Ann Oper Res (2017) 259:217–239
DOI 10.1007/s10479-017-2542-z

ORIGINAL PAPER

Quantifying sustainable control of inventory systems with non-linear backorder costs

Lina Johansson[1] · **Fredrik Olsson**[1]

Published online: 30 May 2017

Abstract Traditionally, when optimizing base-stock levels in spare parts inventory systems, it is common to base the decisions either on a linear shortage cost or on a certain target fill rate. However, in many practical settings the shortage cost is a non-linear function of the customer waiting time. In particular, there may exist contracts between the spare parts provider and the customer, where the provider is obliged to pay a fixed penalty fee if the spare part is not delivered within a certain time window. We consider a two-echelon inventory system with one central warehouse and multiple local sites. Focusing on spare parts products, we assume continuous review base stock policies. We first consider a fixed backorder cost whenever a customer's time in backorder exceeds a prescribed time limit, second a general non-linear backorder cost as a function of the customer waiting time, and third a time window service constraint. We show from a sustainability perspective how our model may be used for evaluating the expected CO_2 emissions associated with not satisfying the customer demands on time. Finally, we generalize some known inventory models by deriving exact closed form expressions of inventory level distributions.

Keywords Inventory · Multi-echelon · Non-linear backorder cost · Time window service constraint · Sustainability

1 Introduction and literature review

In many industries spare parts and aftersales in general are big business (Cohen et al. 2006). For example, the US automotive aftermarket was estimated to be worth $188.6 billion in 2007 (US Automotive Parts Industry Annual Assessment 2009). Another example is the

✉ Fredrik Olsson
fredrik.olsson@iml.lth.se

Lina Johansson
lina.johansson@iml.lth.se

[1] Department of Industrial Management and Logistics, Lund University, P.O. Box 118, 221 00 Lund, Sweden

aviation industry, which stocks spare parts for several billion US-dollars (Harrington 2007). However, although it may be very costly to invest in expensive spare parts, it is crucial to have spare parts available when needed. Obviously, delays and downtimes of bottleneck production equipment may be very costly. A particularly striking example is in the oil rig industry, where the production downtime on an oil rig may incur a cost of $200,000 per day (Turban 1988). For a general overview of models for spare parts inventory control see, e.g., Basten and van Houtum (2014).

We consider a two-echelon inventory system with N locations and one central warehouse. All locations and the warehouse apply $(S-1, S)$ policies (i.e., continuous review base stock control), which is reasonable for low demand items such as spare parts. This paper extends the literature on spare parts inventory control in several new directions. Three different backorder/service level structures are investigated in this paper. In more precise terms, we consider: (1) a model with piecewise constant backorder costs, (2) a model with general non-linear backorder costs, and finally (3) a model with time window service constraints. The motivation for considering these three cases stems from collaboration and discussions with industry, and the state of the current research frontier. In connection with case (1), one main focus is also to develop an approach that can be used in order to explicitly evaluate the expected CO_2 emissions associated with not satisfying the customer demands on time. We provide an exact analysis for all cases considered.

Considering case (1), all unsatisfied demands are backordered and the customers at each location i are satisfied if their demands are met directly or after an acceptable waiting time ω_i. That is, if a customer receives the requested unit within this acceptable waiting time, no penalty cost is incurred. If a customer, on the other hand, has to wait longer than this given time limit, a considerable fixed backorder/penalty cost has to be paid. This backorder cost is independent of the additional waiting time exceeding ω_i. This particular scenario is quite common in many practical settings. One example is Tetra Pak Technical Services which provides customers in the dairy industry with spare parts for packaging machines. In such a case, having to wait for a critical spare part means that the production process is halted. While a short downtime may be acceptable, it is more crucial if the customer at the production site has to wait for the spare part longer than a certain critical time limit. Then, if the critical time limit is violated, the whole batch of the dairy product must be discarded due to the perishable nature of the product. In situations like this one, it is quite common that there exists a service agreement (or contract) in which it is stipulated that the service provider should pay a fixed penalty if the spare part is not delivered within a certain time window. In this case, the fixed penalty cost is directly associated with the cost of discarding the batch of the dairy product (and to some extent also the set up cost for a new batch). Other similar examples can be found in the agriculture sector, where a whole (or parts of) harvest may be lost if the harvest-machines are down due to missing critical spare parts. Similarly as in the case with dairy products, a lost harvest incur a large, well defined, fixed cost.

Another main contribution of this paper is to take a first step in quantifying how decision rules for the logistics system affect the expected CO_2 emissions from a production waste perspective. As noted in Marklund and Berling (2017), there are few models that include emissions associated with not satisfying customer demand in a timely fashion. Marklund and Berling (2017) argue that this is particularly accentuated in the distribution of spare parts, and they explicitly state: “The parts are often quite small suggesting quite limited emissions associated with transportation, inventory holding and warehousing. However, not delivering them promptly may have serious consequences on costs and emissions”. In fact, most literature concerning how to control supply chain systems in order to reduce (or at least quantify) CO_2 emissions have primarily focused on transportation issues, see e.g. McKinnon

(2010) and Alkawaleet et al. (2014). However, it is important to notice that supply chain policy decisions also affect CO_2 emissions related to waste at the customer sites. For example, as noted above, a whole production batch of a dairy product may be wasted if the coordination of supply and demand for critical spares is not aligned.

Of course, there may exist cases where other types of backorder cost structures are more plausible than a model with piecewise constant backorder costs. Therefore, in case (2) we generalize the structure of the backorder cost and focus on general non-linear backorder costs as a function of the waiting time. For many companies a quite long waiting time for a critical spare part may be severe, while a relatively short waiting time may not incur very large costs. Hence, although our modeling technique can handle general non-linear backorder costs, we focus on exponentially increasing backorder costs due to its intuitive and appealing features from a practical point of view. Another possible situation may be when the backorder cost is so called S-shaped. This means that the backorder cost as a function of longer waiting times is concave (instead of convex as in the exponential case). We provide a model that can handle all such possible backorder cost structures (as mentioned, the exponential case should be viewed as a concrete example).

In case (3) we consider a service level instead of a model with backorder costs. This service level is defined as the probability of satisfying a customer demand within a certain time window. In all three cases we utilize information about the timing of outstanding orders for the central warehouse and the downstream locations, respectively.

Early studies on continuous review multi-echelon inventory models include, e.g., Sherbrooke (1968), Graves (1985) and Axsäter (1990). Sherbrooke (1968) considers a two-echelon inventory system with multiple local retailers and one central warehouse, all applying $(S-1, S)$ ordering policies where unsatisfied demands are backordered. Given this inventory system, he develops an approximate method (the METRIC approximation), where the real stochastic leadtimes for the retailers are approximated by their mean values. Graves (1985) extends the results from Sherbrooke (1968) by deriving an exact solution procedure. However, one of the main problems is that the solution provided in Graves (1985) is not in closed form, which means that the evaluation of the inventory level probabilities will be approximate (due to necessary truncation of infinite series). In order to ease the computational burden, Graves (1985) also presents corresponding approximate distributions. Given the same inventory system as in Graves (1985), Axsäter (1990) develops a different exact solution procedure. However, similar to Graves (1985), Axsäter (1990) does not either present a closed form solution. Moreover, he does not derive the distribution of the inventory levels for the local retailers. Instead, he derives a recursive solution procedure in order to obtain the total average cost. In this paper, we extend the analysis of Graves (1985) and Axsäter (1990) by deriving closed form solutions of the probability distributions of the inventory levels and the customer waiting times. For more information concerning continuous review multi-echelon inventory systems see, e.g., Axsäter (1993).

The literature on multi-echelon inventory models with time window service levels is relatively limited. Ettl et al. (2000) consider a multi-echelon inventory system with similar service requirements as in our model. They model their inventory system as a $M^X/G/\infty$ queueing system, and consider both assembly and distribution structures. Although Ettl et al. (2000) in some aspects consider a more general model than we do, their analysis require assumptions like leadtimes based on the exponential distribution. Another related model is presented in Caggiano et al. (2007), where the problem in Ettl et al. (2000) is extended to a multi-item setting. A limitation of this work is that the acceptable customer waiting times at different sites (ω_i in our case) are only allowed to be multiples of the transportation times between different echelons. Another study on multi-echelon inventory models with time window service levels is Wong et al. (2007), which deals only with average time window constraints.

One category of papers in the literature that is quite related to our work is inventory models with emergency supply. Two papers that fall into this category are Moinzadeh and Schmidt (1991) and Moinzadeh and Aggarwal (1997). The former studies a single location system, where the location applies an $(S-1, S)$ policy and has the option to choose either a normal order or an emergency order under a given ordering policy. This policy takes all available information regarding the inventory level and timing of outstanding orders into account. Moinzadeh and Aggarwal (1997) extend the work done by Moinzadeh and Schmidt (1991) by considering a base-stock two-echelon inventory system with one warehouse, multiple retailers and an outside supplier with the possibility of emergency orders. Huang et al. (2011) introduces a committed service time, in which it is acceptable for the customer to be served. After this time has passed, the retailers face a backorder cost (per unit and time unit) and has the ability to fill the demand with an emergency order. In a more recent paper, Howard et al. (2015) evaluate an approximate two-echelon spare parts inventory model using pipeline information. In more detail, they consider an inventory system that includes a central warehouse acting as a supplier and a so called support warehouse. The purpose of the support warehouse is to provide emergency orders to the local warehouses, and as a last resort emergency transshipments can be sent directly from the central warehouse. All of these papers mentioned above consider a standard unit backorder cost per time unit. However, in our model we consider general non-linear backorder costs, and in particular piecewise constant backorder costs.

