Designing replenishment rules in a two-echelon supply chain with a flexible or an inflexible capacity strategy

Robert N. Boute [1,2], Stephen M. Disney [3], Marc R. Lambrecht [2] and Benny Van Houdt [4]

[1] *Operations & Technology Management Center, Vlerick Leuven Gent Management School. Vlamingenstraat 83, 3000 Leuven, Belgium. E-mail: robert.boute@vlerick.be*

[2] *Research Center for Operations Management, Katholieke Universiteit Leuven, Naamsestraat 69, 3000 Leuven, Belgium. E-mail: robert.boute@econ.kuleuven.be, marc.lambrecht@econ.kuleuven.be*

[3] *Logistics Systems Dynamics Group, Cardiff Business School, Cardiff University, Aberconway Building, Colum Drive, Cardiff, CF10 3EU, UK. E-mail: disneysm@cardiff.ac.uk*

[4] *Department of Mathematics and Computer Science, University of Antwerp, Middelheimlaan 1, 2020 Antwerpen, Belgium. Email: benny.vanhoudt@ua.ac.be*

ABSTRACT

We consider a two echelon supply chain where a single retailer holds an inventory of finished goods to satisfy an i.i.d. customer demand, and a single manufacturer produces the retailer's replenishment orders on a make-to-order basis. The objective of this paper is to analyse the impact of the retailer's replenishment policy on total supply chain performance. We consider two strategies with regard to the production capacity. In a flexible capacity strategy, the manufacturer invests in excess capacity to guarantee constant lead times in order to keep inventories low. The amount of investment depends on the retailer's order pattern. In an inflexible capacity strategy, the capacity is limited and independent of the retailer's replenishment decision. This results in stochastic lead times, thereby inflating the retailer's inventory requirements. We treat the variability of the order rate of the retailer as the primary decision variable to minimise total supply chain costs. The objective is to find the value of the replenishment parameter β (parameter to tune the order variability) that minimises total supply chain costs in a flexible and inflexible capacity scenario.

Keywords: *production and inventory control, supply chain performance, bullwhip, queueing, capacity-inventory trade-off*

1. INTRODUCTION

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

The manufacturer does not hold a finished goods inventory but produces the retailer's orders on a make-to-order basis. The manufacturer's production system is characterized by a single server queueing model that sequentially processes the ordered units one by one on a first-come-first-served basis. When the production is busy, the orders join a queue of unprocessed orders. Once the complete replenishment order is produced, it replenishes the retailer's inventory. The time from the moment an order is placed to the moment that it replenishes the retailer's inventory, is the replenishment lead time, T_p. The production process at the

manufacturer implies that the retailer's replenishment lead times are stochastic and correlated with the order quantity.
We examine two important problems in the two echelon system described above. First, we examine the order variability at the retailer (dampening or amplification). Second we examine the capacity strategy of the manufacturer (flexible or inflexible). It is clear that both subsystems interact through the stochastic nature of the lead times and consequently impacts the customer service of the retailer. The major contribution of this paper is the simultaneous treatment of both subsystems so that total supply chain costs are minimised.
Let's briefly introduce the two problems mentioned above.

First we have the order variability at the retailer level. Lee et al. (1997) describe a problem frequently encountered in supply chains, called the bullwhip effect: demand variability increases as one moves up the supply chain. This *amplified* order variability can have large upstream cost repercussions. Balakrishnan et al. (2004) emphasize the opportunities to reduce supply chain costs by *dampening* order variability. However, despite the fact that the manufacturer benefits from smooth production, retailers, driven by the goal of reducing inventory costs, prefer to use replenishment policies that chase demand rather than dampen customer demand variability. Dampening variability in orders may have a negative impact on the retailer's customer service due to inventory variance increases (Disney and Towill 2003). In this paper we analyse the impact of order variability amplification vs. dampening on the performance of a two-echelon supply chain.

Second we have the capacity structure of the manufacturer. The retailer's replenishment orders load the manufacturer's production system. We consider two strategies with regard to the production capacity. The first is a *flexible capacity* strategy. This means that the manufacturer invests in excess capacity in order to produce each order within the period after it was placed. It is clear that when the orders fluctuate wildly, the capacity investments will be larger compared to the situation where the order pattern is flat. At the same time the inventory costs for the retailer are in this scenario low since every order is replenished in the period after it was placed (zero lead times).

The second strategy is an *inflexible capacity* strategy, i.e., the manufacturer's capacity remains at a fixed level, irrespective of the retailer's order pattern. The manufacturer's capacity level may be lower than the maximum possible order quantity. As a result, when the available capacity in a period is insufficient to complete production of an order, then the next period's capacity is used to continue production of this order. The manufacturer delivers the retailer's orders as soon as the total order is produced, implying that lead times are variable and can be strictly positive. Moreover, when the retailer sends a volatile order pattern to the production queue, production (and delivery) lead times will be longer and more variable than when the retailer sends a constant order pattern to production. This in turn affects the retailer's inventory requirements.

In this paper we treat the variability of the order rate of the retailer as the primary decision variable to minimise total supply chain costs. The paper is organized as follows. In the remainder of this section we introduce an example, we give a legend of variables/parameters used in the text and we provide a summary of the assumptions of the model. In section 2, we discuss in greater detail the flexible/inflexible capacity scenarios. Section 3 is devoted to the downstream inventory policy and its impact on order variance. In section 4 we examine the lead time distribution and the net stock distribution. Section 5 describes the trade off by means of a total cost function, which we illustrate with a numerical example in section 6. Section 7 concludes.

1.1. An example

The primary purpose of this paper is to offer managerial insight into a supply chain coordination problem. The situation we have in mind is in the fast moving consumer goods industry. We focus on products requiring short lead times from the manufacturer because of the short life time of the product. Boute et al. (2008) describe the case of a bakery company focusing on authentic specialties in the biscuit and cake market. We have retailers on the one hand and an industrial bakery on the other hand. Given the specific packaging requirements of retailers, the bakery employs a make-to-order policy. For new product introductions (e.g. biscuit pasta) the bakery has to install new machinery and has to decide on the capacity level. We are interested in the interaction between capacity, lead-time distribution, replenishment rules and customer service. This situation does not only arise in the fast moving consumer goods industry but is quite common in many other industrial settings especially when capacity expansion decisions have to be made because of new product introductions. There are many examples of incorrect estimation of the capacity to be installed for new product introductions.

