Ann Oper Res (2013) 203:33–54
DOI 10.1007/s10479-011-0871-x

Optimal allocation of stock levels and stochastic customer demands to a capacitated resource

Shuang Chen · Joseph Geunes

Published online: 2 April 2011

Abstract This paper considers a new class of stochastic resource allocation problems that requires simultaneously determining the customers that a capacitated resource must serve and the stock levels of multiple items that may be used in meeting these customers' demands. Our model considers a reward (revenue) for serving each assigned customer, a variable cost for allocating each item to the resource, and a shortage cost for each unit of unsatisfied customer demand in a single-period context. The model maximizes the expected profit resulting from the assignment of customers and items to the resource while obeying the resource capacity constraint. We provide an exact solution method for this mixed integer nonlinear optimization problem using a Generalized Benders Decomposition approach. This decomposition approach uses Lagrangian relaxation to solve a constrained multi-item newsvendor subproblem and uses CPLEX to solve a mixed-integer linear master problem. We generate Benders cuts for the master problem by obtaining a series of subgradients of the subproblem's convex objective function. In addition, we present a family of heuristic solution approaches and compare our methods with several MINLP (Mixed-Integer Nonlinear Programming) commercial solvers in order to benchmark their efficiency and quality.

Keywords Stochastic resource allocation · Generalized Benders decomposition · Lagrangian relaxation · Mixed integer nonlinear optimization

1 Introduction

The optimal deployment of constrained resources lies at the heart of nearly all operations problems. Many operations settings require the assignment of uncertain customer demands to a resource with limited capacity. One such problem is faced by sales personnel who travel

S. Chen · J. Geunes (✉)
Department of Industrial and Systems Engineering, University of Florida, P.O. Box 116595,
Gainesville, FL 32611, USA
e-mail: geunes@ise.ufl.edu

S. Chen
e-mail: scljj@ufl.edu

to customers with a stock of goods for sale. That is, given a set of customers on whom a sales force may call, the sales manager must determine how to allocate customer visits to individual sales personnel, and which items each sales person should stock in their corresponding vehicles prior to making sales calls. Because customers' purchase preferences are not known with 100% certainty prior to a visit, the required items to stock are not known with certainty prior to dispatching the sales person. Although the degree of uncertainty in demand can be decreased via an initial telephone- or Internet-based pre-sales screening, the demand uncertainty cannot be completely resolved prior to the sales visit. Combined demand allocation and multi-item part stocking decisions are therefore required for each sales vehicle prior to departure on a sequence of sales calls. Similar challenges arise in determining a subset of customers that a service technician will visit on a given service call, when the technician carries a stock of spare parts, and missing items needed for repair either require a special unplanned trip or can be obtained quickly from a high-cost supplier.[1]

These customer assignment and part-stock-level decision problems lead to extremely complex optimization problems. Because of the level of complexity associated with this problem class, the literature on the exact solution of problems of realistic size is reasonably limited. In this paper, we consider an important subproblem that arises in this context—the allocation of customers and items to a single vehicle in a single-period context. As we will see, this subproblem is both interesting and challenging by itself, and may arise as a standalone problem when only a single resource is available. Moreover, the ability to solve this class of subproblems exactly will facilitate effective decomposition approaches for handling the multi-resource planning problem.

Demand allocation problems have been widely studied in the literature in the context of make-to-order systems under the assumption that demand streams can be split among the available facilities (resources). In situations where the demand for items from a customer cannot be split and the demand from each customer must therefore be assigned to a single resource, the problem becomes combinatorial in nature and the decision variables are discrete. The literature on this class of problems under stochastic demand is limited, as this class of problems is generally quite difficult to solve. For the problem class we consider, we will assume that a customer cannot be assigned to multiple vehicles, since customers generally prefer a single service visit or a visit by a single sales person. This class of problems will therefore contain a substantial combinatorial component. From the point of view of allocating a customer subset to a capacity-constrained vehicle, this problem is closely related to certain classes of stochastic knapsack problems (SKP) (see, e.g., Barnhart and Cohn 1998; Kleywegt and Papastavrou 2001; Dean et al. 2004; Merzifonluoğlu et al. 2009, and Ağralı and Geunes 2009; Kosuch and Lisser 2010), with each customer's (uncertain) demand quantity corresponding to the typical "item" size in these knapsack problems. However, our problem generalizes this problem class, as it involves the additional dimension of limited item availability as a result of the item-to-vehicle stock level decisions. That is, the vehicle's capacity constrains the allocation of item stock levels to the vehicle, and the item stock levels within the vehicle constrain the ability to meet customer requirements during sales visits. As in several of these cited works on stochastic knapsack problems, we will focus on the case where item sizes (or customer parts requirements) are independent but not necessarily identically distributed normal random variables

[1] As we later discuss, our model applies a penalty cost to each item demanded that is not in stock. Thus, our model might also apply, for example, to service repair contexts in which each customer owns multiple, independent pieces of equipment, each of which may require at most a single replacement part or module.

(the normal distribution parameters we use are such that the probability of negative demand is negligible).

To determine item stock level decisions for a given vehicle, observe that if we fix the customer-to-vehicle assignment decisions, we then face a constrained multiple-item newsvendor problem involving uncertain shortage costs, expected customer revenues from sales visits, and expected variable operating costs. Newsvendor problems have of course been well studied in the literature. Hadley and Whitin (1963) discussed the constrained multi-item newsvendor model in detail. Another class of closely-related problems to the one we consider is known as the class of repair kit problems (e.g., Smith et al. 1980; Teunter and Haneveld 2002a, 2002b, and Gorman and Ahire 2006). Most of these repair kit problems are aimed at minimizing expected holding costs under given service level requirements. In contrast, we consider the maximization of expected profit under an additional vehicle capacity constraint, using a shortage (penalty) cost in the objective (which, in turn, can be used to ensure desired service levels). The most closely related repair kit problem that considers multiple types of items and vehicle capacity is the one proposed by Gorman and Ahire (2006), who developed a single-pass greedy heuristic approach for the multiple vehicle planning problem.

Within the single-period context we consider, demand allocation and vehicle stocking decisions must be determined prior to actual customer demand realizations. Hence, we face a joint demand assignment and stock level problem with uncertain customer demands. Our goal is to simultaneously set stock levels for the multiple items (parts) within the vehicle and determine customer demand assignments in order to maximize expected profit, equal to the expected revenue from customer visits less expected shortage and variable costs. To the best of our knowledge, no existing work considers an exact solution for combined demand allocation and vehicle stocking problems for multiple items under uncertain demand, as we address in this paper. Thus, our contributions include providing a new model for this problem class, as well as exact and heuristic solution procedures, which we demonstrate to be extremely effective when compared with a state-of-the-art mixed integer nonlinear optimization solver.

Our class of stochastic resource allocation problems has several variants relevant to various operations planning contexts. If we restrict our attention to only one type of item, we will have the static stochastic knapsack problem (Barnhart and Cohn 1998; Kleywegt and Papastavrou 2001; Merzifonluoğlu et al. 2009; Ağralı and Geunes 2009; Kosuch and Lisser 2010). If there is no shared resource (vehicle) capacity constraint, we will have a stochastic multidimensional knapsack problem (MKP) (Kellerer et al. 2004; Vasquez and Vimont 2005; Akçay et al. 2007). The MKP is a well-known NP-hard problem, which implies the NP-hardness of our problem under general demand distributions as well. Further, if we consider the multiple vehicle version of our problem with one item type and no shared resource capacity, the problem takes the form of the joint facility assignment and capacity acquisition problem (Taaffe et al. 2008), where the optimal stock level for the part corresponds to the facility capacity. Similarly, if the customer demand assignments are given, we only need to consider the stock levels for multiple items on the vehicle, which reduces to a set of constrained multi-item newsvendor problems.

Within the context of make-to-stock queues, Benjaafar et al. (2004) appear to be the first to consider the joint demand allocation and inventory control problem. They considered the long-run fraction of demand for each product i (in a set of products) allocated to each facility (from a set of facilities, analogous to the vehicles in the context we discussed above). They considered the minimum long-run expected cost per unit time under continuous assignment variables (the demand allocation problem, or DAP), and the demand partitioning problem

(DPP) with binary assignment variables. They used convex optimization algorithms to solve the DAP and a branch and bound algorithm for the DPP.