Another related stream of literature concerns lateral transshipments between warehouses, where a local warehouse with no stock on hand can request an item from another local warehouse if needed. Yang et al. (2013) assumes, just as we do, that customers have a certain pre-specified acceptable waiting time limit. Within this time limit, the local warehouse will wait for an incoming uncommitted item. If the waiting time is too long, it will request a lateral transshipment having a shorter, but positive, leadtime. Olsson (2015) studies a similar model as Yang et al. (2013), but uses a backorder cost per unit and time unit instead of a time limit. For a review of papers studying lateral transshipments see, e.g., Paterson et al. (2011).

The literature discussed so far is, in general, based on one-for-one ordering policies, which is also the case we consider in this paper. When instead considering batch-ordering policies in multi-echelon inventory settings with time window service levels, the model complexity will increase considerably. In a single-echelon setting Axsäter (2003b) considers an inventory system with lateral transshipments where the locations apply (R, Q) policies. In Axsäter (2003b) an approximate method is developed that uses information about the residual leadtimes of the items in the system. Other relevant literature concerning batch-ordering policies include, e.g., Axsäter (2000) and Katehakis and Smit (2012).

In the following section we formulate our model in detail. Section 3 presents the solution procedure with the derivation of exact closed form expressions for various performance characteristics. In Sect. 4, cost structures are presented together with cost evaluations. In Sect. 5, we present cost optimization procedures for all three cases considered, and in Sect. 6 numerical examples are presented and discussed. In Sect. 6, we also give an application of how to use the theory developed in order to quantify sustainability measures, such as expected CO_2 emissions related to production waste. Some concluding remarks are given in Sect. 7.

2 Model formulation

We consider a two-echelon continuous review inventory model with one central warehouse and N local sites. All transportation times are positive and constant. Customer demands follow independent Poisson processes and occur only at the local sites. Since we focus on spare parts products, we assume that replenishments are made according to base stock ordering policies. All unsatisfied demands are backordered. Furthermore, we assume that backorders at the sites and at the central warehouse are filled according to the FCFS (first come - first served) rule.

In similar models like this one, it is commonly assumed that base stock levels are based on a backorder cost per unit and time unit or a prescribed service level. In those cases a target service level is used, the service level definition is very often the so called fill rate (fraction of demand that can be satisfied directly from stock on hand). However, in this paper, we consider one step function and one general non-linear backorder cost structure, where the cost is a function of the customer's time in backorder. In addition, we also consider a time window based service level.

Let us introduce some notations for parameters and decision variables:

λ_i – customer arrival rate at local site$i$,

L_i – transportation time from the central warehoue to local site$i$,

L_0 – transportation time from the supplier to the central warehouse,

h_i – holding cost per unit and time unit at local site$i$,

h_0 – holding cost per unit and time unit at the warehouse,

N – number of local sites,

S_i – base-stock level at local site$i$,

S_0 – base-stock level at the warehouse,

We proceed by defining the two cost structures and the time window based service level in more detail. In the subsequent analysis, let Y_i denote the time a customer demand is backordered at site i. Note that $0 \leq Y_i \leq L_0 + L_i$ is stochastic and depends on the base-stock levels at site i and the central warehouse.

3 Performance characteristics

Define X_0 as the limiting (i.e., the stationary case) age of the oldest unit at the warehouse not assigned to any waiting customer, where the age is assumed to start when the unit is ordered from the outside supplier. Similarly, let X_i be the limiting age of the oldest unit at site i $(i = 1, \ldots, N)$ not assigned to any waiting customer, where the age of an item is assumed to start when the unit is ordered from the warehouse. Note that, if the warehouse has zero stock on hand when a site orders a unit, the ordered unit will arrive at the site after $L_i + Z$ unit of time, where Z is the stochastic delay at the warehouse. Obviously, we must have $Z = L_0 - X_0$, for $0 \leq X_0 \leq L_0$. Hence, given a stochastic delay of Z units of time at the warehouse, the oldest unit at site i is outstanding if $0 < X_i < L_i + Z$, and in stock if $X_i \geq L_i + Z$.

Let us proceed by noting that $X_0 \sim \text{Erlang}(\lambda_0, S_0)$ and $X_i \sim \text{Erlang}(\lambda_i, S_i)$, where $\lambda_0 = \lambda_1 + \cdots + \lambda_N$. This is the case since $(S_i - 1, S_i)$ policies are used and that the demand process is pure Poisson at the sites and at the warehouse, which means that the well known

PASTA-property (Poisson Arrivals See Time Averages) holds. It should be noted that our technique of tracking the ages of the units in the system yields a richer model than the unit tracking methodology first presented in Axsäter (1990). This is the case since the unit tracking methodology in Axsäter (1990) considers only what happens at times of customer arrivals. In our case, we have full information about the ages of the (oldest) units in the system at all times. This means that, unlike Axsäter (1990), it is possible to generalize our model to incorporate decision rules based on events which are not customer arrivals. In any case, we have the density functions, $f_{X_i}(t)$, and the distribution functions, $F_{X_i}(t)$,

$$f_{X_i}(t) = \lambda_i^{S_i} e^{-\lambda_i t} \frac{t^{S_i-1}}{(S_i-1)!}, \quad t \geq 0 \tag{1}$$

$$F_{X_i}(t) = 1 - \sum_{n=0}^{S_i-1} e^{-\lambda_i t} \frac{(\lambda_i t)^n}{n!}, \quad t \geq 0 \tag{2}$$

for $i = 0, 1, \ldots, N$. Using (1) and $Z = L_0 - X_0$, it is easy to obtain the density of the stochastic delay, Z, as

$$f_Z(t) = f_{X_0}(L_0 - t) = \lambda_0^{S_0} e^{-\lambda_0(L_0-t)} \frac{(L_0-t)^{S_0-1}}{(S_0-1)!}, \quad \text{for } 0 \leq t \leq L_0. \tag{3}$$

The probability mass in the point $Z = 0$ is found by using (2),

$$\mathbf{P}\{Z = 0\} = \mathbf{P}\{X_0 > L_0\} = 1 - F_{X_0}(L_0) = \sum_{n=0}^{S_0-1} e^{-\lambda_0 L_0} \frac{(\lambda_0 L_0)^n}{n!}. \tag{4}$$

Let us continue by deriving the probability function for the inventory level at the local site i, IL_i. Now, given the delay $Z = z$, the conditional stationary probability function for IL_i follows as

$$\mathbf{P}\{IL_i = k | Z = z\} = \frac{(\lambda_i(L_i + z))^{S_i-k}}{(S_i-k)!} e^{-\lambda_i(L_i+z)}, \tag{5}$$

Using (3)–(5), we may remove the condition on Z as follows

$$\mathbf{P}\{IL_i = k\} = \frac{(\lambda_i L_i)^{S_i-k}}{(S_i-k)!} e^{-\lambda_i L_i} \mathbf{P}\{Z = 0\} + \int_0^{L_0} \mathbf{P}\{IL_i = k | Z = z\} f_Z(z) dz. \tag{6}$$

In order to be able to evaluate inventory level probabilities exactly, we provide the following closed form expression of the inventory level probability function:

Proposition 1 *For $k \leq S_i$, $S_0 > 0$, $n := S_i - k$, $m := S_0 - 1$, and $\mu := \lambda_0 - \lambda_i \geq 0$, the closed form expression of* $\mathbf{P}\{IL_i = k\}$ *is given by*

$$\mathbf{P}\{IL_i = k\} = \frac{(\lambda_i L_i)^{S_i-k}}{(S_i-k)!} e^{-\lambda_i L_i} \mathbf{P}\{Z = 0\} \\ + A \sum_{k_1=0}^{n} \sum_{k_2=0}^{m} (-1)^{m-k_2} \binom{n}{k_1} \binom{m}{k_2} L_i^{k_1} L_0^{k_2} \Psi(k_1, k_2), \tag{7}$$

where

$$\Psi(k_1,k_2) = e^{\mu L_0}(m+n-k_1-k_2)! \sum_{j=0}^{m+n-k_1-k_2}\left[\frac{(-1)^{m+n-k_1-k_2-j}}{\mu^{m+n-k_1-k_2-j+1}}\cdot\frac{L_0^j}{j!}\right] - \frac{(-1)^{m+n-k_1-k_2}(m+n-k_1-k_2)!}{\mu^{m+n-k_1-k_2+1}},$$

$$A = \frac{\lambda_i^{S_i-k}}{(S_i-k)!}\cdot\frac{\lambda_0^{S_0}}{(S_0-1)!}e^{-\lambda_i L_i-\lambda_0 L_0}.$$

Proof See “Appendix”. □

Notice that, Proposition 1 is only defined for $S_0 > 0$. However, for $S_0 = 0$ the problem degenerates to a single-echelon inventory system, i.e., $Z \equiv L_0$.

