Managing supply chain inventories:
A multiple retailer, one warehouse, multiple supplier model

Ram Ganeshan*

Department of Quantitative Analysis and Operations Management, The University of Cincinnati, 537 Lindner Hall, ML #0130, Cincinnati, OH 45221-0130, USA

Abstract

Inventories exist throughout the supply chain in various forms for various reasons. Since carrying these inventories can cost anywhere from 20 to 40% of their value a year, managing them in a scientific manner to maintain minimal levels makes economic sense. This paper presents a near-optimal \((s, Q)\)-type inventory policy for a production/distribution network with multiple suppliers replenishing a central warehouse, which in turn distributes to a large number of retailers. The model is a synthesis of three components: (i) the inventory analysis at the retailers, (ii) the demand process at the warehouse, and (iii) the inventory analysis at the warehouse. The key contribution of the model is the seamless integration of the three components to analyze simple supply chains. The decisions in the model were made through a comprehensive distribution-based cost framework that includes the inventory, transportation, and transit components of the supply chain. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Supply chain; Inventory policy

1. Introduction

Most manufacturing enterprises are organized into networks of manufacturing and distribution sites that procure raw material, process them into finished goods, and distribute the finish goods to customers. The terms “multi-echelon” or “multi-level” production/distribution networks are also synonymous with such networks (or supply chains), when an item moves through more than one step before reaching the final customer.

Inventories exist throughout the supply chain in various forms for various reasons. At any manufacturing point, they may exist as raw materials, work in progress, or finished goods. They exist at the distribution warehouses, and they exist in-transit, or “in the pipeline”, on each path linking these facilities. All these are related in the sense that [1]:

- The downstream sites create demands on the upstream inventories.
- These uncertain demands, combined with uncertain production and/or transit times largely determine the inventory at a given site.

*Tel.: (513) 556 7174; fax: (513) 556 5499; e-mail: ram.ganeshan@uc.edu; Web: http://www.econga.cba.uc.edu/~ganeshr

1 The author would like to thank Professor Jack Hayya of Penn State University, and two anonymous referee’s for their comments on earlier drafts of this paper.
Ballou [2] estimates that having these inventories can cost anywhere between 20 and 40% of their value per year. Although carrying inventories is essential to enhance customer service and reduce distribution costs (through multiple warehouses), managing these inventories in a scientific manner to maintain minimal levels makes economic sense. Lee and Billington [3] site several opportunities that exist in managing supply chain inventories. Among them are making coordinated decisions between the various echelons, incorporating sources of uncertainty, and designing proper supply chain performance measures. The central premise here is that the lowest inventories result when the entire supply chain is considered as a single system. Such coordinated decisions have produced spectacular results at Xerox [4], and at Hewlett Packard [5], which were able to reduce their respective inventory levels by over 25%.

2. Motivation and problem setting

Determination of optimal inventory policy for a multi-echelon inventory system is made difficult by the complex interaction between the different levels. Although current research in multi-echelon inventory theory covers a wide variety of problems, it suffers from a few important shortcomings. First, contrary to most extant distribution systems, almost all published research in this area assumes an arborescence (“tree-like”) structure [6]. Under such an assumption, every site in the system receives resupply from only one higher level site, but can distribute to several lower levels. Such a single sourcing scenario seems tenable especially in the last decade, which has seen a significant shift in the sourcing strategy of many firms, moving from the traditional concept of having many suppliers to rely largely on one source. However, researchers have very convincingly argued (see ensuing review) that the use of multiple suppliers, in a majority of cases, reduces the overall inventory and distribution system costs. The obvious question then is how the use of multiple suppliers can be linked to the traditional one-warehouse, multiple-retailers scenario. Such a question can be answered through a model that links the multiple supplier and the one-warehouse, multiple-retailer components together.

A second shortcoming is that the recommendations for distribution inventory policies are made from an “inventory-only” viewpoint. Since production, and transportation are also key determinants in distribution effectiveness [2], costs that reflect production and/or transportation economies need to be incorporated into multi-echelon inventory policy considerations.

Third, research that recommends inventory policy decisions in multi-echelon inventory systems, for the most part, is restricted to either constant or well-known distributions of demand and lead time. Such assumptions are typically made to simplify the inventory analysis at each echelon in the system. Therefore, there is a need for methodologies that relax these restrictive assumptions.

This paper is an attempt to overcome some of the limitations cited above. We consider a “two-echelon” production/distribution system that operates via many identical retail outlets through the lower echelon (see Fig. 1). As random external demand reduces their inventory level, the retail outlets draw their supplies from a central warehouse or (distribution center), which in turn procures the item from many identical suppliers (or manufacturers, and hence the term “production/distribution system”). The problem is to find near-optimal stocking policies (reorder point, order quantity) at both the retailer and the DC level, under stochastic demand and lead time conditions, subject to customer service constraints.

We contribute to the literature in four important ways. First, and perhaps most importantly, we show how well-known concepts in the inventory theory literature can be used to analyze a simple supply chain. Although extensive research exists to analyze individual nodes in the supply chain, these results have not successfully extended to multiple echelons. Second, we use a network that is not arborescent (many suppliers, one warehouse, many retailers). This necessitates a conceptually different model from the ones published in the past. Third, the inventory policy decision in the multi-level system is made from an inventory-logistics framework where both inventory
and transportation components are included in the total cost function. Fourth, we show how the Tyworth et al. [7] procedure can be adapted for evaluating the inventory policy at each of the two echelons. The primary advantage of the procedure is that it relaxes the current restrictions on demand and lead time distributions.

