EASILY IMPLEMENTABLE INVENTORY CONTROL POLICIES

WALLACE J. HOPP
Northwestern University, Evanston, Illinois

MARK L. SPEARMAN
Georgia Institute of Technology, Atlanta, Georgia

RACHEL Q. ZHANG
University of Michigan, Ann Arbor, Michigan
(Received August 1994; revision received October 1995; accepted April 1996)

This work was initiated and supported by a manufacturer of mail processing equipment, which stocks 30,000 distinct parts in a distribution center to support field maintenance of their equipment. To find an effective stocking policy for this system we formulate a constrained optimization model with the objective of minimizing overall inventory investment at the distribution center subject to constraints on customer service and order frequency. Because size, integrality, and nonconvexity make this problem intractable to exact analysis, we develop three heuristic algorithms based on simplified representations of the inventory and service expressions. These lead to what we call easily implementable inventory policies, in which the control parameters for a newly introduced part can be computed in closed form without reoptimizing the rest of the system. Numerical comparisons against a lower bound on the cost function show that even our simplest heuristic works well when a high service level is required. However, we show that a more sophisticated heuristic is more robustly accurate. We also compare our heuristics to methods previously in use by the firm whose system motivated this research and show that they are more efficient in the sense of attaining the same customer service level with a 20–25% smaller inventory investment. Finally, we discuss implementation issues related to the specific needs of the client firm, such as how to handle parts with no or low recent usage and dynamically changing demand for parts.

1. MOTIVATION AND BACKGROUND

This work was initiated and supported by a manufacturer of mail processing equipment. The firm offers service contracts on its equipment and stocks spare parts to support this maintenance function. A large fraction of the stock of spare parts is held in a distribution center (DC). The remainder is distributed among regional facilities and on-site at customer locations. Because service contracts guarantee that outages due to failures will not exceed a certain number of hours per month, it is essential that the likelihood of a part being out of stock at the DC when required be kept low. For this reason, fill rate (defined by our client as the fraction of orders to the DC that are filled within 24 hours) is a key measure used by the firm to evaluate DC performance. However, because inventory is expensive and can become obsolete as equipment models change, the firm does not want to hold excessive amounts of stock. Thus, understanding the inventory versus customer service tradeoff and striking a reasonable balance were our client’s primary concerns.

This inventory system presented a number of modeling and analysis challenges:

1. Large Number of Parts: There are almost 30,000 distinct part numbers in the system with a great deal of variability in cost, replenishment lead time, and demand rate. The ability of a technician to complete a repair is a function of the availability of the parts needed for that repair job. In theory, any part could be critical to a customer, although in practice technicians can sometimes work around the lack of a part (e.g., by getting a substitute from the hardware store, adapting another part, or temporarily repairing the old part). Because customer service depends on part availability in a complex way, we approached the problem as one of minimizing total inventory investment subject to constraints on average fill rate and order frequency. (However, we used the conventional definition of fill rate, the fraction of orders filled from stock, instead of the 24-hour fill rate used by our client, because for purposes of inventory control we are interested in stockouts due to...
lack of inventory, not stockouts due to delivery delays on the part of the DC. Therefore, we expected, and found, the fill rates predicted by our model to be higher than those measured in practice.) Although this turned out to be a reasonable first cut, there are problems with using average fill rate at the DC, as we will discuss below. Furthermore, while an occasional demand is satisfied from outside the system (e.g., from a hardware store), most customer demands are backordered (i.e., a temporary repair is eventually done properly and requires the part). For modeling purposes, we assumed full backordering.

2. Random Demand: Because these are repair parts, demand is largely stochastic, although some parts are used in scheduled maintenance functions. Because (1) even scheduled usage is subject to some randomness due to technicians’ schedules, (2) demand seen by the DC is the superposition of independent demand processes from many facilities and customer sites, and (3) we had no information whatever concerning variance of demand, we chose to model the demand processes as Poisson. However, this assumption is not critical to our approach, which actually winds up approximating the Poisson by the normal distribution for computational reasons anyway.

3. Batch Demand: Frequently, parts are used in batches or “kits” (e.g., machines use 12 sucker cups, which are always replaced simultaneously). Although randomness in the batch size is possible for some parts, we assumed that batches are constant. We did this because (1) data for determining second moments of batch sizes was poor, and (2) underestimating the variance of batch size tends to counteract our overestimate of variance of the demand process by not considering scheduled demand.

4. Dynamic Demand: Demand for repair parts varies with the age of specific models of machines (e.g., some mechanical parts are rarely needed until their machine has been out in the field for several years). For modeling purposes, we made the assumption that a steady-state model using predicted demands for the next year is reasonable. Our rationale for this was that since parts are ordered on average twice per year or so, the effects of this year’s stocking policy will not persist inordinately far into the future. If the policy is revised on the basis of new demand predictions, the system will adapt rapidly. Of course, this is only approximately true and we will mention this issue again in our discussion of future research topics.

5. Multiple Supply Sources: Repair parts are obtained from a variety of sources, including the firm’s own production facilities and outside suppliers. In some cases (e.g., for older machines), parts must be custom made in job shops, which implies that they are expensive and can have long lead times. The DC had estimates of procurement lead times in their data base, but admitted that they could be in error. For modeling purposes, we assumed that these lead times were fixed, but we reviewed and revised the data to adjust what we and our client agreed were overly optimistic lead times.

6. Implementability: Our client’s inventory control system at the time of our study computed reorder points and order quantities solely on the basis of average annual demand. Although probably ineffective, this approach had the benefit of being very simple to implement in a mainframe-based information system. A large-scale optimization procedure was simply not feasible without major changes to our client’s system. Therefore, we took it as a constraint that anything we proposed be easily implementable in the sense of leading to separable closed-form expressions for computing stocking parameters. Of course, as we will see, the parts must be considered together for the purposes of ensuring satisfaction of constraints (Lagrange multipliers). However, as long as this computation can be done outside the system and updated infrequently, we consider it to be within the spirit of easily implementable.

Except for these issues, the inventory problem faced by our client represented a fairly standard continuous review system. We selected the well-known \((Q, r)\) model as appropriate for depicting the behavior of individual stocks. This model dates back at least to Wilson (1934) and has been studied extensively since that time (see Lee and Nahmias 1989 for a survey). The following are a few of the results most directly relevant to our problem.

Hadley and Whitin (1963) formalized the \((Q, r)\) model in a cost framework, where total cost is made up of ordering cost, inventory holding cost, and backorder cost. They derived the exact cost function under the assumption of Poisson demand. However, because the exact cost function is complex and is not even convex without additional restrictions on the demand distribution and the range of \(r\) (Zipkin 1986), simpler approximations have been suggested and a large number of heuristics have been proposed over the years (see Lee and Nahmias 1989).

Recently, some breakthroughs have revised our understanding of the \((Q, r)\) model. First, Federgruen and Zheng (1992) presented a surprisingly simple and efficient algorithm for computing an optimal \((Q, r)\) policy that is linear in \(Q^*\). Second, Zheng (1992) gave some important insights into the effectiveness of heuristics by proving that using the EoQ formula instead of the optimal order quantity, \(Q^*\), and then finding the best reorder point, \(r\), causes an increase in the ordering costs of no more than \(\frac{1}{3}\). He also showed by means of examples that when the inventory cost savings are taken into account, the relative increase in total cost is generally much less than the worst-case bounds.

The literature on implementing \((Q, r)\)-type inventory systems has frequently reported that shortage cost can be difficult to specify. One approach for dealing with this is to use a constrained model that minimizes holding plus ordering cost subject to a certain level of service. Under these conditions, it is a common practice to compute order quantity and reorder point separately. The approximate independence of the optimal order quantity from the service level requirement has been empirically supported (see Brown 1967 and Peterson and Silver 1979), as has the
2.1. Single-Product Case

For purposes of modeling, we assume demand is Poisson with mean rate $\lambda$ units per year, and procurement lead time is a constant $l$. We let $p(x)$ and $P(x)$ represent the probability mass function and cumulative distribution function of demand during replenishment lead time. We let $\theta$ denote the expected leadtime demand, so that $\theta = \lambda l$. Because we are considering repair parts, demands that cannot be filled out of stock must be backordered. Hence, a $(Q, r)$ policy, which continuously monitors inventory and orders $Q$ units each time inventory position reaches $r$, results in profile of inventory level like that shown in Figure 1. Finally, we represent the unit cost of the item by $c$, where $c$ is a constant and is independent of order quantity.

