2019

Naval Wholesale Inventory Optimization

Salmerón, Javier; Craparo, Emily

Springer

Salmeron J., Craparo E.M. (2019) Naval Wholesale Inventory Optimization. In: Fathi M., Khakifirooz M., Pardalos P. (eds) Optimization in Large Scale Problems. Springer Optimization and Its Applications, vol 152. Springer, Cham.

http://hdl.handle.net/10945/67415

Naval Wholesale Inventory Optimization

Javier Salmeron and Emily M. Craparo

Abstract The U.S. Naval Supply Systems Command (NAVSUP), Weapon Systems Support, manages an inventory of approximately 400,000 maritime and aviation line items valued at over $20 billion. This work describes NAVSUP's Wholesale Inventory Optimization Model (WIOM), which helps NAVSUP's planners establish inventory levels. Under certain assumptions, WIOM determines optimal reorder points (ROPs) to minimize expected shortfalls from fill rate targets and deviations from legacy solutions. Each item's demand is modeled probabilistically, and negative expected deviations from target fill rates are penalized with nonlinear terms (conveniently approximated by piecewise linear functions). WIOM's solution obeys a budget constraint. The optimal ROPs and related expected safety stock levels are used by NAVSUP's Enterprise Resource Planning system to trigger requisitions for procurement and/or repair of items based on forecasted demand. WIOM solves cases with up to 20,000 simultaneous items using both a direct method and Lagrangian relaxation. In particular, this proves to be more efficient in certain cases that would otherwise take many hours to produce a solution.

1 Introduction

Wholesale inventory management is broadly concerned with finding strategies to balance customer demand satisfaction with inventory cost. Many different inventory modeling strategies have been proposed; we focus on a framework known as the order-point, order-quantity (s, Q) system ([18], pp. 237–238). In this system, stock replenishment decisions are based on two parameters: the reorder point, s, and the order quantity, Q. As an item's stock level decreases, a reorder is triggered once the item's inventory position is less than or equal to the reorder point s. Inventory position is defined as the quantity on hand plus the quantity on order minus the

J. Salmeron (✉) · E. M. Craparo
Operations Research Department, Naval Postgraduate School, Monterey, CA, USA
e-mail: jsalmero@nps.edu

M. Fathi et al. (eds.), *Optimization in Large Scale Problems*, Springer Optimization and Its Applications 152, https://doi.org/10.1007/978-3-030-28565-4_15

quantity backordered (i.e., owed to customers). When a reorder is triggered, an order of quantity Q is placed. The time it takes for this order to arrive is known as the lead time.

A key feature of an (s, Q) system is that each reorder is triggered by a low inventory position, not low inventory on hand. This prevents the system from placing extra orders during the lead time, when there is already an order due in that will sufficiently replenish the stock level. Silver et al. provide an apt analogy ([18], p. 238):

> A good example of ordering on the basis of inventory position is the way a person takes aspirin to relieve a headache. After taking two aspirin, it is not necessary to take two more every five minutes until the headache goes away. Rather, it is understood that the relief is "on order"— aspirin operates with a delay.

We describe an (s, Q) inventory optimization model known as the Wholesale Inventory Optimization Model (WIOM). This model was developed at the request of the Naval Supply Systems Command (NAVSUP) to provide decision support for approximately 400,000 line items, sometimes referred to as National Item Identification Numbers. The order quantities of these items are predetermined by NAVSUP; thus, WIOM's primary goal is to optimally select reorder points for these items in such a way as to maximize customer demand satisfaction while adhering to a monetary budget. Our figure of merit for customer demand satisfaction is (expected) fill rate, which is defined as the (expected) fraction of customer demand that is satisfied immediately with on-hand inventory, i.e., not backordered. A secondary goal allows the user to optionally discourage deviations from an incumbent solution; this is known as encouraging persistence or reducing "churn" [3]. WIOM attains reorder points that are globally optimal for all items simultaneously considered, i.e., without subordinating certain item decisions to decisions previously made for other items. WIOM is developed as a mixed-integer problem (MIP) that includes the following features, as required by NAVSUP:

- Intrinsic demand uncertainty modeling via probability distribution fitting (parametric) or empirical probability distributions;
- Closed-form approximation of expected fill rate for each item;
- Minimization of weighted, nonlinear penalties due to expected deviation from target fill rates. (Nonlinearities are approximated via piecewise linear functions.)
- Minimization of weighted deviations with respect to legacy levels of safety stock (to be defined later);
- Maximum budget for the expected cost of all items' safety stocks;
- Bounds on decision variables for reorder points.

Section 2 provides a brief overview of the relevant literature. Section 3 describes mathematical constructs that will be used in our formulation. Section 4 describes the WIOM formulation, while Section 5 provides a reformulation of the WIOM model using a Lagrangian relaxation approach. Section 6 compares the two approaches via computational experiments.

2 Literature Review

Given the broad applicability of inventory management systems, it is not surprising that a variety of mathematical models exist to simulate, optimize and provide insights into the behavior of these systems. We provide a brief review of a subset of these models.

Chandra [4] describes a distribution model designed to meet warehouse and customer replenishment requirements, with the goal of minimizing costs incurred from transportation, storage, and orders. Lee [11] expands on work previously done by Sherbrooke [17] to develop a multi-echelon model for repairable items that captures lateral transshipments between customers. Pirkul and Jayaraman [15] develop a MIP to minimize total transportation, distribution, and plant and warehouse costs tri-echelon network with multiple commodities. Axsater [1] considers lateral supply under stochastic demand and develops decision rules to minimize expected costs. Graves [10] considers repairable items in a multi-echelon inventory system and develops an exact model for finding the steady-state distribution of net inventory levels and the number of outstanding orders for each site. Tsiakis et al. [19] use integer programming to determine optimal sizes and locations of warehouses and distribution centers, the resulting transportation links, and the subsequent material flows needed to meet uncertain demands. Ganeshan [9] considers a single item in a multiple-retailer, single-warehouse, multiple-supplier setting and finds near-optimal reorder points and order quantities model to minimize inventory and transportation costs. The model accounts for stochastic demands lead times, as well as customer service constraints. Finally, Ettl et al. [5] formulate a nonlinear optimization problem to minimize the average dollar value of inventory in a supply network, subject to customer service constraints.

