ARTICLE IN PRESS

Int. J. Production Economics ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Int. J. Production Economics

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

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

ARTICLE INFO

Article history:
Received 30 May 2013
Accepted 28 April 2014

Keywords:
Production/inventory system
Lot sizing
Markov chain analysis

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. This procedure allows to determine the optimal inventory parameters within the class of (s, S) policies that minimize the expected ordering and inventory related costs over time. 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.

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

Arrow et al. (1951) introduced a slightly modified version of this model, i.e. the (s, S) inventory policy, in which a reorder 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 position 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, assuming constant lead times, that an (s, S) policy is optimal when a fixed order cost is present (Scarf, 1960; Iglehart, 1963; Veinott, 1966; Porteus, 1971). If orders do not cross, the (s, S) optimality result under constant lead times carries over to stochastic, non-crossing lead times (Muharremoglu and Tsitsiklis, 2008). To our knowledge, there is no analytical work that shows the optimality of the (s, S) policy in finite capacity production/inventory systems. The (s, S) inventory policy is still of main importance today 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

* Corresponding author.
E-mail addresses: ann.noblesse@kuleuven.be (A.M. Noblesse), robert.boute@kuleuven.be (R.N. Boute), marc.lambrecht@kuleuven.be (M.R. Lambrecht), benny.vanhoudt@ua.ac.be (B. Van Houdt).

http://dx.doi.org/10.1016/j.ijpe.2014.04.027
0925-5273/

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. Multiple small batches 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 (Karmarkar, 1987). At the other extreme, if lot sizes are very large, lead times approach an increasing function of the lot size. 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 affect the parameters of the inventory policy. We show that the inclusion of endogenous lead times (as opposed to assuming lead times are exogenous to the inventory model) 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 that 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, 1999).

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.

We restrict our search to the class of (s, S) policies and we optimize the values of the reorder point s and the lot sizing decision $(S-s)$ that 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 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. We derive the lead time distribution for a given set of (s, S) parameters in 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. 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 to the inventory parameters.

Throughout this paper we will adopt the following notations:

- The compound Poisson demand has arrival rate λ, 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 ϕ_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 (γ_p, U_p), and the setup time has an order n_s phase-type representation (γ_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

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

  with λ the arrival rate of customers, $\sum_{i=1}^{m} i d_i$ the expected demand size per customer, and $(\gamma_p(-U_p)^{-1} e_{n_p})$ the expected time to produce one unit. Based on ρ_{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 μ_{ot} the average time between two orders placed.
- 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/μ_{ot}) refers to the expected total ordering cost in a time unit, which is expressed by means of the renewal reward theorem, with μ_{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