From (7) the average stock on hand at site i is obtained as

$$\mathbf{E}\left\{IL_i^+\right\} = \sum_{k=1}^{S_i} k\mathbf{P}\{IL_i = k\}, \tag{8}$$

and the average stock on hand at the central warehouse becomes

$$\mathbf{E}\left\{IL_0^+\right\} = \sum_{k=1}^{S_0} k\frac{(\lambda_0 L_0)^{S_0-k}}{(S_0-k)!}e^{-\lambda_0 L_0}. \tag{9}$$

Let us continue by deriving an expression for the probability that an arriving customer at site i has to wait longer than ω_i units of time. In this case it is reasonable, from a practical point of view, to assume that $0 \le \omega_i \le L_i$. By conditioning on $Z = z$, we obtain the following conditional probability

$$\begin{aligned}\mathbf{P}\{Y_i > \omega_i | Z = z\} &= \mathbf{P}\{L_i + z - X_i > \omega_i\} = \mathbf{P}\{X_i < L_i + z - \omega_i\} = F_{X_i}(L_i + z - \omega_i)\\ &= 1 - \sum_{n=0}^{S_i-1} e^{-\lambda_i(L_i+z-\omega_i)}\frac{(\lambda_i(L_i+z-\omega_i))^n}{n!}.\end{aligned} \tag{10}$$

Hence, by using (3), (4), (10) and the law of total probability we obtain

$$\mathbf{P}\{Y_i > \omega_i\} = \mathbf{P}\{X_i < L_i - \omega_i\}\mathbf{P}\{Z = 0\} + \int_0^{L_0}\mathbf{P}\{X_i < L_i + z - \omega_i\}f_Z(z)dz. \tag{11}$$

Interestingly enough, similar to Proposition 1, it is possible to derive a closed form expression of the probability in (11).

Proposition 2 *The probability that an arriving customer at site i has to wait longer than ω_i units of time is obtained, in closed form, as*

$$\mathbf{P}\{Y_i > \omega_i\} = \mathbf{P}\{X_i < L_i - \omega_i\}\mathbf{P}\{Z = 0\} + \mathcal{I}, \tag{12}$$

where

$$\begin{aligned}
\mathcal{I} &= 1 - \mathbf{P}\{Z = 0\} - \sum_{n=0}^{S_i-1} \Theta(n) \sum_{k_1=0}^{n} \sum_{k_2=0}^{S_0-1} (-1)^{S_0-1-k_2} \binom{n}{k_1} \binom{S_0-1}{k_2} \\
&\quad (L_i - \omega_i)^{k_1} L_0^{k_2} \Omega(n, k_1, k_2), \\
\Omega(n, k_1, k_2) &= e^{(\lambda_0-\lambda_i)L_0} (n + S_0 - 1 - k_1 - k_2)! \sum_{j=0}^{n+S_0-1-k_1-k_2} \left[\frac{(-1)^{n+S_0-1-k_1-k_2-j}}{(\lambda_0 - \lambda_i)^{n+S_0-k_1-k_2-j}} \cdot \frac{L_0^j}{j!} \right] \\
&\quad - \frac{(-1)^{n+S_0-1-k_1-k_2} (n + S_0 - 1 - k_1 - k_2)!}{(\lambda_0 - \lambda_i)^{n+S_0-k_1-k_2}}, \\
\Theta(n) &= \frac{\lambda_i^n}{n!} \cdot \frac{\lambda_0^{S_0}}{(S_0 - 1)!} e^{-\lambda_i(L_i-\omega)-\lambda_0 L_0}.
\end{aligned}$$

Proof See “Appendix”. □

4 Cost structures

4.1 Piecewise constant backorder costs

Assume that a customer arriving at site i incurs a backorder cost that depends on the time the demand is backordered. In more detail, denote $B_i(Y_i)$ as the backorder cost, as a function of Y_i, incurred by a customer arriving at site i.

Assume that a customer arriving at site i has an acceptable waiting time of ω_i units of time for a demanded item. As a start, in view of the discussion about service contracts in the introduction, let us consider a particular simple and important case concerning the backorder cost structure:

$$B_i(Y_i) = \begin{cases} 0 & \text{if } \; Y_i \leq \omega_i, \\ b_i & \text{if } \; Y_i > \omega_i. \end{cases} \tag{13}$$

Hence, if an arriving customer has to wait longer than ω_i for a spare part, then the service providing company is obliged to pay a fixed penalty cost b_i. Otherwise, no backorder cost is incurred. Notice that, in this setting, the assumption of the FCFS-rule is not optimal but reasonable. For example, say that there are two waiting customers at location i and the customer first in line already has been waiting for more than ω_i units of time, while the second has waited less than ω_i. Then, when an item then arrives at location i from the warehouse, it would be more cost efficient to assign the incoming item to the second customer (instead of applying the FCFS-rule). Also, in practice, using a different rule than FCFS could mean that a customer that already has exceeded the waiting time limit would not be prioritized and might end up with a very long waiting time (which in practice would mean that the company soon would be out of business). Similarly, the FCFS-rule is, of course, also an issue for the cases presented in Sects. 4.2 and 4.3.

In a more general case, assume that there are K different time limits, $\omega_i^{(j)}$, $j = 1, 2, \ldots, K$. We consider the following backorder cost (or penalty cost) structure:

$$B_i(Y_i) = \begin{cases} 0 & \text{if } \; Y_i \leq \omega_i^{(1)}, \\ b_i^{(j)} & \text{if } \; \omega_i^{(j)} < Y_i \leq \omega_i^{(j+1)}, \; j = 1, \ldots, K-1, \\ b_i^{(K)} & \text{if } \; Y_i > \omega_i^{(K)}. \end{cases} \tag{14}$$

That is, if an arriving customer has to wait longer than a prescribed time limit $\omega_i^{(j)}$, the service providing company has to pay a fix penalty cost $b_i^{(j)}$. Here, we assume that $\omega_i^{(1)} < \omega_i^{(2)} < \cdots < \omega_i^{(K)}$ and $b_i^{(1)} < b_i^{(2)} < \cdots < b_i^{(K)}$, i.e. that the cost is increasing with the customer waiting time.

This cost structure can be extended to also include a standard linear backorder cost per unit of time. However, in this paper we focus on the non-linear cost expressions.

4.2 Exponential backorder costs

Here we consider a more general type of non-linear backorder cost structure that depends on the time an arriving customer has to spend in backorder. As mentioned in the introduction, although our solution procedure can handle general non-linear backorder costs, we focus on exponentially increasing backorder costs as a function of the customer waiting time. That is, for $Y_i > 0$, let us define:

$$B_i(Y_i) = c_i \cdot a_i^{Y_i}, \tag{15}$$

where $a_i > 1$ and $c_i > 0$, $i = 1, 2, \ldots, N$, are constants. In this case, the backorder cost grows exponentially with the customer's waiting time. This means that the backorder cost rapidly gets large when the customer's waiting time gets longer.

4.3 Time window service constraint

Here we consider a case where we replace the backorder cost by a time window service constraint. In more detail, assume that there is an agreement that requires that the service level should be at least ℓ_i within ω_i units of time. In other words, the time window service level at location i is defined as

$$\beta_i = \mathbf{P}\{Y_i \leq \omega_i\} \geq \ell_i, \quad \forall i \in \{1, \ldots, N\}.$$

Notice that, in view of (14), this time window service level can be extended to a more general service level definition where different target service levels may be defined for different intervals of waiting times. For example, immediate service could be 95%, while service within 4 h could be 97%, etc.

4.4 Evaluation of costs

For the case where there is a backorder cost, let EB denote the total system expected backorder cost, per unit of time. Then, the expected total cost, EC, is obtained as

$$EC = h_0 \mathbf{E}\left\{IL_0^+\right\} + \sum_{i=1}^{N} h_i \mathbf{E}\left\{IL_i^+\right\} + EB, \tag{16}$$

where $\mathbf{E}\left\{IL_i^+\right\}$ and $\mathbf{E}\left\{IL_0^+\right\}$ are defined in (8) and (9), respectively.

In the following analysis we state explicitly how to evaluate EB for piecewise constant and exponential backorder costs. The cost minimization problem for the case with a time window service constraint is defined and analyzed in the optimization section, see Sect. 5.2.