We formulate the single-product problem as the following constrained optimization model:

\begin{align*}
\text{Minimize} & \quad \text{inventory investment} \quad (2.1) \\
\text{Subject to:} & \quad \text{order frequency} \leq F, \quad (2.2) \\
& \quad \text{service level} \geq S, \quad (2.3)
\end{align*}

where $F$ and $S$ are target order frequency and service level specified by the user. We chose a constrained model because (1) specifying service and order frequency targets was more natural to our client than specifying costs of backorders and purchase orders, and (2) we did not have part-specific information with which to estimate purchase order costs and therefore would have had to use a uniform cost across all parts, which would be equivalent to a constraint on average order frequency. However, it is generally not obvious how to choose appropriate values for $S$ and $F$ a priori, so it is important to perform sensitivity analysis on $S$ and $F$ to examine their effect on the cost function.

We can make the objective function precise by using the result from Hadley and Whitin (1963) that average on-hand inventory (i.e., inventory position minus on-order inventory plus back-orders) at any point in time is given by

\begin{equation}
\frac{Q}{2} + \frac{1}{2} + r - \theta + B(r, Q),
\end{equation}

where $B(r, Q)$ is the expected number of back-orders at any point in time, or, equivalently, the average unit years of shortage incurred per year, which can be computed as

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{inventory_level.png}
\caption{Inventory level under a $(Q, r)$ policy.}
\end{figure}
\[ B(r, Q) = \frac{1}{Q} \left[ \beta(r) - \beta(r + Q) \right], \quad (2.5) \]

where
\[ \beta(v) = \sum_{u=v+1}^{\infty} \left( u - v - 1 \right) \left[ 1 - P(u - 1) \right] \]
\[ = \frac{\theta^2}{2} \left[ 1 - P(v - 2) \right] - \theta v \left[ 1 - P(v - 1) \right] + \frac{v(v + 1)}{2} \left[ 1 - P(v) \right]. \quad (2.6) \]

Note that \( \beta(v) \) represents the time-weighted back-orders arising from lead time demand in excess of \( v \).

An order of quantity \( Q \) implies that the average number of orders per year is \( \lambda / Q \), so the order frequency constraint is straightforward to express as \( \lambda / Q \leq F \).

We can express the service constraint mathematically by noting that average service is \( 1 - A(r, Q) \) where \( A(r, Q) \) is the probability of stockout at any point in time. \( A(r, Q) \) can be computed as (see Hadley and Whitin 1963)
\[ A(r, Q) = \frac{1}{Q} \left[ \alpha(r) - \alpha(r + Q) \right], \quad (2.7) \]

where
\[ \alpha(v) = \sum_{u=v+1}^{\infty} \left[ 1 - P(u - 1) \right] \]
\[ = \theta \left[ 1 - P(v - 1) \right] - v \left[ 1 - P(v) \right]. \quad (2.8) \]

Note that \( \alpha(v) \) can be viewed as the expected lead time demand in excess of \( v \).

We can now express the exact formulation of the single-product problem as follows:

Minimize \( c \left( r - \theta + \frac{Q}{2} + \frac{1}{2} + B(r, Q) \right) \) \quad (2.9)

Subject to:
\[ \frac{\lambda}{Q} \leq F, \quad (2.10) \]
\[ 1 - A(r, Q) \geq S, \quad (2.11) \]
\[ r \geq r_i, \quad Q \geq 1, \quad (2.12) \]
\[ r, Q : \text{integer}. \quad (2.13) \]

We have added constraints (2.12) to prevent unreasonable values of \( Q \) and \( r \). Notice that, depending on the situation, the minimum reorder point, \( r_i \), could be negative. For instance, to rule out the possibility of a demand waiting for longer than the replenishment lead time, we would set \( r = -1 \). We will discuss other possible restrictions when we discuss implementation later.

### 2.2. Multiproduct Case

We now move to the more interesting and realistic multiproduct case. The key modeling extension is how to represent the interaction between products. In our constrained optimization framework, the natural way to do this is to specify constraints on average order frequency and service level as follows:

Minimize \( \text{inventory investment} \) \quad (2.14)

Subject to:
\[ \text{average order frequency} \leq F, \quad (2.15) \]
\[ \text{average service level} \geq S. \quad (2.16) \]

Notice that a given average level of service can be attained by choosing the service of some parts to be high and other parts to be low. As we will discuss later in this paper, large discrepancies in the service level for individual parts may not always be appropriate and therefore additional constraints may be needed. However, for purposes of developing the model, we will adhere to this straightforward approach.

To make formulation (2.14)–(2.16) precise, we require the following notation:

\[ N = \text{number of items}, \]
\[ c_i = \text{unit cost for item } i, \]
\[ C = \sum_{i=1}^{N} c_i, \]
\[ \lambda_i = \text{expected demand for item } i \text{ per year}, \]
\[ \Lambda = \sum_{i=1}^{N} \lambda_i, \]
\[ l_i = \text{replenishment lead time for item } i \text{ (assumed constant)}, \]
\[ \theta_i = \lambda_i l_i, \text{ expected demand for item } i \text{ during lead time } l_i, \]
\[ Q_i = \text{order quantity for item } i, \]
\[ r_i = \text{reorder point for item } i, \]
\[ A_i(r_i, Q_i) = \text{probability of stockout for item } i \]
\[ = \frac{1}{Q_i} \left[ \alpha_i(r_i) - \alpha_i(r_i + Q_i) \right], \]
\[ B_i(r_i, Q_i) = \text{expected number of back-orders for item } i \text{ at any time} \]
\[ = \frac{1}{Q_i} \left[ \beta_i(r_i) - \beta_i(r_i + Q_i) \right], \]

where \( \alpha_i(v) \) and \( \beta_i(v) \) can be computed using (2.8) and (2.6) for all \( i \).

Using the exact expressions for inventory and service level in formulation (2.14)–(2.16) yields:

Minimize \[ \sum_{i=1}^{N} c_i \left( r_i - \theta_i + \frac{Q_i}{2} + \frac{1}{2} + B_i(r_i, Q_i) \right) \] \quad (2.17)

Subject to:
\[ \frac{\Lambda}{N \sum_{i=1}^{N} Q_i} \leq F, \quad (2.18) \]
\[ \sum_{i=1}^{N} \lambda_i \left( 1 - A_i(r_i, Q_i) \right) \geq S, \quad (2.19) \]
\[ r_i \geq r_i, \quad Q_i \geq 1, \quad (2.20) \]
\[ r_i, Q_i : \text{integers}. \quad (2.21) \]

In practice, however, it is useful to replace objective (2.17) by
\[ \frac{1}{C} \sum_{i=1}^{N} c_i \left( r_i - \theta_i + \frac{Q_i}{2} + \frac{1}{2} + B_i(r_i, Q_i) \right). \quad (2.22) \]
That is, to use average inventory investment instead of total inventory investment as the criterion. Clearly the two are equivalent since we are simply dividing by a constant. But, since average investment is impacted less by addition and subtraction of parts to the system, the resulting Lagrange multipliers will be more stable to such changes. As we will discuss below, this enables us to add and subtract parts without having to reoptimize the stocking policy parameters. (Of course, when enough changes are made to substantially affect average inventory investment, reoptimization will become necessary.) Because this stability enhances the implementability of the overall system, all further references to formulation (2.17)–(2.21) are assumed to use objective (2.22).

3. A LOWER BOUND

There is no practical way to solve (2.17)–(2.21) exactly. Since the decision variables are integer, branch and bound search could be used in theory. However, for an inventory system with 30,000 parts (60,000 variables), this is not feasible, let alone easily implementable. Lagrangian relaxation might appear attractive at first glance because there are only two constraints, (2.18) and (2.19), to be relaxed. However, we found the following problems with this approach: (1) the K-K-T conditions are not necessary conditions for optimality due to integrality, (2) integer variables can cause the constraints to be slack at the optimum, (3) a duality gap can occur, (4) Lagrangian relaxation is slow for large problems, and (5) nonconvexity of the relaxation makes it difficult to rule out the existence of local minima which differ from the global minimum.