Federgruen and Zipkin (1984) developed a combined routing and inventory allocation model for a single item under stochastic customer demands (Federgruen et al. 1986, generalized this to account for perishable items when "fresh" and "old" stock of the item is available and out-of-date costs may exist). In their model, customer inventory allocation decisions are made prior to assignment and routing decisions. Thus, customer-related inventory costs are separable. In our model, however, inventory is not allocated prior to customer assignment decisions. Therefore, individual demands for customers who share the resource capacity (i.e., customers visited by the same truck) create a single, *pooled* demand distribution for each item, and inventory costs are not separable by customer. Moreover, we consider a multi-item setting in which resource capacity is shared by different items. We note, however, that their model considered a multiple vehicle context, whereas we consider a single-resource problem.

As we later discuss, under a normal demand distribution for each customer's parts requirements, a generalized Benders decomposition (GBD) approach is quite effective for this nonlinear mixed integer problem. GBD techniques, developed by Geoffrion (1972), can be used to efficiently solve nonlinear programming problems with complicating variables by temporarily fixing these variables and dealing with the remaining problem, which is generally much easier to solve. To extend the applicability of the GBD approach, Geromel and Belloni (1986) studied the differentiability of a set of related perturbation functions and proposed a method to handle problems of a more general class than Geoffrion (1972) considered. We will employ these generalized methods of Geromel and Belloni (1986) to generate Benders cuts. Many practical problems, such as multi-commodity network flows, quadratic assignment problems, and combined location-inventory problems have been efficiently solved using GBD techniques. In our approach, we first fix the complicating integer variables; that is, for a given feasible demand allocation, we solve the remaining multi-item constrained newsvendor subproblem using Lagrangian relaxation (Hadley and Whitin 1963). We then use the solution of this subproblem to add the corresponding support function (Benders cut) to the so-called master problem and solve a relaxed master problem. We repeat this process by using the demand allocation decision from the solution of the relaxed master problem to solve a new version of the newsvendor subproblem. Benders cuts are then iteratively added to the master problem until an optimal solution is found.

The remainder of this paper is organized as follows. In Sect. 2 we define and formulate our stochastic demand allocation and inventory stock-level problem. We then analyze important properties of optimal solutions for our problem and develop a generalized Benders decomposition approach in Sect. 3. Section 4 presents a heuristic solution approach for solving large size problem instances that may preclude the use of exact methods. In Sect. 5 we discuss the results of a computational study used to validate our solution methods and compare them with the results of three benchmark commercial solvers, GAMS/LINDOGlobal, GAMS/SBB and GAMS/CoinBonmin. Finally, concluding remarks are provided in Sect. 6.

2 Problem definition and formulation

In this section, we present a single-period model to solve a joint customer demand allocation and multiple-item stock level problem for a resource that must respond to uncertain customer demands. For example, the resource may correspond to a vehicle (and corresponding sales person) on a given working day, and the items may correspond to items that are potentially needed on sales calls. We assume that customer demands are statistically independent and that the demands for items are revealed upon being visited by the sales person. While there are m potential customers, the supplier must choose which subset of customers to serve using a capacity constrained resource (e.g., a vehicle) and the optimal resource stock level for each item. This individual resource level problem may appear as a stand-alone problem for a manager of a single resource, or as a subproblem for a larger multiple resource assignment problem with multiple customer/item demands.

We assume that during a given visit, any item requested that is not available results in an item-specific penalty cost for an inability to complete service (thus, individual item demands are independent of one another). For each item carried on the vehicle, a variable cost is incurred (e.g., for loading/unloading and/or transporting the item). Because the vehicle stocking decisions must be determined prior to actual customer demand realizations, it may be practical in some contexts to consider a salvage value for each unused item carried on the vehicle (this might correspond to a reduction in future loading/unloading costs; alternatively, a negative value would correspond to an opportunity cost of the vehicle capacity usage). A customer-specific expected revenue is gained for each customer visit. The objective is to maximize the expected profit, or equivalently, to minimize the expected cost (equal to the negative of expected profit). Henceforth, we state our objective using this minimization form and refer to this objective as the *expected cost*.

To formalize our model, we define the following notation:

Inputs and Parameters

i: item (part) index, $i = 1, \dots, m$.
j: customer index, $j = 1, \dots, n$.
$\hat{\pi}_j$: expected revenue gained by serving customer j, i.e., by allocating customer j to the vehicle.
$\hat{e}_i$: unit shortage cost incurred for not satisfying a unit of demand for item i.
g_i: unit salvage value incurred for an unused item i.
$\hat{c}_i$: unit variable cost for carrying item i in the vehicle, where $\hat{e}_i > \hat{c}_i > g_i$.
s_i: unit size of item i.
V: vehicle capacity.
d_{ij}: random variable denoting the demand for item i by customer j, with mean and standard deviation μ_{ij} and σ_{ij}, respectively.

Decision Variables

x_j: binary decision variable, equal to 1 if customer j is assigned to the vehicle, 0 otherwise. The vector ("column") $X = [x_1, \dots, x_n]^T$ characterizes the assignment of customers to the vehicle.
y_i: nonnegative decision variable equal to the number of units of item i carried in the vehicle. The vector $Y = [y_1, \dots, y_m]^T$ determines the assignment of items to the vehicle.

We define $D_i = \sum_{j=1}^{n} d_{ij} x_j$ as the random variable for the aggregate demand of item i by all customers assigned to the vehicle (note that D_i is a random variable with distribution parameters determined by the assignment variable values and the random variables for individual customer and item demands). The expected cost as a function of the part stock levels and customer assignments can be expressed as:

$$\begin{aligned} P(Y,X) &= \sum_{i=1}^{m} \hat{e}_i E\left[\left(\sum_{j=1}^{n} d_{ij}x_j - y_i\right)^{+}\right] - \sum_{i=1}^{m} g_i E\left[\left(y_i - \sum_{j=1}^{n} d_{ij}x_j\right)^{+}\right] \\ &\quad + \sum_{i=1}^{m} \hat{c}_i y_i - \sum_{j=1}^{n} \hat{\pi}_j x_j \\ &= \sum_{i=1}^{m} e_i E\left[\left(\sum_{j=1}^{n} d_{ij}x_j - y_i\right)^{+}\right] + \sum_{i=1}^{m} c_i y_i - \sum_{j=1}^{n} \pi_j x_j, \end{aligned} \tag{1}$$

where $[x]^+ = \max\{x, 0\}$, $e_i = \hat{e}_i - g_i$, $c_i = \hat{c}_i - g_i$, and $\pi_j = \hat{\pi}_j - \sum_{i=1}^{m} g_i \mu_{ij}$ denote the *net* penalty cost, variable cost, and revenue, respectively ($\hat{e}_i > \hat{c}_i > g_i$ implies that both e_i and c_i are nonnegative for all i). Observe that, for any customer j, if the coefficient of the third term is such that $\pi_j = \hat{\pi}_j - \sum_{i=1}^{m} g_i \mu_{ij} \le 0$, it is straightforward to show that we can set $x_j = 0$ without loss of optimality. We therefore assume without loss of generality that $\pi_j > 0$ for all j.

We can now formulate our stochastic resource allocation problem with a resource capacity constraint as:

$$\begin{aligned} \text{(P)} \qquad & \min \ P(Y,X) \\ & \text{subject to:} \quad \sum_{i=1}^{m} s_i y_i \le V, \\ & \qquad\qquad\quad x_j \in \{0,1\}, \quad j = 1,\dots,n, \\ & \qquad\qquad\quad y_i \ge 0, \quad i = 1,\dots,m. \end{aligned}$$

The constraint ensures that the vehicle's capacity is not violated. Observe that we permit the stock levels to take continuous values. For a given vector of customer assignments, X, however, we can easily determine optimal integer stock levels for the associated multi-item constrained newsvendor subproblem (see Hadley and Whitin 1963).