Piecewise constant backorder costs For the special case of a backorder cost with just one acceptable time limit, ω_i, we obtain EB as

$$EB = \sum_{i=1}^{N} \lambda_i b_i \mathbf{P}\{Y_i > \omega_i\}. \tag{17}$$

For the more general cost structure in (14) it is reasonable to assume that $0 \leq \omega_i^{(1)} < \omega_i^{(2)} < \cdots < \omega_i^{(K)} \leq L_i$. Hence, using (11), the probability for a customers time in backorder to be in the interval between two consecutive time limits can be written as

$$\mathbf{P}\left\{\omega_i^{(j)} < Y_i \leq \omega_i^{(j+1)}\right\} = \mathbf{P}\left\{Y_i < \omega_i^{(j+1)}\right\} - \mathbf{P}\left\{Y_i < \omega_i^{(j)}\right\}. \tag{18}$$

Given these probabilities, the expected backorder cost per unit of time follows as

$$EB = \sum_{i=1}^{N} \lambda_i \left(\sum_{j=1}^{K-1} b_i^{(j)} \mathbf{P}\left\{\omega_i^{(j)} < Y_i \leq \omega_i^{(j+1)}\right\} + b_i^{(K)} \mathbf{P}\left\{Y_i > \omega_i^{(K)}\right\} \right). \tag{19}$$

Exponential backorder costs Given the cost structure in (15), we derive the expected backorder cost EB. The time in backorder, Y_i, depends both on the age of the oldest item, X_i, and the delay at the warehouse, Z. For a given delay $Z = z$ and age of the oldest item $X_i = t$, we have $Y_i = L_i + z - t$. Hence, the conditional backorder cost for location i becomes

$$B_i(Y_i | X_i = t, Z = z) = B_i(L_i + z - t) = c_i \cdot a_i^{L_i + z - t}. \tag{20}$$

Using (1) and (20), the conditional expected backorder cost, per unit of time, for location i follows as

$$EB_i(Z = z) = \lambda_i \int_0^{L_i + z} B_i(L_i + z - t) f_{X_i}(t) dt. \tag{21}$$

Further, using (3), (4) and (21), we get the (unconditional) expected backorder cost per unit of time for location i as

$$EB_i = EB_i(Z = 0) \cdot \mathbf{P}\{Z = 0\} + \int_0^{L_0} EB_i(Z = z) f_Z(z) dz, \tag{22}$$

which in turn gives us the total expected backorder cost,

$$EB = \sum_{i=1}^{N} EB_i.$$

Remark 1 Observe that when the backorder cost has a linear structure, we have the special case $B_i(Y_i | X_i = t, Z = z) = B_i(L_i + z - t) = b_i(L_i + z - t)$, which is the backorder cost structure studied in Graves (1985) and Axsäter (1990). This directly implies that

$$EB_i(Z = z) = \lambda_i b_i \int_0^{L_i + z} (L_i + z - t) f_{X_i}(t) dt.$$

5 Optimization

5.1 Piecewise constant and general non-linear backorder costs

For a standard two-echelon inventory system with linear holding and backorder costs, the optimization procedure is relatively simple, see e.g. Axsäter (1990). For example, with linear holding and backorder costs it is easy to show that the total cost function is convex in S_i, $i = 1, \ldots, N$, for a given S_0. This appealing property is, however, lost when considering non-linear costs, as in our case. Therefore, in our optimization procedure we will use other characteristics of the system when optimizing the base-stock levels. One such characteristic is described in the following remark, which is easy to prove (we omit the details):

Remark 2 If the leadtime for location i is constant, the holding cost at location i, $H_i(S_i)$, is increasing in S_i. Furthermore, $H_i(S_i) \rightarrow \infty$ when $S_i \rightarrow \infty$.

Our optimization procedure is based on the property described in Remark 2. Due to the non-convex behavior of the total cost function, we derive upper and lower bounds for the optimal values of S_0 and S_i, $i = 1, \ldots, N$, where we denote the optimal value of S_i for a given S_0 as $S_i^*(S_0)$. Similarly, the overall optimal values of S_0 and S_i are denoted as S_0^* and S_i^*, respectively. Then, obviously, we have $\underline{S_i^*} \leq S_i^*(S_0) \leq \overline{S_i^*}$, where $\underline{S_i^*}$ and $\overline{S_i^*}$ are the lower and upper bounds, respectively. In complete analogy, we also derive an upper bound, $\overline{S_0^*}$, and a lower bound, $\underline{S_0^*}$, for the optimal value of S_0.

Hence, S_0^* is found by choosing $S_0 \in \{\underline{S_0^*}, \ldots, \overline{S_0^*}\}$ such that

$$EC(S_0) = C_0(S_0) + \sum_{i=1}^{N} C_i\left(S_0, S_i^*(S_0)\right) \tag{23}$$

is minimized (where $C_0(\cdot)$ is the average holding costs per time unit at the warehouse, and $C_i(\cdot, \cdot)$ is the average holding and backorder costs at the local site i).

The computation time regarding the optimization procedure for finding optimal base-stock levels was, on average, quite moderate (in general, less than 1 min).

Lower and upper bounds for S_i^*

Notice that, when $S_0 = \infty$, the leadtime for location i becomes the shortest possible, i.e., L_i. Hence, in order to obtain a lower bound for S_i^*, the general idea here is to minimize $C_i(S_0, S_i)$ with respect to S_i given that $S_0 = \infty$. Now, since $C_i(S_0, S_i)$, given $S_0 = \infty$, is not convex in S_i we will use Remark 2 for obtaining $\underline{S_i^*} := S_i^*(\infty)$. Recall that when $S_0 = \infty$, the leadtime is constant ($= L_i$) and therefore the characteristics of $H_i(S_i)$ in Remark 2 may be used. The method for finding $S_i^*(\infty)$ can be found below in Algorithm 1. The intuition behind Algorithm 1 is that we continue increasing S_i until the holding cost becomes greater than the minimum cost so far. Then, from Remark 1, we know that we can stop the search.

In order to obtain an upper bound for S_i^*, we use exactly the same method as when deriving the corresponding lower bound. Notice that, the longest possible leadtime for location i, $L_0 + L_i$, is obtained for $S_0 = 0$. Hence, the difference compared to the derivation of the lower bound is that we, in Algorithm 1, set $S_0 = 0$ when deriving the upper bound for S_i^*. Furthermore, instead of starting the optimization algorithm with $S_i = 0$, we can here use the lower bound, $\underline{S_i^*}$, as a starting point for S_i.

Optimization of S_0

In view of Algorithm 1, the problem of finding the optimal S_0 is relatively straight-forward. As mentioned, the total cost function is, in general, not convex in S_0. Therefore, we derive lower and upper bounds for the optimal S_0. A lower bound for S_0^* is found by optimizing $EC\left(S_0, \overline{S_1^*}, \overline{S_2^*}, \ldots, \overline{S_N^*}\right)$ with respect to S_0. That is, if all S_i, $i = 1, \ldots, N$, are fixed and chosen as large as possible, a lower bound for S_0^* is obtained. The optimization procedure is presented in Algorithm 2 below, and is quite similar to Algorithm 1, with only a few modifications. In complete analogy, an upper bound for S_0^* is found by optimizing $EC\left(S_0, \underline{S_1^*}, \underline{S_2^*}, \ldots, \underline{S_N^*}\right)$ with respect to S_0.

Algorithm 1 Computation of $\underline{S_i^*} := S_i^*(\infty)$

```
procedure
    $S_0 = \infty$; $S_i = 0$; $C_{\min} = \infty$;
    while $H_i(S_i) < C_{\min}$ do
        if $C_i(S_0, S_i) < C_{\min}$ then
            $C_{\min} = C_i(S_0, S_i)$;
            $S_{\min} = S_i$;
        end if
        $S_i = S_i + 1$;
    end while
    $\underline{S_i^*} = S_{\min}$;
end procedure
```

Algorithm 2 Computation of $\underline{S_0^*}$

```
procedure
    $S_0 = 0$; $S_i = \overline{S_i^*}\ \forall i$; $C_{\min} = \infty$;
    while $H_0(S_0) < C_{\min}$ do
        if $EC(S_0, S_1, S_2, \ldots, S_N) < C_{\min}$ then
            $C_{\min} = EC(S_0, S_1, S_2, \ldots, S_N)$;
            $S_{\min} = S_0$;
        end if
        $S_0 = S_0 + 1$;
    end while
    $\underline{S_0^*} = S_{\min}$;
end procedure
```

5.2 Time window service constraint

The optimization problem, when considering a time window service constraint instead of backorder costs, becomes

$$\min_{S_0 \geq 0, S_i \geq 0\ \forall i \in \{1,\ldots,N\}} EC(S_0, S_1, \ldots, S_N) = \sum_{i=0}^{N} h_i \mathbf{E}\left\{IL_i^+\right\} \tag{24}$$

$$\text{subject to} \quad \beta_i = \mathbf{P}\{Y_i \leq \omega_i\} = 1 - \mathbf{P}\{Y_i > \omega_i\} \geq \ell_i, \quad \forall i \in \{1, \ldots, N\}, \tag{25}$$

where ℓ_i is the target time window service level. The optimization procedure is quite similar as in Sect. 5.1. First, notice that from (10) and (11) it follows that the time window service level β_i is strictly monotonic in S_i, $i \in \{1, \ldots, N\}$, i.e., $\mathbf{P}\{Y_i \leq \omega_i | S_i\} > \mathbf{P}\{Y_i \leq \omega_i | S_i - 1\}$. Hence, for a given S_0, the minimum value of S_i which satisfies the time window constraint can easily be found.