1.2. Legend of frequently used variables and parameters

- D *:* random variable describing the customer demand, with $f_D(\cdot)$ the corresponding discrete probability function, $E(D)$ the long term average demand, and D_{min} and D_{max} the resp. minimum and maximum demand size
- C_h *:* inventory holding cost per unit, per period*;* C_b *:* per unit shortage cost
- $C(K)$ *:* the linear capital expenditure function; K the size of the capacity investment
- C_0 *:* the fixed capacity investment cost; C_K *:* marginal capacity investment cost; C_P *:* cost per unit overtime production
- M *:* the production time per unit
- ρ : average utilisation rate of the manufacturer's production system
- T_p *:* the replenishment lead time
- β *:* smoothing parameter in the replenishment rule
- O_t *:* order quantity placed at the end of period t
- NS_t *:* on hand inventory at the end of period t
- IP_t *:* inventory position at the end of period t
- *SS :* safety stock
- *DIP* : desired inventory position
- S : base-stock level

1.3. Assumptions

- The sequence of events in a period is as follows. First receive goods from the upstream partner, then observe and satisfy demand and finally place a replenishment order.
- Customer demand D is independently and identically distributed (i.i.d.) over time with an arbitrary, finite, discrete probability distribution function $f_D(\cdot)$.
- If the inventory on hand at the end of the period is positive $(NS_t > 0)$, a holding cost C_h per unit is incurred to carry inventory to the next period. If the inventory on hand is negative $(NS_t < 0)$, a backlog cost C_b per unit shortage is incurred.
- The production ("service") time M of a single unit is deterministic. To ensure stability (of the queue), we assume that the utilization of the production facility (average batch production time divided by average batch interarrival time) is strictly smaller than one.
- Define the capacity K as the number of units that can be produced in a period. The capacity investment cost function is given by $C(K) = C_0 + C_K \cdot K$, where C_0 represents the fixed capacity investment cost and C_K is a constant, marginal capacity investment cost.

When the installed capacity is insufficient, a unit can be produced in overtime capacity at extra cost C_P. We assume that $C_K < C_P$, otherwise it would never be optimal to invest in capacity. The capital expenditure function will be discussed in detail in section 2.

- The manufacturer operates a make-to-order policy and does not incur a setup time or cost. We assume highly automated equipment where setup times are non-existing. This assumption eliminates the batching decision at the manufacturing level.

In Fig. 1 we graphically represent the cost functions.

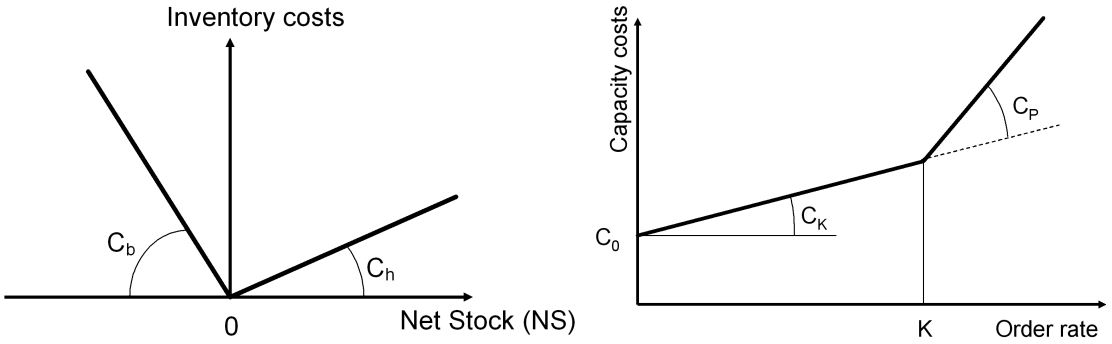

Figure 1: Cost structure of our model

2. FLEXIBLE / INFLEXIBLE CAPACITY

In this section, we will discuss the capacity scenarios in greater detail. In our model a key role is played by the capital expenditure function $C(K) = C_0 + C_K \cdot K$. A good summary of the consequences of this function can be found in Van Mieghem (2008). This cost function allows us to model economies of scale; this means that costs grow sub-linearly, either due to the presence of a fixed cost component or due to decreasing marginal costs. We use the fixed cost model in this paper. An extension to the decreasing marginal cost model (by using power functions) is straightforward. As is indicated by Van Mieghem (2008), C_0 refers to all costs independent of the size of the capacity (costs of planning a capacity expansion, the selection process, real estate, administrative overhead,…). C_K refers to the marginal cost or the cost to add one unit of capacity. In our bakery example the marginal cost depends on the size of the oven and/or packaging machines. The capacity unit may be expressed in tons per time unit in our example.

2.1. Flexible Capacity – impact on capacity investment

Suppose the retailer wants the manufacturer to deliver the replenishment orders within the period after the order was placed (i.e., $T_p = 0$), then the production capacity has to be large enough to complete the production of each replenishment order within one time period. A key trade-off in capacity strategy is balancing the marginal cost of installed capacity C_K with the cost of capacity shortage (Van Mieghem 2008). In our case a capacity shortage implies a unit production in overtime capacity at cost C_P.

The installed capacity K is the number of units that can be produced in a period, and M is the production time of a single unit, expressed as a fraction of a period, or $K = M^{-1}$. The *capacity shortfall* in a given period measures how much of the period's order quantity exceeds available capacity, or equivalently, the number of units that are produced in overtime in that period.

When the installed capacity is equal to the average order quantity, $K = E(O)$, the manufacturer experiences capacity shortfalls half of the time, resulting in frequent production

in overtime if the order pattern is volatile. Therefore, it may be worth to invest in extra capacity above the average order quantity, in order to counter the negative impact of volatility. The purpose of the "excess" capacity is to provide a *safety capacity* to capture higher-than-expected orders. When the order volatility increases, the expected capacity shortfall will increase, but an investment in safety capacity can strongly reduce this capacity shortfall (Van Mieghem 2008).

An alternative strategy is to set the capacity equal to the maximum order quantity, $K = O_{max}$, so that the capacity shortfall is zero and there is no production in overtime. This would be a plausible strategy when the cost of production in overtime is extremely large or when no overtime capacity is available. However, if for instance the order quantity reaches its maximum only occasionally, it may turn out cheaper to install a capacity $K < O_{max}$ and occasionally produce in overtime capacity at cost C_P.

It is clear that the decision to determine the optimal capacity size K^* depends both on the relative cost of invested capacity versus the cost of overtime production, and the distribution of the replenishment orders placed by the retailer.

2.2. Inflexible Capacity – impact on lead times

The situation is totally different in the inflexible capacity scenario; when the available capacity in a period is insufficient to complete production of an order, then the next period's capacity is used to continue the production of this order. There is no production in overtime and the production of an order may be spread over several periods, so that lead times are variable and can be strictly positive.

As the retailer's replenishment orders load the manufacturer's production, the nature of this loading process relative to the available capacity and the variability it creates determine the (production/replenishment) lead times. We actually extend a pure inventory system with *exogenous* lead times to a production-inventory system with *endogenous* lead times. The retailer's inventory replenishment lead times are "endogenously" determined by the manufacturer's production with limited capacity.

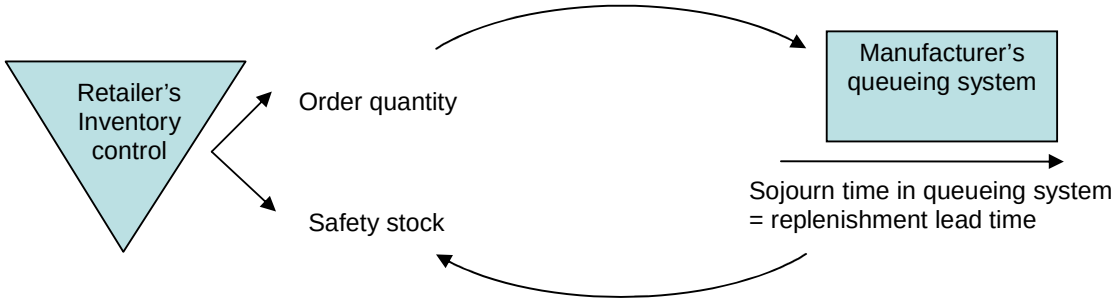

Figure 2: Interaction between retailer's inventory and manufacturer's production

In Fig. 2 the interaction between the retailer's replenishment policy and the manufacturer's production system is illustrated: the replenishment policy generates orders that define the arrival process at the manufacturer's queue. The time until the order is produced (the sojourn time in the queueing system), is the time to replenish the order. Hence, when the retailer amplifies the order variability, this implies a more variable arrival pattern at the production queue, leading to longer and more variable lead times according to the laws of factory physics (Hopp and Spearman 2001). Dampening the variability in the order pattern results in shorter and less variable lead times. This replenishment lead time is a prime determinant in setting the safety stock requirements for the retailer.