3 Generalized Benders decomposition for (P)

For a given vector X of customer assignments, problem (P) becomes a standard multi-item newsvendor problem with a single constraint:

$$\begin{aligned} \text{(MINV)} \qquad & \min \sum_{i=1}^{m} \left\{ c_i y_i + e_i E\left[(D_i - y_i)^+\right]\right\} \\ & \text{subject to:} \quad \sum_{i=1}^{m} s_i y_i \le V, \\ & \qquad\qquad\quad y_i \ge 0. \end{aligned} \tag{2}$$

It is straightforward to solve (MINV) using Lagrangian relaxation (Appendix 1 discusses a solution approach for MINV based on results from Hadley and Whitin 1963). Moreover, for a given vector of item stock levels Y, we have a 0-1 integer programming problem without explicit constraints. We thus consider a generalized Benders decomposition approach for solving this resource allocation problem. We first streamline the notation for problem (P) by defining:

$$H(X,Y)=\sum_{i=1}^{m} e_i E\left[\left(\sum_{j=1}^{n} d_{ij}x_j-y_i\right)^{+}\right];$$

$$G_1(Y)=\sum_{i=1}^{m} c_i y_i;$$

$$Q(X)=\sum_{j=1}^{n} \pi_j x_j;$$

$$G_2(Y)=\sum_{i=1}^{m} s_i y_i - V.$$

Using this notation, problem (P) can be formulated as:

$$(\mathrm{P}') \qquad \min\ H(X,Y)+G_1(Y)-Q(X) \tag{3}$$

$$\text{subject to:}\quad G_2(Y)\le 0, \tag{4}$$

$$X\in\{0,1\}, \tag{5}$$

$$Y\ge 0. \tag{6}$$

Alternatively, we can formulate problem (P′) in the space of the binary variables as follows:

$$\begin{aligned}(\mathrm{P}'') \qquad &\min\ v(X)-Q(X)\\ &\text{subject to:}\quad X\in\{0,1\},\end{aligned}$$

where, for a given vector X, the value function $v(X)$ is determined by solving the subproblem (SP), formulated as

$$\begin{aligned}(\mathrm{SP}) \qquad & v(X)=\min\ H(X,Y)+G_1(Y)\\ &\text{subject to:}\quad G_2(Y)\le 0,\\ &\qquad\qquad\qquad Y\ge 0.\end{aligned} \tag{7}$$

Because the subproblem (SP) is convex in Y, a dual representation of $v(X)$ can be formulated, and its optimal dual solution value equals the optimal primal solution value (see Theorem 2.3 in Geoffrion 1972). Defining the dual variable $w\ge 0$ corresponding to the constraint in (SP), we can write the Lagrangian dual as

$$v(X)=\max_{w\ge 0}\left\{\min_{Y\ge 0}\left[H(X,Y)+G_1(Y)+wG_2(Y)\right]\right\}. \tag{8}$$

Problem (P′) is therefore equivalent to the following Master Problem (MP):

$$\begin{aligned}(\mathrm{MP}) \qquad &\min\ \theta-Q(X)\\ &\text{subject to:}\quad \theta\ge \min\nolimits_{Y\ge 0}\left[H(X,Y)+G_1(Y)+wG_2(Y)\right],\quad \forall w\ge 0,\\ &\qquad\qquad\qquad X\in\{0,1\}.\end{aligned}$$

Clearly we cannot enumerate all possible constraints in the above formulation for all possible values of w. Therefore, we begin with a relaxed formulation of the master problem and iteratively add violated Benders cuts. In order to find violated Benders cuts, we need to first characterize the support function of $v(X)$, which we denote by $\xi(X)$. For a given vector X^k and the optimal dual solution w^k of (SP) associated with the given X^k, we know that

$$v(X^k) = \min_{Y \geq 0} \left[H(X^k, Y) + G_1(Y) + w^k G_2(Y) \right],$$

and by the definition of the support function of $v(X)$ at the point X^k,

$$\xi(X^k) = v(X^k) \quad \text{and} \quad \xi(X) \leq v(X), \quad \forall X \neq X^k.$$

The following proposition and theorem will aid us in characterizing the support function.

Lemma 1 *$P(Y, X)$ in (1) is a jointly convex function on $\mathbb{R}_n^+ \times \mathbb{R}_m^+$.*

Proof To show that (1) is a jointly convex function, we only need to show that $(\sum_{j=1}^n d_{ij} x_j - y_i)^+$ is a jointly convex function by limiting our attention to a given realization of the demand d_{ij}, because the other terms in (1) are all linear, and the expectation of a convex function is a convex function. Note that $(\sum_{j=1}^n d_{ij} x_j - y_i)^+ = \max(0, \sum_{j=1}^n d_{ij} x_j - y_i)$ is a maximum of two jointly convex functions and is, hence, jointly convex. Then, by the fact that the summation of jointly convex functions is jointly convex, the proof is complete. □

Theorem 1 *$v(X)$ is a convex function, and a linear support function $\xi_k(X)$ can be readily calculated as*:

$$\xi_k(X) = v(X^k) + \delta(X^k)(X - X^k),$$

where $\delta(X^k) \in \partial v(X^k)$ is any available subgradient of $v(.)$ at X^k. One such subgradient of $v(.)$ is:

$$\delta(X^k) = \frac{\partial}{\partial X} \left\{ H(X, Y^k) + G_1(Y^k) + w^k G_2(Y^k) \right\}_{X = X^k}.$$

Proof Using results from Heyman and Sobel (1984, p. 525), we can show that $v(X)$ in (7) is convex in x, since $H(X, Y)$ and $G_1(Y)$ are convex in both x and y. Then, because $v(X)$ is convex, it is well known that $\xi_k(X)$ serves as a linear support function at X^k. Details regarding the correctness of this support function can be found in Geromel and Belloni (1986) using Theorem 3 and formulas (31)–(34). We omit the details for the sake of brevity. □

The details of how we compute $\xi_k(X)$ are provided in Appendix 2. We can now represent the Relaxed Master Problem (assuming K linear support functions have been generated) as:

$$\begin{aligned} \text{(RMP)} \qquad & \min\ \theta - Q(X) \\ & \text{subject to:} \quad \theta \geq \xi_k(X), \quad k = 1, 2, \ldots, K, \\ & \qquad\qquad\quad X \in \{0, 1\}, \\ & \qquad\qquad\quad \theta \in \mathbb{R}. \end{aligned}$$

Here $\theta \geq \xi_k(X)$ is a Benders cut for given values of X^k and w^k. Observe that the (RMP) is a mixed 0-1 linear program with a single continuous variable. While such problems are not generally considered easy (they are, of course, NP-Hard), they are much more tractable than our original nonlinear mixed 0-1 program, and problems of reasonable size can be solved in acceptable time using CPLEX. The generalized Benders decomposition approach (and variants thereof) can be shown to be convergent when X is a finite discrete set.

We next formalize our algorithm as follows:

Step 1: Choose an initial vector X^0 that ensures a feasible solution for the subproblem and select an optimality tolerance ϵ. Solve subproblem (SP) at X^0, which is a multi-item constrained newsvendor problem, obtaining Y^0 and a corresponding optimal dual solution w^0 (Appendix 1 discusses the solution approach for the MINV problem). Set the upper bound $UB = v(X^0) - Q(X^0)$ and let $(\overline{X}, \overline{Y}) = (x^0, Y^0)$ denote the initial solution. Generate the support function of $v(.)$ at X^0 as the Benders cut using Theorem 1 and the results in Appendix 2.

Step 2: Solve the relaxed master problem with all previously generated cuts. Let (X^*, θ^*) denote an optimal solution to (RMP) and set the lower bound $LB = \theta^* - Q(X^*)$. If $(UB - LB) < \epsilon$, stop; otherwise, go to Step 3.

Step 3: Solve the subproblem at $X = X^*$, denoting Y^* as the optimal solution vector and $v(X^*)$ as the optimal solution value. If $v(X^*) - Q(X^*) < UB$, set $UB = v(X^*) - Q(X^*)$ and update the incumbent solution, i.e., let $(\overline{X}, \overline{Y}) = (X^*, Y^*)$. If $(UB - LB) < \epsilon$, stop with $(\overline{X}, \overline{Y})$ as an ϵ-optimal solution. Otherwise, recover the optimal dual multiplier w^*, add the corresponding Benders cut to the (RMP) formulation, and return to Step 2.

4 Heuristic solution approach for (P)

The Master Problem in our Benders decomposition approach is a mixed 0-1 linear programming problem. As a result, solution via a commercial solver for extremely large problem instances will likely be impractical. We, therefore, propose a heuristic solution approach for very large problem sizes. To obtain a fast heuristic for solving this problem, we temporarily ignore the vehicle capacity constraint when identifying which customers to serve. Then, for a given vector X, the problem is a multi-item capacitated newsvendor problem, which can be solved efficiently. Thus, our solution approach identifies potential values for the vector X, and computes the corresponding optimal stock levels $Y^*(X)$. The resulting solution $(X, Y^*(X))$ is a feasible solution for problem (P).

