost will not be included in the cost function, as it will not influence the policy parameters (eventually all demand is met). The cost function for a given set of $(s, S)$ parameters is then defined as:

$$C(s,S) = \frac{K}{\mu_{ot}} + h\,[\Phi]^{+} + p\,[\Phi]^{-}\,. \tag{3}$$

The first term $(K/\mu_{ot})$ refers to the expected total ordering cost in a time unit, which is expressed by means of the renewal reward theorem, with $\mu_{ot}$ the average time between orders (which we will define in Section 4.2). The expected holding and penalty cost per time unit are based on $[\Phi]^{+}$, which refers to the expected number of units on hand per time unit, and $[\Phi]^{-}$, which denotes the expected number of units backlogged per time unit (see Section 5). It is worth noting that the above definition of the cost function $C(s, S)$ coincides with the cost function obtained by directly applying the renewal reward theorem on all three components that contribute to the cost:

$$C(s,S) = \frac{K + E[\text{inventory cost per cycle}] + E[\text{penalty cost per cycle}]}{\mu_{ot}}. \tag{4}$$

## 3. Literature review

There are generally two streams of literature on production/inventory systems, where the interaction between the inventory policy and the production system is taken explicitly into account. On the one hand, production/inventory models exist where inventory is managed using a base-stock policy and lead times are determined by a queueing model. On the other hand, production/inventory models exist where the lead time varies with the lot size without making use of a queueing analysis.

Jemaï and Karaesmen (2005) study a single-item make-to-stock production/inventory system, where inventory is managed by a continuous review base-stock policy. The production system produces as soon as the inventory level gets under the base-stock level and stops whenever the inventory level reaches the base-stock level. The arrival process of demand is assumed to be a general renewal process with single units demands, and production times are exponentially distributed. This results in a GI/M/1 make-to-stock queue.

Benjaafar et al. (2004) extend to a multi-item make-to-stock production/inventory system, controlled by a continuous review base-stock policy. Demand arrives in single units according to a renewal process and unit production times and setup times are i.i.d. generally distributed random variables. In this paper, the effect of product variety on inventory costs is examined, given that the production facility undergoes a setup time when it switches between different product types. Because of this setup time, replenishment orders are processed in batches. Therefore they are accumulated until the number of orders reaches the batch size. The system forms a GI/G/1 queue.

Benjaafar et al. (2005) study inventory pooling in a make-to-stock production/inventory system. Inventory is managed using a continuous review base-stock policy, and the production system is an M/M/1 queue. Extensions to a GI/M/1 queue and an M/G/1 queue are analyzed to study the impact of demand and process variability on the value of pooling.

Boute et al. (2007b) study a production/inventory model with a general demand size distribution, and lead times endogenously determined by a periodic review base-stock policy. In this setting, highly variable demand sizes increase the variability of the order process at the production queue, resulting in long lead times. They show that ignoring the impact of endogenous lead times results in a significant underestimation of the required safety stock and therefore leads to lower fill rates. In Boute et al. (2007a), the impact of order smoothing on lead times is studied; they show that safety stocks can be reduced when orders were smoothed.

Another group of literature focuses on continuous review $(r, Q)$ inventory policies and their impact on lead times. One of the first models to study the impact of order quantities on lead times and, subsequently, on safety stock requirements is due to Kim and Benton (1995). Assuming a normally distributed demand, they study a continuous review $(r, Q)$ inventory policy, with fixed order quantities $Q$ and a production lead time which varies linearly with the fixed order quantity. Waiting times in queue are assumed to account for a fixed portion of the lead time. Kim and Benton (1995) develop an iterative algorithm based on an adjusted economic order quantity to find the near-optimal order quantity $Q$ and safety stock. The authors indicate that significant savings can be realized if one takes the interrelationship between the order quantity $Q$ and the lead time into account. Some years later, a correction to the adjusted economic order quantity was made by Hariga (1999). Glock (2012b) also studies a continuous review $(r, Q)$ inventory policy with lot size-dependent lead times. He extends the work of Kim and Benton (1995) and Hariga (1999) by considering the impact of different lead time reduction strategies on total expected costs.

Whereas Kim and Benton (1995) assume the production time per unit to be constant, Cakanyildirim et al. (2000) assume a random production time per unit ($T$). Just like Kim and Benton (1995), they study the impact of lead times in a continuous review $(r, Q)$ policy, when lead times are contingent on the fixed order quantity $Q$. The possibility of economies of scale and/or learning effects is taken into account through a parameter $\theta$, by setting the processing time of $Q$ units equal to $Q^{\theta}T$. Again, waiting times are assumed to be part of the portion of the lead time which is independent of the order quantity.

Ben-Daya and Hariga (2004) consider a continuous review $(r, nQ)$ policy, where order quantities can be a multiple of $Q$, but shipments are of size $Q$. Lead times per shipment are proportional to the lot size $Q$ in addition to a fixed delay which is due to waiting times and transportation. They assume that a new order can be placed only after receiving the $n^{th}$ shipment of the previous order and that the inventory level did not cross the reorder point at the time of receiving every shipment. Demand during lead time is assumed to be normally distributed. An iterative solution procedure is suggested to find approximate solutions.

In contrast to Kim and Benton (1995) and Hariga (1999) (who assume a normally distributed demand), Al-Harkan and Hariga (2007) assume that demand has a general distribution. In this paper, the same relationship between lead time and lot sizes is adopted as in Kim and Benton (1995). Therefore, also in this paper, the waiting time is a fixed part of the total lead time and lead times are fixed for a given order quantity $Q$. The authors provide a solution method based on a combination of simulation and a search procedure.