The Lagrangian relaxation approach [6] takes advantage of the fact that, in many applications, a small set of constraints complicates an otherwise simple optimization problem. Based on this observation, it reformulates the problem to remove those constraints and instead include additional terms in the objective function to penalize their violations. This approach has some history in inventory systems. For example, Sherbrooke [17] formulates a nonlinear model that calculates stock levels by minimizing total backorders across all customers. He uses a marginal analysis technique to arrive at optimal solutions. Muckstadt [13] then modifies and expands upon Sherbrooke's formulation using Lagrangian relaxation.

3 Fill Rate Calculation

3.1 Overview

Consider a generic inventory item i, whose random demand X_i is known in terms of a probability distribution function (PDF) with density $f_{X_i}(x)$, if continuous, or an analogous probability mass function, if discrete. A PDF is typically estimated by either one of two methods: (a) parametric fit to a standard PDF (e.g., Poisson, binomial, generalized negative binomial, and normal distributions are frequently used in inventory models) using mean and standard deviation estimates from observations or engineering data; or, (b) non-parametric fit by observing demand in lead-time intervals for an empirical distribution. The specific method used in each case may depend on the item characteristics and the number of observations available, among others.

For the items considered by NAVSUP, the order quantities Q_i are also provided, therefore we will consider them fixed. WIOM's primary figure of merit depends on an expected fill rate calculation. Thus, we require a closed-form formula to approximate each item's steady-state, expected fill rate, f_i, as a function of the inputs, $f_{X_i}(x)$ and Q_i, and the chosen reorder point s_i.

The difficulty in calculating the steady-state, expected fill rate (except in trivial cases) stems from the massive number of potential realizations of random demand X_i over a long period. Each realization results in a different pattern of orders met (i.e, with on-hand stock available), and backorders, which must be averaged to estimate the expected fill rate. This inherent difficulty can be dealt with via simulation of random demand arrivals, and subsequent order placement (upon reaching a given reorder point). However, our goal is to develop a closed-form approximation of expected fill rate that enables us to incorporate the reorder point as a decision variable in an optimization model where items share other constraints.

The baseline, closed-form calculation of expected fill rate used in this work is based on the well-known approximation described in Silver et al. ([18], pp. 258, 299):

$$1 - f_i = \frac{1}{Q_i} \int_{s_i}^{\infty} (x - s_i)\, f_{X_i}(x)\, dx. \tag{1}$$

We note the right-hand side in Eq. (1) attempts to estimate expected backorders during a lead-time period. In the case of normal demand, $X_i \sim N\left(\hat{\mu}_{X_i}, \hat{\sigma}_{X_i}\right)$, the suggested formula to calculate the reorder point is $s_i = \hat{\mu}_{X_i} + k_i \hat{\sigma}_{X_i}$, where k_i (known as the safety factor) satisfies:

$$1 - f_i = \frac{1}{Q_i} \hat{\sigma}_{X_i} \int_{k_i}^{\infty} (u - k_i)\, f_{N(0,1)}(u)\, du. \tag{2}$$

Note: The equivalence with (1) follows after substituting $k_i = \left(s_i - \hat{\mu}_{X_i}\right)/\hat{\sigma}_{X_i}$ into (2) and a variable change $u = \left(x - \hat{\mu}_{X_i}\right)/\hat{\sigma}_{X_i}$. Since $\int_{k_i}^{\infty} (u - k_i)\, f_{N(0,1)}(u)\, du$ is tabulated (see, e.g., [18], pp. 724–734), identity (2) becomes very practical for normally distributed demand. Of course, in practice, $s_i = \hat{\mu}_{X_i} + k_i \hat{\sigma}_{X_i}$ must be rounded to an integer.

Equation (1) requires several assumptions ([18], p. 253), including "no crossing orders" and "average level of backorders negligibly small when compared with the average level of on-hand stock." We note that a large percentage of the items NAVSUP handles have expected demand during the lead time that largely exceed

Q_i and, therefore, may not fully comply with the above (depending on the chosen reorder point). In addition, when Q_i is less than the expected demand during the lead time, we can expect multiple, simultaneous orders during a lead-time period.

As a consequence, for many modeled items, Eq. (1) suffers from incorrectly assessing the shortage from the reorder point s_i. A better reference would be the expected on-hand inventory for item i at the time of placing an order (less than s_i in cases in the cases when there is more expected demand during the lead time than the order quantity, Q_i). Accordingly, we should adjust the fill rate estimation in Eq. (1) as follows:

- Define $\tilde{c}_i := \max\left\{1, \frac{\hat{\mu}_{X_i}}{Q_i}\right\}$ as the expected number of cycles (expected orders per lead-time period). When Q_i is large (no simultaneous orders expected), $\tilde{c}_i = 1$ and a cycle's length matches the lead time. When Q_i is small, $\tilde{c}_i = \frac{\hat{\mu}_{X_i}}{Q_i} > 1$, and the lead time has many cycles.
- Replace lead-time demand X_i by "cycle-time demand" Y_i. As for X_i, a PDF $f_{Y_i}(.)$ for Y_i can be estimated with: (a) parametric methods, using mean cycle demand $\hat{\mu}_i := \hat{\mu}_{Y_i} = \hat{\mu}_{X_i}/\tilde{c}_i$ and standard deviation $\hat{\sigma}_i := \hat{\sigma}_{Y_i} = \hat{\sigma}_{X_i}/\tilde{c}_i$; or, (b) via an empirical distribution from observations drawn in intervals of cycle-time length.
- Replace s_i by $s_i' := s_i - (\tilde{c}_i - 1)\, Q_i$, which is the expected on-hand at the beginning of a cycle.

The adjusted fill rate estimate for item i, with cycle-time demand Y_i, is given by the following equation:

$$1 - f_i = \frac{1}{Q_i} \int_{s_i'}^{\infty} \left(y - s_i'\right) f_{Y_i}(y) dy. \tag{3}$$

Roth [16] performed simulations on thousands of parts with parametrically fit demand distributions. He concluded that "the majority of WIOM estimated fill rates are within 2% of the simulated fill rates," with a slight tendency to over-estimate fill rate.

3.2 Properties of the Adjusted Fill Rate Approximation and Modeling Considerations

We first note the following property:

Property 1 The fill rate approximation established by Eq. (3) satisfies $1 - \tilde{c}_i \leq f_i \leq 1$.