3. DOWNSTREAM INVENTORY POLICY

3.1. Replenishment rule

Given the common practice in retailing to replenish inventories frequently (e.g. daily) and the tendency of manufacturers to produce to demand, we focus on periodic review, base-stock or order-up-to replenishment policies.

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

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

The base-stock level *S* is the inventory required to ensure a given customer service level. Orders are placed every *R* periods and after an order is placed, it takes T_p periods for the replenishment to arrive. Hence the risk period (the time between placing a replenishment order until receiving the subsequent replenishment order) is equal to the review period plus the replenishment lead time $R + T_p$. Since customer demand is i.i.d., the best estimate of all future demands is simply the long term average demand, $E(D)$. Consequently, the base-stock level equals

$$S = [E(T_p) + R] \cdot E(D) + SS, \tag{2}$$

with *SS* denoting the retailer's safety stock.

In the remainder of this paper we assume that the review period *R* is one base period, i.e., we place an order at the end of every period, similar to the standard Beer Game setup (Sterman 1989). Substituting (2) into (1) we obtain

$$\begin{aligned} O_t &= E(D) + E(T_p) \cdot E(D) + SS - IP_t \\ &= E(D) + [DIP - IP_t], \end{aligned} \tag{3}$$

where $E(T_p) \cdot E(D) + SS$ can be seen as the *desired* inventory position *DIP*, which is the sum of the desired pipeline stock and desired net stock. The difference between the desired and actual inventory position $[DIP - IP_t]$ is denoted as the *inventory position deficit*.

Magee (1958) and Forrester (1961) introduce a proportional controller β into the inventory deficit, resulting in the following *generalised* order-up-to policy:

$$O_t = E(D) + \beta \cdot [DIP - IP_t], \tag{4}$$

with $0 < \beta < 2$. Forrester (1961) refers to $1/\beta$ as the "adjustment time". When $\beta < 1$ he explicitly acknowledges that the deficit recovery should be spread out over time, whereas $\beta > 1$ implies an overreaction to the inventory deficit. This replenishment rule is particularly powerful (Disney and Towill 2002) as it encompasses e.g. the way people play the Beer Game (Sterman 1989, Naim and Towill 1995), a general case of order-up-to policies and many variants of it (Dejonckheere et al. 2003), and with fine tuning it can reflect Materials Requirements Planning (Disney 2001). This "proportional order-up-to" policy is also

equivalent to the "full-state order-to-up" policy (Gaalman and Disney, 2006), assuming, as we do, an i.i.d. demand process.

3.2. Order variance amplification/dampening

When customer demand is i.i.d., the generalised replenishment policy generates an auto-correlated order pattern (see appendix A), given by

$$O_t = (1 - \beta) \cdot O_{t-1} + \beta \cdot D_t \,. \tag{5}$$

From this order "path" over time we can derive the steady state distribution of the order quantities given the finite, discrete demand distribution $f_D(\cdot)$. Let us denote the order distribution by $f_O(\cdot)$ and its corresponding cumulative order distribution by $F_O(\cdot)$.

Observe that when $\beta > 1$, the order pattern is negatively correlated and the generalised order-up-to policy may generate negative order quantities. Since in our model it is not possible to send negative orders to production, we have to preclude the possibility of negative orders. The following restriction on beta given the minimum and maximum demand ensures that $O_t \geq 1$ (see appendix B):

$$D_{min} + (1 - \beta) \cdot D_{max} \geq 2 - \beta. \tag{6}$$

To examine the variability in orders created by the generalised order-up-to policy, we look at the ratio of the variance of the orders over the variance of demand (in the literature this variance ratio is commonly used as a measure for the bullwhip effect). This can be easily derived from Eqn. (5):

$$\frac{Var(O)}{Var(D)} = \frac{\beta}{2 - \beta} \,. \tag{7}$$

Hence, if we do not smooth, i.e. if $\beta = 1$, these expressions reduce to the standard base-stock policy, where $O_t = D_t$: we chase sales and thus there is no variance amplification. For $1 < \beta < 2$ we create bullwhip (variance amplification) and for $0 < \beta < 1$ we generate a smooth replenishment pattern (dampening order variability).

4. DETERMINATION OF LEAD TIMES AND INVENTORY

4.1. Determination of lead time distribution

The replenishment orders loading the production system are characterised by Eqn. (5). By analysing the characteristics of these replenishment orders, we implicitly analyse the characteristics of the production orders that arrive at the manufacturer's production system. As we can see from Eqn. (5), the generalised order-up-to policy generates batch arrivals with a fixed interarrival time (equal to the review period, $R = 1$) and with variable (auto-correlated) batch sizes.

Based on *matrix analytic methods* (Neuts 1981, Latouche and Ramaswami 1999), Boute et al. (2006) developed a discrete time queueing model to estimate the lead time distribution given a batch arrival process with a fixed interarrival time and positively correlated batch sizes. In their queueing model, production times are phase type (PH) distributed. We can use their methodology to find the lead time distribution in our production model, since a PH distribution can also be used to model deterministic production times, as we assume here. In addition we extend their model for negatively correlated batch sizes, which is the case when

$\beta>1$ (see Eqn. (5)). We do take restriction (6) into account in order to avoid negative batch sizes.

This queueing analysis returns the lead time distribution $f_{Tp}(\cdot)$ for each value of β. In other words, we use the methodology for determining the lead time distribution, described in Boute et al. (2006), and we use this result to incorporate it in a supply chain coordination mechanism in a flexible or inflexible capacity scenario.

4.2. Determination of inventory distribution

When demand is probabilistic, there is a definite chance of not being able to satisfy some of the demand directly out of stock. Therefore, a buffer or safety stock is required to meet unexpected fluctuations in demand. We characterize the retailer's inventory random variable and use it to find its safety stock requirements. Due to the production process, lead times are stochastic and as a consequence we do not know exactly when a replenishment occurs.