Our paper contributes to the existing literature, since we analyze the relationship between lot sizes and lead times, but we relax the assumption that waiting times are a fixed part of the total lead time. Instead we determine waiting times endogenously by means of a queueing analysis. This explicitly takes the characteristics of the production system into account. We consider a continuous review $(s, S)$ policy and we develop a methodology to determine the parameters in an $(s, S)$ inventory policy that minimize expected total costs per time unit, taking the impact of endogenous production lead times into account. Another distinction to the existing literature is that in an $(s, S)$ policy, the order quantities are random due to the random overshoot, which contrasts the $(r, Q)$ policy, where the order quantity is constant.

## 4. Queueing analysis to determine the lead time distribution

We start by characterizing the distribution of lead times that correspond to a given set of $(s, S)$ inventory parameters. To do so, we develop a queueing model. The arrival process of orders at the queue is characterized by stochastic batch sizes and stochastic inter-arrival times, and a correlation between order quantities and the time between orders might exist[1] (depending on the chosen demand size distribution).

Because of this arrival process, the distribution of lead times is not straightforward and standard queueing formulas are not available. We therefore develop a four dimensional Markov process that characterizes this production/inventory system and can be used to derive the distribution of the time an order spends in the production system (recall that this time corresponds to a queueing time, a setup time and an effective production time).

*4.1. Markov process to characterize the production/inventory system*

Define the four dimensional continuous-time Markov process $(Y_t, W_t, R_t, Z_t)_{t\geq 0}$ with:

- $Y_t$ the time spent in the system (in queue, setup, and production) of the order currently in production at time $t$ ($Y_t \geq 0$),
- $W_t$ the overshoot of the order in production at time $t$ ($0 \leq W_t \leq m-1$),
- $R_t$ the number of units of the current order that still need to start/complete production at time $t$ ($0 \leq R_t \leq S-s+m-1$),
- $Z_t$ the current phase of the unit in production (or setup) at time $t$.

Note, we allow the random variable $R_t$ to attain the value of zero to incorporate the setup times. More specifically, the Markov process is defined such that the setup is performed after producing all units of an order. Whether the setup is performed first or last has no impact on the lead time and inventory level distributions, while it simplifies the description of the Markov process a little. Thus, when $R_t = 0$ the range of $Z_t$ is $\{1, \dots, n_s\}$, while for $R_t > 0$ the range of $Z_t$ is $\{1, \dots, n_p\}$. Further, the value of $R_t$ is clearly limited by $S - s + W_t$ for all $t$. Nevertheless, we set the range of $R_t$ equal to $\{0, \dots, S-s+m-1\}$. This implies that some of the states of the process are in fact transient, but this has no impact on the results as the corresponding steady state probabilities of these transient states will be equal to zero. It is possible to exclude these transient states at the expense of complicating the description of the Markov process.

The three dimensional Markov process $(Y_t, R_t, Z_t)$ is in fact sufficient to derive the distribution of the lead times, but in Section 5, we will use the same Markov process to compute

[1]To illustrate this possible correlation, assume for instance that $S - s = 3$ and demands arrive every time unit with a discrete demand size distribution with $d_2 = 0.5$ and $d_3 = 0.5$. In that case there is a positive correlation between order quantities and time between orders: if the order quantity equals 3, then the time between orders equals one time unit. If the order quantity is 4 or 5, the time between orders equals two time units.

the probability distribution of the inventory levels, for which it is helpful to keep track of the order size (which is determined by $W_t$) in combination with its time spent in the system ($Y_t$).

If there is no order in production, $Y_t$, $W_t$, $R_t$ and $Z_t$ are all set equal to zero. As long as an order is being produced, $Y_t$ increases linearly over time. At some point in time, a transition in $Z_t$ or $R_t$ can occur, i.e., when resp. the phase of the unit in production changes (recall that we assume a phase-type distribution for the production time per unit and for the setup time of the order) or when one unit of the order completes production (and the next unit of the same order starts production). When production of the entire order is completed, the next order starts production, and the value of $Y_t$ decreases to the time that this new order has spent so far in the system. Indeed, as the orders are produced in sequence of arrival (FIFO), no crossovers exist and the time spent in queue of the new order will be less than the lead time of the previous order (and thus a decrease in $Y_t$ occurs). The amount of the decrease is actually equal to the inter-arrival time between this order and the previous one. If production is completed and the facility is empty (there is no queue of orders), $Y_t$ decreases to zero.

From the Markov process $(Y_t, W_t, R_t, Z_t)_{t\geq 0}$, which observes the production system at any point in time, we define another Markov process $(X_t = Y_t, L_t = (W_t, R_t, Z_t))_{t\geq 0}$ which observes the production facility only when there is a unit in production (or in setup). This new Markov process skips all the time intervals during which $Y_t$, $W_t$, $R_t$ and $Z_t$ are zero in the original Markov process $(Y_t, W_t, R_t, Z_t)$. The resulting Markov process is a bivariate process $(X_t, L_t)_{t\geq 0}$, with $X_t \geq 0$ and $L_t \in \mathcal{L} = \{(w, 0, z)|w = 0, \ldots, m-1, z = 1, \ldots, n_s\} \cup \{(w, r, z)|w = 0, \ldots, m-1, r = 1, \ldots, S-s+m-1, z = 1, \ldots, n_p\}$. Denote $l = |\mathcal{L}| = m\left[(S-s+m-1)\,n_p + n_s\right]$. The process $(X_t, L_t)_{t\geq 0}$ belongs to the class of Markov processes with a matrix exponential distribution of order $l$ (Sengupta 1989, 1990), where the class of matrix exponential distributions extends the class of phase-type distributions.