This paper is organized in the following way. Section 3 highlights the relevant literature. Section 4 introduces the model and the solution procedure. Section 5 describes a simulation model that is used to verify the model in Section 4, and Section 6 presents our conclusions and directions for future research.

3. Literature review

The model in this paper borrows key ideas from two streams of literature: the multi-echelon inventory systems and the order-splitting systems, i.e., the splitting of an order among several suppliers or manufacturers.

3.1. Multi-echelon inventory systems

Since a supply chain consists of various levels or echelons, a key question that needs to be addressed is how partners at these various levels interact with each other. The literature bears evidence of two major but contrasting philosophies: the “push” and the “pull” system. In a distribution system such as the one described in this paper, a push philosophy would mean that there is a central decision maker, say a warehouse manager, who has access to information about inventory levels at all the concerned facilities; all inventory decisions are then made centrally based on this information. In the pull system, however, the inventory decisions are made by local managers based on their local conditions [8]. In both the push and the pull systems, the decision variables (order quantity, reorder point, etc.) are determined so that the overall system costs are minimized. Since this paper deals with a continuous review pull system, we will only review the literature concerning such systems. For a comprehensive discussion of the push system, the reader is directed to Federgruen [8].
One of the earliest continuous review multi-echelon inventory model was due to Sherbrooke [9]. He considers a two-echelon system with several retailers at the lowest echelon and a warehouse that supplies to these retailers. In determining the optimal level of inventory in the system, he introduced the now classic METRIC approximation. When the demand distribution is Poisson, the METRIC approximation essentially characterizes the outstanding retailer orders as Poisson. Graves [10] extends the METRIC approximation by estimating two parameters, the mean and the variance, to describe the outstanding retailer orders. He fits the negative binomial distribution to these parameters to determine the optimal inventory policy. Axsaäter [11] later provides an exact solution to the problem and shows that the METRIC approximation underestimates, whereas the Graves two-parameter approximation overestimates the retailer backorders. Muckstadt [12], and later Lee [13] also provide simple extensions to the basic Sherbrooke model.

All the above studies use a one-for-one ordering policy, i.e., an order is placed as soon as a demand has occurred. Axsaäter [11] showed how the methods for the one-for-one ordering policy can be extended when there is only one retailer. Analysis of batch-ordering policies in arborescent systems (when number of retailers > 1) can be done similar in spirit to Sherbrooke [9]. Deuermeyer and Schwarz [14] were perhaps the first to analyze such a system. They estimated the mean and variance of the lead time demand to obtain average inventory levels and backorders at the warehouse, assuming that lead-time demand is normally distributed. The retailer lead-time demand was also approximated using a normal distribution. (Since this paper is largely inspired by [14], the format of analysis used is similar in style to [14].) In addition to reviewing the literature in the area, Moinzadeh and Lee [15], Lee and Moinzadeh [16], and Svoronos and Zipkin [17] also provide several extensions to the Deuermeyer and Schwarz [19] model. The reader is directed to Axsaäter [11] for a review.

3.2. Order splitting among suppliers/manufacturers

Sculli and Wu [20] first showed that splitting an order between two suppliers with normal lead times reduces the effective lead times (time the first portion arrives). Hayya et al. [21], Sculli and Shum [22], Kelle and Silver [23,24], and Guo and Ganeshan [25] all use principles of order statistics to show that splitting orders between suppliers does reduce the effective lead times under gamma, normal, uniform, or Weibull lead time resulting in lesser amount of safety stocks for the same level of service.

The above studies analyze the use of many suppliers from a methodological perspective. Ramasesh et al. [26] were the first to explicitly consider the costs and benefits associated with the use of multiple suppliers. Under the assumption of constant demand and either exponential or uniform lead times, they showed that dual sourcing reduces inventory costs. Ramasesh et al. [27] came to a similar conclusion even when the exponential lead times of the two suppliers in use did not have the same parameters. More recently, Lao and Zhao [28] and Chiang and Benton [29] extended this dual sourcing study to incorporate random demand, an unequal split between the suppliers, and any stochastic lead time distribution. Chiang and Benton concluded that when lead times were stochastic, dual sourcing consistently performed better (in terms of inventory system costs) than single sourcing.

4. Model development and solution

The purpose of our model is to provide a near-optimal reorder point, order quantity \((s, Q)\)-type inventory policy for the retailers and the warehouse that minimizes the total logistics costs, subject to customer service constraints. In designing the model, three “sub-systems” need to be analyzed (see also Deuermeyer and Schwarz [14]): (i) the inventory at each of the retailers, (ii) the demand process at the warehouse, and (iii) the inventory at the warehouse. The model gives an exact representation of the retailer and warehouse inventory position and an approximation to the demand process at the warehouse. A synthesis of these three subsystems is used to arrive at the final cost-minimizing model.
4.1. Assumptions and notation