We monitor the inventory on hand at the end of every period, after customer demand is observed and after a replenishment order has been placed. At the end of period t, there may be $k \geq 0$ orders waiting in the production queue and there is always 1 order in service (since the observation moment is immediately after an order placement) which is placed k periods ago (O_{t-k}). Note that k is a function of t, but we write k as opposed to $k(t)$ to simplify the notation. In appendix C we show that the net stock distribution can then be written as

$$NS_t = SS - Z_t. \tag{8}$$

$$with\ Z_t = \sum_{i=0}^{k-1} D_{t-i} - E(T_p)\cdot E(D) + \sum_{i=k}^{t-1} (1-\beta)^{i-k}\cdot (D_{t-i} - E(D)). \tag{9}$$

The evolution of Z_t determines the evolution of the net stock NS_t. Since $E(Z) = 0$, $E(NS)=SS$. By means of the Markov process of the above mentioned queueing model, Boute et al. (2006) develop an algorithm to find the steady state distribution of Z_t, denoted by $f_Z(\cdot)$. The exact analysis is not straightforward due to the correlation between the different terms that make up Z_t. The value of D_{t-k} influences the age k of the current order in service: the larger the demand size, the larger the order size and consequently the longer it takes to produce the order. Moreover, since the order quantity is also affected by previously realised demand terms (see Eqn. (5)), the demand terms D_{t-i}, $i \geq k + 1$ also influence the order's age, k.

Given the distribution of Z, the amount of safety stock SS determines the corresponding inventory distribution $f_{NS}(\cdot)$. The value of SS is a decision variable and depends on the cost structure and the distribution of Z (see section 5). Since Z is function of β, SS is also impacted by the value of β.

In the flexible capacity scenario each replenishment order is produced within the period after it is placed, so that the production queue is always empty when an order is sent to production, or $k = 0$ in Eqn. (9). Moreover, since the lead time $T_p = 0$, Z_t simplifies to

$$Z_t = \sum_{i=0}^{t-1} (1-\beta)^i (D_{t-i} - E(D)), \tag{10}$$

and its steady state distribution $f_Z(\cdot)$ can be found from the compound demand distribution.

5. SUPPLY CHAIN PERFORMANCE

In this section we measure the impact of the retailer's order decision (order variance amplification/dampening) on total supply chain performance. We consider the inventory costs at the retailer and the capacity costs at the manufacturer, and search for the value of the replenishment parameter β that minimises total supply chain costs for the flexible and inflexible capacity scenarios. In the next section we illustrate our analysis with a numerical example.

5.1. Cost function

The capacity costs include the capacity investment cost given by $C(K)$, and the number of units that are produced in overtime in a period (which is zero in the inflexible capacity scenario). The inventory costs per period consist of a holding cost to keep a unit in inventory for a unit of time and a backlog cost for every unit of demand that can not be immediately fulfilled from the inventory on hand. Hence the inventory costs equal $C_h \cdot NS$ if $NS \geq 0$, and $C_b \cdot (-NS)$ if $NS < 0$. It is however more elegant to write the net stock NS as a function of the safety stock SS and the distribution of Z: $NS = SS - Z$. Inventory and capacity costs are minimised by finding the optimal values for the safety stock SS^* and the installed capacity K^*:

$$C_{INV}(SS^*, Z) = \min_{SS^*} \{ C_h \cdot E[(SS^* - Z)^+] + C_b \cdot E[(SS^* - Z)^-] \} \quad (11)$$

$$C_{CAP}(K^*, O) = \min_{K^*} \{ C(K^*) + C_p \cdot E[(O - K^*)^+] \} \quad \text{when capacity is flexible,}$$

$$= \min_{K^*} C(K^*) \quad \text{when capacity is inflexible.} \quad (12)$$

The inventory and capacity costs depend on the distribution functions of resp. Z and O, which are both function of the replenishment parameter β. The cost-minimisation problem can then be formulated as finding the optimal value of β which minimises the sum of total inventory and capacity costs:

$$\min_{\beta} \{ C_{INV}(SS^*, Z) + C_{CAP}(K^*, O) \}. \quad (13)$$

5.2. Flexible Capacity Strategy

a) Optimal safety stock SS^* that minimises inventory costs for a given β

The inventory cost function

$$C_{INV} = C_h \cdot E[(SS - Z)^+] + C_b \cdot E[(SS - Z)^-] \quad (14)$$

is minimised by the critical fractile value, which provides the optimal stock out probability (Zipkin 2000):

$$Pr(NS < 0) = C_h / (C_h + C_b). \quad (15)$$

The safety stock that corresponds to this stock out probability minimises the inventory costs:

$$Pr(Z \leq SS^*) = C_b / (C_h + C_b)$$
$$SS^* = F_Z^{-1}(C_b / (C_h + C_b)), \quad (16)$$

where $F_Z(\cdot)$ denotes the cumulative distribution function of Z. Substituting SS^* into Eqn. (14) provides the lowest inventory cost for a given value of β.

Clearly, as Z becomes more volatile, the optimal safety stock value SS^* increases, and the inventory costs increase as well. From the steady state distribution of Z_t, given by Eqn. (10), we find that

$$Var(Z) = Var(D) \cdot 1 / \beta(2 - \beta) . \quad (17)$$

Hence, Z has a higher variance as we dampen the order pattern ($\beta < 1$) or as we amplify the orders ($\beta > 1$), compared to a pure chase sales policy ($\beta = 1$). As a result the inventory costs increase as we dampen or amplify the order variance, and are minimal when $\beta = 1$.

b) Optimal capacity size K^* that minimises capacity costs for a given β

In order to produce each order within one time period, the manufacturer has to invest in capacity. The objective is to determine the installed capacity K, defined as the number of units that can be produced per period, which minimises the capacity cost function, given by

$$C_{CAP} = C_0 + C_K \cdot K + C_P \cdot E[(O - K)^+]. \quad (18)$$

The optimal capacity size K^* that minimises this capacity cost function, satisfies a newsvendor solution. Van Mieghem (2008) shows that the optimal *capacity sizing condition* is given by:

$$Pr(O > K^*) = C_K / C_P, \quad (19)$$

which in turn defines the optimal capacity size as

$$K^* = F_O^{-1}((C_P - C_K) / C_P), \quad (20)$$

with $F_O(\cdot)$ the cumulative order distribution function.

When the order sizes fluctuate wildly, it is preferable to invest in more capacity since there is more need for production in overtime, which is much more expensive than a capacity investment itself. When the order pattern is flat, the optimal capacity size K^* will be lower since there is less need to produce in overtime. The optimal capacity size therefore depends on the retailer's ordering decision to amplify or dampen the order variance. Since the order pattern increases in variability as β increases, the optimal capacity investment K^* and its corresponding capacity costs C_{CAP} increase as β increases.

c) Value of β that minimises total supply chain costs

For a given value of the replenishment parameter β we described how to find the values of K^* and SS^* that minimise resp. the capacity and inventory costs. In order to find the value β that minimises total supply chain costs, we add up the inventory and capacity costs corresponding to the optimal values of K^* and SS^*. Note that there is no interaction between inventory and capacity costs. Changing the capacity investment has no impact on lead times in a flexible capacity strategy, since every order needs to be produced within the order after it was placed. Hence safety stocks are not affected by capacity investments and can be treated independent of capacity investment decisions.

If we add up capacity and inventory costs, we obtain the following dynamics in the total cost function. On the one hand, inventory costs show a U-shaped convex function of the parameter β with a minimum in $\beta=1$; both order variance amplification and dampening increase inventory costs compared to the chase sales policy. The capacity costs, on the other hand, increase as β increases; compared to the chase sales policy, the capacity costs are lower when order variance is dampened and higher when the order variance is amplified.