In this Markov process $(X_t, L_t)$, the time spent in the system $X_t$ increases linearly over time unless one of the following three events occurs (starting from $(x, i)$, with $i \in \mathcal{L}$):

1. The current order remains in service, but the production or setup phase changes or a unit of the order completes production and, when this is not the last unit of the order that needs to be produced, the production of the next unit of the same order (or setup) starts. In that case, a transition to $(x, j)$ occurs; we denote its rate matrix as $(A_0)_{i,j}$ (for $i \neq j \in \mathcal{L}$).
2. The entire order is produced and completes its setup (meaning the order will be replenished in inventory). When the inter-arrival time of the subsequent order is at most $u$ time units, a downward jump in $X_t$ (the time spent in the system) occurs to some

value in the interval $([x-u,x),j)$, for $0<u<x$. We denote its rate as $A_{i,j}(u)$ and let $dA_{i,j}(u)$ reflect its density function. The matrix $A(u)$, containing the rates $A_{i,j}(u)$ of a transition from $i$ to $j$, takes the correlation between $j$ (which contains the order size of the new order) and $u$ (the inter-arrival time between this new order and the previous one) into account. Given that $u<x$, the next order has spent at least some time $x-u$ in queue and a jump in the interval $[x-u,x)$ occurs.

3. The entire order is produced and completes its setup and after production of the order, the queue is empty. In that case a downward jump in $X_t$ to $(0,j)$ occurs and we denote its rate as $\int_{u=x}^{\infty} dA_{i,j}(u)$. $\int_{u=x}^{\infty} dA_{i,j}(u)$ defines the rate to jump from state $(x,i)$ to $(0,j)$, which occurs if the inter-arrival time between the current and the next order is larger than or equal to the lead time of the current order (which implies that the queue is empty at the time of production completion of the current order).

We define the (negative) diagonal entries of $A_0$ such that $(A_0+\int_{u=0}^{\infty} dA(u))e_l=0$. The production rate matrix $A_0$ is the transition matrix when the current order remains in service, and is defined by the sequence of production times of every single unit in the order, and a setup time once the entire order is completed. Therefore, per unit in the order, we have transitions in the production phase, characterized by $U_p$ when the same unit continues its service, and $u_p\gamma_p$, when a unit ends production and the next unit of the batch starts production (with $u_p=-U_p e_{n_p}$ denoting the unit completion rate), and for the last unit in production we start the setup time (characterized by $\gamma_s$ and $U_s$). As we additionally want to keep track of the original order size in $A_0$ (which is required to determine the inventory levels in a later phase), we use the Kronecker product to multiply these transitions with the $(m\times m)$ identity matrix $I_m$, as we have $m$ possible order sizes (ranging from $S-s$ to $S-s+m-1$). This way we obtain a matrix filled with zeros, except for the $m$ sub-matrices on the diagonal; each of those $m$ sub-matrices defines the transitions in production and setup of an order with a specific order quantity. Hence the matrix $A_0$ is defined as an $l\times l$ rate matrix, with every of the $m$ sub-matrices an $[(S-s+m-1)\,n_p+n_s]\times[(S-s+m-1)\,n_p+n_s]$ matrix:

$$A_0 = I_m \otimes \begin{bmatrix} U_s & & & & 0 \\ u_p\gamma_s & U_p & & & \\ & u_p\gamma_p & \ddots & & \\ & & \ddots & \ddots & \\ 0 & & & u_p\gamma_p & U_p \end{bmatrix}. \tag{5}$$

We define $\eta$ as the rate of production completion of an order:

$$\eta = e_m \otimes u_{fin}, \text{ with } u_{fin} = \begin{bmatrix} u_s \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{6}$$

and $u_s = -U_s e_{n_s}$ the completion rate of the setup time. Similar to the production matrix $A_0$, consisting of $m$ sub-matrices each referring to the production of a specific order quantity, the $l \times 1$ production completion vector $\eta$ can also be seen as a superposition of $m$ different vectors, $u_{fin}$, which defines the production completion rate for a given order quantity. This explains the Kronecker product in (6).

The matrix $dA(u)$ is the density function of the matrix $A(u)$, which characterizes an order completion: it holds the rates to go from one state, just before production completion of the order, to another state observing the start of production of the subsequent order, with an inter-arrival time up to $u$ between both orders. The matrix $dA(u)$ is defined as

$$dA(u) \quad = \quad \eta\alpha \exp(C_0 u) C_1, \tag{7}$$

with $\eta$ the rate of production completion of an order (Eq. (6)); the term $\alpha \exp(C_0 u)$ denotes the density of the inter-arrival time $u$ (which corresponds to the time between two subsequent orders); and the matrix $C_1$ defines the transitions to the next order that starts production. We now describe the time between orders (defined by the vector $\alpha$ and the matrix $C_0$) and the matrix $C_1$ in the following subsection.

### *4.2. Characterization of the time between orders*

To determine the time between subsequent orders, we characterize an $(s, S)$ inventory policy by a continuous time Markov chain with the states $j \in \{1, \ldots, S - s - 1\}$ referring to the inventory positions ranging from $S$ to $s + 1$. We define state $j$ as the *modified* inventory position which corresponds to the inventory position $S + 1 - j$. The time between subsequent orders is equivalent to the time it takes for the Markov chain to decrease from inventory position $S$ until absorption in the order point $s$ or below. As at the start of a cycle (when an order is placed), the inventory position is raised to $S$, the initial vector $\alpha$ of this Markov chain is defined as $[1, 0, \ldots, 0]$.