The model assumes that a single product, with a constant unit price \( v \), weighing one pound (for convenience) is stocked in a distribution system that consists of \( N_r \) identical retailers, one central warehouse or distribution center, and \( n \) identical suppliers. The daily demand at retailer \( r \) (\( r = 1, \ldots, N_r \)) is stochastic, assumed to be Poisson-distributed with a mean \( \lambda_r \). Whenever the inventory level depletes to a level \( s_r \), the retailer places an order \( Q_r \) to the warehouse. The lead times of each retailer are assumed to be the sum of three components: a constant order processing time at each retailer, waiting time at the warehouse in the event it is out of stock, and the transit time from the warehouse to the retailer which is assumed to be a random variable with a given pdf. We approximate the waiting time of a given retailer order at the warehouse by its mean using the METRIC approach. Any demand that is not met at the retailer is backordered. The warehouse’s inventory, meanwhile, depletes with the retailer orders and whenever the inventory falls below \( s_w \), it places an order of \( Q_w/n \) (i.e., the warehouse lot size is split equally between the suppliers) simultaneously to each of the \( n \) identical suppliers. To facilitate the analysis, it is assumed that \( Q_w \) is an integer multiple of \( Q_r \). It takes a constant amount of time to process and place the order to these suppliers. The transit times from each supplier, however, is stochastic, described by a random variable of a given pdf. It is assumed that the warehouse does not place an order until all the previous orders are received from the suppliers. Any retailer order not satisfied by the warehouse is backordered. Since transportation rates are dependent on the retailer lot size, it is assumed that there is no lot-splitting at the warehouse (if the warehouse has less than \( Q_r \) units in stock, the entire order of size \( Q_r \) will be backordered). Also, it is assumed that the first portion received from a supplier primarily determines the service level at the warehouse. The objective is to minimize the total inventory-logistics costs subject to customer service constraints.

The following summarizes the notation used in this paper. The definitions of some are deferred until they are used. The subscripts \( r \) and \( w \) are used to identify retailer and warehouse.

- \( N_r \) = number of retailers,
- \( n \) = number of suppliers,
- \( s_r, s_w \) = reorder point at the retailer and at the warehouse,
- \( Q_r, Q_w \) = order quantity at each retailer and at the warehouse,
- \( D_r, D_w \) = daily demand at each retailer (with mean \( \mu_{Pr} \)) and at the warehouse (with mean \( \mu_{Pw} \)),
- \( R_r, R_w \) = expected annual demand at each retailer and the warehouse, \( R_i = 360\mu_{P_i}, i = r, w \),
- \( T_r, T_w \) = the lead time for the retailer and the “first lead time” of the warehouse,
- \( L_r, L_w \) = the transit times from the warehouse to the retailer and from each of the suppliers to the warehouse with means \( \mu_{L_r} \) and \( \mu_{L_w} \),
- \( Y_r, Y_w \) = the demand during lead time at each retailer (with mean \( \mu_{Yr} \)) and at the warehouse (with mean \( \mu_{Yw} \)),
- \( \rho_r, \rho_w \) = the order fill rates at each retailer and at the warehouse.

4.2. The retailer inventory analysis

For notational ease, we assume that there is only one retailer. The following analysis will be identical for all retailers.

The retailer inventory position depletes with demand, and when the level of inventory is \( s_r \), the retailer places an order of \( Q_r \) units at the warehouse. The warehouse then ships to the retailer the order, which takes \( L_r \) periods (discrete random variable with any pmf, that includes a discrete approximation of a continuous random variable) to arrive at the retailer. In the event the warehouse is out of stock, the retailer waits an additional time, \( W_r \), the time until the warehouse is replenished by one of its suppliers (in reality \( W_r \) is zero for a retailer who is not backlogged. Like the METRIC approach, this model approximates \( W_r \) by its expected value). Therefore, the total lead time of the retailer, \( T_r \), is

\[
T_r = \gamma_r + L_r + W_r, \tag{1}
\]
where \( \gamma_r \) is a fixed order processing time. Since \( \gamma_r \) and \( W_r \) are assumed to be constants, \( T_r \) will have a pmf similar to \( L_r \).

The target number of units short per replenishment cycle can be represented as

\[
(1 - \rho_r)Q_r. \tag{2}
\]

Let \( ES_r \) represent the expected units short per replenishment cycle. The service objective, for given values of \( Q_r \) and \( \rho_r \), is to find \( s_r \) such that \( ES_r \leq (1 - \rho_r)Q_r \).

As stated earlier, period demand \( D_t \) is distributed \( \text{Poisson}(\lambda_r) \). Let the lead time have a pmf \( P(T_r = t_r) = P_{t_r} \). The conditional distribution of demand given \( T_r = t_r \), for \( t_r = 1, 2, \ldots, m \), is therefore Poisson \( (\lambda_r t_r) \) \[30\]. Since backorders occur as soon as the number of outstanding orders exceeds \( s_r \), the expected units short per replenishment cycle for each conditional distribution, \( ES_{t_r} \) (i.e., expected shortages given \( T_r = t_r \)) is given by

\[
ES_{t_r}(s_r) = \sum_{i=s_r}^{\infty} (i-s_r)e^{-\lambda_r t_r}(\lambda_r t_r)^i/i! \tag{3}
\]

Eq. (3) can be calculated as

\[
\lambda_r t_r(1 - F_{\lambda_r}(s_r - 1) - s_r(1 - F_{\lambda_r}(s_r)), \tag{4}
\]

where \( F_{\lambda_r}(\cdot) \) is the cdf of the Poisson distribution with mean \( \lambda_r t_r \). The implicit assumption in calculating Eqs. (3) and (4) is that the average backorders are relatively small to the average on-hand stock (see Ref. \[31\], pp. 273–275). Using Eqs. (3) and (4), the expected units short over the entire stochastic lead time can easily be calculated as

\[
ES_r = \sum_{t_r=1}^{m} ES_{t_r}(s_r)P_{t_r}. \tag{5}
\]

4.3. Demand process at the warehouse

Although the demand size is fixed at the warehouse \( (Q_r) \), the interval between any two orders is stochastic and depends on \( Q_r \). We characterize the demand process at the warehouse as a superposition of the retailer ordering processes.