Hence, dampening the orders ($\beta<1$) may reduce total supply chain costs in case the decrease in capacity costs outweighs the increase in inventory costs. If dampening the orders leads to an increase in inventory costs, which is larger than the decrease in capacity costs, it is preferable not to dampen any further. In other words, the extent to which we should smooth the order pattern depends on the relative costs of capacity and inventory. Note that amplification, $\beta>1$, always leads to higher inventory and capacity costs, irrespective of the cost parameters.

5.3. Inflexible Capacity Strategy

a) Optimal safety stock SS^* that minimises inventory costs for a given β

Analogous to the flexible capacity strategy, the safety stock SS* that minimises inventory costs, is given by

$$SS^* = F_Z^{-1}(C_b / (C_h + C_b)). \qquad (21)$$

In this case however, Z is the steady state distribution of Z_t given by Eqn. (9), which has a more complex function than Eqn. (10). The distribution of Z is now affected by β in two ways. First, similar to the flexible capacity strategy, the order variance has an impact on the variance of Z. Fluctuations are minimal in a pure chase policy $(\beta = 1)$, and variability increases when orders are dampened $(\beta < 1)$ or amplified $(\beta > 1)$. But in the inflexible capacity strategy there is also a second factor that impacts the distribution of Z. The value of β also affects the lead time distribution; lead times increase as β increases due to the increased variability in the order pattern. As a consequence, order variance dampening leads to lower and less variable lead times, exercising a compensating effect on the required safety stock. At the same time, order variance amplification increases the inventory variability ánd increases lead times, reinforcing the increased safety stock requirements.

b) Optimal capacity size K^* that minimises capacity costs for a given β

The capacity level remains fixed in the inflexible capacity strategy, independent of the order decision. Since there is no production in overtime, the capacity cost function, reduced to

$$C_{CAP} = C_0 + C_K \cdot K \qquad (23)$$

is minimised when the installed capacity K is as small as possible. However, in order to obtain a stable system, the capacity investment K has to be larger than the average order quantity $E(O)$. This ensures that the average utilisation rate of the manufacturer's production system, ρ, is smaller than one.

c) Value of β that minimises total supply chain costs

For a given value of the replenishment parameter β we described how to find the value SS^* that minimises inventory costs. Capacity costs are minimised when the installed capacity is as small as possible, provided that it exceeds the average order quantity. However, in an

inflexible capacity strategy there is an interaction between the capacity investment and inventory costs. The installed capacity determines the production load, which has an impact on lead times. A large capacity investment reduces the production load, so that production (queueing) lead times are shorter. These lead times in turn determine safety stocks and corresponding inventory costs.

Hence, in order to find the value of β that minimises total supply chain costs, we may not simply add up the inventory and capacity costs that correspond to K^* and SS^*, due to the interaction between both. We need to trade-off capacity and waiting, which is in this case a *capacity-inventory trade-off*. For instance, as inventory costs are relatively cheap, it is preferable not to invest in too much capacity and instead hold more inventory. A high cost of inventory on the contrary increases the need for capacity investment in order to keep inventory holdings low.

In order to seek the lowest total supply chain costs, we assume a capacity size K and measure the impact of β on the inventory costs. Order variance amplification increases inventory variability and lead times, blowing up the inventory costs. Order variance dampening result in shorter and less variable lead times compared to the chase sales policy, which may compensate the increase in inventory variability. Hence, depending on the lead time impact, inventory costs may be lowered by smoothing the replenishment orders to some extent. If we smooth too much however, the lead time reduction may not compensate the increase in inventory variability anymore.

To trade-off the cost of capacity against the cost of inventory, we change the capacity level K and measure its impact on inventory costs. It is clear that lead times (and inventory costs) decrease as the capacity investment K increases, since this decreases the utilization rate. However, due to the complexity of our queueing model we cannot quantify the exact relation between the utilization rate and lead times analytically. Hence by means of a search procedure we determine the optimal capacity size K^* that minimises total supply chain costs. Obviously, the value of K^* depends on the relative costs of capacity and inventory.

6. NUMERICAL EXAMPLE

To illustrate our analysis, we consider the following numerical example. A retailer daily observes a customer demand which is randomly distributed between 21 and 40 units with an average of 30.5 units and a standard deviation of 7.5. The retailer replenishes his inventory with the generalised replenishment rule, i.e., he places orders at the end of every day equal *to* $O_t = E(D) + \beta \cdot [DIP - IP_t]$ (see Eqn. (4)).

When the replenishment parameter $\beta < 1$, the retailer sends a smooth, positively correlated order pattern to the manufacturer (dampening scenario). When $\beta > 1$, the order pattern is negatively correlated with a larger variance than the observed customer demand (bullwhip scenario). In order to exclude the possibility of negative order quantities, we limit the replenishment parameter to $\beta < 1.525$ (larger values of β may theoretically generate negative order quantities, see Eqn. (6)).

We assume the following cost components. A holding cost $C_h = 1$ is incurred per unit per day and a backlog cost $C_b = 20$ is incurred per unit that cannot be immediately satisfied from the inventory on hand. There is a fixed capacity investment cost $C_0 = 2$ and an additional cost per unit of installed capacity $C_K = 2$. A unit can be produced in overtime capacity at extra cost $C_P = 5$.

6.1. Flexible Capacity Strategy

In Fig. 3 we plot the optimal safety stock SS^* that is required to maintain a 95.24% customer service level (the optimal stock-out probability equals $C_h/(C_b+C_h) = 0.0476$)). We observe that the safety stock increases as the order variance is dampened $(\beta < 1)$ or amplified $(\beta > 1)$, and the minimal safety stock is found in a pure chase sales policy $(\beta = 1)$. The corresponding inventory costs C_{inv} show a similar pattern. Overall, we observe that inventory costs are relatively low due to the zero lead times.

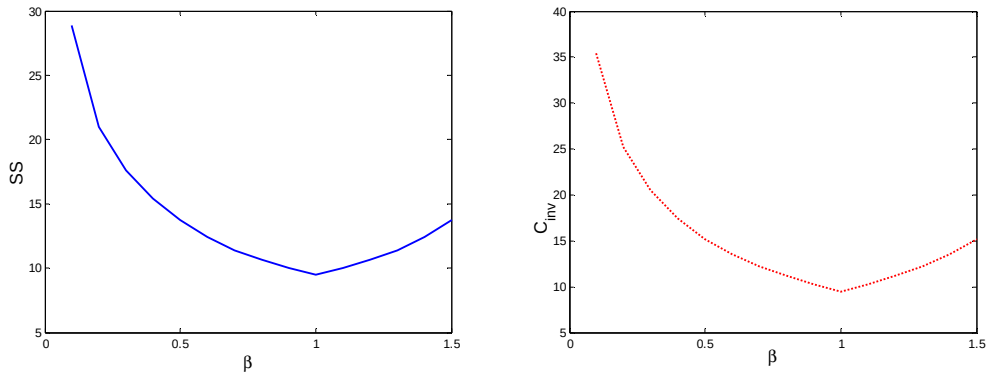

Figure 3 : Flexible capacity strategy: Impact of β on optimal safety stock SS^* and corresponding inventory costs C_{inv}

In Fig. 4 we present the impact of the replenishment parameter β on the capacity costs. As intuitively expected, capacity costs (C_{CAP}) increase as the order pattern becomes more volatile (i.e., as β increases). When we look at the total supply chain costs $(C_{INV} + C_{CAP})$, we observe that order variance amplification $(\beta > 1)$ clearly increases total supply chain costs due to the combined increase in inventory and capacity costs. When we smooth the orders $(\beta < 1)$, the interplay between inventory and capacity reveals that dampening the orders to a certain extent decreases total supply chain costs, but if we dampen the order variance too much, the decrease in capacity costs cannot compensate for the increase in inventory costs, and total supply chain costs increase. The optimal value of β depends on the relative size of capacity and inventory costs. For our numerical example, the optimal value of β equals 0,6.