Define $C_0$ as the $(S - s) \times (S - s)$ rate matrix of this Markov chain, which contains the rates of inventory position changes when no order is placed. A transition from state $j$

to state $j' > j$ occurs when demand depletes inventory (which implies that the inventory position decreases from $S+1-j$ to $S+1-j'$); its rate $(C_0)_{j,j'}$ is defined by the demand rate $\lambda$ and the probability $d_{j'-j}$ of observing a demand of size $j'-j$. As inventory positions cannot increase, unless an order is placed, the rates below the diagonal of matrix $C_0$ are zero. Hence,

$$C_0 = \lambda \begin{bmatrix} -1 & d_1 & \dots & \dots & d_{S-s-1} \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ & & & \ddots & d_1 \\ 0 & & \dots & 0 & -1 \end{bmatrix}. \tag{8}$$

With an $(s,S)$ policy, an order is placed after a number of demands have arrived (i.e., when the aggregated sum of demand sizes exceeds the value of $S-s$). Given that demand follows a compound Poisson distribution, the time between (batch) demand arrivals is exponentially distributed. As the number of batch arrivals prior to order placement is variable, the time between subsequent orders is characterized by a phase-type distribution. For more information on phase-type distributions, we refer to the work of Latouche and Ramaswami (1987). The density of time between subsequent orders is defined by the density function $f(u) = \alpha \exp(C_0 u)(-C_0 e_{S-s})$. The average time between two orders, $\mu_{ot}$, can then be found using the average of this phase-type distribution: $\mu_{ot} = -\alpha (C_0)^{-1} e_{S-s}$.

The $(S-s) \times l$ matrix $C_1$ describes the transition rates of going from inventory position $i \in \{S, ..., s+1\}$ just before an order is placed (with $i$ the $S-i+1^{th}$ row of matrix $C_1$), to state $j \in \mathcal{L}$ when production of an order starts (the $j^{th}$ column of matrix $C_1$). It is defined as

$$C_1 = \sum_{k=1}^{m} \lambda \begin{bmatrix} 0 \\ \vdots \\ 0 \\ d_m \\ \vdots \\ d_k \end{bmatrix} v_{k-1}, \tag{9}$$

where $\lambda$ refers to the arrival rate of demand (triggering the order), the demand size distribution $d_i$ determines the different order quantity probabilities (and the overshoot $W_t$), and the vector $v_{k-1}$ of length $l$ defines the phase in which the first unit in production starts and corresponds to the first unit which is going to be produced of an order with size $S-s+k-1$

with $1 \leq k \leq m$. As $v_{k-1}$ refers to one unit of a specific order quantity (i.e., the first unit of an order of size $S - s + k - 1$), all its entries are equal to zero, except for the $n_p$ entries $\{(k-1)\, l/m + n_s + (S - s + k - 2)n_p + 1, \dots, (k-1)\, l/m + n_s + (S - s + k - 1)n_p\}$, which equal the initial vector $\gamma_p$, as those entries correspond to the states of the first unit of the order quantity $S - s + k - 1$. (Observe that the zero-entries correspond to the states of the remaining $(S-s+k-2)$ units of this order that need to be produced (and its $n_p$ phases), the $n_s$ phases of the setup time, and the states of the order sizes, different from $S - s + k - 1$.)

The vector $v_{k-1}$ ensures that a transition occurs to the right sub-matrix of the production matrix $A_0$, i.e., the sub-matrix that corresponds to the size of the order placed. For example, the transition probabilities in production of an order with size $S - s + 1$ is found in the second sub-matrix of the production matrix $A_0$ (the first sub-matrix refers to orders of size $S - s$, whereas the last sub-matrix refers to orders of size $S - s + m - 1$). Thus, the first $(S - s + m - 1)\, n_p + n_s$ entries of the vector $v_1$ (referring to orders of size $S - s$) are zero; the subsequent $n_s$ entries of $v_1$ are also zero, as they refer to the setup time of the order; the following $(S - s)\, n_p$ entries are as well zero, as they refer to the $S - s$ remaining units that need to be produced; the subsequent $n_p$ entries are non-zero and contain the probability in which phase the first unit in production starts (which is given by $\gamma_p$). The remaining $(m-2)\,((S - s + m - 1)\, n_p + n_s)$ entries of vector $v_1$ are zero, as they refer to larger order quantities ranging from $S - s + 2$ to $S - s + m - 1$.

### *4.3. Distribution of the time in the system*

Due to Sengupta (1989), the length $l$ vector $\delta(x)$ (for $x \geq 0$) holding the steady-state density of the states $\{(x, i) | i \in \mathcal{L}\}$ for any time spent in the production system $x \geq 0$ exists if and only if $\rho < 1$ and can be written as

$$\delta(x) = \delta(0) exp(Tx), \tag{10}$$

where the $l \times l$ matrix $T$ is the smallest non-negative solution to

$$T = A_0 + \int_{x=0}^{\infty} exp(Tx) dA(x), \tag{11}$$

and the initial vector $\delta(0) = \tau(-T)$, with $\tau$ the unique invariant vector of $A = A_0 + \int_{u=0}^{\infty} dA(u)$, i.e., $\tau A = 0$ and $\tau e_l = 1$. From (10), the steady state densities for the lead time $\pi(x)$ can then be derived as

$$\pi(x) = \frac{\delta(x)\eta}{\int_{y=0}^{\infty} \delta(y)\eta dy} = \frac{\tau(-T)exp(Tx)\eta}{\tau\eta}, \tag{12}$$

and the cumulative distribution function $\Pi(x)$ of the lead time is given by

$$\Pi(x) = \frac{\int_{y=0}^{x} \delta(y)\eta dy}{\int_{y=0}^{\infty} \delta(y)\eta dy} = 1 - \frac{\tau e^{Tx}\eta}{\tau\eta}. \tag{13}$$