We begin by expressing problem (P) as a function of the vector X only, when the capacity constraint (2) is ignored. In the absence of this constraint, for a given vector X, we can solve the associated multi-item newsvendor problem and obtain the optimal stock value $y_i^*(X)$ for each item i, given by $y_i^*(X) = F_{X,i}^{-1}(\rho_i)$, where $\rho_i = \frac{e_i - c_i}{e_i}$, and $F_{X,i}^{-1}$ is the inverse cumulative distribution function (CDF) of the demand distribution for item i implied by the vector X. Since the demands are normally distributed, we can write the stock levels as $y_i(X) = \sum_j \mu_{ij} x_j + z(\rho_i)\sqrt{\sum_j \sigma_{ij}^2 x_j}$, where $z(\rho_i) = \Phi^{-1}(\rho_i)$ is the standard normal variate value associated with the fractile ρ_i, and Φ^{-1} is the inverse CDF of the standard unit normal distribution. In addition, we define $\Lambda_i(y_i(X))$ as the loss function for a given stock level $y_i(X)$ and customer assignment vector X, i.e., $\Lambda_i(y_i(X)) = \int_{y_i(X)}^{\infty}(D_i - y_i(X)) f_i(D_i) dD_i = E[(\sum_j d_{ij} x_j - y_i(X))^+]$. Introducing the standard normal loss function $L(z) = \int_z^{\infty}(u - z)\phi(u)du$, where $\phi(u)$ is the probability density function (pdf) of the standard normal distribution with CDF $\Phi(u)$, we can write the loss function $\Lambda_i(y_i(X))$ in terms of the standard normal loss function, $\Lambda_i(y_i(X)) = \sqrt{\sum_j \sigma_{ij}^2 x_j} L(z(\rho_i))$.

Substituting these expressions for $y_i(X)$ and the loss function into $P(Y, X)$ in (1), with a slight abuse of notation, the expected cost can be written in the following form:

$$P(X) = P(Y^*(X), X) = -\sum_j \Big[\pi_j - \sum_i c_i \mu_{ij}\Big] x_j + \sum_i [c_i z(\rho_i) + e_i L(z(\rho_i))] \sqrt{\sum_j \sigma_{ij}^2 x_j}. \tag{9}$$

Let $K_i = c_i z(\rho_i) + e_i L(z(\rho_i))$ represent the i^{th} coefficient value of the square root terms and denote $r_j = -\pi_j + \sum_i c_i \mu_{ij}$ as the negative of the expected net revenue for serving customer j. We then need to solve the following problem in the case of normally distributed demands:

$$\text{(RP)} \qquad \min \sum_j r_j x_j + \sum_i K_i \sqrt{\sum_j \sigma_{ij}^2 x_j}$$
$$\text{subject to:} \quad x_j \in \{0, 1\}, \quad j = 1, \dots, n.$$

Note that we can, without loss of optimality, set $x_j = 0$ for any j such that $r_j \geq 0$, i.e., any j such that $c_{ij}\mu_{ij} \geq \pi_j$ (which implies that the expected customer cost outweighs the customer's associated expected revenue). This relaxed problem RP, obtained by dropping the capacity constraint and expressing Y in terms of X has some special properties:

Proposition 1 *$K_i = c_i z(\rho_i) + e_i L(z(\rho_i))$ is a nonnegative constant under our assumption that $e_i > c_i > 0$ for all i.*

Proof For ease of exposition, we represent the standard normal variate value $z(\rho_i)$ using z. To show $K_i \geq 0$ when assuming that $e_i > c_i > 0$, we only need to show $z + L(z) \geq 0$, because the loss function $L(z)$ is nonnegative for any z. Using $L(z) = \phi(z) - z(1 - \Phi(z))$ we have

$$\begin{aligned} z + L(z) &= \phi(z) + z\Phi(z) \\ &= \phi(-z) - (-z)(1 - \Phi(-z)) \\ &= L(-z) \\ &\geq 0. \end{aligned}$$

The second equality follows from the symmetry of the normal distribution, which completes the proof. □

Proposition 2 *The objective function of RP is concave, and thus for the continuous relaxation of RP obtained by replacing each $x_j \in \{0, 1\}$ with $0 \leq x_j \leq 1$, the linear relaxation of RP is a concave minimization problem with an optimal solution at one of the integral extreme points of $[0, 1]^n$.*

Proof See Geunes, Shen, and Romeijn (2004). □

Proposition 3 *The objective function of RP is a submodular function. Minimizing a rational submodular function is solvable by strongly polynomial combinatorial algorithms.*

Proof Properties B.1 through B.4 and Lemma B.1 of Shen et al. (2003) imply that $P(X)$ is a submodular function. □

As discussed in Shen et al. (2003), Grotschel et al. (1981) showed that a submodular function can be minimized in polynomial time. Strongly polynomial time algorithms are available for this problem class as a result of work by Iwata et al. (2000) and Schrijver (2000).

We are interested in exploiting the particular special structure of our problem in order to obtain a fast heuristic approach for problem (P). We first discuss a solution method for a special case of our problem. This special case arises when either $\sigma_{ij} = \sigma_j$, for all j (the equal-item-variance case) or when only one type of item $m = 1$ is considered. Because these special cases are mathematically equivalent, we illustrate the formulation for the former case below.

$$\text{(RPS)} \qquad \min \sum_j r_j x_j + \sum_i K_i \sqrt{\sum_j \sigma_j^2 x_j}$$

$$\text{subject to:} \quad x_j \in \{0, 1\}, \quad j = 1, \dots, n.$$

We can solve this special case using a simple sorting scheme as in Taaffe et al. (2008). We first sort customers in nonincreasing order of the ratio of expected net revenue to the uncertainty in that customer's demand. This results in indexing customers such that

$$-\frac{r_1}{\sigma_1^2} \geq -\frac{r_2}{\sigma_2^2} \geq \cdots \geq -\frac{r_n}{\sigma_n^2}. \tag{10}$$

Proposition 4 *After indexing customers in nonincreasing order of the above ratio, an optimal solution to (RPS) exists such that if $x_k^* = 1$ for some $k \in \{1, \dots, n\}$, then $x_l^* = 1$ for all $l \in \{1, 2, \dots, k-1\}$.*

Proof See Shen et al. (2003). □

For the general case with unequal item variances, we cannot show that a sorting scheme as in (10) is directly available. We can, however, utilize the insight from (10) to arrive at a heuristic ranking scheme for customers. Since we would like to capture the tradeoff between revenues and variance-related cost, we replace the denominator with a weighted-average variance across items for each customer. That is, we use a weighted-average variance value, where customer j's product i variance is weighted by a factor γ_{ij}. We then define $\overline{\sigma}_j^2(\gamma^j) = \frac{\sum_i \gamma_{ij}\sigma_{ij}^2}{\sum_i \gamma_{ij}}$ (where γ^j is an m-vector of γ_{ij} values). Our heuristic approach is motivated by the fact that there exist weight values γ_{ij} for all i, j such that, after indexing items in nonincreasing order of $-\frac{r_j}{\overline{\sigma}_j^2(\gamma^j)}$, an optimal selection of customers exists of the form defined in Proposition 4 (this follows because any one of the $n!$ possible ordered vectors of customer indices can be obtained through an appropriate choice of γ_{ij} values). One option is to set $\gamma_{ij} = \mu_{ij}$, in which case each variance value is weighted by the corresponding expected demand value. Another option is to set $\gamma_{ij} = K_i$, weighting the variance by the corresponding cost coefficient. To account for the influence of both of the above two factors, we might also set $\gamma_{ij} = \mu_{ij} K_i$. Alternatively, when $\gamma_{ij} = \gamma_j$ for all i, we obtain the arithmetic average of variance, i.e., $\overline{\sigma}_j^2 = \frac{\sum_i \sigma_{ij}^2}{m}$. This approach, therefore, provides a family of heuristic solutions, with each member of the family defined by the choice of γ_{ij} values.

Our heuristic approach therefore sorts customers in nonincreasing order of $-\frac{r_j}{\sigma_j^2(\gamma^j)}$ for some choice of each of the n vectors γ^j, and considers each solution containing customers $\{1, \dots, k\}$ for $k = 1, \dots, n$ (in our computational tests, we consider the cases with $\gamma_{ij} = \mu_{ij}$, $\gamma_{ij} = K_i$ and $\gamma_{ij} = \mu_{ij} K_i$, as well as the simple arithmetic average case). We summarize our heuristic solution algorithm as follows.