Proof:
To prove $f_i \leq 1$ we note that $\int_{s_i'}^{\infty} \left(y - s_i'\right) f_{Y_i}(y) dy \geq 0$. Given that $Q_i \geq 1$, it immediately follows that $f_i = 1 - \int_{s_i'}^{\infty} \left(y - s_i'\right) f_{Y_i}(y) dy / Q_i \leq 1$.

The lower bound, $1 - \tilde{c}_i \leq f_i$, is derived as follows:

$$f_i = 1 - \int_{s_i'}^{\infty} \left(y - s_i'\right) f_{Y_i}(y) dy / Q_i = 1 - \int_{s_i'}^{\infty} y \, f_{Y_i}(y) dy / Q_i + \int_{s_i'}^{\infty} s_i' f_{Y_i}(y) dy / Q_i$$

$$\geq 1 - \hat{\mu}_i / Q_i + (s_i - (\tilde{c}_i - 1)\, Q_i) \underbrace{\Pr\{Y_i \geq (s_i - (\tilde{c}_i - 1)\, Q_i)\}}_{p} / Q_i$$

$$= 1 - \underbrace{\left(\hat{\mu}_{X_i} / \tilde{c}_i\right) / Q_i}_{=1} + s_i \, p / Q_i - (\tilde{c}_i - 1)\, p \geq p \underbrace{(1 - \tilde{c}_i)}_{\leq 0} \geq 1 - \tilde{c}_i .$$

■

Property 1 ensures that, when $\tilde{c}_i = 1$, we can ensure $0 \leq f_i \leq 1$. However, like Eq. (1), the adjusted fill rate Eq. (3) could still produce a "negative" fill rate estimate, as the lower bound is only $1 - \tilde{c}_i$. In such cases we wish the optimization to use an estimated fill rate of zero in the calculations; thus, we need to decompose the fill rate approximation into its positive and negative components, f_i^+ and f_i^-, respectively. More specifically, consider the following indicator:

$$\Delta_i^{\tilde{c}} = 1 \text{ if } \tilde{c}_i > 1, \text{and zero otherwise } \left(\text{if } \tilde{c}_i = 1\right),$$

and set $f_i = f_i^+ - \Delta_i^{\tilde{c}} f_i^-$. The restated approximation to be used in the optimization model is:

$$1 - \left(f_i^+ - \Delta_i^{\tilde{c}} f_i^-\right) = \int_{s_i'}^{\infty} \left(y - s_i'\right) f_{Y_i}(y) dy / Q_i, \tag{4}$$

which reduces to Eq. (3) when $\Delta_i^{\tilde{c}} = 0$, and otherwise carries out the desired decomposition of fill rate as $f_i = f_i^+ - f_i^-$. Still, in order to make Eq. (4) work, we need to ensure its right-hand side is met in a manner that prevents f_i^+ and f_i^- from becoming positive simultaneously. Thus, we use binary variables, $\tilde{f}_i^+$ and $\tilde{f}_i^-$ to control if the estimation in the right-hand side of Eq. (4) is positive or negative. This is accomplished by the three following constraints:

$$\tilde{f}_i^+ + \tilde{f}_i^- = 1; \quad \tilde{f}_i^+ \geq f_i^+; \quad \text{and,} \quad \tilde{f}_i^- \geq f_i^- / \tilde{M}_i, \tag{5}$$

where $\tilde{M}_i$ could be any sufficiently large constant. Note that Constraints (5) ensure:

$$\tilde{f}_i^+ = 1, \quad \tilde{f}_i^- = 0, \quad f_i^- = 0 \text{ if } f_i^+ > 0, \quad \text{and}$$

$$\tilde{f}_i^+ = 0, \quad \tilde{f}_i^- = 1, \quad f_i^+ = 0 \text{ if } f_i^- > 0.$$

These constraints appear in WIOM when $\tilde{c}_i > 1$. For numerical computation purposes, it is a good idea to make $\tilde{M}_i$ as small as possible. Based on Property 1, we set $\tilde{M}_i = |1 - \tilde{c}_i| = \tilde{c}_i - 1 > 0$.

4 Optimization Model

WIOM is established a MIP with piece-wise, linear approximations of nonlinear penalties for deviations from target fill rates. The mathematical formulation of WIOM follows:

4.1 Indices and Index Sets

i, item, for $i \in I$;
n, demand-level index for item $i \in I$, for $n \in N_i$;
m, penalty segment index for piece-wise linearization of nonlinear penalties (applied to deviations from target fill rates), for $m \in M$.

4.2 Input Data [with Units, if Applicable]

t_i, lead-time for item i [quarters];
X_i, lead-time demand random variable [units of issue per lead-time period];
$\hat{\mu}_{X_i}$, expected value of X_i [units of issue per lead-time period];
Q_i, order quantity for item i [units of issue per order];
$\overline{f}_i$, desired (target) fill rate for item i [fraction];
w_i, weight for meeting required fill rate for item i [weight units];
c_i, cost per unit in safety level [\$/unit of issue];
b, safety stock budget for all items [\$];
$\underline{s}_i, \overline{s}_i$, lower and upper bounds on reorder point for item i [units of issue], [units of issue];
$\hat{z}_i^{\mathrm{SS},0}$, legacy (i.e., initial) safety stock used to encourage persistence for item i [units of issue];
δ_i^{P}, relative penalty for (lack of) persistence with respect to legacy reorder point for item i [fraction]. Note: assume, without loss of generality, that $\sum_{i \in I} \delta_i^{\mathrm{P}} = 1$;
γ^{P}, persistence penalty [relative weight of persistence with respect to fill rate, e.g., fill rate penalty/unit of persistence deviation].