We could, in principle, compute $T$ iteratively using Eq. (11) by setting $T_0 = 0$ and

$$T_{n+1} = A_0 + \int_{x=0}^{\infty} \exp(T_n x) dA(x), \tag{14}$$

in order to obtain the distribution of the time in the system $\delta(x)$ and the lead time distribution $\pi(x)$. However this method to compute $T$ results in linear convergence only, making it impractical for high loads. To obtain an algorithm with quadratic convergence, we construct a fluid queue (Latouche 2006) as follows. As long as an order is in production, the time spent in the system ($X_t$) increases linearly; as soon as production completion of this order has occurred, the time spent in the system decreases either to zero (if the queue is empty at the time of production completion of the current order) or to the time spent waiting in queue of the subsequent order (if the queue is non-empty at the time of production completion of the current order, the decrease in the time spent in the system is then equal to the inter-arrival time between the current and the next order). In order to obtain a fluid queue, we define $X_t$ as the fluid and replace these (instant) decreases by intervals of the appropriate length during which the fluid decreases linearly (i.e., to zero for empty queues or to the waiting time for non-empty queues). This way, we obtain a fluid queue with $l$ phases in which the fluid increases (our original $l$ states, as units of the same order are produced) and $(S - s)$ phases in which the fluid decreases (the time between orders placed corresponds to the decrease in inventory positions from $S$ to $s + 1$).

Let $F_{++}$ hold the transition rates at which the phase changes while the fluid increases (i.e., production of the order continues, therefore time spent in the system increases linearly), $F_{+-}$ the rates when the fluid changes from an increase to a decrease (i.e., upon production completion of the order), $F_{-+}$ the rates when the fluid changes from a decrease to an increase (a new order starts production) and $F_{--}$ the rates when the fluid continues to decrease (i.e., when demands deplete inventory, but the order point $s$ is not yet reached, so the time between orders increases). Then, $F_{++} = A_0$, $F_{--} = C_0$, $F_{-+} = C_1$ and

$$F_{+-} = \eta\alpha. \tag{15}$$

The matrix $F$ defined as

$$F = \begin{bmatrix} F_{++} & F_{+-} \\ F_{-+} & F_{--} \end{bmatrix}, \tag{16}$$

is the $(S-s+l) \times (S-s+l)$ rate matrix of the underlying continuous-time Markov chain of the fluid queue.

If we take the expression for the steady state of a fluid queue (Latouche 2006) and observe the queue only when the level increases, one finds that its steady state $\delta(x)$ has a matrix exponential form $\delta(x) = \delta(0)\exp(Tx)$, with $T = F_{++} + \Psi F_{-+}$ where $\Psi$ is the minimal nonnegative solution to an algebraic Riccati equation (Latouche 2006). Thus, to compute $T$, it suffices to determine $\Psi$ and this can be done in a very efficient way as we employ the Structure-preserving Doubling Algorithm (SDA), which is discussed in Appendix A. Finally, to compute $\tau$, the invariant vector of $A$, we note that $A$ can be expressed as $A = A_0 + F_{+-}(-C_0)^{-1}C_1$.

## 5. The probability distribution of inventory levels at a random point in time

The Markov process defined in Section 4 enables us to compute the distribution of the inventory levels at an arbitrary point in time, which is needed to determine the expected inventory holding and shortage costs. We denote $\phi_i$ as the probability of having $S-i$ units on hand, where $i \geq 0$. We make a distinction between two different situations: either the production is busy, or it is idle.

In case the production is busy and the current order in production is of size $k$ (with $k \in \{S-s, ..., S-s+m-1\}$), we know that the current inventory level is at most $S-k$. More specifically, the inventory level equals exactly $S-k$ in case no demand has occurred since this order of size $k$ was placed; if one customer would have arrived with a demand size of $k_1$ units, then the current inventory level would be depleted to $S-k-k_1$. In other words, to determine the current inventory level when production is busy, we need to know the size of the order in production and the total number of units demanded since the order in production was placed (i.e., the demand during the time that the order has so far spent in the production system).

We start by computing the joint probabilities $q_{k,n}$ that the current order in production is of size $k$ (with $k$ ranging from $S-s$ to $S-s+m-1$) and $n$ Poisson demand arrivals have occurred since this order was placed, given that production is busy. The time spent in the production system is calculated based on the matrix exponential form of the steady state of $(X_t, L_t)_{t\geq 0}$, established in (10), and the probability of $n$ demands arriving during this period $x$ is given by $\frac{(\lambda x)^n}{n!}\exp(-\lambda x)$. Then, the vector $q_n = (q_{S-s,n}, \ldots, q_{S-s+m-1,n})$ can be expressed as:

$$\begin{aligned} q_n &= \delta(0)\int_{x=0}^{\infty}\exp(Tx)\frac{(\lambda x)^n}{n!}\exp(-\lambda x)dx(I_m\otimes e_{l/m}) \\ &= \delta(0)\lambda^n(\lambda I_l - T)^{-(n+1)}(I_m\otimes e_{l/m}). \end{aligned} \tag{17}$$

The Kronecker product $I_m \otimes e_{l/m}$ ensures that all entries which relate to the same order quantity are summed. For example, the first $(S-s+m-1)\,n_p+n_s$ entries of the $1\times l$ vector are summed: the sum of them refers to the probability of having $n$ demands during the time that the order currently in production has so far spent in the system and the current order quantity equals $S-s$. The Kronecker product also sums the second up to the $m^{th}$ group of $(S-s+m-1)\,n_p+n_s$ entries, such that $q_n$ is a $1\times m$ vector with an entry for the joint probabilities of $n$ demands and all $m$ possible order quantities.