Given these difficulties, Lagrangian relaxation makes sense as a means for finding a lower bound but not for finding an optimal solution. We develop such a lower bound in the appendix and will make use of it to evaluate the effectiveness of the (easily implementable) heuristics we describe below.

4. HEURISTICS

The basic idea behind all of our heuristics for addressing formulation (2.17)–(2.21) is to approximate the expressions for inventory and service with simpler formulas so that, given suitable Lagrange multipliers, the $Q_i$ and $r_i$ values can be computed in simple form. We present three approaches along these lines, the difference being the manner in which the service expression is approximated.

In all three heuristics we approximate inventory by assuming that $r - \theta \geq 0$, so that on average, on-hand inventory is positive when orders arrive (see Figure 2). Therefore we can express average inventory as

$$h(r, Q) = r - \theta + \frac{Q}{2}.$$  \hspace{1cm} (4.23)

Of course, if $r - \theta < 0$, this expression is not even valid in expectation. However, we have found it useful to ignore this case because (1) it gives a simple inventory formula that allows direct solution of the Lagrangian, and (2) parts for which $r - \theta < 0$ tend to be those with low demand and therefore modeling them inaccurately does not greatly affect the overall quality of the solution.

We can approximate “exact” service (i.e., the long run probability of stockout with two types of service, labeled Type I and Type II (Nahmias 1993). Type I service is the fraction of cycles (intervals between orders) in which we are able to fill all orders or, equivalently, the probability that there is no stockout during the replenishment leadtime. If demand is modeled as discrete, then service is simply given by $P(r)$ under this definition. If demand is modeled as a continuous random variable, where $G(x)$ and $g(x)$ denote the cdf and pdf of demand during replenishment leadtime, then average service is given by $G(r)$.

Type II service is the proportion of demands met from stock or $1 - \alpha(r)/Q$, where $\alpha(r)$ is the average number of stockouts during the replenishment leadtime. In the discrete case, $\alpha(r)$ is given by expression (2.8), which can be written as:

$$\alpha(r) = \sum_{u=r+1}^{\infty} \{1 - P(u - 1)\} = \sum_{u=r}^{\infty} (u - r) p(u).$$ \hspace{1cm} (4.24)

In the continuous case,

$$\alpha(r) = \int_{r}^{\infty} [1 - G(t)] \, dt = \int_{r}^{\infty} (t - r) g(t) \, dt.$$ \hspace{1cm} (4.25)

4.1. A Type I Heuristic

We can derive simple formulas for $Q_i$ and $r_i$ by using the following approximations:

1. Inventory for product $i$ is given by (4.23), so $h_i(r_i, Q_i) = r_i - \theta_i + Q_i/2$.

2. Service for product $i$ is given by the Type I formula, $G_i(r_i)$ (the cdf of lead time demand for part $i$).

3. Demand for product $i$ during the replenishment lead time is approximated by the normal distribution with mean $\theta_i$ and standard deviation $\sqrt{\beta_i}$ (i.e., to match two moments with the Poisson distribution).

Under these assumptions, multiproduct formulation (2.17)–(2.21) becomes:

Minimize

$$\frac{1}{C} \sum_{i=1}^{N} c_i \left( r_i - \theta_i + \frac{Q_i}{2} \right)$$ \hspace{1cm} (4.26)

Figure 2. Average inventory when $r - \theta = 0$. 


Subject to: \[ \frac{1}{N} \sum_{i=1}^{N} \frac{\lambda_i}{Q_i} \leq F, \] \[ \sum_{i=1}^{N} G_i (r_i) \geq S, \] \[ r_i \geq r, \quad Q_i \geq 1, \quad i = 1, 2, \ldots, N, \] \[ r_i, Q_i : \text{integer.} \] (4.27) (4.28) (4.29) (4.30)

The Lagrangian for this problem is
\[ L = \frac{1}{C} \sum_{i=1}^{N} c_i \left( r_i - \theta_i + \frac{Q_i}{2} \right) + \lambda \left( \frac{1}{N} \sum_{i=1}^{N} \frac{\lambda_i}{Q_i} - F \right) \] \[ - \mu \left( \sum_{i=1}^{N} \frac{\lambda_i}{Q_i} - S \right). \] (4.31)

Differentiating \( L \) with respect to \( Q_i \) and solving for \( Q_i \) yields
\[ \frac{\partial L}{\partial Q_i} = \frac{c_i - \frac{\nu \lambda_i}{NQ_i^2}}{2C} = 0, \quad i = 1, 2, \ldots, N, \] \[ Q_i = \frac{2\nu \lambda_i C}{c_i N}, \quad i = 1, 2, \ldots, N. \] (4.32) (4.33)

Differentiating \( L \) with respect to \( r_i \) and solving for \( r_i \) yields
\[ \frac{\partial L}{\partial r_i} = \frac{c_i - \mu \lambda_i}{C} \phi_i \left( \frac{r_i - \theta_i}{\sqrt{\theta_i}} \right) \frac{1}{\sqrt{\theta_i}} = 0, \] \[ r_i = \theta_i + \sqrt{-2 \theta_i \ln \left( \frac{\nu \lambda_i}{\sqrt{2 \pi} \theta_i} \right)}, \quad i = 1, 2, \ldots, N. \] (4.34) (4.35)

Since we want to restrict \( Q_i \geq 1 \) and \( r_i \geq r \), we modify these formulas to:
\[ Q_i = \max \left\{ \frac{2\nu \lambda_i C}{c_i N}, 1 \right\}, \] \[ r_i = \begin{cases} \theta_i + \sqrt{-2 \theta_i \ln \left( \frac{\nu \lambda_i C}{\sqrt{2 \pi} \theta_i} \right)}, & \text{if } \sqrt{-2 \theta_i \ln \left( \frac{\nu \lambda_i}{\sqrt{2 \pi} \theta_i} \right)} \leq 1 \\ r, & \text{otherwise}. \end{cases} \] (4.36) (4.37)

Furthermore, we round \( Q_i \) and \( r_i \) to integers.

Notice that in the Type I heuristic, order quantities and reorder points are determined separately. This allows us to use the following procedure for finding \( v \) and \( \mu \): First, we adjust \( v \) and compute \( Q_i \) using (4.36) to achieve the highest order frequency that satisfies the order frequency constraint. Then we adjust \( \mu \) and compute \( r_i \) using (4.37) to achieve the lowest service level that satisfies the service constraint. Finally, in order to ensure a feasible solution to the original problem, we use the exact service expression (not the simplified version) when searching for appropriate Lagrange multipliers.

4.2. A Type II Heuristic

We can derive alternate heuristic formulas by using the same approximations as in the previous section but with Type II service instead of Type I service. Under these assumptions, the multiproduct model becomes:

Minimize \[ \frac{1}{C} \sum_{i=1}^{N} c_i \left( r_i - \theta_i + \frac{Q_i}{2} \right) \] (4.38)

Subject to: \[ \frac{1}{N} \sum_{i=1}^{N} \frac{\lambda_i}{Q_i} \leq F, \] \[ \sum_{i=1}^{N} \frac{\lambda_i}{Q_i} \left( 1 - \frac{1}{Q_i} \int_0^{\theta_i} \Phi(t - \theta_i) \, dt \right) \geq S, \] \[ r_i \geq r, \quad Q_i \geq 1, \quad i = 1, 2, \ldots, N, \] \[ r_i, Q_i : \text{integer.} \] (4.39) (4.40) (4.41) (4.42)

The Lagrangian for this problem is
\[ L = \frac{1}{C} \sum_{i=1}^{N} c_i \left( r_i - \theta_i + \frac{Q_i}{2} \right) + \lambda \left( \frac{1}{N} \sum_{i=1}^{N} \frac{\lambda_i}{Q_i} - F \right) \] \[ - \mu \left( \sum_{i=1}^{N} \frac{\lambda_i}{Q_i} - S \right). \] (4.43)