Let \( N_r(t) \) be the orders placed by retailer \( r \) up to time \( t \). Since a retailer on average places an order once every \( Q_r \) units of demand, the times between the orders will have the Erlang-\( Q_r \) (see also Ref. \[14\]) distribution with mean \( Q_r/\lambda_r \). Let \( N_w(t) \) be the number of orders at the warehouse in time \([0, t] \), then

\[
N_w(t) = \sum_{r=1}^{N_r} N_r(t). \tag{6}
\]

It can be established that the times between orders of retailer \( r \) form a renewal process (i.e., the times between orders are iid random variables). \( N_w(t) \), therefore, is a superposition of \( N_r \) renewal processes. \( N_w(t) \) is however nonstationary \[32\]. There are two cases when \( N_w(t) \) is stationary. One, when \( Q_r = 1 \), in which case \( N_w(t) \) will be Poisson. Second, when \( N_r \) and \( t \) are sufficiently large, \( N_w(t) \) will be approximately Poisson. Although \( N_w(t) \) is Poisson for \( N_r \to \infty \) (see Ref. \[32\]), Kim \[33\] shows that when \( N_r \geq 20 \), the Poisson process is an excellent approximation to \( N_w(t) \). This is not a restrictive assumption. For example, the Clearfield, PA distribution center of Wal-Mart serves over 75 stores. Therefore, the order arrival process at the warehouse would be approximately Poisson with rate \[32\]

\[
\lambda_w = N_r \lambda_r / Q_r. \tag{7}
\]

This is an important result. Often, in solving multi-echelon problems, one of the most difficult tasks is to determine effective demand at the upstream sites \[18\]. It has now been established that when \( N_r \) is large \( (\geq 20) \), the orders, each of size \( Q_r \), arrive according to a Poisson process. This allows us to use a modification of Section 4.2 to analyze the inventory position at the warehouse.

4.4. Inventory analysis at the warehouse

The analysis of warehouse inventory for the most part is similar to the retailer analysis, with two modifications. One, we have to account for the fact that each order is of size \( Q_r \), and second, we have to consider that the warehouse uses multiple suppliers.

For a given order fill rate \( \rho_w \), the target number of orders not satisfied (recall no lot splitting at the
Let ES\(_w\) represent the expected orders not satisfied. As before, for a given value of \(Q_w\), the task is to find \(s_w\) such that \(ES_w \leq Q_w(1 - \rho_w)/Q_r\). The choice of \(Q_w\), and therefore \(s_w\), depends on the overall objective, cost minimization.

The period demand at the warehouse, \(D_w\), as seen in Section 4.3 is Poisson(\(\lambda_w\)) (note that each “demand” at the warehouse is \(Q_r\) units). Let \(T^i_w\) be the lead time of the \(i\)th supplier, \(i = 1, \ldots, n\). Since the warehouse places \(n\) orders of size \(Q_w/n\) simultaneously, the first portion of the order will have a lead time equivalent to the shortest of those \(n\) lead times. The lead time of the warehouse, \(T_w\) can therefore be written as (see also Ref. [7])

\[
T_w = \text{Min}(T^1_w, \ldots, T^n_w) + \gamma_w. \tag{9}
\]

Let \(H(t)\) be the cmf of each of the \(T^i_w\). The cmf of \(T_w\), \(I(t, n)\), therefore, can be represented as \(P(T_w \leq t)\), which in turn reduces to

\[
I(t, n) = 1 - [1 - H(t)]^n. \tag{10}
\]

Let \(ES_{\gamma_w}\) represent the expected shortages for a given effective lead time, \(t_w\). As before this can be represented as

\[
ES_{\gamma_w}(s_w) = \sum_{i=s_w}^{\infty} (i - s_w)e^{-\lambda_w t_w}(\lambda_w t_w)^i/i!. \tag{11}
\]

This expression reduces to

\[
ES_{\gamma_w}(s_w, Q_r) = \lambda_w t_w (1 - F_{\lambda_w t_w}(s_w/Q_r) - 1) - s_w/Q_r (1 - F_{\lambda_w t_w}(s_w/Q_r)), \tag{12}
\]

where \(F_{\lambda_w t_w}(.)\) is the cdf of the Poisson distribution with mean \(\lambda_w t_w\).

Therefore, \(ES_w\) can now be easily computed over the entire stochastic lead time as

\[
ES_w = \sum_{t_w=1}^{l} ES_{\gamma_w}(s_w, Q_r) P_{t_w}, \tag{13}
\]

where \(P_{t_w}\) is the probability that \(T_w = t_w\), \(t_w = 1, \ldots, l\).

### 4.5. Synthesis

This section presents a model that minimizes the total logistics system cost. Interrelationships between the various submodels described above need to be understood to build this cost model. First, the retailers have an effect on the warehouse; second, the warehouse has an effect on the retailers; and third, the supplier lead times have an effect on the warehouse.

The retailers have an effect on the warehouse through their demand process. We have established that the order arrival process at the warehouse is approximately Poisson, if the number of retailers is greater than or equal to 20. The ordering policy of the retailers depends on the local inventory position, or alternatively the demand at each retailer location. Section 4.2 described a way to calibrate \(s_r\) for given values of \(Q_r\) and \(D_r\) at each of the retailers. Given \(Q_r\), Section 4.3 described the order arrival process at the warehouse.