Given the joint probability $q_{k,n}$ we can compute the probability $\phi_i$ that the number of units on hand equals $S-i$, for $i\geq 0$, at an arbitrary point in time as follows. When the server is busy (with probability $\rho$), we determine $\phi_i$ using $q_{k,n}$, and additionally take the demand size into account for each of the $n$ Poisson demand arrivals (which is the $n$-fold convolution of the i.i.d. demand size distribution).

When the server is idle (with probability $1-\rho$), we can compute the probabilities of having $S-i$ units on hand (for $i\in\{0,...,S-s-1\}$) as the steady-state vector $g$ of the fluid queue, given that the amount of fluid is zero. By definition, if the server is empty, the next order arriving at the production queue does not have to wait in queue and the fluid is zero. This stochastic vector $g$ can be computed as $g(F_{--}+F_{-+}\Psi)=0$ (Latouche 2006) .

This leads to the probability distribution of the inventory levels: for $i\geq 0$,

$$\phi_i = (1-\rho)g_{i+1} + \rho\sum_{k\leq i, n\leq i-k} q_{k,n}\sum_{\substack{k_1,\ldots,k_n>0\\ k_1+\ldots+k_n=i-k}}\left(\prod_{s=1}^{n} d_{k_s}\right). \tag{18}$$

From (18), we can derive the expected number of units in stock and the expected number of units backordered at a random point in time:

$$[\Phi]^+ = \sum_{i=0}^{S}\phi_i\,(S-i)\,, \tag{19}$$

$$[\Phi]^- = \sum_{i=S+1}^{\infty}\phi_i\,(i-S)\,, \tag{20}$$

which enables us to determine the expected total cost per time unit for a given $(s,S)$ inventory policy (Eq. 3).

## 6. Numerical illustration

In this section we show how our methodology can be applied to find the optimal inventory parameters within the class of continuous review $(s, S)$ policies that minimize the expected ordering and inventory related costs per time unit. We numerically illustrate the impact of the inventory parameters on the lead times in a production/inventory setting, which will in turn affect total inventory costs. More specifically, the lot sizing decision (i.e., the value of $S - s$) determines the arrival process of orders at the queue and thus the lead times. The order point $s$ by itself does not impact this arrival process (and thus lead times), but it does affect the safety stock and the inventory costs. As we will show, ignoring this endogenous lead time effect may lead to inappropriate lot sizing decisions and excessive costs.

In our numerical experiment, we assume demand arrives according to a compound Poisson process with on average 0.2 customer arrivals per hour. Each customer has a random demand size which is distributed according to a zero-truncated binomial distribution ($B \sim (20, 0.5)$) with on average 10 units per customer and a maximum demand size $m = 20$. The probability distribution of demand sizes is plotted in Figure 1. We assume that the setup times per order are exponentially distributed with an average of two hours per order, and production times per unit are also exponentially distributed with on average 15 minutes per unit. The average production load $\rho$ depends on the number of orders placed (and thus on the number of setups), and equals 90% for $S - s = 1$, and it decreases for larger values of $S - s$ (e.g. for $S - s = 50$, $\rho = 57\%$).

In Figures 2 and 3 we show how the average lead time and the lead time variance are impacted by the value of $S - s$ (note that our procedure computes the entire lead time distribution, but we only illustrate its first two moments). In Figure 2, we observe two effects: on the one hand, as lot sizes increase, expected production times increase linearly as a function of the lot size; on the other hand, reducing $S - s$ causes an increase in the ordering frequency, which causes an increase in the utilization rate and waiting times at the production queue, as every order needs to undergo a setup time. Figure 3 shows that the variance in lead times decreases as $S - s$ increases. (Observe that, for $S - s < 4$, reducing $S - s$ only has a marginal impact on lead times, since the probability of demand sizes smaller than four units is almost zero, and the $(s, S)$ policy acts very similar to an order-up-to policy for $S - s < 4$.)

Traditionally, the batching parameter $S - s$ is often set equal to the economic order quantity. We provide a small example which is based on the following cost assumptions: the holding cost per unit per hour $h = 1$, the penalty cost per unit per hour $p = 9$ and the fixed distr_demand.jpg

Figure 1: Probability distribution of demand sizes

exp_lt.jpg

Figure 2: Expected lead time in function of the batching parameter $S - s$

var_lt.jpg

Figure 3: Variance of lead times in function of the batching parameter $S - s$

ordering cost $K = 10$ (this cost example refers to case 1 in Table 1). In this example, the EOQ equals 6.32. Therefore, we set $S - s$ equal to 6 or 7. The lead time distribution that corresponds to $S - s = 7$ is plotted in Figure 4.

lt_eoq.jpg

Figure 4: Probability density function of the lead time distribution

The order point $s$ is set to minimize the expected total inventory cost per hour. When we treat the lead time distribution of Figure 4 as exogenous (and assuming no crossover in orders), the distribution of inventory levels ($\Phi$) can be computed from the distribution of inventory positions ($IP$) and demand during lead time ($DDLT$) (Zipkin 2000):

$$\Phi = IP - DDLT. \tag{21}$$

For $S - s = 7$, we find that $(s, S)$ equal to $(108, 115)$ minimizes expected total cost per hour. From Figures 2 and 3, we can see that for $S - s = 6$ or 7, the average lead time and its variance are quite high, which results in high safety stocks and a high order point $s$.