In order to find the optimal value of S_0, again a very similar procedure as in Sect. 5.1 can be developed. In short, given $\overline{S_i^*} = \arg\min_{S_i} C_i(0, S_i)$ such that $\beta_i \geq \ell_i$, a lower bound of the optimal S_0 becomes 0 since minimizing $EC\left(S_0, \overline{S_1^*}, \overline{S_2^*}, \ldots, \overline{S_N^*}\right)$ with respect to S_0, such that (25) is satisfied, gives the lower bound $\underline{S_0^*} = 0$. This is the case since $EC\left(S_0, \overline{S_1^*}, \overline{S_2^*}, \ldots, \overline{S_N^*}\right)$ is obviously increasing in S_0. The upper bound $\overline{S_0^*}$ can be found in a similar manner by minimizing $EC\left(S_0, \underline{S_1^*}, \underline{S_2^*}, \ldots, \underline{S_N^*}\right)$ with respect to S_0, such that (25) is satisfied. That is, we start with $S_0 = 0$ and increase S_0 by one unit at a time until (25) is satisfied.

6 Applications and numerical experiments

In this section we first evaluate the three cases considered in Sect. 4 for a number of test problems. In all test problems we consider an inventory system with two downstream sites, i.e., $N = 2$. For the sake of simplicity let us also assume that the local sites are identical (although this is not necessary from a modeling point of view). Secondly, we consider how our model may be used in order to quantify environmental effects associated with not satisfying customer demands in a timely fashion. We provide here a small numerical study concerning sustainability aspects in connection with CO_2 emissions.

6.1 Numerical evaluation of different backorder cost structures

For the first cost structure with a piecewise constant backorder cost function, we let $K = 1$, i.e., there is a single acceptable customer waiting time limit. The customer arrival intensity is either $\lambda_i = 0.1$ or $\lambda_i = 0.5$. The holding costs are, for simplicity, assumed to be the same at all locations. In more detail, we consider the settings $h_0 = h_i = 0.5$ and $h_0 = h_i = 1$. Moreover, we let the backorder cost take values in $b_i \in \{10, 100, 500, 1000\}$. The main purpose is to study the effect of the backorder cost in the form of a step function. Notice that, in this numerical study we consider several different ratios of b_i and h_i. A low ratio corresponds to a situation with relatively expensive spare parts. When this ratio is relatively high, the backorder costs are considerably larger than the holding costs. This may be the case if the spare parts are relatively inexpensive and/or it is very costly if the spare part is not available when needed. In all problems the transportation time to the central warehouse is $L_0 = 10$. In Table 1 we have $L_i = 1$, and in Table 2 we set $L_i = 5$. The customer's acceptable waiting time is either 10, 30 or 50% of the transportation time to the local sites.

As we can see in Tables 1 and 2, the optimal base-stock levels at the local sites in many cases tend to increase when ω_i gets lower. On the other hand, the optimal base-stock level at the central warehouse tends to, for most cases, decrease with lower ω_i. Hence, when the acceptable time limit, ω_i, is rather low, inventory should be pushed to the downstream locations. The intuitive explanation is that when ω_i is low, it is relatively likely that an arriving customer's waiting time will exceed ω_i if there is no stock on hand at the downstream sites. Therefore, the base-stock levels at the local sites should be relatively high in these cases. On the other hand, if ω_i is relatively long it is not that crucial to keep stock close to the customers. As expected, the total expected cost increases with higher backorder costs and lower acceptable waiting times.

The results from Tables 1 and 2 also reflect the difficulty of finding structural optimization properties of the system. For example, in Table 1, it is important to notice that the optimal S_0 is in general not monotonic in ω_i. Such characteristics also make it hard for practitioners to develop heuristic optimization procedures which often rely on convexity results. Hence, since the optimization results might be counterintuitive for practitioners, it is important to use the exact optimization procedure developed in Sect. 5. However, it is interesting to note that the optimal *total amount of items* in the system, $S_0^* + \sum_{i=1}^N S_i^*$, is non-increasing in ω_i for all cases considered in Tables 1 and 2. This result is in line with intuition since when ω_i increases, customers acceptable waiting times increase, and we may lower the total amount of stock in the system and still provide sufficiently high customer service. However, due to the non-monotonic property discussed above, it is very difficult to argue exactly *how* the total stock is allocated among the local sites and the central warehouse.

Apart from the test problems where $b_i = 10$ in Tables 1 and 2, the probability of exceeding the acceptable time limit is quite small. In these problems the backorder cost, b_i, is large

Table 1 Optimal base-stock levels, expected total cost and probability of exceeding the acceptable time limit

ω_i	b_i	$h_0 = h_i = 0.5$						$h_0 = h_i = 1$					
		$\lambda_i = 0.1$			$\lambda_i = 0.5$			$\lambda_i = 0.1$			$\lambda_i = 0.5$		
		S_0^*, S_i^*	EC	$\mathbf{P}\{Y_i > \omega_i\}$	S_0^*, S_i^*	EC	$\mathbf{P}\{Y_i > \omega_i\}$	S_0^*, S_i^*	EC	$\mathbf{P}\{Y_i > \omega_i\}$	S_0^*, S_i^*	EC	$\mathbf{P}\{Y_i > \omega_i\}$
$0.1 \cdot L_i$	10	1, 1	1.51	0.4513	10, 3	3.92	0.1311	0, 1	1.99	0.6638	8, 3	6.16	0.2618
$0.3 \cdot L_i$	10	1, 1	1.49	0.4402	10, 3	3.78	0.1173	0, 1	1.98	0.6570	10, 2	5.97	0.2476
$0.5 \cdot L_i$	10	1, 1	1.47	0.4289	10, 3	3.66	0.1050	0, 1	1.97	0.6501	10, 2	5.70	0.2212
$0.1 \cdot L_i$	100	3, 2	2.96	0.0277	14, 4	6.24	0.0073	2, 2	5.21	0.0688	12, 4	11.21	0.0219
$0.3 \cdot L_i$	100	3, 2	2.91	0.0251	13, 4	6.05	0.0104	2, 2	5.13	0.0648	12, 4	10.88	0.0186
$0.5 \cdot L_i$	100	3, 2	2.86	0.0228	15, 3	5.87	0.0086	2, 2	5.06	0.0611	14, 3	10.47	0.0144
$0.1 \cdot L_i$	500	3, 3	3.88	0.0048	15, 5	7.57	0.0011	4, 2	6.94	0.0114	14, 5	14.12	0.0022
$0.3 \cdot L_i$	500	5, 2	3.83	0.0043	16, 4	7.34	0.0017	4, 2	6.75	0.0095	15, 4	13.56	0.0031
$0.5 \cdot L_i$	500	4, 2	3.70	0.0079	16, 4	7.09	0.0012	4, 2	6.60	0.0079	15, 4	13.16	0.0023
$0.1 \cdot L_i$	1000	4, 3	4.19	0.0014	16, 5	8.09	0.0006	3, 3	7.77	0.0048	15, 5	15.14	0.0011
$0.3 \cdot L_i$	1000	4, 3	4.15	0.0013	15, 5	7.88	0.0009	5, 2	7.67	0.0043	16, 4	14.75	0.0017
$0.5 \cdot L_i$	1000	5, 2	4.01	0.0031	17, 4	7.61	0.0006	4, 2	7.39	0.0079	16, 4	14.18	0.0012

$L_0 = 10$ and $L_i = 1$

Table 2 Optimal base-stock levels, expected total cost and probability of exceeding the acceptable time limit