Derived data:

$\tilde{c}_i$, number of cycles during a lead time for item i [orders per lead time]: $\tilde{c}_i := \max\{1,\ \hat{\mu}_{X_i}/Q_i\}$;
$\Delta_i^{\tilde{c}}$, one if $\tilde{c}_i > 1$ (i.e., if $\hat{\mu}_{X_i} > Q_i$), and zero otherwise;

$\hat{\mu}_i$, expected value of demand during a cycle [units of issue per cycle]: $\hat{\mu}_i := \hat{\mu}_{X_i}/\tilde{c}_i$;

d_{in}, p_{in}, n-th level of cycle demand and its probability, respectively, for item i [units of issue], [fraction]. For most of the items, if a parametric fit to a discrete distribution or an empirical PDF is used, we allow up to a maximum of 100 demand levels (e.g., $d_{in} := n, \forall\, n = 0...99$) and calculate the associated p_{in} at each level using the mass function; for continuous parametric fits, we set $p_{in} := 0.005 + 0.01n, \forall\, n = 0...99$ and derive d_{in} as the value of the inverse cumulative probability distribution level at p_{in}. In some cases, special modifications to these rules are also considered;

w_{im}, penalty for deviation from target fill rate for item i within bracket m [penalty units]: $w_{im} := m\, w_i$. (Note: the first bracket $m = 1$ has the lowest penalty rate, creating an incentive to avoid subsequent brackets as the penalty becomes steeper);

$\overline{f}^-_{im}$, maximum deviation below target for item i within bracket m [fraction]: $\overline{f}^-_{im} = \overline{f}_i \;\; m^2/\sum_{j \in M} j^2$. (Note: this divides the maximum fill rate shortage $\overline{f}_i$ into segments where the first bracket $m = 1$ is the shortest);

$\tilde{M}_i$, large number greater (in magnitude) than any possible "negative" fill rate estimation for item i, if $\tilde{c}_i > 1$ [fraction]: $\tilde{M}_i := |1 - \tilde{c}_i| = \tilde{c}_i - 1$ is used in our models.

4.3 Decision Variables

s_i, reorder point for item i [units of issue];

$z_i^{SS,+}$, $z_i^{SS,-}$, deviations below and above, respectively, with respect to initial safety stock for item i [units of issue];

f_i^+, f_i^-, positive and negative components, respectively, for the expected fill rate for item i [fraction]. (The negative component is only applicable if $\tilde{c}_i > 1$);

$\tilde{f}_i^+$, $\tilde{f}_i^-$ binary variables to record the expected fill rate sign (only applicable if $\tilde{c}_i > 1$);

f^-_{im}, expected fill rate shortage (with respect to target) for item i within penalty segment m [fraction];

z_{in}^{SO}, ancillary variable for expected stockouts for item i if demand level n occurs [units of issue]:

$$z_{in}^{SO} = \max\left\{d_{in} - (s_i - (\tilde{c}_i - 1)\, Q_i)\,, 0\right\}; \tag{6}$$

z_i^{SS}, ancillary variable for "planned safety stock" for item i [units of issue]:

$$z_i^{SS} = \max\left\{s_i - \hat{\mu}_i, 0\right\}; \tag{7}$$

f, **z**, **s**, decision vectors for all *f*-, *z*-, and *s*-termed decision variables, respectively;
$\mathcal{W}(\mathbf{f};\mathbf{z};\mathbf{s})$, WIOM's objective function value assessed at decision vectors **f**, **z**, **s**.

4.4 Formulation

We formulate WIOM as the following MIP:

$$\text{WIOM}: \min_{\mathbf{f};\mathbf{z};\mathbf{s}} \mathcal{W}(\mathbf{f};\mathbf{z};\mathbf{s}) = \sum_i \sum_m w_{im}\ f_{im}^- + \gamma^{\mathrm{P}} \sum_i \delta_i^{\mathrm{P}} \frac{z_i^{\mathrm{SS},+} + z_i^{\mathrm{SS},-}}{\left|\hat{z}_i^{\mathrm{SS},0}\right| + 1}, \tag{8}$$

subject to:

$$Q_i\left(1 - \left(f_i^+ - \Delta_i^{\tilde{c}} f_i^-\right)\right) = \sum_{n \in N_i} p_{in} z_{in}^{\mathrm{SO}} \quad \forall i, \tag{9}$$

$$z_{in}^{\mathrm{SO}} \geq d_{in} - \left(s_i - \left(\tilde{c}_i - 1\right) Q_i\right) \quad \forall i, n \in N_i, \tag{10}$$

$$z_{in}^{\mathrm{SO}} \geq 0 \quad \forall i, n \in N_i, \tag{11}$$

$$\tilde{f}_i^+ \geq f_i^+ \quad \forall i \mid \Delta_i^{\tilde{c}} = 1, \tag{12}$$

$$\tilde{f}_i^- \geq f_i^- / \tilde{M}_i \quad \forall i \mid \Delta_i^{\tilde{c}} = 1, \tag{13}$$

$$\tilde{f}_i^+ + \tilde{f}_i^- = 1 \quad \forall i \mid \Delta_i^{\tilde{c}} = 1, \tag{14}$$

$$f_i^+ \geq \overline{f}_i - \sum_m f_{im}^- \quad \forall i, \tag{15}$$

$$z_i^{\mathrm{SS}} \geq s_i - \hat{\mu}_i \quad \forall i, \tag{16}$$

$$z_i^{\mathrm{SS}} \geq 0 \quad \forall i, \tag{17}$$

$$\sum_i c_i z_i^{\mathrm{SS}} \leq b, \tag{18}$$

$$s_i - \hat{\mu}_i = \hat{z}_i^{SS,0} + z_i^{SS,+} - z_i^{SS,-} \quad \forall i, \tag{19}$$

$$\underline{s}_i \leq s_i \leq \overline{s}_i \quad \forall i, \tag{20}$$

$$s_i \geq 0 \text{ and integer} \quad \forall i, \tag{21}$$

$$z_i^{SS,+}, z_i^{SS,-} \geq 0 \quad \forall i, \tag{22}$$

$$0 \leq f_{im}^- \leq \overline{f}_{im}^- \quad \forall i, m, \tag{23}$$

$$\tilde{f}_i, \tilde{f}_i^- \geq 0 \text{ and integer} \quad \forall i \mid \Delta_i^{\tilde{c}} = 1, \tag{24}$$

$$f_i^+, f_i^- \geq 0 \quad \forall i. \tag{25}$$

4.5 *Description of the Formulation*

The objective function (8) has two goals: (a) minimizing weighted deviations from target fill rates across all items (with steeper penalties applied as we move away from the target fill rate for each item); and (b) minimizing weighted penalties for lack of persistence (relative deviations from legacy safety stocks). The persistence term can be voided altogether by setting $\gamma^P = 0$.

Constraints (9) capture the (discretized) approximation of expected fill rate in Eq. (4). In particular, the stockouts at each demand level, as specified in Eq. (6), are implemented in the model as linear constraints (10) and (11).