The economic order quantity makes a trade-off between the fixed ordering cost and the inventory costs, but does not explicitly take its impact on lead times into account. As one can see from Figures 2 and 3, increasing the value of $S - s$ could, in this example, decrease the expected lead time and the variance of lead times, and therefore result in lower inventory costs. When we use the procedure developed in this paper to include the impact of the lot sizing decision on lead times and inventory levels, we find that the inventory related costs can indeed be reduced by increasing the batching parameter $S - s$ (in our example to a value around 25 to 30). Figure 5 shows the contour plot of the expected inventory related cost (holding and penalty cost) per hour. It contains the contour lines, which connect the combinations of inventory parameters where the expected inventory cost has the same value. When the lines are close together, the magnitude of the gradient is large: a small deviation in the inventory parameters then causes a steep change in the expected inventory cost. Figure 5 shows that the gradient is particularly large for $S - s < 10$, which indicates that high costs are observed for values of $S - s < 10$, due to the high utilization rate in production for these small lot sizes, resulting in long and variable lead times and thus high inventory related costs.

The ordering costs are decreasing in the lot size $S - s$. Adding the inventory related costs and the ordering costs, we can then determine the expected total cost per hour for a given $(s, S)$ policy (Eq. (3)). Based on a search procedure, we find that setting the $(s, S)$ parameters equal to $(36, 62)$ minimizes the expected total cost per hour. This policy has an expected total cost of 42.94. Observe that in this case, $S - s = 26$, which results in a shorter average lead time and a smaller lead time variance (see Figures 2 and 3), compared to the lot size based on the EOQ. As lead times are shorter and less variable, the order point $s$ is considerably lower (36 compared to 108), leading to a major reduction in safety stocks and inventory costs. The cost difference with the traditional approach based on the EOQ is significant: if we evaluate the policy $(108, 115)$ in our production/inventory setting, the resulting expected total cost per hour is 116.88.

We extend our numerical example to different cost scenarios (see Table 1). Table 2 shows the results. The first two columns provide the optimal $(s, S)$ parameters, and their corresponding total costs, when the endogeneity in lead times is taken into account in finding the $(s, S)$ parameters that lead to the lowest costs. The third and fourth column provide the $(s, S)$ parameters, and their corresponding total costs, when $S - s$ is set equal to the EOQ (rounded to the integer which results in the lowest total costs), and $s$ is set to minimize holding_case1.jpg

Figure 5: Expected holding cost and penalty cost per hour in function of the order point $s$ and the batching parameter $S - s$

| | h | p | K |
|---|---|---|---|
| Case 1 | 1 | 9 | 10 |
| Case 2 | 1 | 19 | 10 |
| Case 3 | 1 | 9 | 50 |
| Case 4 | 1 | 19 | 50 |

Table 1: Different cost scenarios.

| | $(s, S)$ Endogenous | Costs Endogenous | $(s, S)$ EOQ | Costs EOQ |
|---|---|---|---|---|
| Case 1 | (36,62) | 42.9430 | (108,115) | 116.8844 |
| Case 2 | (46,74) | 53.1630 | (143,150) | 151.3624 |
| Case 3 | (35,63) | 45.4753 | (41,56) | 51.4359 |
| Case 4 | (46,76) | 55.5563 | (54,69) | 64.2507 |

Table 2: $(s, S)$ parameters and corresponding costs when lead times are considered endogenous, versus the traditional EOQ approach.

expected inventory costs, assuming the lead time distribution that corresponds to the EOQ.

Clearly, setting the inventory parameters based on the EOQ, and ignoring its impact on lead times, may result in significantly higher expected total costs per hour when the EOQ is not desirable for the production environment. However, the more the EOQ approaches the optimal production lot size (see Figures 2 and 3), the smaller its cost difference will be. For instance, in case 3 and 4, EOQ=15, which results in relatively short and smooth lead times, and thus the discrepancy between the EOQ approach and the endogenous lead time approach is smaller than in case 1 and 2.

## 7. Conclusion

In this paper, we study the lot sizing decision in a production/inventory system with endogenous lead times. Lot sizes are often set equal to the economic order quantity, in order to make a trade-off between fixed ordering costs and inventory costs. However, the batching parameter $S - s$ determines the lead time distribution in production, which in turn impacts inventory costs. Therefore, when minimizing total (inventory and ordering) costs, setting $S - s$ equal to the economic order quantity can be far from optimal.

We provide a method based on a Markov chain approach, which allows to compute the distribution of the time an order spends in the production system (i.e., in queue, in setup and in production) and the distribution of inventory levels. In a numerical analysis, we show that ignoring the endogeneity of lead times in a production/inventory system may lead to inappropriate lot sizing decisions and higher expected total costs. This cost discrepancy is very outspoken if EOQ values deviate significantly from desirable production lot sizes. In these cases, the endogenous treatment of lead times is of particular importance.

## Appendix

In order to obtain the minimal non-negative solution $\Psi$ of an algebraic Riccati equation, we employ the Structure-preserving Doubling Algorithm (SDA) of Guo et al. (2007) outlined below. First define $A = -F_{++}, B = F_{+-}, C = F_{-+}$ and $D = -F_{--}$. Next, set $\gamma = \max\{\max_i a_{ii}, \max_i d_{ii}\}$ and let $A_\gamma = A + \gamma I$ and $D_\gamma = D + \gamma I$. Further, let $W_\gamma = A_\gamma - BD_\gamma^{-1}C$ and $V_\gamma = D_\gamma - CA_\gamma^{-1}B$.