ω_i	b_i	$h_0 = h_i = 0.5$						$h_0 = h_i = 1$					
		$\lambda_i = 0.1$			$\lambda_i = 0.5$			$\lambda_i = 0.1$			$\lambda_i = 0.5$		
		S_0^*, S_i^*	EC	$\mathbf{P}\{Y_i > \omega_i\}$	S_0^*, S_i^*	EC	$\mathbf{P}\{Y_i > \omega_i\}$	S_0^*, S_i^*	EC	$\mathbf{P}\{Y_i > \omega_i\}$	S_0^*, S_i^*	EC	$\mathbf{P}\{Y_i > \omega_i\}$
$0.1 \cdot L_i$	10	0, 2	1.63	0.4253	8, 6	4.38	0.1659	0, 1	1.98	0.7654	7, 5	6.64	0.3513
$0.3 \cdot L_i$	10	0, 2	1.56	0.3908	8, 6	3.85	0.1131	0, 1	1.93	0.7408	7, 5	5.83	0.2701
$0.5 \cdot L_i$	10	0, 2	1.49	0.3554	9, 5	3.30	0.0996	0, 1	1.87	0.7134	8, 4	5.06	0.2577
$0.1 \cdot L_i$	100	2, 3	3.44	0.0461	11, 8	7.28	0.0126	2, 3	5.95	0.0461	10, 8	13.00	0.0196
$0.3 \cdot L_i$	100	2, 3	3.21	0.0349	12, 7	6.50	0.0098	1, 3	5.62	0.0769	10, 7	11.53	0.0244
$0.5 \cdot L_i$	100	2, 3	3.03	0.0258	12, 6	5.71	0.0116	3, 2	5.22	0.0577	11, 6	10.02	0.0188
$0.1 \cdot L_i$	500	3, 4	4.44	0.0044	13, 9	8.95	0.0019	2, 4	8.09	0.0108	12, 9	16.53	0.0031
$0.3 \cdot L_i$	500	2, 4	4.27	0.0077	13, 8	8.07	0.0021	3, 3	7.61	0.0160	12, 8	14.81	0.0036
$0.5 \cdot L_i$	500	4, 3	3.97	0.0046	14, 7	7.19	0.0014	3, 3	7.04	0.0103	12, 7	13.13	0.0042
$0.1 \cdot L_i$	1000	3, 4	4.89	0.0044	12, 10	9.59	0.0011	3, 4	8.89	0.0044	13, 9	17.90	0.0019
$0.3 \cdot L_i$	1000	3, 4	4.56	0.0028	13, 9	8.70	0.0007	2, 4	8.55	0.0077	13, 8	16.14	0.0021
$0.5 \cdot L_i$	1000	3, 4	4.35	0.0017	13, 8	7.77	0.0008	4, 3	7.93	0.0046	14, 7	14.38	0.0014

$L_0 = 10$ and $L_i = 5$

Table 3 Optimal base-stock levels and expected total cost

a_i	$h_0 = h_i = 0.5$				$h_0 = h_i = 1$			
	$\lambda_i = 0.1$		$\lambda_i = 0.5$		$\lambda_i = 0.1$		$\lambda_i = 0.5$	
	S_0^*, S_i^*	EC	S_0^*, S_i^*	EC	S_0^*, S_i^*	EC	S_0^*, S_i^*	EC
1.1	1, 0	0.50	6, 2	1.48	1, 0	0.57	6, 1	1.75
1.5	2, 1	1.48	9, 3	2.64	3, 0	2.40	9, 2	3.93
2	4, 1	2.15	11, 3	3.51	3, 1	3.91	10, 3	5.88
4	7, 1	3.64	15, 3	5.38	6, 1	6.75	14, 3	9.85

$L_0 = 10$ and $L_i = 2$

compared to the holding cost h_i. Moreover, as expected, we see that the probability of exceeding the acceptable time limit is significantly higher when the backorder cost tends to be relatively low. When the items are relatively expensive to keep in stock we note that the optimal base-stock levels are low, especially when the customer arrival rate is low.

For our second cost structure with an exponentially increasing backorder cost function, we evaluate our model for a set of test problems where we let $c_i = 1$ and a_i in (15) be one of four values, $a_i \in \{1.1, 1.5, 2, 4\}$. As before, the customer arrival rate is either $\lambda_i = 0.1$ or $\lambda_i = 0.5$. The holding costs are the same at all locations and either $h_0 = h_i = 0.5$ or $h_0 = h_i = 1$. The transportation times are $L_0 = 10$ and $L_i = 2$. We consider different ratios of a_i and h_i where a lower ratio corresponds to relatively expensive spare parts and vice versa.

As seen in Table 3, the optimal base-stock levels at the central warehouse increase rather rapidly with increasing values of a_i. It is also interesting to notice that the base-stock levels at the local sites are kept at relatively low levels, although a_i increases significantly. The intuition behind this behavior is that when a_i is relatively large, long customer waiting times will be very costly, while short waiting times are not so expensive. Notice that, in order to suppress long customer waiting times it may be wise to allocate stock to the central warehouse instead of the local sites. That is, instead of allocating relatively large amount of stock to *each* local site, the optimal stock policy will suggest that stock should be kept upstream, since *long* waiting times can be suppressed just as well from stock at the central warehouse. It is interesting to note that this result runs absolutely counter to traditional multi-echelon models with fill rate constraints, see e.g., Muckstadt and Thomas (1980), Axsäter (2003a), and Hausman and Erkip (1994). Optimal solutions of such inventory systems very often show that inventory should be prioritized downstream, while the fill rate at the central warehouse may be only about 50%.

For the case with time window service level constraints $\beta_i \geq \ell_i$, we evaluate our model for test problems with three different levels of the time window service level that must be achieved, that is $\ell_i \in \{0.90, 0.95, 0.98\}$. As before, the customer arrival rate is either $\lambda_i = 0.1$ or $\lambda_i = 0.5$. The holding costs are the same at all locations and either $h_0 = h_i = 0.5$ or $h_0 = h_i = 1$. The transportation times are $L_0 = 10$ and $L_i = 2$ and the acceptable waiting time ω_i is either 0, 10, 30 or 50% of the transportation time L_i.

In Table 4, we see that there often is little or no change in the optimal base-stock levels depending on the value of ω_i. With increasing ℓ_i, the base-stock levels also increase, as expected. Notice that, a time window service level constraint where $\omega_i = 0$ is the same as the traditional fill rate defined as the fraction of demand that can be satisfied immediately from stock on hand. Comparing the test problems in Table 4 where ω_i is zero to the the ones where ω_i has a larger value, it can be seen that optimizing the system using the traditional fill rate even though customers are willing to wait for a certain amount of time may lead to increased costs, especially for higher holding costs or higher customer arrival rates. For

Table 4 Optimal base-stock levels, expected total cost and time window service level

ω_i	ℓ_i	$h_0 = h_i = 0.5$						$h_0 = h_i = 1$					
		$\lambda_i = 0.1$			$\lambda_i = 0.5$			$\lambda_i = 0.1$			$\lambda_i = 0.5$		
		S_0^*, S_i^*	EC	β_i	S_0^*, S_i^*	EC	β_i	S_0^*, S_i^*	EC	β_i	S_0^*, S_i^*	EC	β_i
0	0.90	2, 2	1.83	0.9058	11, 4	3.55	0.9217	2, 2	3.65	0.9058	11, 4	7.11	0.9217
$0.1 \cdot L_i$	0.90	2, 2	1.83	0.9108	11, 4	3.55	0.9314	2, 2	3.65	0.9108	11, 4	7.11	0.9314
$0.3 \cdot L_i$	0.90	2, 2	1.83	0.9203	10, 4	3.09	0.9200	2, 2	3.65	0.9203	10, 4	6.17	0.9200
$0.5 \cdot L_i$	0.90	2, 2	1.83	0.9291	11, 3	2.63	0.9053	2, 2	3.65	0.9291	11, 3	5.26	0.9053
0	0.95	3, 2	2.31	0.9532	13, 4	4.52	0.9598	3, 2	4.62	0.9532	13, 4	9.04	0.9598
$0.1 \cdot L_i$	0.95	3, 2	2.31	0.9572	12, 4	4.03	0.9531	3, 2	4.62	0.9572	12, 4	8.06	0.9531
$0.3 \cdot L_i$	0.95	3, 2	2.31	0.9646	12, 4	4.03	0.9665	3, 2	4.62	0.9646	12, 4	8.06	0.9665
$0.5 \cdot L_i$	0.95	3, 2	2.31	0.9709	11, 4	3.55	0.9606	3, 2	4.62	0.9709	11, 4	7.11	0.9606
0	0.98	3, 3	3.30	0.9918	13, 5	5.51	0.9868	3, 3	6.60	0.9918	13, 5	11.01	0.9868
$0.1 \cdot L_i$	0.98	3, 3	3.30	0.9926	12, 5	5.01	0.9826	3, 3	6.60	0.9926	12, 5	10.02	0.9826
$0.3 \cdot L_i$	0.98	4, 2	2.80	0.9825	12, 5	5.01	0.9876	4, 2	5.61	0.9825	12, 5	10.02	0.9876
$0.5 \cdot L_i$	0.98	4, 2	2.80	0.9875	13, 4	4.52	0.9859	4, 2	5.61	0.9875	13, 4	9.04	0.9859

$L_0 = 10$ and $L_i = 2$

example, among the test problems, the cost increase can be as high as 35% (see the problems where $h_i = 1$, $\lambda_i = 0.5$, and $\omega_i = 0.5L_i$).

6.2 Application: sustainable inventory control

Let us in this section demonstrate how our model may be used for quantifying CO_2 emissions emanating from the customer production site. To this end, we focus on the practical case concerning packaging of dairy products, as mentioned in Sect. 1. To be more specific, let us assume that the dairy product is milk. Quite a few previous studies have investigated the carbon footprint related to milk production. Thoma et al. (2013) conclude that the range of CO_2 emissions is approximately 0.75–1.5 kg, per kg milk produced. Hence, we assume that, on average, there is a one to one correspondence between the amount of milk produced and the amount of CO_2 emissions (in units of weight).