Constraints (12)–(14) simply restate (5) by decomposing the closed-form fill rate calculation into positive and negative components.

Constraints (15) allocate the shortfall of the achieved fill rate with respect to target fill rate into different penalty brackets (in increasing order per the objective function).

Constraints (16) and (17) calculate NAVSUP's so-called "planned safety stock," as specified in Eq. (7). The combined cost of all planned safety stocks is limited by a budget in Constraint (18).

Constraints (19) calculate the deviations up or down from given initial safety stocks, for the purpose of calculating persistence penalties. We note that the deviation is with respect to the "unconstrained" safety stock $s_i - \hat{\mu}_i$ (which could be a negative value), and not necessarily with respect to the planned safety stock $z_i^{SS} \geq 0$ which is used for cost in Constraint (18).

Constraints (20)–(25) establish additional bounds and domain constraints for the decision variables.

4.6 Alternative Formulations and Generalizations

In some settings, expected fill rate requirements are established by groups of items. That is, if g is a group comprised of items $i \in I_g$, Constraints (15) are replaced by:

$$\frac{\sum_{i \in I_g} \hat{\mu}_i^{\text{Year}} f_i^+}{\sum_{i \in I_g} \hat{\mu}_i^{\text{Year}}} \geq \overline{f}_g - \sum_m f_{gm}^- \quad \forall g, \tag{26}$$

where:

- $\hat{\mu}_i^{\text{Year}}$ represents the yearly (or any other fixed-time reference) expected demand for item $i \in I_g$;
- $\overline{f}_g$ is the overall target on expected fill rate for the group's items; and,
- f_{gm}^- is the decision variable for expected fill rate shortage in group g and penalty bracket m.

Note that Constraints (26) assess a group's overall expected fill rate by factoring in the items' demands. This ensures that an item that is seldom ordered contributes less to the group's overall expected fill rate than another item that is frequently ordered.

Of course, for those items in groups, WIOM also replaces objective function terms $\sum_i \sum_m w_{im} f_{im}^-$ by $\sum_g \sum_{i \in I_g} \sum_m w_{im} f_{gm}^-$.

The original WIOM formulation (without groups) is slightly simpler and more frequently used by NAVSUP. We note, however, that the formulation by groups can be seen a generalization of the original WIOM, which considers single-item groups.

Although NAVSUP defines safety stock as in (7), other variations are conceivable. In some settings, z_i^{SS} could take into account the specific probability of each level of safety stock:

$$z_i^{\text{SS}} = \sum_{n \in N_i} p_{in} \max\{s_i - d_{in}, 0\}. \tag{27}$$

We note that $\sum_{n \in N_i} p_{in} \max\{s_i - d_{in}, 0\} \geq \sum_{n \in N_i} p_{in} (s_i - d_{in}) = s_i - \hat{\mu}_i$. Strict inequality occurs frequently, for example, by simply taking two equally likely demand levels: $d_{i1} = 0$, $d_{i2} = 2$, $p_{i1} = p_{i2} = 0.5$. Then, if the chosen reorder point is $s_i = \hat{\mu}_i = 1$, the definition in (7) leads to $z_i^{\text{SS}} = 0$; but, the alternative definition in (27) yields $z_i^{\text{SS}} = 0.5$. Thus, it is important to establish the desired interpretation of safety stock beforehand. Of course, in order to express (27) as linear constraints in WIOM, we would need to add ancillary variables z_{in}^{SS}, similarly to how we used z_{in}^{SO} for stockouts.

5 Lagrangian Approach

5.1 Lagrangian Model

The instances provided by NAVSUP vary widely in size and complexity. WIOM can solve some instances in seconds, while others take hours using a commercial MIP solver. For this reason, we have developed an alternative approach via Lagrangian relaxation (see, e.g., [2], pp. 257–300, [6, 14], pp. 323–337).

The only coupling (i.e., complicating) constraint in WIOM is (18), which limits the planned safety stock cost to a given budget. The removal of the budget constraint requires penalizing its violation in the objective function and creating the Lagrangian version of WIOM (LWIOM):

$$\text{LWIOM} : \max_{\theta \geq 0} \mathcal{L}(\theta) := \min_{(\mathbf{f};\mathbf{z};\mathbf{s})} \sum_i \sum_m w_{im} f_{im}^- + \gamma^{\text{P}} \sum_i \delta_i^{\text{P}} \frac{z_i^{\text{SS},+} + z_i^{\text{SS},-}}{\left|\hat{z}_i^{\text{SS},0}\right| + 1} + \theta \left(\sum_i c_i z_i^{\text{SS}} - b \right) \tag{28}$$

subject to (9)–(17), (19)–(25)

$\mathcal{L}(\theta)$ reduces to a number of separable subproblems, either by item, if Constraints (15) are used, or by groups of items, if Constraints (26) replace (15). In either case, those subproblems are notably simpler to solve than the full WIOM. Note that although an instance of WIOM containing n items can, in principle, be separated into n single-item subproblems using Lagrangian relaxation, this may not be the most computationally efficient approach. Our experience indicates that it is preferable to formulate subproblems containing dozens or perhaps hundreds of items instead, due to the overhead involved in formulating the subproblems.

In what follows, $(\mathbf{f};\mathbf{z};\mathbf{s})^*$ and θ^* will denote the optimal solutions to WIOM and LWIOM, respectively, and $\mathcal{W}^* = \mathcal{W}(\mathbf{f};\mathbf{z};\mathbf{s})^*$ and $\mathcal{L}^* = \mathcal{L}(\theta^*)$ their optimal objective function values.

By weak duality, for any $\theta \geq 0$ and WIOM-feasible solution, $(\mathbf{f};\mathbf{z};\mathbf{s})$, we have $\mathcal{L}^* \leq \mathcal{L}(\theta) \leq \mathcal{W}(\mathbf{f};\mathbf{z};\mathbf{s}) \leq \mathcal{W}^*$. However, the presence of integer variables in WIOM creates a potential duality gap $\varepsilon \geq 0$ between $\mathcal{L}^*$ and $\mathcal{W}^*$. That is, $\mathcal{L}(\theta^*) + \varepsilon = \mathcal{W}(\mathbf{f};\mathbf{z};\mathbf{s})^*$ for some $\varepsilon \geq 0$. This gap depends on the relative sizes of (i) the convex hull of the full set of constraints and (ii) the intersection of the convex hull of the non-complicating constraints and the set of complicating constraints ([14], pp. 329). In addition, the gap also depends on the objective coefficients. A proof to guarantee that a certain problem type has no duality gap is, in general, complicated. In most other cases, a counterexample can be easily found. LWIOM incurs duality gaps in some of our cases, as shown by our computational results.