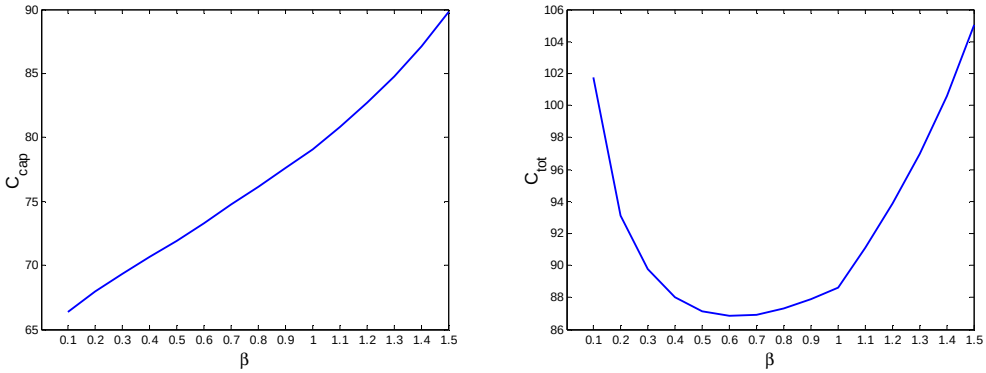

Figure 4: Flexible capacity strategy: Impact of β on capacity costs and total supply chain costs

.

6.2. Inflexible Capacity Strategy

Suppose we assume a daily capacity equal to 32.5 units (at a capacity cost of C_{CAP} = 67). This implies an average production load of $\rho = 30.5/32.5 = 0.9385$. The impact of β on the average lead time $E(T_p)$ and the optimal safety stock SS^* is shown in Fig. 5. The optimal safety stock reveals a different trend compared to the flexible strategy (Fig. 3). This is due to the stochastic lead times, which depend on the arrival pattern at the production queue. We observe in Fig. 5 that lead times increase with β due to the increased variability in order sizes. This lead time effect has an impact on the optimal safety stock. The optimal safety stock increases as the order variance is amplified $(\beta > 1)$, but decreases when the order variance is dampened to some degree, in this case up to $\beta = 0.7$. When the order variance is dampened to a large extent $(\beta < 0.7)$, the decrease in lead times cannot compensate the increase in inventory variability and safety stocks increase sharply.

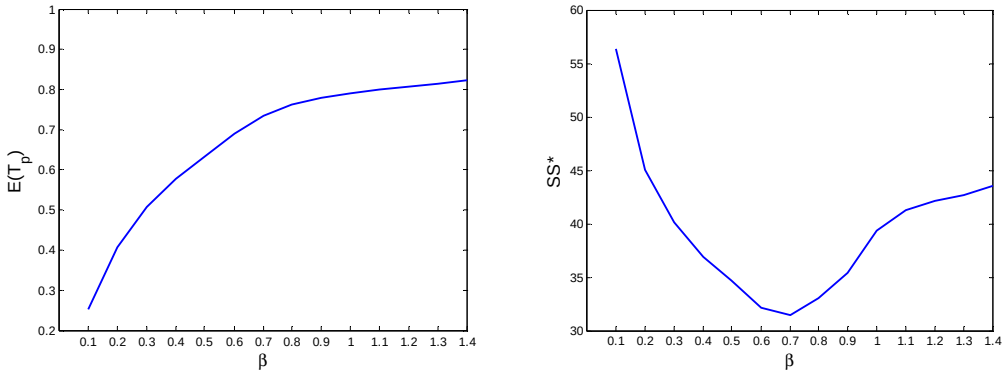

Figure 5: Inflexible capacity strategy: Impact of β on average lead time $E(T_p)$ and optimal safety stock SS*

The corresponding inventory costs show a similar trend (Fig. 6). Since capacity costs remain fixed, independent of β, total supply chain costs are obtained by adding the capacity cost of $C(K)=67$ to the inventory costs.

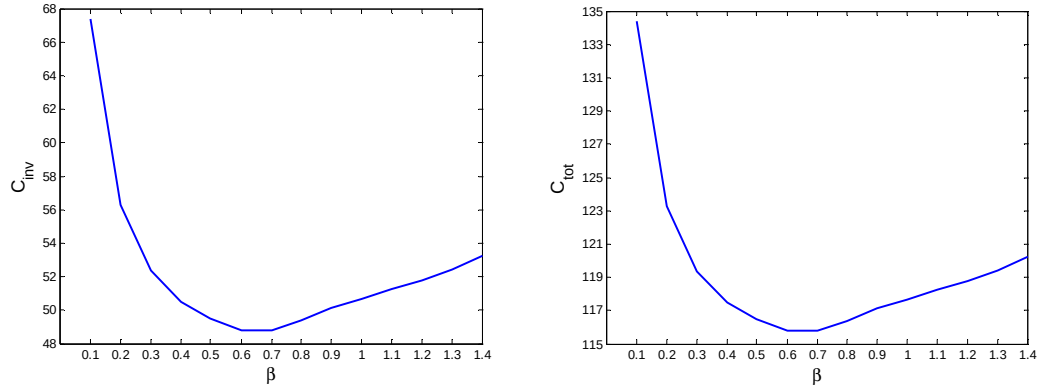

Figure 6: Inflexible capacity strategy: Impact of the replenishment parameter β on inventory costs and total supply chain costs when K = 32.5 (C_{CAP}=67)

Suppose we increase the installed capacity slightly to $K = 33$ (at a total capacity cost of $C_{CAP} = 68$). This extra capacity investment decreases the average production load to $\rho = 30.5/33 = 0.9242$, which in turn causes lead times to decrease. Since lead times determine the

optimal safety stocks, an investment in excess capacity will reduce the corresponding inventory costs.

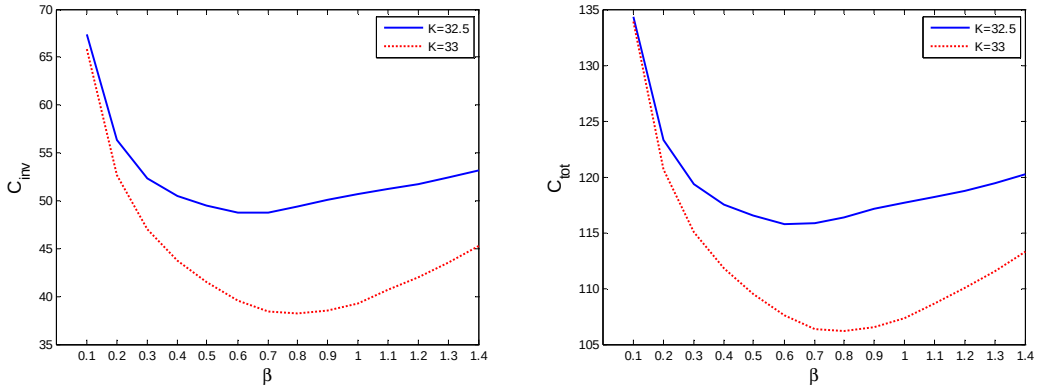

Figure 7: Inflexible capacity strategy: Impact of the replenishment parameter β on inventory costs when K = 33 (C_{CAP}=68)

In Fig. 7 we plot the inventory costs when we increase the capacity to $K = 33$, and compare it with the case where $K = 32{,}5$. We observe that the inventory costs are indeed lower when we increase capacity. Moreover, adding the capacity cost of $C(K)=68$ to these inventory costs, we obtain lower total supply chain costs: the decrease in inventory costs compensates the increase in capacity costs. Hence, in this case, it is beneficial to increase capacity (at extra cost) since it improves total supply chain performance.