Next, the SDA algorithm initializes $E_0$, $F_0$, $G_0$ and $H_0$ as $E_0 = I - 2\gamma V_\gamma^{-1}$, $F_0 = I - 2\gamma W_\gamma^{-1}$, $G_0 = 2\gamma D_\gamma^{-1} C W_\gamma^{-1}$ and $H_0 = 2\gamma W_\gamma^{-1} B D_\gamma^{-1}$. Finally, the iteration

$$
\begin{aligned}
E_{k+1} &= E_k(I - G_k H_k)^{-1} E_k, \\
F_{k+1} &= F_k(I - H_k G_k)^{-1} F_k, \\
G_{k+1} &= G_k + E_k(I - G_k H_k)^{-1} G_k F_k, \\
H_{k+1} &= H_k + F_k(I - H_k G_k)^{-1} H_k E_k,
\end{aligned}
$$

guarantees that $G_k$ converges quadratically[2] to $\Psi$. The iteration is repeated until $\min(\|E_k\|_1, \|F_k\|_1) < 10^{-15}$. The computation time of SDA can be further reduced by means of the ADDA algorithm Wang et al. (2011), which uses the same iteration as SDA, but initializes $E_0, F_0, G_0$ and $H_0$ using two parameters $\alpha = \max_i a_{ii}$ and $\beta = \max_i d_{ii}$.


Al-Harkan, I., M. Hariga. 2007. A simulation optimization solution to the inventory continuous review problem with lot size dependent lead time. *Arabian Journal for Science and Engineering* **32**(2) 327.

Arrow, K.J., T. Harris, J. Marschak. 1951. Optimal inventory policy. *Econometrica* **19**(3) 250–272.

Ben-Daya, M., M. Hariga. 2004. Integrated single vendor single buyer model with stochastic demand and variable lead time. *International Journal of Production Economics* **92**(1) 75–80.

Benjaafar, S., W.L. Cooper, J.-S. Kim. 2005. On the benefits of pooling in production-inventory systems. *Management Science* **51**(4) 548–565.

Benjaafar, S., J.-S. Kim, N. Vishwanadham. 2004. On the effect of product variety in production–inventory systems. *Annals of Operations Research* **126**(1-4) 71–101.

Boute, R.N., S.M. Disney, M.R. Lambrecht, B. Van Houdt. 2007a. An integrated production and inventory model to dampen upstream demand variability in the supply chain. *European Journal of Operational Research* **178**(1) 121–142.


[2]Except for the *null-recurrent* case, which never occurs in our case as $\rho < 1$.

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

Cakanyildirim, M., J.H. Bookbinder, Y. Gerchak. 2000. Continuous review inventory models where random lead time depends on lot size and reserved capacity. *International Journal of Production Economics* **68**(3) 217–228.

Caplin, A.S., J. Leahy. 2010. Economic theory and the world of practice: a celebration of the $(s,S)$ model. *The Journal of Economic Perspectives* **24**(1) 183–202.

Glock, C.H. 2012a. The joint economic lot size problem: A review. *International Journal of Production Economics* **135**(2) 671–686.

Glock, C.H. 2012b. Lead time reduction strategies in a single-vendor–single-buyer integrated inventory model with lot size-dependent lead times and stochastic demand. *International Journal of Production Economics* **136**(1) 37–44.

Guo, C.-H., B. Iannazzo, B. Meini. 2007. On the doubling algorithm for a (shifted) nonsymmetric algebraic Riccati equation. *SIAM Journal on Matrix Analysis and Applications* **29**(4) 1083–1100.

Hariga, M.A. 1999. A stochastic inventory model with lead time and lot size interaction. *Production planning & control* **10**(5) 434–438.

Harris, F.W. 1913. How many parts to make at once. *Factory, the Magazine of Management* **10 (2)** 135–136, 152.

Hoque, M.A. 2013. A vendor–buyer integrated production–inventory model with normal distribution of lead time. *International Journal of Production Economics* **144**(2) 409–417.

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

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

Karmarkar, U.S. 1987. Lot sizes, lead times and in-process inventories. *Management Science* **33**(3) 409–418.

Kim, J.S., W.C. Benton. 1995. Lot size dependent lead times in a Q, R inventory system. *The International Journal of Production Research* **33**(1) 41–58.

Latouche, G. 2006. Structured Markov chains in applied probability and numerical analysis. *Markov Anniversary Meeting*. 69–78.

Latouche, G., V. Ramaswami. 1987. *Introduction to matrix analytic methods in stochastic modeling*, vol. 5. Society for Industrial and Applied Mathematics.

Porteus, E.L. 1971. Optimality of generalized $(s,S)$ policies. *Management Science Series A-theory* **17**(7) 411–426.

Scarf, H. 1960. The optimality of $(s,S)$ policies in the dynamic inventory problem. *Mathematical Methods in the Social Sciences* 196–202.

Sengupta, B. 1989. Markov processes whose steady state distribution is matrix-exponential with an application to the GI/PH/1 queue. *Advances in Applied Probability* 159–180.

Sengupta, B. 1990. The semi-Markovian queue: theory and applications. *Stochastic Models* **6**(3) 383–413.

Veinott, A.F. 1966. On the optimality of $(s,S)$ inventory policies: new conditions and a new proof. *SIAM Journal on Applied Mathematics* **14**(5) pp. 1067–1083.

Wang, W.-G., W.-C. Wang, R.-C. Li. 2011. ADDA: Alternating-Directional Doubling Algorithm for M-matrix algebraic Riccati equations. Tech. Rep. 2011-04, The University of Texas Arlington.

Zipkin, P.H. 2000. *Foundations of inventory management*, vol. 2. McGraw-Hill New York.