5.2 Lagrangian Algorithm

The Lagrangian function $\mathcal{L}(\theta)$ is concave (but not necessarily differentiable), so solving for $\max_{\theta} \mathcal{L}(\theta)$ can be carried out via subgradient optimization. Here, at each iteration k, the incumbent solution $\hat{\theta}_k$ is updated to a new solution along the direction of unitary vector $\xi_k/\|\xi_k\|$, where $\xi_k = \xi_k\left(\hat{\theta}_k\right)$ is a subgradient of $\mathcal{L}(\theta)$ at $\theta = \hat{\theta}_k$. A so-called step size $\lambda_k \geq 0$ dictates the amount of change along the direction, where the choice of λ_k must satisfy certain conditions in order to ensure asymptotic convergence (see e.g. [2], pp. 446–441). Given those conditions may lead to very slow convergence, in practice, they are replaced by alternative (heuristic) rules that have proven empirically efficient. Given that LWIOM contains a unique dualized constraint, θ is a single, real-valued variable. A subgradient ξ at $\theta = \hat{\theta}$ is given by $\xi\left(\hat{\theta}\right) = \sum_i c_i z_i^{SS}\left(\hat{\theta}\right) - b$, which describes the amount by which the incumbent expected safety stock under- or over-expends the given budget. Thus, $\xi_k/\|\xi_k\|$ can only become ± 1, respectively, and the update step is simply $\hat{\theta}_{k+1} := \hat{\theta}_k \pm \lambda_k$, respectively.

Given that θ is real-valued, θ^* can be found more efficiently than by subgradient methods using univariate search algorithms, such as binary search or dichotomous search. These methods use an initial interval of uncertainty, $\theta \in [\theta_{\min}, \theta_{\max}]$, where $\theta_{\min} = 0$ and $\theta_{\max}$ is specified below (according to Eq. (29)). We next outline these two algorithms:

- Binary search, as inspired by the bisection method: Given $[\theta_{\min}, \theta_{\max}]$, and subgradient function $\xi\left(\hat{\theta}\right) := \sum_i c_i z_i^{SS}\left(\hat{\theta}\right) - b$, verify $\xi(\theta_{\min}) > 0$, and $\xi(\theta_{\max}) < 0$; Main Step: update $\hat{\theta} := (\theta_{\min} + \theta_{\max})/2$ and evaluate $\xi\left(\hat{\theta}\right)$; If $\xi\left(\hat{\theta}\right) > 0$, update $\theta_{\min} := \hat{\theta}$; If $\xi\left(\hat{\theta}\right) < 0$, update $\theta_{\max} := \hat{\theta}$; If $\xi\left(\hat{\theta}\right) \approx 0$ or $\theta_{\min} \approx \theta_{\max}$, STOP (otherwise, return to Main Step).
- Dichotomous search, using golden section: Given $[\theta_{\min}, \theta_{\max}]$, $\alpha := \left(1 + \sqrt{5}\right)/2$, $\theta_a := \theta_{\min} + (1 - \alpha)(\theta_{\max} - \theta_{\min})$, $\theta_b := \theta_{\min} + \alpha(\theta_{\max} - \theta_{\min})$, evaluate $\mathcal{L}(\theta_a)$ and $\mathcal{L}(\theta_b)$; Main Step: If $\mathcal{L}(\theta_a) \geq \mathcal{L}(\theta_b)$, then update $\theta_{\max} := \theta_{\mathrm{b}}$, $\theta_{\mathrm{b}} := \theta_a$, $\mathcal{L}(\theta_b) := \mathcal{L}(\theta_a)$, $\theta_a := \theta_{\min} + (1 - \alpha)(\theta_{\max} - \theta_{\min})$, and evaluate $\mathcal{L}(\theta_a)$; If $\mathcal{L}(\theta_a) < \mathcal{L}(\theta_b)$ then update $\theta_{\min} := \theta_a$, $\theta_a := \theta_b$, $\mathcal{L}(\theta_a) := \mathcal{L}(\theta_b)$, $\theta_b := \theta_{\min} + \alpha(\theta_{\max} - \theta_{\min})$, and evaluate $\mathcal{L}(\theta_b)$; If $\theta_{\min} \approx \theta_{\max}$, STOP (otherwise, return to Main Step).

We have implemented both the binary search and dichotomous search algorithms with very similar computational performance. Note that, whilst the former is driven by the sign of the Lagrangian's subgradient, the latter relies on assessments of the original Lagrangian function. Of course, either method should keep track of the best incumbent lower and upper bounds on $\mathcal{W}^*$. If we let $(\mathbf{f};\mathbf{z};\mathbf{s})^*_{/\hat{\theta}}$ denote the optimal solution to the inner problem in (28) for (primal) variables $(\mathbf{f};\mathbf{z};\mathbf{s})$ given $\hat{\theta}$, that is:

$$\mathcal{L}\left(\hat{\theta}\right)=\sum_i\sum_m w_{im}f^{-*}_{im/\hat{\theta}}+\gamma^{\mathrm{P}}\sum_i\delta_i^{\mathrm{P}}\frac{\mathrm{z}^{\mathrm{SS},+*}_{i/\hat{\theta}}+\mathrm{z}^{\mathrm{SS},-*}_{i/\hat{\theta}}}{\left|\hat{\mathrm{z}}_i^{\mathrm{SS},0}\right|+1}+\hat{\theta}\left(\sum_i c_i z^{\mathrm{SS}*}_{i/\hat{\theta}}-b\right),$$

then, a lower bound is given by $\mathcal{L}\left(\hat{\theta}\right)$ (at any iteration), and an upper bound is given by iterations where $(\mathbf{f};\mathbf{z};\mathbf{s})^*_{\hat{\theta}}$ is WIOM-feasible.