6.3. Impact of the cost parameters on the replenishment policy

As previously mentioned, the value of the replenishment parameter β that minimises total supply chain costs depends heavily on the relative costs of inventory and capacity. Consider in our numerical example a higher capacity costs of $C_K = 4$ for a unit produced with the installed capacity and $C_P = 10$ for a unit produced in overtime capacity.

In case capacity is flexible, we obtain total capacity costs as shown in Fig. 8. As capacity costs are more expensive, it is preferable to dampen the orders to a larger extent. In the considered numerical example, it is optimal to smooth orders with a value of $\beta=0{,}4$.

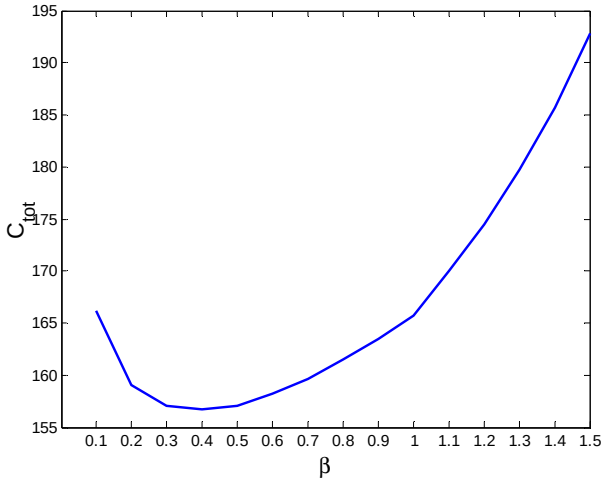

Figure 8: Flexible capacity strategy: Impact of the replenishment parameter β on total supply chain costs with increased capacity costs

In case capacity is inflexible, the curve of the total cost function will remain unchanged as capacity is fixed, independent of the replenishment parameter. Obviously, total costs will be higher as capacity is more expensive (see Fig. 9).

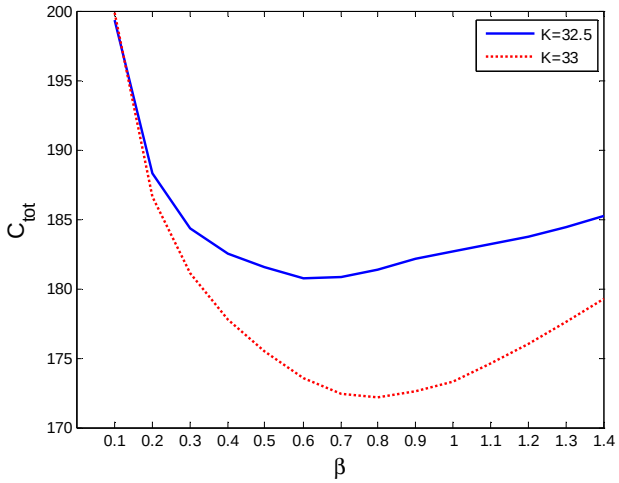

Figure 9: Inflexible capacity strategy: Impact of the replenishment parameter β on total supply chain costs with increased capacity costs

6.4. Summary

This numerical example well illustrates the dynamics resulting from the retailer's inventory decision and the manufacturer's strategy of a flexible or an inflexible capacity. Both in the flexible and inflexible capacity scenarios, order variance amplification increases total supply chain costs, and order variance dampening may lead to lower supply chain costs. Consequently, order smoothing is preferable. The degree to which we should smooth, depends on the observed customer demand pattern and the cost components in the supply chain.

7. CONCLUSIONS

In this paper we analyse the impact of the replenishment rule at the retailer on the performance of two-echelon retailer-manufacturer supply chain. We treat the variability of the order rate of the retailer as a primary decision variable to minimise total supply chain costs. The manufacturer prefers a dampened or smooth order pattern from his retailer, as this enables him to minimise his own capacity costs. The retailer, however, is not inclined to do so since a reduction in his order variance comes at the cost of an increased inventory. Both order variance amplification and dampening increase the retailer's inventory variability, inflating his safety stock requirements.

We propose a coordinative supply chain approach, thereby considering two strategies with regard to the capacity strategy. Both capacity scenarios reveal different dynamics with regard to the inventory and capacity costs in the supply chain. However, when considering a total supply chain perspective, we find that in both scenarios dampening the order variability at the retailer may lead to lower total supply chain costs. The degree to which we should smooth depends on the observed customer demand pattern and the cost components in the supply chain. At the same time we find that order variance amplification increases total supply chain costs, both in the flexible and inflexible capacity scenario.

ACKNOWLEDGMENTS
We gratefully acknowledge the valuable comments received from the anonymous reviewers and the participants of the IGLS Conference 2006 in Innsbruck. Their constructive remarks enabled us to substantially improve on earlier versions of the paper. This research contribution is supported by the contract grants G.0051.03 and G.0547.09 of the Research Programme of the Fund for Scientific Research — Flanders (Belgium) (F.W.O.-Vlaanderen).

APPENDIX A: ORDER PATTERN GENERATED BY THE GENERALISED ORDER-UP-TO POLICY

In this appendix we show that the generalised order-up-to policy given by Eqn. (4) generates an auto-correlated order pattern given by

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

Proof. The generalised order-up-to policy generated orders according to

$$O_t = E(D) + \beta \cdot [DIP - IP_t].$$

Then,

$$\begin{aligned} O_t - O_{t-1} &= E(D) + \beta \cdot [DIP - IP_t] - E(D) - \beta \cdot [DIP - IP_{t-1}] \\ &= \beta \cdot (IP_{t-1} - IP_t). \end{aligned} \quad \text{(A1)}$$

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

$$\begin{aligned} IP_t &= IP_{t-1} + O_{t-1} - D_t \\ IP_{t-1} - IP_t &= D_t - O_{t-1} . \end{aligned} \quad \text{(A2)}$$

Substituting (A2) into (A1) results in

$$\begin{aligned} O_t - O_{t-1} &= \beta \cdot (D_t - O_{t-1}). \\ O_t &= (1 - \beta) \cdot O_{t-1} + \beta \cdot D_t . \end{aligned}$$ ■