The warehouse, however, has an effect on the retailers also. In theory, the warehouse may be understocked when a retailer order arrives. Therefore, that particular order has to wait a random amount of time (this is the time until the first split order from one of its suppliers arrives) until it is satisfied. This random wait time depends on several factors: the number of backorders at the warehouse at the instant order \(m\) is placed by retailer \(r\) (it is noted here that to determine this random time, every order from every retailer needs to be tracked), the number of warehouse orders outstanding, the number of retailer demands prior to this demand, the number of suppliers who have not yet delivered their portions, etc. Therefore, it is through this random time that all the retailers are related to each other (violation of the assumption that the retailers are independent). The random wait time is practically implemented in this paper by assuming that all retailers wait a constant amount of time equal to the expected value of \(W_r\). It is not a restrictive assumption, since most retail warehouses have high order fill rates (the Clearfield, PA, distribution center of Wal-Mart has a 97% fill rate). Therefore the amount of time the warehouse is backordered is often a very small portion of the retailer lead time. The average delay, \(E(W_r)\), can be
determined as
\[ Q_w(1 - \rho_w)/(Q_r)^2. \]  
(14)

Finally, the warehouse lead time depends on the minimum of the supplier lead times. A stream of research (see Section 3) has established that order splitting reduces the risk of shortage. Alternatively, the warehouse can achieve a higher level of service with the same level of safety stock by using more suppliers. On the other hand, since transportation costs are sensitive to the weight shipped, economies cannot be achieved by splitting the order among too many suppliers. Also, in calculating the warehouse lead time, it is assumed (to facilitate the analysis) that the chance of orders crossing is negligible. Such an assumption is tenable when \( Q_\infty \) is large with respect to mean retailer demands during lead times. Mathematically,
\[ Q_w/\mu_{D_w} > B, \]  
(15)
where \( B \) is a positive real number whose value depends on the probability with which the orders to the \( n \) suppliers do not crossover. For a required probability \( \tau \) \[ 23],
\[ B > F^{-1}_l(\tau^{1/n}). \]  
(16)

4.6. The total cost model

The expected total annual cost for the system, \( \text{ETAC} \), can be expressed as
\[ \text{ETAC}(Q_w, s_w, Q_r, s_r) = c_o + c_h + c_s, \]  
(17)
where \( c_o \) is the ordering cost, \( c_h \) the holding costs, and \( c_s \) the transportation cost. We can express each of these costs as follows (the first term in each of these expressions represents cost at the warehouse, and the second term the aggregated cost at all the retailers):

The total ordering cost is
\[ c_o = (a + bn)/Q_w + R_rN_rA_r/Q_r, \]  
(18)
where \( a \) and \( b \) are setup cost parameters at the warehouse, and \( A_r \) is the ordering cost at the retailer.

The holding cost is
\[ c_h = (\mu_{L_w} + Q_w/2 + s_w - \mu_{Y_w})vh \]  
+ (\mu_{L_r} + Q_r/2 + s_r - \mu_{Y_r})vhN_r,  
(19)
where \( \mu_{L_w}, \mu_{D_w}, Q_w/2, \) and \( s_i - \mu_{Y_i}, i = w, r, \) represent the in-transit, cycle, and safety stocks receptively at the warehouse and each of the retailers. The symbol \( h \) is the cost of holding one unit for one year.

The transportation cost is
\[ c_s = g(Q_w/n)R_w + g(Q_r)N_rR_r, \]  
(20)
where \( g() \) represents the functional relationship between the transportation rate and lot size (or weight of each lot). Recall that, without loss of generality, we have assumed that the weight of each unit is one pound). The total transportation cost is therefore the sum of shipping \( R_w \) units from the suppliers to the warehouse and \( R_r \) units to each of the \( N_r \) retailers.

Therefore, the problem is to find \( Q_w, s_w, Q_r, s_r \) that minimize \( \text{ETAC} \) subject to
\[ \text{ES}_w \leq Q_w(1 - \rho_w)/Q_r, \]  
(21)
\[ \text{ES}_r \leq Q_r(1 - \rho_r), \]  
(22)
\[ Q_w = kQ_r, \]  
(23)
\[ Q_w \geq B\mu_{D_w}, \]  
(24)
\[ s_w, s_r \geq 1. \]  
(25)
Eqs. (21) and (22) represent customer service constraints (Sections 4.2 and 4.4). Constraint (23) assumes that \( Q_w \) is an integer multiple of \( Q_r, k \) being any positive integer. This ensures that the methodology described in Section 4.2 can be used for the distribution center also. Constraint (24) assures that the warehouse orders to the suppliers do not crossover with a probability \( \tau \). Eq. (17) can be optimized by using numerical search techniques such as the Newton or the conjugate gradient method.

5. Model verification

5.1. A simulation model

The model presented in Section 4 is based on a number of simplifying assumptions. This section
describes a computer simulation model that was used to verify the approximate model.

The simulation of the model presented in Section 4 was constructed using the SLAM II [34] simulation language. The inputs to the simulation are the number of retailers, the number of suppliers, the demand distribution at each retailer, the distribution of the retailer–warehouse transit times, the supplier–warehouse transit times, and the lot-sizes and the reorder points at each retailer and the warehouse.

The lot sizes and the reorder points of each retailer and the warehouse that were input to the simulation were determined through the analytical model (for some prespecified level of retailer and warehouse fill rates). The purpose of the simulation was to obtain point and interval estimates of the retailer and the warehouse backorder levels (or fill rates) using these analytically determined lot sizes and reorder points. These will be compared to the actual fill rate to arrive at the accuracy of the model.