Both methods require us to establish an initial interval of uncertainty for θ, that is $\theta\in[\theta_{\min},\theta_{\max}]$. We first set $\theta_{\min}=0$, noting that, if $(\mathbf{f};\mathbf{z};\mathbf{s})^*_{/\hat{\theta}=0}$ is a feasible solution to WIOM (i.e., if it satisfies budget Constraint (18)), then $\theta^*=0$ constitutes an optimal solution to LWIOM, $(\mathbf{f};\mathbf{z};\mathbf{s})^*_{/\hat{\theta}=0}$ is an optimal solution to the original WIOM, $(\mathbf{f};\mathbf{z};\mathbf{s})^*=(\mathbf{f};\mathbf{z};\mathbf{s})^*_{/\hat{\theta}=0}$, and $\mathcal{L}^*=\mathcal{W}^*$ (no duality gap). This is not a common case but, if the budget is not constraining the optimal solution, trying $\theta^*=0$ before initiating a local search on θ will save unnecessary iterations: the r golden search methods would converge to $\theta^*=0$, but only asymptotically.

We implement the binary or dichotomous search in the more interesting scenario where $\theta_{\min}=0$ does not produce an optimal solution. If WIOM is a feasible problem, by concavity of $\mathcal{L}(\theta)$, we know that $\mathcal{L}(\theta)$ must be monotonically non-decreasing over $[0,\theta^*]$ and monotonically non-increasing for $\theta\geq\theta^*$. Thus, $\theta_{\max}$ should be sufficiently large to ensure $\theta^*\in[0,\theta_{\max}]$. Because θ can be interpreted as the rate of change in $\mathcal{W}(\mathbf{f};\mathbf{z};\mathbf{s})^*$ per unit of change in b, a trivial upper bound on θ can be computed independently of b because:

(i) $\mathcal{W}(\mathbf{f};\mathbf{z};\mathbf{s})\geq 0$ for all feasible $(\mathbf{f};\mathbf{z};\mathbf{s})$;
(ii) all coefficients in the objective function are non-negative; and,
(iii) upper bounds on all objective variables exist as follows:

$$f^-_{im}\leq\overline{f}^-_{im};$$

$$\mathrm{z}_i^{\mathrm{SS},-}\leq\max\left\{\hat{\mathrm{z}}_i^{\mathrm{SS},0}-\left(\underline{s}_i-\hat{\mu}_i\right),0\right\}:=\overline{\mathrm{z}}_i^{\mathrm{SS},-};\quad\text{and,}$$

$$\mathrm{z}_i^{\mathrm{SS},+}\leq\max\left\{\left(\overline{s}_i-\hat{\mu}_i\right)-\hat{\mathrm{z}}_i^{\mathrm{SS},0},0\right\}:=\overline{\mathrm{z}}_i^{\mathrm{SS},+},$$

so we can derive $\theta_{\max}$ by simply substituting those bounds into $\mathcal{W}(\mathbf{f};\mathbf{z};\mathbf{s})$:

$$\theta_{\max}:=\sum_i\sum_m w_{im}\,\overline{f}^-_{im}\;+\;\gamma^{\mathrm{P}}\sum_i\delta_i^{\mathrm{P}}\frac{\overline{\mathrm{z}}_i^{\mathrm{SS},+}+\overline{\mathrm{z}}_i^{\mathrm{SS},-}}{\left|\hat{\mathrm{z}}_i^{\mathrm{SS},0}\right|+1}\tag{29}$$

We note that, even though a tighter (i.e., smaller) $\theta_{\max}$ can be derived based on the (linear programming) interpretation of the dual variable for a constraint, such derivation should be done carefully given that WIOM is a MIP. For example, a unitary increase in budget b can produce a decrease in expected fill rate shortfall even if budget Constraint (18) is met with strict inequality. This, in fact, occurs in our computational experience where, in some cases, the expected cost of safety stock is close (but not equal) to budget b simply because reorder points s_i (and thus safety stocks z_i^{SS}) can only be modified in increments of one full unit, making the added cost of an extra item violate the budget. This MIP duality issue can be realized with a simple academic example: Consider $\max_{z \in \{0,1,2,\dots\}} z$ subject to $2z \leq b$ for $b = 1$. Obviously, the MIP optimal solution is $z^* = 0$ (which under spends the given budget). However, for $b = 2$, $z^* = 1$. This shows how an apparently non-binding budget is actually binding in the MIP. Moreover, a unitary increase of b causes the objective to increase by 1 unit, when the constraint coefficient for z is 2, which (under a purely linear programming reasoning) would make us think that the objective increase should be only 0.5 (as it is in the continuous solution, from $z = 0.5$ to $z = 1$).

6 Computational Results

The test cases presented here represent realistic NAVSUP scenarios, which may consist of consumable items, repairable items, or a mix. In practice, a demand for which there exists a carcass available for repair will trigger a repair order. The carcass will eventually be fixed (unless found unrepairable) and become a ready-for-issue item that is delivered to the customer. That process incurs a different lead time than a regular purchase order. WIOM does not model these two streams for repairable items, but uses NAVSUP's estimates on the fraction of surviving carcasses to approximate a lead time for a "generic" order. This combined lead time does not distinguish if the demand will be fulfilled with a purchase or a repair. Similarly, NAVSUP also provides the order quantity as a combined figure. Finally, NAVSUP uses WIOM's calculated safety stock levels into their Enterprise Resource Planning system. This includes more specific forecasts and algorithms that trigger actual requisitions for item procurement and/or repair based on current data.

WIOM has been developed in the Windows 7 operating system and requires (as additional software) the General Algebraic Modeling System (GAMS) optimization environment with the GAMS/CPLEX solving engine [7, 8, 12].

We compare the performance of the Lagrangian relaxation approach and the full MIP using several instances derived from realistic NAVSUP problems. All instances use the concept of groups introduced in Section 4.6, along with notional budget values. We set a time limit of 2 hours for each method, and a stopping criterion if the incumbent solution is proven within 1% from optimal. Our results appear in Table 1.