Step 1: Choose a value for each of the n values of m-vectors γ^j. Compute the ratio value $R_j = -\frac{r_j}{\sigma_j^2(\gamma^j)}$ for each customer j. Re-index all customers in nonincreasing order of R_j values. Define each of the n assignment vectors $\{1, \dots, k\}$ for $k = 1, \dots, n$; that is, assignment X^j is such that $x_1 = x_2 = \cdots = x_j = 1$, $x_{j+1} = \cdots = x_n = 0$. Set $p = 1$.

Step 2: Solve the multi-item constrained news vendor problem with the assignment vector X^p to determine $Y_p^*(X^p)$ and record the optimal objective value $v(X^p) - Q(X^p)$.

Step 3: Let $p = p + 1$; if $v(X^p) - Q(X^p) < v(X^{p-1}) - Q(X^{p-1})$ and $p \le n$, return to Step 2. Otherwise, let $j^* = p - 1$; the heuristic solution consists of the assignment vector X^{j^*} and stock levels $Y^*(X^{j^*})$ with objective function value $v(X^{j^*}) - Q(X^{j^*})$.

Note that the heuristic solution is integral and feasible. In addition, in Step 3, instead of enumerating the n assignment vectors $\{1, \dots, k\}$ for $k = 1, \dots, n$, we stop once the objective value begins increasing in order to speed up the heuristic. We did this because, after testing instances with the data provided in the next section, we observed that the solutions obtained using this stopping criterion were the same as those obtained after enumerating and comparing all n assignments.

5 Computational results

In this section, we present computational results for our generalized Benders decomposition algorithm and heuristic approach for solving the stochastic resource allocation problem (P). We will demonstrate the benefits of our solution approach when compared to three commercial nonlinear integer solvers. Another goal is to analyze the effect of different parameters on the results, such as the total number of served customers and the number of master and subproblem iterations (which also corresponds to the number of Benders cuts in the relaxed master problem). We implemented our algorithm in C++, with the 0-1 linear integer relaxed master problem solved using ILOG's CPLEX 12.1 solver with Concert Technology. We performed all tests on a computer with an Intel® Pentium 4 CPU, 3.4 GHz processor with 1.99 GB of RAM. In all of our experiments, we used a relative optimality tolerance of 10^{-5}. In order to avoid a "divide by zero" error, both in our algorithm and in GAMS, we added a constraint to ensure $\sum_{j=1}^n x_j \ge 1$. Then, if the optimal objective value for the minimization problem is larger than 0, we know it is better not to deliver items to any customers since $X = 0$ and $Y = 0$ are always feasible to the problem at a total cost of 0.

Table 1 summarizes the common data used in our computational study. For each problem instance, the demand, revenue and all cost data were generated from uniform distributions. We let $U(\ell, u)$ denote the continuous uniform distribution with lower bound ℓ and upper bound u.

To benchmark the performance of our algorithm against LINDOGlobal, SBB, and Coin-Bonmin, we tested our solution method for 16 problem sets and computed the running time and optimality performance (these 16 problem sets use vehicle capacity level 3 and customer revenue level 2 shown in Table 2). Each of these problem sets is characterized by a unique combination of the number of customers and the number of items, (n, m), where we considered $n \in \{5, 10, 30, 50\}$ and $m \in \{3, 10, 20, 50\}$. For each combination of (n, m)

Table 1 Parameter distributions used in computational tests

Resource Data			
Unit Shortage Cost, $\hat{e}_i$	Unit Salvage Value, g_i	Unit Handling Cost, $\hat{c}_i$	Capacity, V
$U(2.5, 3.6)$	$U(0.1, 0.5)$	$U(1, 1.5)$	$U(5mn, 80mn)$
Customer Data			
Expected Demand, μ_{ij}	Standard Deviation, σ_{ij}	Unit item size, s_i	Fixed Revenue, $\hat{\pi}_j$
$U(50, 100)$	$U(14, 50)$	$U(0.5, 1)$	$U(100m, 220m)$

Table 2 Levels used for analysis of the impact of capacity level V and customer revenue value $\hat{\pi}_j$

Level	1	2	3	4
Capacity distribution	$U(5mn, 10mn)$	$U(10mn, 20mn)$	$U(20mn, 40mn)$	$U(40mn, 80mn)$
Revenue distribution	$U(100m, 140m)$	$U(140m, 180m)$	$U(180m, 220m)$	$U(220m, 260m)$

values, we tested 10 randomly generated problem instances, for a total of 160 test cases. For comparison purposes, we also solved each problem instance using GAMS/LINDOGlobal, GAMS/SBB and GAMS/CoinBonmin, three commercial integer nonlinear solvers guaranteeing eventual convergence to globally optimal solutions for general nonlinear problems with continuous and/or discrete variables (assuming memory is not exhausted before convergence). We set the relative optimality tolerance to 10^{-5}, the iteration limit to 200,000, and the time limit to 5,000 seconds in GAMS.

In addition to the initial set of 160 problem instances described above, we also considered four levels of resource capacity and customer-specific unit revenue, respectively, (shown in Table 2) in order to study the impact of different values of these two important parameters. To this end, we will consider an additional 16 problem sets corresponding to each pair of distributions used for these two parameters, with the number of customers fixed at 10 and the number of items fixed at 3. We tested 10 randomly generated problem instances for each of these problem sets, which constitutes another 160 test cases with $n = 10$ and $m = 3$. For this latter set of 160 test problems, which were used to evaluate the impacts of capacity and revenue levels, we used our generalized Benders decomposition approach to obtain optimal solutions (although we did not compare the results of these cases with those using any of the GAMS solvers). We therefore tested a total of 320 problem instances, of which 160 cases were solved using both our algorithm and the three different GAMS solvers, and an additional 160 test problems were solved using our algorithm only.

GAMS modeling

In order to solve the full problem (P) using GAMS solvers, we needed to encode the equation for $E[(D_i - y_i)^+]$ for each i, where $D_i \sim N(\tilde{U}_i, \Theta_i^2)$. For ease of explanation, we suppress the index i in describing how we did this. We used the *errorf*(x) function (integral of the standard normal distribution from $-\infty$ to x) in GAMS modeling, with the equation

$$E[(D - y)^+] = \Theta L(z),$$

where $z = \frac{y-\tilde{U}}{\Theta}$, and $\tilde{U}$ and Θ are the mean and standard deviation of D, respectively (in terms of our model, $\tilde{U}_i = \sum_{j=1}^{n} \mu_{ij} x_j$ and $\Theta_i^2 = \sum_{j=1}^{n} \sigma_{ij}^2 x_j^2$). The standard normal loss function, $L(z)$, is given by

$$L(z) = \int_z^{\infty} (u - z)\phi(u)du = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} - z(1 - \Phi(z)),$$

where $\phi(.)$ is the pdf of the standard normal distribution and $\Phi(.)$ is the associated CDF. In GAMS, we can express $\Phi(.)$ using $\Phi(z) = errorf(z)$.

Comparison with GAMS solvers

To compare our algorithm with the commercial solvers, we consider running time and optimality performance. As discussed previously, we first fixed the resource capacity and customer revenue levels to levels 3 and 2, respectively, and then tested the 16 different problem sets and studied the results. The reason for choosing capacity level 3 and revenue level 2 is that our initial computational testing showed that these settings led to a good balance between revenues and overflow costs, and the resulting problems were among the more computationally intensive problems we tested. The results of our tests, averaged over the 10 random problem instances for each given set of parameter levels, are presented in Table 3. In the table, GBD represents our generalized Benders decomposition approach. The table shows that our approach is generally at least 30 times faster than GAMS/LINDOGlobal, and 2 times faster than GAMS/SBB and GAMS/CoinBonmin. All of the problems we tested were solved within 17 seconds using our algorithm. However, GAMS/LINDOGlobal takes around 30 seconds for smaller size problems, while the majority of problems were solved in 2 to 10 minutes, with the largest size problem requiring nearly 1 hour.