The simulation was run for 10 years (a year is assumed to have 360 days). The simulation statistics were collected over the last nine years to account for the initialization bias. Several pilot runs were conducted to make sure that this length was enough to eliminate the initialization bias and was adequate to provide independence (of the performance measures in question) between successive replications of a simulation run. Statistics (mean and standard deviation) are collected on the demand rate and backorder levels at the retailer, the number of retailer orders, the proportion backordered at the warehouse, and the average delay at the warehouse.

### 5.2. Model verification

After verifying that the simulation model was providing statistically useful measures of system performance, we used the output of the simulation to test the accuracy of the analytical model. The cost structure used in all the simulation experiments is given in Table 1. These parameters are fairly representative of the food industry (see, for example, Ref. [35]). For a given shipment size, $x$, the transportation rates (per unit) from the warehouse to the retailer, and from each of the suppliers to the warehouse were estimated to be (for the specifics of the estimation procedures see Ref. [7])

$$g(x) = 67.33 - 5.97\ln(x).$$

For the purposes of verification, systematic tests were conducted for $N_r = 20$. The parameters of interest were varied as follows:

1. Daily Demand at Retailer($D_r$): Poisson distributed with mean 10 and Poisson distributed with mean 20
2. The warehouse–retailer transit time ($L_r$): Gamma distributed with mean 2 and variance 1 and Gamma distributed with mean 2 and variance 2
3. The supplier–warehouse transit time ($L_w$): Gamma distributed with mean 3 and variance 2 and Gamma distributed with mean 3 and variance 4
4. Retailer fill-rate ($\rho_r$): 99%, 95%
5. Warehouse fill-rate ($\rho_w$): 99%, 98%, 97%, 96%

A set of 32 unique test problems were constructed using combinations of the above parameter values (see Table 2). Each of these 32 test problems were analytically solved using Eqs. (17)–(25) to determine the optimal lot size and reorder points at each retailer and the warehouse. These (lot sizes

### Table 1

<table>
<thead>
<tr>
<th>Distribution system structure</th>
<th>Number of retailers</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of suppliers</td>
<td>3</td>
</tr>
<tr>
<td>Inventory</td>
<td>Product value</td>
<td>$5</td>
</tr>
<tr>
<td></td>
<td>Holding (%)</td>
<td>50%</td>
</tr>
<tr>
<td>Procurement</td>
<td>Setup parameter (retailer)</td>
<td>$90</td>
</tr>
<tr>
<td></td>
<td>Setup parameter (warehouse)</td>
<td>$90 + 10n, $</td>
</tr>
<tr>
<td></td>
<td>Order Processing Periods</td>
<td>2 days</td>
</tr>
<tr>
<td></td>
<td>Periods/Year</td>
<td>360 Days</td>
</tr>
<tr>
<td>Transportation</td>
<td>Transportation costs/unit $G(x)$:</td>
<td>$67.33 - 5.97\ln(x)$</td>
</tr>
</tbody>
</table>
Table 2
The verification experiment

<table>
<thead>
<tr>
<th>Experiment #</th>
<th>Retailer fill-rate (%)</th>
<th>Demand</th>
<th>Ret–ware LT</th>
<th>Ware–sup LT</th>
<th>Ware fill rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>99</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>99</td>
</tr>
<tr>
<td>2</td>
<td>99</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>98</td>
</tr>
<tr>
<td>3</td>
<td>99</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>97</td>
</tr>
<tr>
<td>4</td>
<td>99</td>
<td>H</td>
<td>H</td>
<td>L</td>
<td>99</td>
</tr>
<tr>
<td>5</td>
<td>99</td>
<td>H</td>
<td>L</td>
<td>H</td>
<td>98</td>
</tr>
<tr>
<td>6</td>
<td>99</td>
<td>H</td>
<td>L</td>
<td>H</td>
<td>97</td>
</tr>
<tr>
<td>7</td>
<td>99</td>
<td>H</td>
<td>L</td>
<td>L</td>
<td>96</td>
</tr>
<tr>
<td>8</td>
<td>99</td>
<td>L</td>
<td>L</td>
<td>H</td>
<td>99</td>
</tr>
<tr>
<td>9</td>
<td>99</td>
<td>L</td>
<td>L</td>
<td>H</td>
<td>98</td>
</tr>
<tr>
<td>10</td>
<td>99</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>97</td>
</tr>
<tr>
<td>11</td>
<td>99</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>96</td>
</tr>
<tr>
<td>12</td>
<td>99</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>99</td>
</tr>
<tr>
<td>13</td>
<td>99</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>98</td>
</tr>
<tr>
<td>14</td>
<td>99</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>97</td>
</tr>
<tr>
<td>15</td>
<td>99</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>96</td>
</tr>
<tr>
<td>16</td>
<td>95</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>99</td>
</tr>
<tr>
<td>17</td>
<td>95</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>98</td>
</tr>
<tr>
<td>18</td>
<td>95</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>97</td>
</tr>
<tr>
<td>19</td>
<td>95</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>96</td>
</tr>
<tr>
<td>20</td>
<td>95</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>99</td>
</tr>
<tr>
<td>21</td>
<td>95</td>
<td>H</td>
<td>L</td>
<td>H</td>
<td>98</td>
</tr>
<tr>
<td>22</td>
<td>95</td>
<td>H</td>
<td>L</td>
<td>H</td>
<td>97</td>
</tr>
<tr>
<td>23</td>
<td>95</td>
<td>H</td>
<td>L</td>
<td>H</td>
<td>96</td>
</tr>
<tr>
<td>24</td>
<td>95</td>
<td>H</td>
<td>L</td>
<td>H</td>
<td>99</td>
</tr>
<tr>
<td>25</td>
<td>95</td>
<td>L</td>
<td>H</td>
<td>L</td>
<td>99</td>
</tr>
<tr>
<td>26</td>
<td>95</td>
<td>L</td>
<td>H</td>
<td>L</td>
<td>98</td>
</tr>
<tr>
<td>27</td>
<td>95</td>
<td>L</td>
<td>H</td>
<td>L</td>
<td>97</td>
</tr>
<tr>
<td>28</td>
<td>95</td>
<td>L</td>
<td>H</td>
<td>L</td>
<td>96</td>
</tr>
<tr>
<td>29</td>
<td>95</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>99</td>
</tr>
<tr>
<td>30</td>
<td>95</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>98</td>
</tr>
<tr>
<td>31</td>
<td>95</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>97</td>
</tr>
<tr>
<td>32</td>
<td>95</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>96</td>
</tr>
</tbody>
</table>