Recall from the discussion in Sect. 1 that we assume a whole milk-batch is wasted if the production downtime exceeds the critical time limit ω_i. Hence, the expected CO_2 emissions of production waste in kg per unit of time becomes

$$\mathbf{E}\{CO_2\} = \sum_{i=1}^{N} \lambda_i \mathbf{P}\{Y_i > \omega_i\} M_i, \tag{26}$$

where M_i is the average batch size (in kg) at the production facility corresponding to the local inventory site i. In practice, M_i ranges from approximately 1000–30,000 kg.

Similar expressions for $\mathbf{E}\{CO_2\}$ for other products than milk may, of course, be evaluated by using exactly the same modeling technique, and the evaluation of $\mathbf{E}\{CO_2\}$ in (26) may be of interest in many other related applications. For example, if a CO_2 tax is introduced by the government for the specific product produced, a corresponding model should take the average CO_2 cost into account. Another related problem is government imposed restrictions on CO_2 emissions for specific products. In such a case, our modeling technique may be used in order to evaluate if these CO_2 restrictions are satisfied or not, when deciding inventory target levels in spare part logistics systems.

In order to set the CO_2 emissions emanating from customer production sites in relation to CO_2 emissions from transportation, let us consider a small set of numerical examples. For simplicity, we consider the same test bed as in Table 1. That is, in Table 1, we present optimal base-stock levels and probabilities $\mathbf{P}\{Y_i > \omega_i\}$ for a system with $N = 2$ locations (for a specific parameter setting). Assuming $M_i = 15{,}000$ as a benchmark, and given the probabilities $\mathbf{P}\{Y_i > \omega_i\}$ in Table 1 (for the case $h_0 = h_i = 1$) we list, in Table 5, the corresponding expected CO_2 emissions of production waste. As mentioned, to set the CO_2 emissions of production waste in relation to transportation, it is interesting to notice that the carbon intensity (expressed as gCO_2 per tonne-km) for heavy trucks is approximately 200 (McKinnon 2010). This corresponds to 5 tonne-km per kg CO_2. To illustrate this relation we also list, in Table 5, the equivalent number of tonne-km of transportation by heavy trucks.

In Table 5, we can conclude that $\mathbf{E}\{CO_2\}$ may be relatively large even for quite low probabilities of exceeding the acceptable waiting time. Of course, it is also clear that when $\mathbf{P}\{Y_i > \omega_i\}$ is very low (which corresponds to a very high fixed penalty cost), then $\mathbf{E}\{CO_2\}$ is also very low. However, it is important to realize that in some situations the CO_2 emissions related to (too long) production downtime may be relatively high, while the downtime cost may be moderate. For example, in order to avoid discarding a dairy product, excessive energy in terms of cooling may be considered (for other types of products, excessive heating has to be initiated in order to avoid obsolescence).

Table 5 Examples of $\mathbf{E}\{CO_2\}$ and the corresponding number of tonne-km of transportation by heavy trucks

$\lambda_i = 0.1$			$\lambda_i = 0.5$		
$\mathbf{P}\{Y_i > \omega_i\}$	$\mathbf{E}\{CO_2\}$ (kg)	Transportation (km)	$\mathbf{P}\{Y_i > \omega_i\}$	$\mathbf{E}\{CO_2\}$	Transportation (km)
0.6638	1991	9955	0.2618	3927	19,635
0.6570	1971	9855	0.2476	3714	18,570
0.6501	1950	9750	0.2212	3318	16,590
0.0688	206	1030	0.0219	329	1645
0.0648	194	970	0.0186	279	1395
0.0611	183	915	0.0144	216	1080
0.0114	34	170	0.0022	33	165
0.0095	28	140	0.0031	47	235
0.0079	24	120	0.0023	36	180
0.0048	14	70	0.0011	17	85
0.0043	13	65	0.0017	26	130
0.0079	24	120	0.0012	18	90

7 Conclusions

In this paper, we have presented an exact analysis of a two-echelon spare part inventory model with new types of backorder cost structures. By using the stationary age distribution of the units in the system we extended the results from Graves (1985) and Axsäter (1990) by deriving exact closed form expressions for the inventory level distributions (Proposition 1), and also for the customer waiting time distributions (Proposition 2).

Furthermore, we analyzed a model with a piecewise constant backorder cost, where a significant fixed cost is incurred if the customer waiting time exceeds a pre-specified threshold value. This backorder cost structure was then generalized to a general non-linear backorder cost as a function of the customer waiting time. As an example, we analyzed a model with an exponentially increasing backorder cost as a function of the waiting time. Moreover, a corresponding model with time window service levels was explored. As a final step, we investigated how policy decisions affect, indirectly, the expected CO_2 emissions related to production waste. These kinds of indirect consequences of policy decisions, in terms of expected CO_2 emissions, have been largely ignored in the literature. Instead, most papers have focused on CO_2 emissions from a transportation point of view.

For all cases considered in this paper, we also developed optimization procedures for the base-stock levels. Using these optimization procedures, we presented a numerical study in order to investigate how the optimal policy (given the base-stock policy structure) behaves. In particular, contrary to most results from multi-echelon inventory models, it is interesting to notice that inventory should be pushed to the central warehouse in cases where the cost for long waiting times are significantly larger than for short ones.

Possible future extensions may be to include some kinds of emergency supply in order to avoid stops in production and production waste. Another line of research would be to generalize to more complex demand structures, such as compound Poisson demand. Such an extension would be rather straight-forward by keeping track of all individual items in an ordered batch. Admittedly, the computational effort associated with obtaining exact inventory level distributions would, however, be significant.

Acknowledgements The authors thank two anonymous referees and Professor Johan Marklund for the useful remarks and suggestions.

Appendix

Proof of Proposition 1 Consider the integral in (6). We have:

$$\int_0^{L_0} \mathbf{P}\{IL_i = k | Z = z\} f_Z(z) dz = \int_0^{L_0} \frac{(\lambda_i(L_i + z))^{S_i - k}}{(S_i - k)!} e^{-\lambda_i(L_i+z)}$$
$$\cdot \lambda_0^{S_0} e^{-\lambda_0(L_0-z)} \frac{(L_0 - z)^{S_0-1}}{(S_0 - 1)!} dz = \frac{\lambda_i^{S_i-k}}{(S_i - k)!} \cdot \frac{\lambda_0^{S_0}}{(S_0 - 1)!} e^{-\lambda_i L_i - \lambda_0 L_0}$$
$$\int_0^{L_0} (L_i + z)^{S_i-k} (L_0 - z)^{S_0-1} e^{(\lambda_0-\lambda_i)z} dz$$

Now, for notational purposes we set $n := S_i - k$, $m := S_0 - 1$, and $\mu := \lambda_0 - \lambda_i \geq 0$. Then, the well known binomial theorem (see any textbook in Calculus) gives,

$$(L_i + z)^n (L_0 - z)^m = \sum_{k_1=0}^{n} \sum_{k_2=0}^{m} \binom{n}{k_1} L_i^{k_1} z^{n-k_1} \binom{m}{k_2} L_0^{k_2} (-z)^{m-k_2}.$$

Hence, we have the integral

$$A \int_0^{L_0} (L_i + z)^{S_i-k} (L_0 - z)^{S_0-1} e^{\mu z} dz$$
$$= A \int_0^{L_0} \left(\sum_{k_1=0}^{n} \sum_{k_2=0}^{m} \binom{n}{k_1} L_i^{k_1} z^{n-k_1} \binom{m}{k_2} L_0^{k_2} (-z)^{m-k_2} \right) e^{\mu z} dz$$
$$= A \sum_{k_1=0}^{n} \sum_{k_2=0}^{m} (-1)^{m-k_2} \binom{n}{k_1} \binom{m}{k_2} L_i^{k_1} L_0^{k_2} \int_0^{L_0} z^{m+n-k_1-k_2} e^{\mu z} dz,$$

where

$$A = \frac{\lambda_i^{S_i-k}}{(S_i - k)!} \cdot \frac{\lambda_0^{S_0}}{(S_0 - 1)!} e^{-\lambda_i L_i - \lambda_0 L_0}.$$

What remains is to calculate the integral $\Psi = \int_0^{L_0} z^{m+n-k_1-k_2} e^{\mu z} dz$. For this task, let us first consider an arbitrary positive integer M and a constant $a \geq 0$. Then, by successive integration by parts we obtain the following indefinite integral

$$\int z^M e^{az} dz = \frac{z^M}{a} e^{az} - \frac{M}{a} \int z^{M-1} e^{az} dz = \cdots = e^{az} M! \sum_{j=0}^{M} \frac{(-1)^{M-j}}{a^{M-j+1}} \cdot \frac{z^j}{j!}. \qquad (27)$$

Hence, by using (27), we have the following definite integral

$$\Psi = \int_0^{L_0} z^{m+n-k_1-k_2} e^{\mu z} dz$$
$$= e^{\mu L_0}(m+n-k_1-k_2)! \sum_{j=0}^{m+n-k_1-k_2} \left[\frac{(-1)^{m+n-k_1-k_2-j}}{\mu^{m+n-k_1-k_2-j+1}} \cdot \frac{L_0^j}{j!} \right]$$
$$- \frac{(-1)^{m+n-k_1-k_2}(m+n-k_1-k_2)!}{\mu^{m+n-k_1-k_2+1}}.$$