Differentiating \( L \) with respect to \( Q_i \) and solving for \( Q_i \) yields
\[ \frac{\partial L}{\partial Q_i} = \frac{c_i - \frac{\nu \lambda_i}{NQ_i^2}}{2C} = 0, \quad i = 1, 2, \ldots, N, \] \[ Q_i = \frac{2\nu \lambda_i C}{c_i N}, \quad i = 1, 2, \ldots, N. \] (4.44)

Differentiating \( L \) with respect to \( r_i \) and solving for \( r_i \) yields
\[ \frac{\partial L}{\partial r_i} = \frac{c_i - \mu \lambda_i}{C} \phi_i \left( \frac{r_i - \theta_i}{\sqrt{\theta_i}} \right) \frac{1}{\sqrt{\theta_i}} = 0, \] \[ r_i = \theta_i + \sqrt{-2 \theta_i \ln \left( \frac{\nu \lambda_i}{\sqrt{2 \pi} \theta_i} \right)}, \quad i = 1, 2, \ldots, N. \] (4.45) (4.46)

Again noting that we must have \( Q_i \geq 1 \) and \( r_i \geq r \), we can modify these expressions as follows:
\[ Q_i = \max \left\{ \frac{2\nu \lambda_i C}{c_i N}, 1 \right\}, \] (4.48)

\[ r_i = \begin{cases} \theta_i + \Phi^{-1} \left( \frac{1 - \frac{\Lambda Q_i c_i}{C \mu \lambda_i}}{\sqrt{\theta_i}} \right), & \text{if } \Lambda Q_i c_i \leq C \mu \lambda_i, \\ r, & \text{otherwise}. \end{cases} \] (4.49)

Hadley and Whitin (1963) suggested that it is usually true that the term \( \alpha_i(r_i + Q_i) \) in the expression for the exact service level
\[ 1 - A_i(r_i, Q_i) = 1 - \frac{1}{Q_i} \left[ \alpha_i(r_i) - \alpha_i(r_i + Q_i) \right]. \] (4.50)
and the term $\beta_i (r_i + Q_i)$ in the expression of the number of back-orders

$$B(r_i, Q_i) = \frac{1}{Q_i} \left[ \beta(r_i) - \beta(r_i + Q_i) \right]$$  \hspace{1cm} (4.51)

are negligible in problems of practical interest (although Zipkin 1986 argues that this is not always the case). These terms are important only when there is a significant probability that the lead time demand will be greater than $r_i + Q_i$. In other words, Type II service will be a very good approximation except when service is low. Since parts with low service are likely to be those with low demand and hence have little weight in the average service expression, one would expect the Type II model to be quite accurate.

However, using the Type II heuristic is more difficult than using the Type I heuristic because the equations for $Q_i$ and $r_i$ are coupled. Hence, we must solve for $Q_i$ and $r_i$ values simultaneously, instead of using closed-form expressions amenable to simple systems (e.g., spreadsheets). This fails to meet the easily implementable criterion and is enough of a drawback to prevent the Type II heuristic from being implemented in our client’s environment.

### 4.3. A Hybrid Heuristic

Although the Type II heuristic is not directly useful for our purposes, its development gives us a way to refine the Type I heuristic. The key advantage of the Type II approach is the enhanced accuracy of Type II over Type I service. We can retain this advantage, at least partially, while keeping the simplicity (i.e., separability of $Q_i$ and $r_i$) of the Type I heuristic by combining the two heuristics. We do this by first computing the order quantities from the Type I model

$$Q_i = \max \left\{ \frac{2r_i \lambda_i C}{c_i N}, 1 \right\}, \hspace{1cm} (4.52)$$

and then using the formula from the Type II model

$$r_i = \begin{cases} \sqrt{\frac{\lambda_i c_i}{C \mu \lambda_i}}, & \text{if } \lambda_i c_i \leq C \mu \lambda_i, \\ \frac{\lambda_i c_i}{C \mu \lambda_i}, & \text{otherwise}, \end{cases} \hspace{1cm} (4.53)$$

to compute reorder points. We call this the Hybrid heuristic since it combines formulas derived from both the Type I and Type II models. Although it is more complex than the Type I heuristic, it is still closed-form if we use the following polynomial approximation for the inverse normal distribution (Abramowitz and Stegun 1964):

$$\Phi^{-1}(p) = t - \frac{a_0 + a_1 t + a_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3} + |\epsilon(p)|, \hspace{1cm} (4.54)$$

such that $\epsilon(p) < 0.00045$ as long as $p > 0.5$, where $t = \sqrt{-2 \ln p}$, and

- $a_0 = 2.515517, \quad d_1 = 1.432788,$
- $a_1 = 0.802853, \quad d_2 = 0.189269,$
- $a_2 = 0.010328, \quad d_3 = 0.001308.$

### Table I

Data for Six Realistic Cases

<table>
<thead>
<tr>
<th>Case</th>
<th>$N$</th>
<th>Target $S$</th>
<th>Target $F$</th>
<th>Current Avg. Inv.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>209</td>
<td>0.9549</td>
<td>1.4911</td>
<td>1.78</td>
</tr>
<tr>
<td>Case 2</td>
<td>100</td>
<td>0.9698</td>
<td>1.5096</td>
<td>1.53</td>
</tr>
<tr>
<td>Case 3</td>
<td>50</td>
<td>0.9648</td>
<td>1.4669</td>
<td>1.23</td>
</tr>
<tr>
<td>Case 4</td>
<td>20</td>
<td>0.9440</td>
<td>1.2126</td>
<td>1.86</td>
</tr>
<tr>
<td>Case 5</td>
<td>10</td>
<td>0.9586</td>
<td>1.7752</td>
<td>4.75</td>
</tr>
<tr>
<td>Case 6</td>
<td>5</td>
<td>0.9551</td>
<td>1.6333</td>
<td>6.37</td>
</tr>
</tbody>
</table>

If $p$ is less than 0.5, we solve for $1 - p$ and reverse the sign.

### 5. NUMERICAL TESTS

To evaluate the above heuristics, we tested them on six data sets composed of randomly chosen subsets of our client’s actual data and seven test examples deliberately chosen to have different characteristics. Table I summarizes the number of parts for each of the six realistic cases, along with the service level, order frequency, and average inventory investment attained by the stocking rules currently in use by our client. Tables II and III give unit cost, $c_i$, replenishment lead time, $l_i$, and demand rate $\lambda_i$ for seven ten-part test examples. These examples were selected to cover a range of possible scenarios with Examples 1, 2, 3 varying $c_i$, $l_i$, $\lambda_i$, respectively, Example 4 matching high cost with long lead times, Example 5 matching long lead times with low demand, Example 6 matching high cost with low demand, and Example 7 matching high cost with long lead times and low demand.

In all tests we assume that the minimum reorder point is $r_i = -1$. We use $S$ and $F$ to represent target service level and order frequency, while $s$, $f$ and $Avg$ Inv represent the service level, order frequency and average inventory investment actually achieved. Since the target service level and order frequency are not achieved exactly due to integrality, we use the Lagrangian relaxation approach described in the appendix to compute lower bounds (LB) on the $Avg$ Inv based on both target service/order frequency (called unadjusted lower bounds) and the service/order frequency actually achieved by the heuristic (called adjusted lower bounds). We report the percentage of the $Avg$ Inv above these bounds as $Unadj$ and $Adj$, respectively.