Key: H = High, L = Low, Number of retailers = 20; number of suppliers = 3.
Demand (at each retailer): H = Poisson with mean 20, L = Poisson with mean 10.
Retailer–warehouse LT: H = Gamma distributed with mean 2 and variance 2, L = Gamma distributed with mean 2 and variance 1.
Warehouse–supplier LT: H = Gamma distributed with mean 3 and variance 2, L = Gamma distributed with mean 3 and variance 1.

and reorder points), along with the parameters 1–3 from above (retailer demand, transit times between warehouse and retailer, and supplier and warehouse) serve as inputs to the simulation. The key outputs of the simulation are the retailer (the backorder levels at the 20 retailers are averaged to arrive at a composite number (see Ref. [14])) and warehouse backorder levels (fill rates). These are then compared to the “true” backorder levels (fill rates) to determine the accuracy of the model. We use two measures of model accuracy. The first is the “percent error”, defined as

$$\text{Percent Error} = \frac{100 \times (\text{mean retailer/warehouse backorder level from simulation} - \text{analytical backorder level})}{\text{mean retailer/warehouse backorder level from simulation}}.$$
The second measure of model accuracy, $\Delta$, is measured as the difference between the point estimate of the mean from the simulation and the model’s prediction, expressed in terms of the standard error of the mean (estimated from the simulation). Specifically,

$$\Delta = \frac{\text{mean retailer backorder level from the simulation} - \text{theoretical backorder level}}{\text{Standard error of the retailer backorder level from the simulation}}.$$

The fit of the model at both the retailer and the warehouse levels seems very good. The percentage errors at the retailer and warehouse levels, $|\Delta W|$ and $|\Delta R|$, are illustrated in Table 3. From the table, the analytical model predicts the retailer backorders within 3.59% on average. The average percent

<table>
<thead>
<tr>
<th>Experiment #</th>
<th>Warehouse</th>
<th>Retailer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\Delta W$</td>
<td>Percent error</td>
</tr>
<tr>
<td>1</td>
<td>0.032989588</td>
<td>0.12%</td>
</tr>
<tr>
<td>2</td>
<td>0.118476588</td>
<td>0.66%</td>
</tr>
<tr>
<td>3</td>
<td>0.515685982</td>
<td>2.04%</td>
</tr>
<tr>
<td>4</td>
<td>0.436937063</td>
<td>2.68%</td>
</tr>
<tr>
<td>5</td>
<td>0.016822337</td>
<td>0.07%</td>
</tr>
<tr>
<td>6</td>
<td>0.123839187</td>
<td>0.68%</td>
</tr>
<tr>
<td>7</td>
<td>0.427199762</td>
<td>2.07%</td>
</tr>
<tr>
<td>8</td>
<td>0.467277227</td>
<td>2.78%</td>
</tr>
<tr>
<td>9</td>
<td>0.034482801</td>
<td>0.18%</td>
</tr>
<tr>
<td>10</td>
<td>0.104347897</td>
<td>0.61%</td>
</tr>
<tr>
<td>11</td>
<td>0.012746569</td>
<td>0.13%</td>
</tr>
<tr>
<td>12</td>
<td>0.163487833</td>
<td>1.54%</td>
</tr>
<tr>
<td>13</td>
<td>0.050406536</td>
<td>0.35%</td>
</tr>
<tr>
<td>14</td>
<td>0.017031482</td>
<td>0.14%</td>
</tr>
<tr>
<td>15</td>
<td>0.032992036</td>
<td>0.33%</td>
</tr>
<tr>
<td>16</td>
<td>0.067853247</td>
<td>0.79%</td>
</tr>
<tr>
<td>17</td>
<td>0.034710709</td>
<td>0.19%</td>
</tr>
<tr>
<td>18</td>
<td>0.188235127</td>
<td>0.87%</td>
</tr>
<tr>
<td>19</td>
<td>0.262623353</td>
<td>1.59%</td>
</tr>
<tr>
<td>20</td>
<td>0.513949782</td>
<td>2.73%</td>
</tr>
<tr>
<td>21</td>
<td>0.123020767</td>
<td>0.79%</td>
</tr>
<tr>
<td>22</td>
<td>0.01692991</td>
<td>0.12%</td>
</tr>
<tr>
<td>23</td>
<td>0.608030209</td>
<td>2.29%</td>
</tr>
<tr>
<td>24</td>
<td>0.310383523</td>
<td>2.16%</td>
</tr>
<tr>
<td>25</td>
<td>0.107594846</td>
<td>0.69%</td>
</tr>
<tr>
<td>26</td>
<td>0.034075866</td>
<td>0.27%</td>
</tr>
<tr>
<td>27</td>
<td>0.248311101</td>
<td>1.49%</td>
</tr>
<tr>
<td>28</td>
<td>0.237400834</td>
<td>1.83%</td>
</tr>
<tr>
<td>29</td>
<td>0.040322528</td>
<td>0.17%</td>
</tr>
<tr>
<td>30</td>
<td>0.156491571</td>
<td>0.88%</td>
</tr>
<tr>
<td>31</td>
<td>0.309021001</td>
<td>1.81%</td>
</tr>
<tr>
<td>32</td>
<td>0.455390053</td>
<td>2.76%</td>
</tr>
</tbody>
</table>