APPENDIX B: BOUNDS ON THE ORDER QUANTITIES GENERATED BY THE GENERALISED OUT POLICY

This section provides upper and lower bounds on the order quantities generated by the generalised order-up-to policy in Eqn. (4).
When $0 < \beta < 1$ the minimal and maximal order quantities are given by

$$O_{min} = D_{min}$$
$$O_{max} = D_{max},$$

since the generated order quantity is a simple exponential smoothing from the observed customer demand.
When $1 < \beta < 2$ we prove that the theoretical minimum and maximum order quantities are respectively given by

$$O_{min} = \frac{D_{min} + (1-\beta)D_{max}}{2-\beta}$$
$$O_{max} = \frac{D_{max} + (1-\beta)D_{min}}{2-\beta}.$$

Proof. Let the order quantity O_t reach its maximal value O_{max} in an arbitrary period t. Then, the order quantity in the next period t + 1 reaches its new minimum value O_{min} when the minimum demand realises, or

$$O_{t+1} = \beta \cdot D_{min} + (1-\beta) \cdot O_t$$
$$= O_{min}.$$

Subsequently, a new maximum O_{max} is reached in the following period when the maximum demand is realised, or

$$O_{t+2} = \beta \cdot D_{max} + (1-\beta) \cdot O_{t+1}$$
$$= O_{max}.$$

Suppose the order pattern successively reaches its new minimum and maximum order quantity. Then, O_{2n} and O_{2n+1} are the respective minimum and maximum order quantities, given by

$$O_{min} = O_{2n} = \beta \cdot D_{min} + (1-\beta) \cdot O_{2n-1} \quad \text{(A3)}$$
$$O_{max} = O_{2n+1} = \beta \cdot D_{max} + (1-\beta) \cdot O_{2n}. \quad \text{(A4)}$$

When $1 < \beta < 2$, we find that the minimum and maximum order quantities are respectively given by

$$O_{min} = \frac{D_{min} + (1-\beta)D_{max}}{2-\beta} \quad \text{(A5)}$$
$$O_{max} = \frac{D_{max} + (1-\beta)D_{min}}{2-\beta} \quad \text{(A6)}$$

Indeed, substituting (A5 – A6) into (A3 – A4) returns (A5 – A6) again. ■

Furthermore, using (A5), the restriction $O_{min} \geq 1$ can then be translated as

$$D_{min} + (1-\beta) \cdot D_{max} \geq 2-\beta.$$

APPENDIX C: DISTRIBUTION OF THE NET STOCK

In this section we derive an expression for the net stock distribution in function of the distribution of customer demand.

The inventory on hand NS_t at the end of period t is equal to the initial inventory on hand plus all replenishment orders received so far minus total observed customer demand. Since at the end of period t, the order O_{t-k} is in service, the orders placed more than k periods ago, i.e. O_{t-i}, $i \geq k+1$, are already delivered in inventory, while customer demand is satisfied up to the current period *t*. For our purposes the initial inventory level is a control variable, equal to the safety stock *SS*, determining the retailer's customer service. Since we assume that $O_t=D_t=E(D)$ *for* $t \leq 0$, the net stock after satisfying demand in period t is equal to

$$NS_t = SS + \left(E(T_p)+1\right)\cdot E(D) + \sum_{i=k+1}^{t-1} O_{t-i} - \sum_{i=0}^{t-1} D_{t-i}. \qquad \text{(A7)}$$

Substituting the auto-correlated order pattern (5) into (A7) gives

$$\begin{aligned} NS_t &= SS + \left(E(T_p)+1\right)\cdot E(D) + \sum_{i=k+1}^{t-1}\left[(1-\beta)\cdot O_{t-i-1} + \beta\cdot D_{t-i} - D_{t-i}\right] - \sum_{i=0}^{k} D_{t-i} \\ &= SS + \left(E(T_p)+1\right)\cdot E(D) + \sum_{i=k+1}^{t-1}\left[(1-\beta)\cdot O_{t-i-1} - (1-\beta)\cdot D_{t-i}\right] - \sum_{i=0}^{k} D_{t-i}. \end{aligned}$$

Since $O_t = D_t = E(D)$ for $t \leq 0$, we find after backward substitution of Eqn. (5) that, for $t > 0$,

$$O_t = (1-\beta)^t \cdot E(D) + \sum_{j=1}^{t} \beta(1-\beta)^{j-1} D_{t-j+1},$$

so that we obtain

$$\begin{aligned} NS_t &= SS + \left(E(T_p)+1\right)\cdot E(D) + \sum_{i=k+1}^{t-1}\left[(1-\beta)^{i-k}\cdot E(D) + \sum_{j=1}^{t-i-1}\beta(1-\beta)^j D_{t-i-j} - (1-\beta)\cdot D_{t-i}\right] - \sum_{i=0}^{k} D_{t-i} \\ &= SS + E(T_p)\cdot E(D) + \sum_{i=k}^{t-1}\left[(1-\beta)^{i-k}\cdot\left(E(D) - D_{t-i}\right)\right] - \sum_{i=0}^{k-1} D_{t-i}. \end{aligned}$$

REFERENCES

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

Boute, R.N., Disney, S.M., Lambrecht, M.R. and Van Houdt, B. (2006). An integrated production and inventory model to dampen upstream demand variability in the supply chain. *European Journal of Operational Research*, Vol.178 (1), pp 121-142.

Boute, R.N., Disney, S.M., Lambrecht, M.R. and Van Houdt, B. (2008). A win-win solution

for the bullwhip problem. *Production Planning & Control*, Vol??, N°??, 2008, pp 1-10

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

Disney, S.M. (2001). *The production and inventory control problem in Vendor Managed Inventory supply chains.* PhD thesis. Cardiff University.

Disney, S.M. and Towill, D.R. (2002). A robust and stable analytical solution to the production and inventory control problem via a z-transform approach. *Proceedings of the 12th International Conference on Production Economics*, Igls, Austria, pp. 37-47.

Disney, S.M. and Towill, D.R. (2003). On the bullwhip and inventory variance produced by an ordering policy. *Omega*, 31, pp. 157-167.

Forrester, J. (1961). *Industrial Dynamics.* MIT Press, Cambridge MA.

Gaalman, G. and Disney, S.M., (2006), "On the bullwhip effect of Order-Up-To policies for ARMA(2,2) demand and arbitrary lead-times", *Proceedings of the 14th International Working Conference on Production Economics*, Innsbruck, Austria, February, Vol. 4, pp. 55-64.

Hopp, W.J. and Spearman, M.L. (2001). *Factory Physics.* 2nd edn. Irwin, McGraw-Hill.

Hosoda, T. (2005). The principles governing the dynamics of supply chains. PhD thesis. Cardiff University.

Latouche, G. and Ramaswami, V. (1999). Introduction to Matrix Analytic Methods and stochastic modeling. SIAM. Philadelphia.

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

Magee, J. F. (1958). *Production Planning and Inventory Control*. McGraw-Hill. New York.

Nahmias, S. (1997). Production and Operation Analysis. 3rd edn. McGraw-Hill.

Naim, M.M. and Towill, D.R. (1995). What's in the pipeline? *Proceedings of the 2nd International Symposium on Logistics*, pp. 135-142.

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

Sterman, J.D. (1989). Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Science*, 35 (3), pp. 321-339.

Van Mieghem, J. A. (2008). *Operations Strategy, principles and practice*. Dynamics Ideas, Belmont, Mass. U.S.A.

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