### Table II

Data for Test Examples 1–3

<table>
<thead>
<tr>
<th>Item</th>
<th>$c_i$</th>
<th>$l_i$</th>
<th>$\lambda_i$</th>
<th>$c_i$</th>
<th>$l_i$</th>
<th>$\lambda_i$</th>
<th>$c_i$</th>
<th>$l_i$</th>
<th>$\lambda_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item</td>
<td>Ex. 1</td>
<td>Ex. 2</td>
<td>Ex. 3</td>
<td>Ex. 2</td>
<td>Ex. 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>250</td>
<td>50</td>
<td>10</td>
<td>20</td>
<td>360</td>
<td>10</td>
<td>20</td>
<td>50</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>50</td>
<td>10</td>
<td>20</td>
<td>300</td>
<td>10</td>
<td>20</td>
<td>50</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>50</td>
<td>10</td>
<td>20</td>
<td>40</td>
<td>10</td>
<td>20</td>
<td>50</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>50</td>
<td>10</td>
<td>20</td>
<td>35</td>
<td>10</td>
<td>20</td>
<td>50</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>50</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>10</td>
<td>20</td>
<td>50</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>50</td>
<td>10</td>
<td>20</td>
<td>20</td>
<td>10</td>
<td>20</td>
<td>50</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>50</td>
<td>10</td>
<td>20</td>
<td>15</td>
<td>10</td>
<td>20</td>
<td>50</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>50</td>
<td>10</td>
<td>20</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>50</td>
<td>10</td>
<td>20</td>
<td>5</td>
<td>10</td>
<td>20</td>
<td>50</td>
<td>200</td>
</tr>
</tbody>
</table>
Since even an optimal algorithm is unlikely to be able to meet the service and order frequency constraints exactly, the error relative to adjusted lower bounds is probably a fairer measure of the accuracy of the heuristics.

### 5.1. Type I Heuristic

The results for the six realistic cases are shown in Table IV. Clearly, the Type I heuristic can perform poorly relative to the lower bound (although we did find it to consistently outperform the policy currently in use by our client). Part of the reason is that the formula for reorder points only allows them to be either greater than or equal to \( \theta_0 \) or equal to their minimum value, \( r_j \) (i.e., no reorder points in the interval \( (r_j, \theta_0) \) are allowed). It is not uncommon to find feasible solutions with some reorder points between \( r_j \) and \( \theta_0 \) that achieve lower inventory investment. The Type I heuristic is incapable of finding such solutions. However, overall the Type I heuristic is often a very good approximation for situations where the target service level is very high, e.g., greater than or equal to 98%.

Table V shows the results of the Type I heuristic on the seven test examples with target service level 0.98 and target order frequencies of one, two, three, and four orders per year. The heuristic performs well in all trials except one, test Example 7 with target order frequency 1. In this example, the approximation is 15.37% above the adjusted lower bound. We will examine this more closely below.

### 5.2. Hybrid Heuristic

Table IV gives the solutions obtained using the Hybrid heuristic for the six realistic cases and shows that this heuristic works better than the Type I heuristic in all cases. We also tested all combinations of target service level 0.85, 0.9, and 0.95 and target order frequency 1, 2, 3, and 4 for the seven test examples, as summarized in Table VI. The Hybrid heuristic is within 10% of the adjusted lower bound on the optimal objective values in 77 out of 84 (i.e., in 92%) of the trials.

Table VII, which shows the errors relative to the adjusted lower bound for test Example 1 with \( F = 2 \) at various levels of \( S \), illustrates that the Hybrid heuristic substantially outperforms the simpler Type I heuristic. Note that although the Hybrid heuristic achieves uniformly smaller errors, the differences become small when service level is

---

### Table III

Data for Test Examples 4–7

<table>
<thead>
<tr>
<th>Item</th>
<th>( c_j )</th>
<th>( l_j )</th>
<th>( \lambda_j )</th>
<th>( c_j )</th>
<th>( l_j )</th>
<th>( \lambda_j )</th>
<th>( c_j )</th>
<th>( l_j )</th>
<th>( \lambda_j )</th>
<th>( c_j )</th>
<th>( l_j )</th>
<th>( \lambda_j )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>250</td>
<td>360</td>
<td>20</td>
<td>360</td>
<td>2</td>
<td>250</td>
<td>50</td>
<td>2</td>
<td>250</td>
<td>360</td>
<td>2</td>
<td>250</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>300</td>
<td>20</td>
<td>300</td>
<td>3</td>
<td>200</td>
<td>50</td>
<td>3</td>
<td>200</td>
<td>300</td>
<td>3</td>
<td>200</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>40</td>
<td>10</td>
<td>40</td>
<td>4</td>
<td>9</td>
<td>50</td>
<td>4</td>
<td>9</td>
<td>40</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>35</td>
<td>10</td>
<td>35</td>
<td>5</td>
<td>8</td>
<td>50</td>
<td>5</td>
<td>8</td>
<td>35</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>30</td>
<td>10</td>
<td>30</td>
<td>6</td>
<td>7</td>
<td>50</td>
<td>6</td>
<td>7</td>
<td>30</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>25</td>
<td>10</td>
<td>25</td>
<td>7</td>
<td>6</td>
<td>50</td>
<td>7</td>
<td>6</td>
<td>25</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>20</td>
<td>10</td>
<td>20</td>
<td>8</td>
<td>5</td>
<td>50</td>
<td>8</td>
<td>5</td>
<td>20</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>15</td>
<td>10</td>
<td>15</td>
<td>9</td>
<td>4</td>
<td>50</td>
<td>9</td>
<td>4</td>
<td>15</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>100</td>
<td>3</td>
<td>50</td>
<td>100</td>
<td>3</td>
<td>10</td>
<td>100</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>200</td>
<td>2</td>
<td>50</td>
<td>200</td>
<td>2</td>
<td>10</td>
<td>200</td>
<td>2</td>
</tr>
</tbody>
</table>

---

### Table IV

Type I and Hybrid Solutions for the Realistic Cases

<table>
<thead>
<tr>
<th>Avg. Inv.</th>
<th>Type I Heuristic</th>
<th>Hybrid Heuristic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LB (Unadj.)</td>
<td>LB (Adj.)</td>
</tr>
<tr>
<td>1</td>
<td>0.6571</td>
<td>0.5239 (25.42%)</td>
</tr>
<tr>
<td>2</td>
<td>0.6571</td>
<td>0.5413 (21.39%)</td>
</tr>
<tr>
<td>3</td>
<td>0.2691</td>
<td>0.2336 (15.20%)</td>
</tr>
<tr>
<td>4</td>
<td>0.6730</td>
<td>0.5115 (31.57%)</td>
</tr>
<tr>
<td>5</td>
<td>1.6464</td>
<td>1.3672 (20.04%)</td>
</tr>
<tr>
<td>6</td>
<td>3.9981</td>
<td>3.2625 (22.55%)</td>
</tr>
</tbody>
</table>

---

### Table V

Type I Solutions for the Seven Test Examples with \( S = 0.98 \)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13.21, 1.65</td>
<td>6.59, 6.38</td>
<td>7.57, 4.76</td>
<td>11.42, 4.82</td>
<td>6.06, 4.47</td>
<td>8.52, 4.42</td>
<td>20.65, 15.37</td>
</tr>
<tr>
<td>2</td>
<td>15.82, 1.02</td>
<td>5.22, 4.56</td>
<td>11.07, 6.15</td>
<td>9.35, 6.00</td>
<td>12.18, 8.22</td>
<td>12.45, 0.87</td>
<td>12.19, 6.12</td>
</tr>
<tr>
<td>3</td>
<td>21.26, 1.54</td>
<td>9.55, 2.90</td>
<td>10.49, 5.43</td>
<td>10.88, 1.86</td>
<td>12.02, 8.68</td>
<td>6.31, 2.06</td>
<td>7.87, 1.18</td>
</tr>
<tr>
<td>4</td>
<td>16.74, 0.08</td>
<td>7.99, 2.72</td>
<td>10.09, 6.65</td>
<td>8.73, 0.28</td>
<td>10.70, 8.48</td>
<td>6.22, 1.98</td>
<td>6.61, 1.45</td>
</tr>
</tbody>
</table>
high. When service level is lower, Type I service is a poor approximation of exact service, and hence the heuristic based on it becomes significantly less accurate than the Hybrid heuristic, as we would expect.

5.3. Type II Heuristic

Although the Hybrid heuristic generally works well, there are still some situations in which it does not perform acceptably (the test examples with a * in Table VI). Recall that the order quantity expression in both the Type I and Hybrid heuristic is actually an EOO formula, which does not take into account any information about the reorder points. The Type I heuristic works well when the target service level is extremely high, because under this condition the optimal order quantity is fairly independent of the target service level (Brown 1967, Peterson and Silver 1979, and Tijms 1986). In the Hybrid heuristic, reorder points are computed with the consideration of order quantities, which enables the heuristic to adjust the inventory investment caused by a poor choice of order quantities. When target service level and order frequency are not too low, that part of the inventory investment can be adjusted to an acceptable level. However, when the target service level and order frequency are too low, the choice of order quantities brings the service to a very high level even if we set reorder points at their minimum levels. (We note, however, that this corresponds to a situation where the user chooses to replenish frequently but wants poor service—hardly a typical combination in practice.) When this occurs, there is little room for adjustment via the reorder points. Therefore, it would appear that high accuracy in these situations requires order quantities and reorder points to be computed simultaneously. This suggests that the Type II model should work well.