| Average      | 0.195922416 | 1.12% | 0.00945667 | 3.59% |
error at the warehouse meanwhile is 1.12%. The quantity \(|\Delta W|\) takes on values from 0.01 to 0.51, with the average being 0.53. Standard deviations of backorders (of retailer orders) at the warehouse ranged from 0.372 to 0.899, with an average of 0.665. The quantity \(|\Delta R|\), on the other hand, ranges from 0.000296 to 0.016, with the average being 0.0094. The magnitude of the standard deviation of backorders at the retailer ranges (over the 32 experiments) from 13.9 to 83, with an average of around 39 units. These numbers indicate that at a 5% level of significance, there is no appreciable difference between the simulated and the analytical values. For each of the 32 test problems, both \(\Delta W\) and \(\Delta R\) were positive, i.e., the analytical model underestimated the number of backorders at both the retailer and the warehouse levels. At the retailer level, this is not unexpected since the METRIC approximation is shown to consistently underestimate the backorder levels (see Ref. [11]). Table 3 shows two other trends. First, the analytical model seems to perform better when the warehouse fill rates are higher. The average percent errors at the warehouse are 0.32, 0.53, 1.47, and 2.16 when the warehouse fill rates are 99%, 98%, 97%, and 96%, respectively. Since the mean delay time at the warehouse is a function of the warehouse fill rate, the model inaccuracy may lie in determining the mean delay at the warehouse (recall that the random wait time is replaced by the constant delay). A second trend that can be observed by inspecting Table 3 is that the analytical model predicts the warehouse dynamics, i.e., backorders and average stock levels better than the retailer inventory dynamics. Since safety stock levels and service levels depend on the retailer lead time, the relatively higher \(|\Delta R|\) could be a result of the METRIC approximation. The analytical model seems to underestimate the “wait time” of the backordered retailer. This leads to the underestimation of the safety stocks, and hence the lower service levels.

At the warehouse level, the discrepancies could be a cause of several assumptions. The inventory policies in the analytical model were calculated so that there is a small (5%) chance of orders crossing over (\(\tau\) was set to 95% in Eq. (16) to calculate \(B\)). This value of \(B\) was then used in Eq. (24)). In the simulation, the cross-over of supplier orders seems to have an effect on the warehouse cycle inventory \((Q_w/2)\), and thus the consequent replenishment of the retailers. The analytical model also assumed that the shortages on the second and subsequent portions are negligible. This might be another source that might lead to a greater \(|\Delta W|\), and hence a greater \(|\Delta R|\).

### 6. Conclusions and directions for future research

This paper has presented a near-optimal \((s, Q)\)-type inventory-logistics cost minimizing model for a production/distribution network with multiple suppliers supplying a distribution center, which in turn distributes to a large number of identical retailers. The model was a synthesis of three components: (i) the inventory analysis at the retailers; (ii) the demand process at the warehouse; and (iii) the inventory analysis at the warehouse. The decisions in the model were made through a comprehensive distribution-based cost framework that includes the inventory, transportation, and transit components of the supply chain. The approximate model was verified using simulation. The simulation indicated that the analytical model was quite good, estimating the retailer backorders within 3.59%, and the warehouse backorder levels within 1.12% on an average (over the 32 experiments that were conducted).

The model has several important managerial implications. A non-exhaustive list of those would include:

- **Inventory and service benchmarking** – Table 3 indicates that the analytical model is reasonably accurate in obtaining service levels within 3.59% at the retailer level, and within 1.12% at the warehouse (the analytical model underestimates the backorders at both the retailer and warehouse level). Management can therefore benchmark its inventory levels with the model results (i.e., inventory policy that determines the levels of cycle, safety, and in-transit stocks) in such a system. This would give a clearer indication of the amount of safety stock that needs to be held at each stocking location.
• What-if analysis – the model is flexible to include changes in number of suppliers, retailers, their lead times, or transportation modes. Such analysis can also be used for future design.

However, the model suffers from a number of limitations. Future research can address some of these issues:

• the model is limited to two echelons and assumes that the suppliers are always in stock. Extensions of the model need to be undertaken beyond two echelons
• the model assumes identical suppliers and retailers for modeling simplicity. Better procedures need to be introduced to tackle the “non-identical” case.
• Although the simulation presented in Section 5 indicates that the analytical model is good at predicting service levels (or inventory levels) at the warehouse and retailer level, it also recognizes the need to model the random delay at the warehouse in a better way. In the current model, the average time is underestimated, thus leading to higher retailer backorder levels.

References


[27] H. Lao, L.G. Zhao, Optimal ordering policies with two suppliers when lead times and demands are all stochastic, European Journal of Operations Research 68 (1993) 120–133.