Table 3 Computational test results for average running time (second) and performance

Data Set		GAMS MINLP Solvers				Our Approach
		LINDOGlobal		CoinBonmin	SBB	GBD
m	*n*	Time	PR	Time*	Time*	Time*
3	5	21.20	100.00%	1.42	4.15	0.53
	10	30.20	100.00%	2.16	6.49	0.75
	30	86.78	100.00%	3.76	9.66	0.81
	50	365.25	100.00%	6.09	17.74	1.35
10	5	28.89	100.00%	2.47	5.85	1.46
	10	46.75	100.00%	2.65	6.57	1.79
	30	394.75	99.91%	7.05	18.47	3.55
	50	697.25	99.96%	9.13	20.94	4.32
20	5	30.90	100.00%	2.98	5.97	2.21
	10	74.75	100.00%	3.48	9.36	3.20
	30	440.25	99.63%	9.12	19.77	6.67
	50	797.29	99.68%	16.39	21.38	7.11
50	5	166.00	99.97%	3.79	6.34	3.57
	10	314.44	87.16%	5.30	9.89	9.10
	30	547.33	98.95%	23.49	31.17	16.70
	50	2916.88	62.14%	45.45	30.73	14.43

*PR = 100.00%

To provide a benchmark for the performance of our algorithm, we use the Performance Ratio (PR) as an index, which corresponds to the average solution value as a percentage of the optimal solution value. Our algorithm, SBB and CoinBonmin found an optimal solution for all problems, while LINDOGlobal works well only for smaller size problems (for larger size problems, LINDOGlobal does not perform quite as well). Much of the time it stops with only locally optimal solutions as a result of the iteration limit, which accounts for simplex iterations, barrier iterations, nonlinear iterations and box iterations. For the problem with $(m,n) = (50, 50)$, for 10 randomly generated instances, three of the cases could not be solved to optimality within the time limit of 5,000 seconds, and in one case it was not able to return any feasible solutions within the 5,000 seconds. The average PR was as low as 62.14% for GAMS/LINDOGlobal.

Based on the above comparison, we found that our algorithm significantly outperformed GAMS MINLP solvers LINDOGlobal, SBB and CoinBonmin across the 160 randomly generated problem instances under vehicle capacity level 3 and customer revenue level 2.

Parameter analysis

In Table 4, we show how the different levels of two important parameters (capacity and revenues) affected the results for the case of $(m,n) = (3, 10)$. The average value (across the 10 randomly generated instances) of expected profit is shown in column 6 of the table. The first column corresponds to the four levels of the resource capacity V and the second column corresponds to levels of customer-specific revenue $\hat{\pi}_j$. The average computing time is shown in the third column, and the average number of iterations is shown under the label "Iterations" in column 4 (representing the number of master/subproblem iterations), which we will discuss in more detail later in this section. The column labeled "# Customers" is the average number of customers served by the vehicle in the optimal solution, or equivalently $\sum_{j=1}^{n} x_j$.

Not surprisingly, the number of master/subproblem iterations is closely related to the required CPU time. The greater the number of iterations, the greater the number of Benders

Table 4 Computational test results for various levels of capacity and revenue

V	$\hat{\pi}_j$	Time	Iterations	# Customers	Exp Profit
1	1	1.08	9.70	1.50	91.21
	2	0.74	6.90	2.00	296.37
	3	0.48	4.70	3.30	605.65
	4	0.24	2.10	6.80	1159.08
2	1	0.87	7.80	2.80	205.76
	2	0.88	7.80	3.50	607.39
	3	0.71	6.75	4.88	1081.97
	4	0.32	3.00	7.88	1835.30
3	1	1.58	13.50	5.13	386.28
	2	1.03	8.50	5.88	1041.90
	3	0.73	6.50	6.63	1647.46
	4	0.20	1.56	9.78	2981.94
4	1	0.85	7.50	8.50	515.38
	2	0.19	1.38	9.88	1764.59
	3	0.21	1.75	9.88	2763.69
	4	0.19	1.30	9.90	3907.31

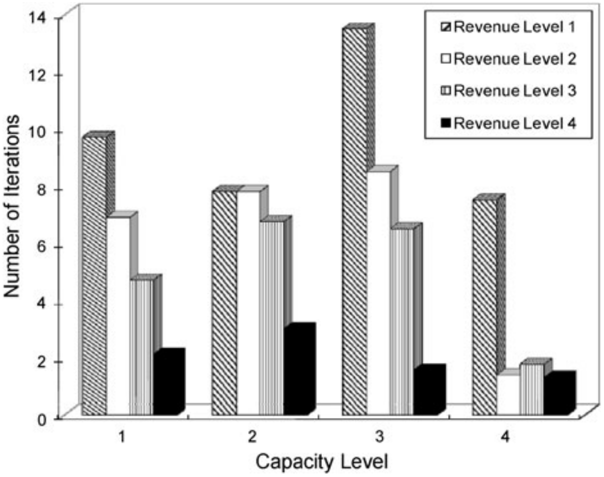


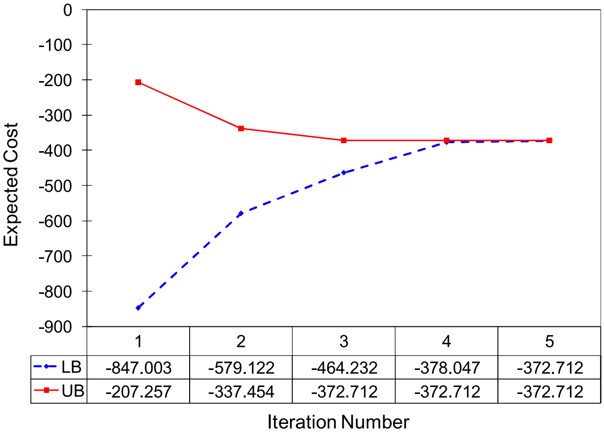
Fig. 1 Number of iterations required for solving problems with different parameter levels

cuts added to the relaxed master problem. This implies a greater number of CPLEX solver calls, and a greater number of lower and upper bounds generated by the Lagrangian relaxation, both of which increase the required running time. The impact of different parameter levels on the running time is illustrated in Fig. 1, again for the case of $(m, n) = (3, 10)$.

In Fig. 1, each bar in the chart corresponds to a customer-specific revenue level, and the results are grouped by vehicle capacity level. The vertical axis shows the average number of iterations required. We can see that for each vehicle capacity level, the heights of the four bars in a row are decreasing as the revenue level increases from level 1 to 4 (recall that the average customer revenue increases in the level number). When all customer revenues increase (all else being equal), we are more likely to assign a higher number of customers to a vehicle, and there will be relatively fewer attractive choices for demand allocation, thus resulting in fewer iterations. At the extreme, for example, when all customers have very high revenues relative to costs, an optimal solution would assign all customers to the vehicle and solve the corresponding multi-item constrained newsvendor problem. As the figure shows, we cannot establish a definite pattern as a function of the capacity level for a given revenue level (the variation in the average number of iterations is reasonably small, and can at least partially be attributed to random variation among problem instances). As we noted previously, the more computationally difficult problems occur when the revenues are closely matched to the overflow costs. As the capacity level increases (in particular at capacity level 4), the expected overflow costs tend to decrease for a given number of customers, leading to less computationally intensive problem instances.

Referring again to Table 4, under each capacity level, the number of customers served and the value of the expected profit both increase as the customer revenue levels increase; similarly, for each revenue level, when the vehicle capacity increases, the number of customers served increases and the expected profit increases. These effects are all quite intuitive and are in accordance with what we would expect in real-world practice. We note that most of the problems were solved with an active vehicle capacity constraint; in the last of the capacity settings (V at level 4), the vehicle capacity constraint was redundant for about half of the instances, which is responsible for the decrease in the number of required iterations at this capacity level.

In order to illustrate the convergence of our generalized Benders decomposition approach for our problem, we illustrate an instance solved within five iterations with $m = 3$, $n = 10$,



	1	2	3	4	5
LB	-847.003	-579.122	-464.232	-378.047	-372.712
UB	-207.257	-337.454	-372.712	-372.712	-372.712

Fig. 2 GBD convergence illustration when $m = 3, n = 10$, V is at level 1, and $\hat{\pi}_j$ is at level 2

V at level 1 and $\hat{\pi}_j$ at level 2, shown in Fig. 2. We can see that the upper bound (UB) converges quickly at the beginning, which we found was the case in general. So the number of iterations required to improve the lower bound (LB) is an important factor with respect to running time.

To summarize, our numerical results showed that our generalized Benders decomposition algorithm can solve the multi-item, multi-customer resource allocation problem much faster than well-known benchmark commercial nonlinear solvers (LINDOGlobal, SBB and Coin-Bonmin), and typically solved the problems we tested within 17 seconds. The optimality performance comparison and parameter analysis provided further evidence of the efficiency and effectiveness of our algorithm.