To test this conjecture, we use the following iterative approach to implement the Type II heuristic. First we obtain \( Q_o, \nu, r_o, \mu, \) and the relative inventory investment from the Hybrid heuristic. Then we apply the following three steps in each iteration:

1. Adjust \( \nu \) and compute \( Q_o \) using (4.48) to achieve the highest order frequency that satisfies the order frequency constraint;
2. Adjust \( \mu \) and compute \( r_o \) using (4.49) to achieve the smallest service level that satisfies the service level constraint; and
3. Compute the current inventory investment and record the best inventory investment so far.

The algorithm stops when the number of iterations reaches a preset maximum number of iterations or the Lagrange multipliers converge.

We applied the Type II heuristic to the seven examples marked with a * in Table 6 and summarize the results in Table VIII. We conclude that the Type II heuristic works significantly better in all trials except those associated with test Example 7. In this example with \( F = 1 \) and \( S = 0.85 \) or 0.9, the algorithm computes \( \mu \) to be near 0 and hence the reorder points are all \(-1\) (i.e., the target order frequency is so low that it pushes the service level to 0.9265 even when we set all reorder points to \(-1\)). Even the Type II heuristic cannot adjust the inventory investment caused by such a poor choice of order quantities. However, as noted above, the situation represented by this example is not terribly realistic and, given this, a 12.13% error is not too bad. Performance improves slightly when \( S \) increases to 0.95 (with \( F = 1 \)) in test Example 7, improves even more when \( S \) is increased to 0.98, and improves substantially when \( S \) is set at 0.99. Thus, while setting \( F \) very low causes order quantities that greatly affect the service level, the Type II model coupled with a high target service can achieve reasonable accuracy.
Table VIII
Type II Solutions to Selected Test Examples

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>F</td>
<td>s</td>
<td>f</td>
<td></td>
</tr>
<tr>
<td>Ex. 1</td>
<td>0.85</td>
<td>3</td>
<td>0.8598</td>
<td>2.9845</td>
</tr>
<tr>
<td>Ex. 4</td>
<td>0.85</td>
<td>1</td>
<td>0.8503</td>
<td>0.9991</td>
</tr>
<tr>
<td>Ex. 4</td>
<td>0.85</td>
<td>2</td>
<td>0.8510</td>
<td>1.9850</td>
</tr>
<tr>
<td>Ex. 4</td>
<td>0.85</td>
<td>3</td>
<td>0.8516</td>
<td>2.9667</td>
</tr>
<tr>
<td>Ex. 7</td>
<td>0.85</td>
<td>1</td>
<td>0.9265</td>
<td>0.9665</td>
</tr>
<tr>
<td>Ex. 7</td>
<td>0.9</td>
<td>1</td>
<td>0.9265</td>
<td>0.9665</td>
</tr>
<tr>
<td>Ex. 7</td>
<td>0.95</td>
<td>1</td>
<td>0.9514</td>
<td>0.9762</td>
</tr>
<tr>
<td>Ex. 7</td>
<td>0.98</td>
<td>1</td>
<td>0.9812</td>
<td>0.9972</td>
</tr>
<tr>
<td>Ex. 7</td>
<td>0.99</td>
<td>1</td>
<td>0.9902</td>
<td>0.9999</td>
</tr>
</tbody>
</table>

We can draw the following conclusions concerning the performance of the three heuristics:

1. The Type I heuristic tends to work well provided the target service is high (above 95%). At extremely high service levels, Type I service is close to the exact service and therefore leads to an accurate model.

2. The Hybrid heuristic works well provided that order frequency is sufficiently high (at least two per year). If order frequency is small (1 or 1.5), the order quantities can be so large that service is high regardless of the reorder points chosen. The heuristic has no flexibility to adjust the $r_i$ values to compensate for such errors in the $Q_i$ values. When order frequency, $F$, is large, this does not happen and the heuristic and the approximation work well.

3. In situations with both low service and low order frequency, the Type II heuristic, which is not easily implementable because $Q_i$ and $r_i$ are coupled, can work better than either of the two simpler methods. If the computer system on which the policy is implemented is sophisticated enough to solve coupled pairs of equations, this method can be a practical all-purpose alternative, since it still provides separation by part number (i.e., the entire system does not need to be regenerated each time a new part is added).

5.4. Stability Tests

As we noted earlier, to stabilize the Lagrange multipliers, we make use of average inventory investment as the objective in the optimization model. To test the effectiveness of this approach, we examined the effect of adding 10 new part numbers to realistic case 1 (which had 209 parts). (Note that this represents an increase of about 5% in the number of parts, which would be equivalent to about 1,500 new parts in the actual 30,000 part system.) We chose these new parts randomly from our client’s part list.

Using a target service of 0.9549 and order frequency of 1.4911 (i.e., the same levels used in realistic case 1) we computed the optimal Lagrange multipliers using the Type I heuristic and compared them to those generated in the original run, that is $\mu = 10.5524$ and $\nu = 0.1860$. The results are reported in Table IX. The multiplier $\mu$ (for the order frequency constraint increased by as much as 1.5% and decreased by as much as 17%, while the multiplier $\nu$ (for the service constraint) increased by as much as 17% and decreased by as much as 16%. The large changes are the result of adding new parts that happen to have demand rates or costs that are very different from the average of the existing parts.

While the changes in Lagrange multipliers may appear substantial, it must be remembered that these appear under a square root in the (Type I) formulas for both $Q_i$ and $r_i$. Therefore, we would expect $Q_i$ and $r_i$ to be much more robust. We verify this by examining test 6 from Table IX, which represents the most severe change in multipliers with a 17% decrease in $\nu$ and a 16% decrease in $\mu$. In Table X we compute the resulting $Q_i$ and $r_i$ values for the ten new parts using both the reoptimized and nonreoptimized multipliers. For many parts, the order quantity and reorder point do not change at all. For those parts that do change, the changes are not large.

From this, we can conclude that the inventory stocking parameters are quite insensitive to moderate changes in the part mix due to adding new parts or deleting existing ones. Of course, over time the changes will eventually alter the system to the point where reoptimization will become necessary. In our client’s system, part turnover is typically less than 5% per year, and therefore semi-annual or annual re-computation of the Lagrange multipliers should be sufficient. Hence, in fairness we must conclude that the method we are proposing here is only almost easily implementable, since an algorithm is periodically required. However, since the optimization model is not required on a day-to-day basis, we feel that this approach is certainly within the spirit of easily implementable.

Table IX
Lagrange Multipliers for Trials with 5% New Parts

<table>
<thead>
<tr>
<th>Test</th>
<th>$\mu$</th>
<th>$\nu$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.2072</td>
<td>0.1866</td>
</tr>
<tr>
<td>2</td>
<td>10.3079</td>
<td>0.1860</td>
</tr>
<tr>
<td>3</td>
<td>10.7135</td>
<td>0.1939</td>
</tr>
<tr>
<td>4</td>
<td>8.8115</td>
<td>0.1933</td>
</tr>
<tr>
<td>5</td>
<td>10.3898</td>
<td>0.1924</td>
</tr>
<tr>
<td>6</td>
<td>8.7591</td>
<td>0.1556</td>
</tr>
<tr>
<td>7</td>
<td>9.2998</td>
<td>0.2167</td>
</tr>
<tr>
<td>8</td>
<td>10.4508</td>
<td>0.2128</td>
</tr>
<tr>
<td>9</td>
<td>9.3976</td>
<td>0.2190</td>
</tr>
<tr>
<td>10</td>
<td>10.6569</td>
<td>0.1942</td>
</tr>
</tbody>
</table>
6. IMPLEMENTATION EXPERIENCE

The above heuristics provide a foundation for addressing the concerns of our client. However, there were several practical issues that needed attention:

1. **Zero Usage Parts**: Many of the 30,000 parts in the system showed no usage in the previous year. At the time of our initial contact, our client considered it reasonable to set \( \lambda_i = 0.5 \) for such parts to represent the possibility that future demand could occur. However, after initial runs of our model (predictably, in hindsight) tended to result in stocking of the cheap parts but not the expensive ones and incorporating their service levels into aggregate service, our client reconsidered. They felt that the data on these parts were too speculative to allow them to have an influence on the other parts. Therefore, we decided to omit these parts from the model and set their stocking policies according to other considerations (e.g., whether or not the part is being phased in or phased out of the system).