To conclude,

$$\mathbf{P}\{IL_i = k\} = \frac{(\lambda_i L_i)^{S_i-k}}{(S_i-k)!} e^{-\lambda_i L_i} \mathbf{P}\{Z=0\} + A \sum_{k_1=0}^{n} \sum_{k_2=0}^{m} (-1)^{m-k_2} \binom{n}{k_1} \binom{m}{k_2} L_i^{k_1} L_0^{k_2} \Psi.$$

□

Proof of Proposition 2 Consider the integral in (11), and let us once again define $\mu := \lambda_0 - \lambda_i \geq 0$. Denoting this integral as $\mathcal{I}$, we get:

$$\mathcal{I} = \int_0^{L_0} \mathbf{P}\{X_i < L_i + z - \omega_i\} f_Z(z) dz$$
$$= \int_0^{L_0} \left(1 - \sum_{n=0}^{S_i-1} e^{-\lambda_i (L_i + z - \omega_i)} \frac{(\lambda_i (L_i + z - \omega_i))^n}{n!} \right) \cdot \lambda_0^{S_0} e^{-\lambda_0 (L_0 - z)} \frac{(L_0 - z)^{S_0-1}}{(S_0-1)!} dz$$
$$= \int_0^{L_0} f_Z(z) dz - \int_0^{L_0} \sum_{n=0}^{S_i-1} \frac{\lambda_i^n}{n!} \cdot \frac{\lambda_0^{S_0}}{(S_0-1)!} e^{-\lambda_i (L_i - \omega) - \lambda_0 L_0} (L_i + z - \omega_i)^n (L_0 - z)^{S_0-1} e^{\mu z} dz.$$

By defining the function $\Theta(n)$ as,

$$\Theta(n) = \frac{\lambda_i^n}{n!} \cdot \frac{\lambda_0^{S_0}}{(S_0-1)!} e^{-\lambda_i (L_i - \omega) - \lambda_0 L_0},$$

we obtain a similar solution as in the proof of Proposition 1:

$$\mathcal{I} = 1 - \mathbf{P}\{Z=0\} - \sum_{n=0}^{S_i-1} \Theta(n) \int_0^{L_0} (L_i + z - \omega_i)^n (L_0 - z)^{S_0-1} e^{\mu z} dz$$
$$= 1 - \mathbf{P}\{Z=0\} - \sum_{n=0}^{S_i-1} \Theta(n) \left(\sum_{k_1=0}^{n} \sum_{k_2=0}^{S_0-1} (-1)^{S_0-1-k_2} \binom{n}{k_1} \binom{S_0-1}{k_2} (L_i - \omega_i)^{k_1} L_0^{k_2} \right.$$
$$\left. \cdot \int_0^{L_0} z^{n+S_0-1-k_1-k_2} e^{\mu z} dz \right). \tag{28}$$

By using (27), the integral in (28) becomes

$$\int_0^{L_0} z^{n+S_0-1-k_1-k_2} e^{\mu z} dz = e^{\mu L_0}(n+S_0-1-k_1-k_2)!$$
$$\sum_{j=0}^{n+S_0-1-k_1-k_2} \left[\frac{(-1)^{n+S_0-1-k_1-k_2-j}}{\mu^{n+S_0-k_1-k_2-j}} \cdot \frac{L_0^j}{j!} \right] - \frac{(-1)^{n+S_0-1-k_1-k_2}(n+S_0-1-k_1-k_2)!}{\mu^{n+S_0-k_1-k_2}}.$$

To conclude,

$$\mathbf{P}\{Y_i > \omega_i\} = \mathbf{P}\{X_i < L_i - \omega_i\}\mathbf{P}\{Z = 0\} + \mathcal{I}.$$

□

References

Alkawaleet, N., Hsieh, Y., & Wang, Y. (2014). Inventory routing problem with CO_2 emissions consideration. In P. Golinska (Ed.), *Logistics operations, supply chain management and sustainability*. Switzerland: Springer.

Axsäter, S. (1990). Simple solution procedures for a class of two-echelon inventory problems. *Operations Research*, *38*, 64–69.

Axsäter, S. (1993). Continuous review policies for multi-level inventory systems with stochastic demand. In S. C. Graves, et al. (Eds.), *Handbooks in OR & MS* (Vol. 4, pp. 175–197). Amsterdam: North Holland.

Axsäter, S. (2000). Exact analysis of continuous review (R, Q) policies in two-echelon inventory systems with compound Poisson demand. *Operations Research*, *48*, 686–696.

Axsäter, S. (2003a). Supply chain operations: Serial and distribution inventory systems. In A. G. de Kok, S. Graves (Eds.), *Handbooks in OR & MS, Vol. 11, supply chain management: Design, coordination and operations*. Amsterdam: Elsevier.

Axsäter, S. (2003b). A new decision rule for lateral transshipments in inventory systems. *Management Science*, *49*, 1168–1179.

Basten, R. J. I., & van Houtum, G. J. (2014). System-oriented inventory models for spare parts. *Surveys in Operations Research and Management Science*, *19*, 34–55.

Caggiano, K. E., Muckstadt, P. L., & Rappold, J. A. (2007). Optimizing service parts inventory in a multiechelon, multi-item supply chain with time-based customer service-level agreements. *Operations Research*, *55*, 303–318.

Cohen, M. A., Agrawal, N., & Agrawal, V. (2006). Winning the aftermarket. *Harvard Business Review*, *84*, 129–138.

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

Graves, S. C. (1985). A multi-echelon inventory model for a repairable item with one-for-one replenishment. *Management Science*, *31*, 1247–1256.

Harrington, L. (2007). From just in case to just in time. *Air Transport World*, *4*, 77–80.

Hausman, W. H., & Erkip, N. K. (1994). Multi-echelon vs. single-echelon inventory control policies for low-demand items. *Management Science*, *40*, 597–602.

Howard, C., Marklund, J., Tan, T., & Reijnen, I. (2015). Inventory control in a spare parts distribution system with emergency stocks and pipeline information. *Manufacturing & Service Operations Management*, *17*, 142–156.

Huang, S., Axsäter, S., Dou, Y., & Chen, J. (2011). A real-time decision rule for an inventory system with committed service time and emergency orders. *European Journal of Operational Research*, *215*, 70–79.

Katehakis, M. N., & Smit, L. C. (2012). On computing optimal (Q, r) replenishment policies under quantity discounts. *Annals of Operations Research*, *200*, 279–298.

Marklund, J., & Berling, P. (2017). Green inventory management. In Y. Bouchery, T. Tan, J. Fransoo, & C. Corbett (Eds.), *Sustainable supply chains*. Berlin: Springer.

McKinnon, A. (2010). Green logistics: The carbon agenda. *LogForum*, *6*, 1–9.

Moinzadeh, K., & Aggarwal, P. K. (1997). An information based multiechelon inventory system with emergency orders. *Operations Research*, *45*, 694–701.

Moinzadeh, K., & Schmidt, C. P. (1991). An $(S-1, S)$ inventory system with emergency orders. *Operations Research*, *39*, 308–321.

Muckstadt, J. A., & Thomas, L. J. (1980). Are multi-echelon inventory models worth implementing in systems with low-demand-rate items? *Management Science*, *26*, 483–494.

Olsson, F. (2015). Emergency lateral transshipments in a two-echelon inventory system with positive transshipment leadtimes. *European Journal of Operational Research*, *242*, 424–433.

Paterson, C., Kiesmüller, G., Teunter, R., & Glazebrook, K. (2011). Inventory models with lateral transshipments: A review. *European Journal of Operational Research*, *210*, 125–136.

Sherbrooke, C. C. (1968). METRIC: A multi-echelon technique for recoverable item control. *Operations Research*, *16*, 122–141.

Thoma, G., Popp, J., Nutter, D., Shonnard, D., Ulrich, R., Matlock, M., et al. (2013). Greenhouse gas emissions from milk production and consumption in the United States: A cradle-to-grave life cycle assessment circa 2008. *International Dairy Journal*, *31*, S3–S14.

Turban, E. (1988). Review of expert systems technology. *IEEE Transactions on Engineering Management*, *35*, 71–81.

US Department of Commerce, Office of Transportation and Machinery (2009). *US automotive parts industry annual assessment*.

Wong, H., Kranenburg, B., van Houtum, G. J., & Cattrysse, D. (2007). Efficient heuristics for two-echelon spare parts inventory systems with an aggregate mean waiting time constraint per local warehouse. *OR-Spectrum*, *29*, 699–722.

Yang, G., Dekker, R., Gabor, A. F., & Axsäter, S. (2013). Service parts inventory control with lateral transshipment and pipeline stock flexibility. *International Journal of Production Economics*, *142*, 278–289.