Table 1 Computational results comparing the performance of the MIP formulation with the Lagrangian relaxation approach

Input characteristics					Results				
Instance	# items	# groups	Fill rate goal (by group)	Budget ($1,000)	Method	$\mathcal{W}$ (**f**;**z**;**s**)	Run time (minutes)	Status	Budget used ($1,000)
Consumables #1	4,494	51	85%	17,190	MIP	0.17	120	Feasible	17,190
					Lagrangian	0.03	95	Feasible	15,625
Consumables #2	19,948	4	95%	66,782	MIP	0	118	Optimal	66,762
					Lagrangian	0.03	115	Feasible	15,365
Consumables #3	2,368	37	85%	7,923	MIP	0	8.5	Optimal	7,913
					Lagrangian	0.03	3.3	Feasible	2,472
Repairables #1	6,431	251	85%	170,020	MIP	0	119	Optimal	170,011
					Lagrangian	0	95	Feasible	111,774
Repairables #2	923	4	95%	8,929	MIP	0.12	120	Feasible	8,928
					Lagrangian	0.39	0.27	Feasible	4,542
Repairables #3	9,946	178	85%	887,486	MIP	0	5.5	Optimal	548,007
					Lagrangian	0	28.7	Feasible	176,887

Our primary figures of merit are the two methods' computation times, their primal objective values $\mathcal{W}$ (**f**; **z**; **s**) (i.e., the sum of expected fill rate and persistence penalties), and budget expenditures. In all instances, budget values are notional

Results demonstrate several phenomena typical of Lagrangian relaxations. First, for smaller, easier instances, the overhead involved in creating the subproblems causes the Lagrangian approach to take longer than the MIP. In instances such as "Repairables #3," the difference is large enough to give the MIP a significant advantage. As the problem instances become larger and more difficult, the Lagrangian approach becomes advantageous. Despite the fact that case "Repairables #2" is of relatively small size, the MIP times out without finding an optimal solution, while the Lagrangian is able to converge quickly, although it converges to a suboptimal solution.

In addition to the differences in computation time, we also observe patterns in the types of solutions produced by both methods. Notably, the MIP is able to prove optimality of its solutions, while the Lagrangian is only able to certify feasibility due to the duality gap. This occurs even in instances where the solution is, in fact, optimal (e.g., "Repairables #1" and "Repairables #3"). For those instances where the MIP and Lagrangian solutions differ in their primal objective value, the Lagrangian's quality is slightly inferior (i.e., its penalty is higher). However, its budget performance is superior across all instances, often significantly so. While minimizing cost is not an objective in the original formulation, it may nevertheless be a desirable side effect of the Lagrangian formulation, which rewards lower-cost solutions. Note that cost-minimizing solutions could, in principle, also be obtained using the MIP formulation. For instance, one might employ a hierarchical approach by first determining the best possible objective value (with respect to fill rate and persistence), then solving a modified version of WIOM designed to minimize cost, subject to a constraint on the fill rate and persistence penalty. However, such a hierarchical approach would incur additional computation time due to the fact that it involves two separate model runs.

7 Conclusions

We have introduced WIOM, a MIP that helps NAVSUP planners to set reorder points for thousands of maritime and aviation line items under uncertain demand. WIOM seeks to minimize weighted, expected shortfalls from fill rate targets and deviations from legacy solutions under a limited safety stock budget. We adjust an existing closed-form approximation of expected fill rate that better captures multiple expected orders per lead time, and incorporate it into the optimization model. We solve realistic instances of WIOM provided by NAVSUP via both a general-purpose MIP solver and by Lagrangian relaxation. Preference for either method depends on the case and metric used: objective value, computational time, or fraction of budget used.

References

1. Axsater, S.: A new decision rule for lateral transshipments in inventory systems. Manag. Sci. **49**, 1168–1179 (2003)
2. Bazaraa, M., Sherali, H., Shetty, C.: Nonlinear Programming. Theory and Algorithms, 3rd edn. Wiley, Hoboken (2006)
3. Brown, G.G., Dell, R.F., Wood, R.K.: Optimization and persistence. Interfaces. **27**, 15–37 (1997)
4. Chandra, P.: A dynamic distribution model with warehouse and customer replenishment requirements. J. Oper. Res. Soc. **44**, 681–692 (1993)
5. Ettl, M., Feigin, G.E., Lin, G.Y., Yao, D.D.: A supply network model with base-stock control and service requirements. Oper. Res. **48**, 216–232 (2000)
6. Fisher, M.L.: The Lagrangian relaxation method for solving integer programming problems. Manag. Sci. **50**, 1861–1871 (2004)
7. GAMS. Online: www.gams.com (2018)
8. GAMS/CPLEX. Online: https://www.gams.com/latest/docs/S_CPLEX.html (2018)
9. Ganeshan, R.: Managing supply chain inventories: a multiple retailer, one warehouse, multiple supplier model. Int. J. Prod. Econ. **59**, 341–354 (1999)
10. Graves, S.C.: A multi-echelon inventory model for a repairable item with onefor-one replenishment. Manag. Sci. **31**, 1247–1256 (1985)
11. Lee, H.L.: A multi-echelon inventory model for repairable items with emergency lateral transshipments. Manag. Sci. **33**, 1302–1316 (1987)
12. Microsoft Corporation. Online: http://www.microsoft.com/en-us/default.aspx (2018)
13. Muckstadt, J.A.: A continuous time, multi-echelon, multi-item system with timebased service level constraints. In: Mikosch, T.V., Robinson, S.M., Resnick, S.I. (eds.) Analysis and Algorithms for Service Parts Supply Chains. Springer, New York (2005)
14. Nemhauser, G., Wolsey, L.: Integer and Combinatorial Optimization. Wiley, New York (1999)
15. Pirkul, H., Jayaraman, V.: Production, transportation, and distribution planning in a multi-commodity tri-echelon system. Transp. Sci. **30**, 291–302 (1996)
16. Roth, G.F.: In: Naval Postgraduate School (ed.) A Simulation of Alternatives for Wholesale Inventory Replenishment, MS in Operations Research Thesis, Monterey (2016)
17. Sherbrooke, C.C.: Introduction. In: Hillier, F.S. (ed.) Optimal Inventory Modeling of Systems: Multi-Echelon Techniques, 2nd edn. Kluwer Academic Publishers, Boston (2004)
18. Silver, E., Pyke, D., Peterson, R.: Inventory Management and Production Planning and Scheduling, 3rd edn. Wiley, New York (1998)
19. Tsiakis, P., Shah, N., Pantelides, C.C.: Design of multi-echelon supply chain networks under demand uncertainty. Ind. Eng. Chem. Res. **40**, 3585–3604 (2001)