Heuristic performance

To test the performance of our heuristic approach described in the previous section, we consider four variants of our heuristic approach, one with the simple arithmetic average case, the second with $\gamma_{ij} = \mu_{ij}$, which we refer to as the demand weighted-average case, the third with $\gamma_{ij} = K_i$ which we refer to as the cost weighted-average case, and the fourth one using $\gamma_{ij} = \mu_{ij} K_i$. We denote these by HS, HD, HC and HDC, respectively, in Table 5. For comparison purposes, we use the same 160 test instances shown in Table 3. For these heuristics, we present the resulting optimality gap in Table 3, equal to one minus the Performance Ratio (PR).

From Table 5, we observe that the computational requirements are substantially lower for our heuristic, resulting in computing times around 200–1000 times faster than LINDOGlobal and 4–16 times faster than our Benders decomposition method. The only time-consuming element of the heuristic is the computation of the expected value term in the objective function $v(X)$ for multiple assignment vectors X^k, $k \in \{1, 2, \dots, n\}$, while in the Benders decomposition algorithm, each iteration requires the computation of an expected value term and the solution of a linear integer programming problem.

The required computation time for HC and HDC is a little longer than for HS and HD as a result of the need to compute the cost parameter K_i. The optimality gap on average is 2.61% for HS, 2.64% for HD, 2.33% for HC, and 2.31% for HDC. However, the solutions obtained by HS, HD, HC, and HDC are typically different, and so we recommend running all of these methods and choosing the best solution among them, since the running time is very fast when compared to the exact algorithm. As the table shows, the performance of the

Table 5 Computational test results for average running time (second) and performance of the heuristic approach

Data Set		HS		HD		HC		HDC	
m	*n*	Time	GAP	Time	GAP	Time	GAP	Time	GAP
3	5	0.09	1.81%	0.09	1.81%	0.14	1.82%	0.14	1.81%
	10	0.15	4.19%	0.15	3.60%	0.19	3.11%	0.18	2.10%
	30	0.28	5.50%	0.28	5.30%	0.30	5.12%	0.30	5.15%
	50	0.29	6.00%	0.30	6.00%	0.35	3.78%	0.32	3.96%
10	5	0.14	3.20%	0.14	3.20%	0.16	3.20%	0.17	2.25%
	10	0.20	2.75%	0.20	3.80%	0.22	2.65%	0.22	2.31%
	30	0.43	2.98%	0.42	3.20%	0.46	3.36%	0.47	3.62%
	50	0.58	2.59%	0.59	2.40%	0.62	3.34%	0.62	3.00%
20	5	0.20	1.66%	0.20	1.66%	0.22	1.27%	0.22	1.65%
	10	0.28	1.33%	0.28	1.20%	0.31	1.01%	0.31	1.01%
	30	0.52	3.24%	0.51	3.15%	0.56	3.50%	0.55	3.21%
	50	0.91	1.75%	0.90	2.36%	1.09	1.68%	1.12	2.39%
50	5	0.40	1.22%	0.40	1.22%	0.43	0.32%	0.43	1.22%
	10	0.55	1.17%	0.56	1.07%	0.61	0.86%	0.60	1.07%
	30	1.20	1.25%	1.18	1.12%	1.40	1.20%	1.38	1.20%
	50	2.20	1.12%	2.22	1.18%	2.34	1.07%	2.36	1.04%

heuristic approach improves as the number of items increases, as has often been shown to be the case for knapsack problem heuristics in the literature. Because the heuristic approach is intended for larger problem instances for which exact approaches are impractical, this is a promising result for handling problems of large size in reasonable computing times.

6 Conclusion

This study considered a stochastic resource allocation problem with normally distributed demands for multiple items and a resource capacity constraint. We simultaneously considered the demand allocation and vehicle stocking problem for multiple items. This problem has a broad set of applications in practice, although we focused on the sales visit context. Moreover, our stochastic resource allocation problem arises as a subproblem for solving a multiple resource allocation problem, which serves as one of our future research directions.

In this study, we provided an exact solution method and a heuristic approach for solving this problem. The exact method we proposed uses Benders decomposition and transforms the nonlinear mixed integer problem into two smaller ones, one of which is a nonlinear continuous problem that can be solved using Lagrangian relaxation relatively easily, and the other of which is a 0-1 linear integer problem which can be handled effectively using the CPLEX solver. After solving these two problems iteratively, we converge to a globally optimal solution because of the problem's joint convexity properties. The heuristic approach expresses the stock level vector Y in terms of the binary X variables and, by sorting customers in nonincreasing order of a ratio of expected net revenue to variance (motivated by the solution for a single-item selective newsvendor problem), we obtain near-optimal solutions (within 2.47% of optimality on average) typically within 1 second.

Based on our numerical study, our exact algorithm has proven to be quite efficient when compared with three advanced commercial nonlinear solvers, GAMS/LINDOGlobal, GAMS/SBB, and GAMS/CoinBonmin. Our heuristic approach performed very well for larger problem sizes, which constitute the class of problems for which a heuristic approach would likely be most beneficial in practice.

Acknowledgement This work was partially supported by NSF Grant #CMMI-0927930.

Appendix 1: Solving the multi-item newsvendor subproblem

Here we discuss the solution of the multi-item constrained newsvendor problem:

$$\begin{aligned} \text{(SP)} \qquad & \min\ H(X,Y)+G_1(Y) \\ & \text{subject to:}\quad G_2(Y)\le 0, \\ & \qquad\qquad\qquad Y\ge 0. \end{aligned}$$

The above Subproblem can be written as

$$\begin{aligned} \text{(SP)} \qquad & \min \sum_{i=1}^{m} e_i E\left[\left(\sum_{j=1}^{n} d_{ij}x_j - y_i\right)^{+}\right] + \sum_{i=1}^{m} c_i y_i \\ & \text{subject to:}\quad \sum_{i=1}^{m} s_i y_i \le V, \\ & \qquad\qquad\qquad Y\ge 0. \end{aligned}$$

For a given X^k, we use the Lagrange multiplier approach to obtain the optimal solution Y^* of this convex program (Hadley and Whitin 1963). We introduce a Lagrange multiplier $w\ge 0$ and form the Lagrangian function $Lg(X^k,Y)$ (note that in terms of our prior notation, $Lg(X^k,Y)$ is equivalent to the Lagrangian relaxation objective of $v(X^k)$ in (8)):

$$Lg(X^k,Y)=\sum_{i=1}^{m} e_i E\left[\left(\sum_{j=1}^{n} d_{ij}x_j^k - y_i\right)^{+}\right] + \sum_{i=1}^{m} c_i y_i + w\left(\sum_{i=1}^{m} s_i y_i - V\right)$$

Then, for a given w, using the necessary and sufficient first-order conditions, the y_i^* values are determined by solving the set of equations

$$\frac{\partial Lg(X^k,Y)}{\partial y_i}=e_i(F_i(y_i)-1)+c_i+ws_i=0,\quad i=1,\dots,m,$$

or

$$F_i(y_i)=\frac{e_i-c_i-ws_i}{e_i},\quad i=1,\dots,m,$$

where $F_i(.)$ is the cumulative distribution function of demand $D_i=\sum_{j=1}^{n} d_{ij}x_j^k$ for the given X^k. Note that we can set a lower bound on w of zero, i.e., $LB_w=0$, and an upper bound of $UB_w=\max_i(\frac{e_i-c_i}{s_i})$ without loss of optimality.

An effective computational procedure for determining the value of w that produces a saddlepoint solution would begin with a value of w between the lower and upper bounds, e.g., $\overline{w} = \frac{LB_w + UB_w}{2}$, and then compute the values of y_i, $i = 1, \ldots, m$, from the above equations; if $y_i < 0$, or $(e_i - c_i - ws_i) < 0$ for some i, let $y_i = 0$. Then, compute $\hat{V} = \sum_{i=1}^{m} s_i y_i$. By using binary search, if $\hat{V} > V$, we select a larger value of w; that is, let $LB_w = \overline{w}$. If $\hat{V} < V$, let $UB_w = \overline{w}$. Repeat this process as needed; normally the procedure is stopped when $|\hat{V} - V| < \epsilon$ for some tolerance $\epsilon > 0$. We will then gain the optimal values w^* and Y^* for the given X^k.