2. **Minimum Service Levels**: To the shock of our client, initial runs of our model (with \( r_i = -1 \)) resulted in solutions that carried no inventory of parts with substantial demand (e.g., it set \( Q_i = 1 \) and \( r_i = 1 \) for parts with \( \lambda_i = 20 \) or more units per year, so that replenishments would be ordered only after a demand has been backordered). The reason, of course, was that these were expensive parts and that the model was able to reduce overall inventory investment by not stocking them and compensating by increasing inventory levels, and service, of less expensive parts. But, as our client firmly pointed out, such a policy would unevenly distribute service delays to customers whose machines happened to require these parts. To avoid this, we changed the lower limits on the reorder points to \( r_i = \theta_i \), so that the model would always order when stock fell below average lead time demand for a part. This would ensure that service for any part would be no less than approximately 50%.

We point out, however, that if one moves to an explicit multi-echelon model (with the DC model treated here as a submodel) in which customer service level is constrained, then restrictions on fill rates at the DC would become unnecessary. However, in the interim, while a multi-echelon approach is being worked out, our client felt that such constraints would be a reasonable proxy for customer service constraints.

3. **Forecasting Demand**: After looking at multiple years of usage data, it became clear that demand rates varied significantly from year to year. Because of this, we realized that using last year's usage to estimate next year's demand is not necessarily appropriate. Because our client felt that usage rates were unpredictable (i.e., subject to considerable noise), our short-term response was to use a three year moving average to smooth out demand estimates. However, because the underlying dynamics are that usage follows a product life cycle, we also suggested a longer term option of using a forecasting model (e.g., exponential smoothing) incorporating a linear trend. It is a relatively simple matter to show that the model with a trend tends to predict future demands more accurately than a simple moving average. But, because of the substantial year-to-year noise, a time series model may not be the best approach. Developing a model that explicitly incorporates the position of products that require the part in their life cycles is a future research challenge, as is the problem of determining whether a dynamic forecasting model coupled with a stationary inventory model is an effective way to capture evolving demand patterns.

With these modifications we applied the Type I and Hybrid heuristics to our client's data and compared these to the policy that was currently in effect. We should note that our client’s policy had recently been improved. Formerly, reorder points and order quantities were computed strictly as functions of annual demand. After our initial discussions, they modified their formulas to make adjustments for procurement leadtime. However, the formulas still did not involve component costs.

Table XI shows the results of these comparisons. This table shows the inventory investment (in millions of dollars), service level, the average order frequency achieved by using our client’s (improved) current policy and the Type I and Hybrid heuristics, where the target service and order frequency were set to match those achieved by the client's policy. We have included both the case where parts with no usage last year, modeled using \( \lambda_i = 0.5 \), are considered as regular parts and the case where these parts are eliminated from consideration altogether.

<table>
<thead>
<tr>
<th>Table X</th>
<th>Comparison of Existing Policy with Heuristics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Part</strong></td>
<td><strong>With Reoptimization</strong></td>
</tr>
<tr>
<td>( \mu = 8.7591, \nu = 0.1556 )</td>
<td><strong>( Q_i )</strong></td>
</tr>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>34</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \lambda_i = 0.5 )</th>
<th><strong>Inventory</strong></th>
<th><strong>Order Freq.</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Current Policy</td>
<td>$3.594M</td>
<td>1.773</td>
</tr>
<tr>
<td>Hybrid Heuristic</td>
<td>$2.820M</td>
<td>1.771</td>
</tr>
<tr>
<td>Type I Heuristic</td>
<td>$3.594M</td>
<td>1.770</td>
</tr>
<tr>
<td>Without Inventory</td>
<td>$3.594M</td>
<td>1.770</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parts</th>
<th><strong>Service</strong></th>
<th><strong>Order Freq.</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Current Policy</td>
<td>0.9581</td>
<td>3.6391</td>
</tr>
<tr>
<td>Hybrid Heuristic</td>
<td>0.9584</td>
<td>3.6505</td>
</tr>
<tr>
<td>Type I Heuristic</td>
<td>0.9577</td>
<td>3.6502</td>
</tr>
</tbody>
</table>

(Values rounded to two decimal places for simplicity.)
From our client’s perspective, the most important result shown in Table XI is the fact that the new stocking policies result in roughly a 20% reduction in inventory investment for comparable service and order frequency, which is worth at least $600,000. Table XI also shows that the Type I heuristic works about as well as the Hybrid heuristic. This is not unexpected because of the high level of target service set by the current policy. In fact, the Type I heuristic achieves slightly lower inventory in the case where the \( \lambda_i = 0.5 \) parts are excluded. However, it also achieves slightly lower service (due to the discreteness of the variables and our stopping criterion), and therefore cannot be taken as evidence that the Type I heuristic is more accurate for this case.

Because our heuristics, unlike the policy currently in use by our client, allow the user to specify target service and order frequency levels, they provide a mechanism for examining the sensitivity of inventory investment to these constraints. Figure 3 shows the sensitivity of inventory investment (under the Type I heuristic) to changes in target service level and order frequency. This figure show that while inventory becomes extremely sensitive to service level when it becomes high, it is relatively insensitive to order frequency. Under the assumptions of the Type I model, there are no large gains to be made by placing additional purchase orders per year. However, it remains an open question of how placing more (and smaller) purchase orders per year will affect the ability of the system to respond to changes in demand for parts over time.

At the time of this writing, the system described here has not been implemented yet. The primary reason has been personnel changes that have severely slowed progress. However, our contact with the client has led to (1) the adoption of a slightly better model that does incorporate replenishment leadtime (instead of simply annual usage) in the formulas, (2) a review and upgrade of the accuracy of the leadtime data, (3) greater use of multiple years’ of past usage in estimating future demand, and (4) tighter coordination of the inventories at the DC and in the field. These changes seem to be paying dividends in terms of improving customer service and (slightly) reducing inventories.

We feel that the policy described here could produce even greater gains. However, an impediment to its implementation is the complexity of the optimization model for computing the Lagrange multipliers. To update the stocking formulas, this optimization must be run periodically. Therefore, to make our system even more easily implementable, we are working on a system to classify parts into a few categories (akin to ABC analysis) and use a spreadsheet to allow the user to specify different service targets for the categories leading to an average fill rate equal to the overall target. While this will not be as effective as a complete optimization model that sets a different service level for every part, careful choice of categories and fill rates may still allow substantial gains over the current policy. This approach, which would require nothing more sophisticated than a spreadsheet, would be truly easily implementable.

7. FURTHER WORK

This work was successful in that it resulted in an (almost) easily implementable inventory policy that performs significantly better than the client’s current policy. Our models are also an improvement over current practice because they provide greater control over the inventory system by allowing the user to vary target service and order frequency. This enabled our client to better understand the economic tradeoffs inherent in the system.

In addition to our short-term focus on replacing the optimization model with a spreadsheet-based search tool, two areas are in need of long-term research attention: (1) a better model for forecasting future demand and (2) an explicit approach for explicitly managing inventories in the multi-echelon system. We are continuing to work on both of these and hope to be able to provide our client with tools for making additional improvements in their system.

APPENDIX

A LOWER BOUND PROCEDURE

Single-Product Case

For clarity of presentation, we first present an approach for computing a lower bound for the single-product model.

The Lagrangian dual to formulation (2.9)–(2.13) is:

\[
\text{Maximize } L(v, \mu) \tag{7.55}
\]

Subject to:

\[
\mu \geq 0, \tag{7.56}
\]

\[
v \geq 0, \tag{7.57}
\]

where

\[
L(v, \mu) = \inf_{Q, r, t} \left( r - \theta + \frac{Q}{2} + \frac{1}{2} B(r, Q) \right)
+ v \left( \frac{A}{Q} - F \right) + \mu (A(r, Q)
- (1 - S)) : Q \geq 1, r \geq r \right). \tag{7.58}
\]
Obviously, for any nonnegative \( v \) and \( \mu \), \( L(v, \mu) \) is a lower bound on the objective of the Lagrangian dual and hence the original problem. Since

\[
\frac{r + Q}{2} + \frac{1}{2} = \frac{1}{Q} \sum_{y=r+1}^{r+Q} y, \tag{7.59}
\]

\[
B(r, Q) = \frac{1}{Q} \sum_{i=0}^{\infty} \frac{r+Q}{y=r+1} p(i + y) = \frac{1}{Q} \sum_{y=r+1}^{r+Q} \sum_{i=0}^{\infty} ip(i + y), \tag{7.60}
\]

\[
A(r, Q) = \frac{1}{Q} \sum_{i=0}^{\infty} \frac{r+Q}{y=r+1} p(i + y) = \frac{1}{Q} \sum_{y=r+1}^{r+Q} \sum_{i=0}^{\infty} p(i + y), \tag{7.61}
\]

we can rewrite \( L(v, \mu) \) as

\[
L(v, \mu) = \inf_{Q, r} \left\{ \frac{1}{Q} \left( v\lambda + \sum_{y=r+1}^{r+Q} \left( cy + \sum_{i=0}^{\infty} (ic + \mu) \cdot p(i + y) \right) \right) : Q \geq 1, r \geq r \right\} - c\theta - vF - \mu(1 - S). \tag{7.62}
\]

To further simplify this, let

\[
H(y) = cy + \sum_{i=0}^{\infty} (ic + \mu)p(i + y) \tag{7.63}
\]

\[
= \mu + c\theta + \sum_{i=0}^{y-1} (c(y - i) - \mu)p(i), \tag{7.64}
\]

so that we can write

\[
L(v, \mu) = \inf_{Q, r} \left\{ \frac{1}{Q} (v\lambda + \sum_{y=r+1}^{r+Q} H(y)) : Q \geq 1, r \geq r \right\} - c\theta - vF - \mu(1 - S). \tag{7.65}
\]

This puts our objective into the form used by Federgruen and Zheng (1992). To use their algorithm, we require the following lemma.

**Lemma 1.** \( H(y) \) is unimodal in \( y \).

**Proof.** Consider

\[
H(y + 1) - H(y) = -\mu p(y) + cP(y). \tag{7.66}
\]

If \( c \geq \mu \), then \( H(y) \) is increasing in \( y \). Otherwise

\[
[H(y + 2) - H(y + 1)] - [H(y + 1) - H(y)] = p(y) \left( \mu + \frac{\theta}{y+1} (c - \mu) \right), \tag{7.67}
\]

can only change sign at most once. Since

\[
H(y) \geq 0, \tag{7.68}
\]

\[
H(1) - H(0) = (c - \mu)p(0) < 0, \tag{7.69}
\]

and

\[
\lim_{y \to \infty} (H(y + 1) - H(y)) = c \geq 0, \tag{7.70}
\]

it follows that \( H(y) \) is unimodal. \( \Box \)

With this, we can use Federgruen and Zheng’s algorithm as an efficient means for solving the following formulation of the problem to compute \( L(v, \mu) \) for fixed \( v \) and \( \mu \):

\[
\text{Minimize } \frac{1}{Q} \left( v\lambda + \sum_{y=r+1}^{r+Q} \left( cy + \sum_{i=0}^{\infty} (ic + \mu)p(i + y) \right) \right) \tag{7.71}
\]

Subject to: \( r \geq r, \quad Q \geq 1, \tag{7.72} \)

\( Q, r : \text{integer.} \tag{7.73} \)

Any search procedure (e.g., Powell’s method, which is described in Press et al. 1986) can be used to adjust \( \mu \) and \( v \) in order to find as tight a lower bound as possible.

**Multiproduct Case**

We can extend the above lower bounding procedure to the multiproduct case by observing that the Lagrangian dual to formulation (2.17)–(2.21) is:

\[
\text{Maximize } L(v, \mu) \tag{7.74}
\]

Subject to:

\[
\mu \geq 0, \tag{7.75}
\]

\[
v \geq 0, \tag{7.76}
\]

where

\[
L(v, \mu) = \inf_{Q_i, r_i} \left\{ \frac{1}{C} \sum_{i=1}^{N} c_i \left( r_i - \theta_i + \frac{Q_i}{2} + \frac{1}{2} + B_i(r_i, Q_i) \right) \right. \tag{7.77}
\]

\[
+ v \left( \frac{1}{N} \sum_{i=1}^{N} \lambda_i \right) - \mu \left( \frac{1}{N} \sum_{i=1}^{N} \lambda_i \right) A_i(r_i, Q_i) \right. \]

\[
- (1 - S) : Q_i \geq 1, r_i \geq r_i \right\} = \sum_{i=1}^{N} L_i(v_i, \mu_i) \]

\[
- \mu(1 - S), \tag{7.78}
\]

and

\[
L_i(v_i, \mu_i) = \inf_{Q_i, r_i} \left\{ \frac{1}{Q_i} \left( v_i\lambda_i + \sum_{y=r_i+1}^{r_i+Q_i} \left( c_iy + \sum_{j=0}^{\infty} (jc_i + \mu_i) \cdot p(j + y) \right) \right) : Q_i \geq 1, r_i \geq r_i \right\} \tag{7.79}
\]

\[
- c_i\theta_i, \tag{7.80}
\]

where

\[
c_i = \frac{c_i}{C}, \tag{7.79}
\]

\[
v_i = \frac{v}{N}, \tag{7.80}
\]

\[
\mu_i = \frac{\lambda_i}{\Lambda} \mu. \tag{7.81}
\]

Solving the subproblem
Minimize \( \frac{1}{C} \sum_{i=1}^{N} c_i \left( r_i - \theta_i + \frac{Q_i}{2} + \frac{1}{2} + B_i(r_i, Q_i) \right) \)
\[ + \nu \left( \frac{1}{N} \sum_{i=1}^{N} \lambda_i - F \right) \]
\[ + \mu \left( \frac{1}{N} \sum_{i=1}^{N} \lambda_i A_i(r_i, Q_i) - (1 - S) \right) \]  
(7.82)

Subject to: \( r_i \geq r_i, \quad Q_i \geq 1, \quad i = 1, 2, \ldots, N, \)  
(7.83)

\( Q_i, r_i : \text{integer}, \)  
(7.84)

for given \( \nu \) and \( \mu \) is equivalent to solving \( N \) subproblems of the form

Minimize \( \frac{1}{Q_i} \left( \nu_1 \lambda_i + \sum_{y=r_i+1}^{r_i+Q_i} \left( c_j y + \sum_{j=0}^{y} \left( j c_{+j} + \mu_1 \right) p(y + j) \right) \right) \)  
(7.85)

Subject to: \( r_i \geq r_i, \quad Q_i \geq 1, \)  
(7.86)

\( Q_i, r_i : \text{integer}, \)  
(7.87)

which can be solved as in the single-product case by using Federgruen and Zheng's algorithm coupled with a search procedure for the multipliers \( \nu \) and \( \mu \).

While the solution to (7.82)–(7.84) for given \( \nu \) and \( \mu \) will yield a lower bound on the optimal objective, it may or may not yield a feasible \( (Q_i, r_i) \) solution. Even if the resulting \( Q_i \) and \( r_i \) values are feasible, we cannot conclude that they are optimal because they generally will not make constraints (2.18) and (2.19) tight. The strongest statement we can make is that the \( Q_i \) and \( r_i \) values are optimal for the \( F \) and \( S \) they produce. If we want to use the lower bounding method to find a \( Q_i, r_i \) solution, we can keep track of the best feasible solution found as we search on \( \nu \) and \( \mu \). However, such an iterative approach is far from easily implementable.

**ACKNOWLEDGMENT**

This work was supported in part by the National Science Foundation under grants SES-9119621 and DMI-9322830.

**REFERENCES**