Appendix 2: Obtaining the subgradient of $v(.)$ at X^k

By Theorem 1,

$$\delta(X^k) = \frac{\partial}{\partial X}\left\{H(X, Y^k) + G_1(Y^k) + u^k G_2(Y^k)\right\}_{X=X^k} = \frac{\partial}{\partial X}\left\{H(X, Y^k)\right\}_{X=X^k}.$$

Under the assumption that $d_{ij} \sim N(\mu_{ij}, \sigma_{ij}^2)$, we let $D_i = \sum_{j=1}^{n} d_{ij} x_j$; thus $E(D_i) = \sum_{j=1}^{n} \mu_{ij} x_j = \tilde{U}_i(X)$, $Var(D_i) = \sum_{j=1}^{n} \sigma_{ij}^2 x_j^2 = \sigma_i^2(X)$, and $D_i \sim N(\tilde{U}_i(X), \sigma_i^2(X))$. We represent $H(X, Y^k)$ using the following equation:

$$\begin{aligned}
H(X, Y^k) &= \sum_{i=1}^{m} e_i E\left[\left(\sum_{j=1}^{n} d_{ij} x_j - y_i^k\right)^+\right] \\
&= \sum_{i=1}^{m} e_i \frac{\sigma_i(X)}{\sqrt{2\pi}} \int_{y_i^k}^{\infty} \frac{(D_i - y_i^k)}{\sigma_i^2(X)} \exp\left\{-\frac{(D_i - \tilde{U}_i(X))^2}{2\sigma_i^2(X)}\right\} d(D_i) \\
&= \sum_{i=1}^{m} e_i \frac{\sigma_i(X)}{\sqrt{2\pi}} \left[-\exp\left\{-\frac{(D_i - \tilde{U}_i(X))^2}{2\sigma_i^2(X)}\right\}\Bigg|_{D_i = y_i^k}^{\infty} \right. \\
&\quad \left. + \frac{(\tilde{U}_i(X) - y_i^k)}{\sigma_i^2(X)} \int_{y_i^k}^{\infty} \exp\left\{-\frac{(D_i - \tilde{U}_i(X))^2}{2\sigma_i^2(X)}\right\} d(D_i)\right] \\
&= \sum_{i=1}^{m} e_i \frac{\sigma_i(X)}{\sqrt{2\pi}} \left[\exp\left\{-\frac{(y_i^k - \tilde{U}_i(X))^2}{2\sigma_i^2(X)}\right\} \right. \\
&\quad \left. + \frac{(\tilde{U}_i(X) - y_i^k)}{\sigma_i(X)} \sqrt{2} \int_{\frac{y_i^k - \tilde{U}_i(X)}{\sqrt{2}\sigma_i(X)}}^{\infty} \exp\left\{-z^2\right\} dt\right]
\end{aligned} \tag{11}$$

Then,

$$\begin{aligned}
\frac{\partial}{\partial X}\{H(X, Y^k)\}_{X=X^k} &= \sum_{i=1}^{m} e_i \left[\frac{1}{\sqrt{2\pi}} \frac{\partial}{\partial X}\left\{\sigma_i(X) \exp\left\{-\frac{(y_i^k - \tilde{U}_i(X))^2}{2\sigma_i^2(X)}\right\}\right\} \right. \\
&\quad \left. + \frac{1}{\sqrt{\pi}} \frac{\partial}{\partial X}\left\{(\tilde{U}_i(X) - y_i^k) \int_{\frac{y_i^k - \tilde{U}_i(X)}{\sqrt{2}\sigma_i(X)}}^{\infty} \exp\{-z^2\} dt\right\}\right]_{X=X^k}
\end{aligned} \tag{12}$$

Given $X = X^k$, define $A_i = \exp\{-\frac{(y_i^k - \tilde{U}_i(X^k))^2}{2\sigma_i^2(X^k)}\}$, $\tilde{U}_i = \tilde{U}_i(X^k)$, and $B_i = \sigma_i(X^k)$, and the first derivative term can be written as:

$$\begin{aligned} T_1^i &= \frac{\partial}{\partial x_j}\left\{\sigma_i(X)\exp\left\{-\frac{(y_i^k - \tilde{U}_i(X))^2}{2\sigma_i^2(X)}\right\}\right\}\Bigg|_{X=X^k} \\ &= \frac{A_i\sigma_{ij}^2 x_j^k}{B_i} - \sqrt{2}A_i(y_i^k - \tilde{U}_i)\left(\frac{y_i^k - \tilde{U}_i(X)}{\sqrt{2}\sigma_i(X)}\right)'\Bigg|_{X=X^k}. \end{aligned} \tag{13}$$

Here $(\frac{y_i^k - \tilde{U}_i(X)}{\sqrt{2}\sigma_i(X)})'|_{X=X^k}$ is the partial derivative of $(\frac{y_i^k - \tilde{U}_i(X)}{\sqrt{2}\sigma_i(X)})$ with respect to x_j at the point $X = X^k$.

Similarly, the second derivative term can be obtained:

$$\begin{aligned} T_2^i &= \frac{\partial}{\partial x_j}\left\{(\tilde{U}_i(X) - y_i^k)\int_{\frac{y_i^k - \tilde{U}_i(X)}{\sqrt{2}\sigma_i(X)}}^{\infty}\exp\{-t^2\}dt\right\}\Bigg|_{X=X^k} \\ &= \mu_{ij}(1 - F_i(y_i^k))\sqrt{\pi} + A_i(y_i^k - \tilde{U}_i)\left(\frac{y_i^k - \tilde{U}_i(X)}{\sqrt{2}\sigma_i(X)}\right)'\Bigg|_{X=X^k}. \end{aligned} \tag{14}$$

Here $F_i(.)$ is the cumulative distribution function of the demand $D_i(X^k) = \sum_{j=1}^n d_{ij}x_j^k \sim N(\tilde{U}_i, B_i^2)$. So

$$\begin{aligned} \frac{\partial}{\partial x_j}\{H(X, Y^k)\}_{X=X^k} &= \sum_{i=1}^m e_i\left[\frac{1}{\sqrt{2\pi}}T_1^i + \frac{1}{\sqrt{\pi}}T_2^i\right] \\ &= \sum_{i=1}^m e_i\left[\frac{A_i\sigma_{ij}^2 x_j^k}{\sqrt{2\pi}B_i} + \mu_{ij}(1 - F_i(y_i^k))\right]. \end{aligned} \tag{15}$$

Note that above equations are based on the assumption $X^k \neq 0$. If $X^k = 0$, then $B_i = 0$ and

$$\frac{\partial}{\partial x_j}\{H(X, Y^k)\}_{X=X^k=0} = \sum_{i=1}^m e_i\mu_{ij}. \tag{16}$$

References

Ağralı, S., & Geunes, J. (2009). A single-resource allocation problem with Poisson resource requirements. *Optimization Letters*, *3*(4), 559–571.

Akçay, H., Li, H., & Xu, S. H. (2007). Greedy algorithm for the general multidimensional knapsack problem. *Annals of Operation Research*, *150*(1), 17–29.

Barnhart, C., & Cohn, A. M. (1998). The stochastic knapsack problem with random weights: a heuristic approach to robust transportation planning. In *Proceedings of Tristan III*, Puerto Rico, 17–23 June.

Benjaafar, S., Elhafsi, M., & Vericourt, F. D. (2004). Demand allocation in multiple-product, multiple-facility, make-to-stock systems. *Management Science*, *50*(10), 1431–1448.

Dean, B. C., Goemans, M. X., & Vondrak, J. (2004). Approximating the stochastic knapsack problem: the benefit of adaptivity. In *Proceedings of the 45th annual IEEE symposium on foundations of computer science*.

Federgruen, A., & Zipkin, P. H. (1984). A combined vehicle routing and inventory allocation problem. *Operations Research*, *32*(5), 1019–1037.

Federgruen, A., Prastacos, G., & Zipkin, P. H. (1986). An allocation and distribution model for perishable products. *Operations Research*, *34*(1), 75–82.

Geoffrion, A. M. (1972). Generalized Benders decomposition. *Journal of Optimization Theory and Applications*, *10*(4), 237–260.

Geromel, J. C., & Belloni, M. R. (1986). Nonlinear programs with complicating variables: theoretical analysis and numerical experience. *IEEE Transactions on Systems, Man, and Cybernetics*, *16*(2), 231–239.

Geunes, J., Shen, Z.-J., & Romeijn, H. E. (2004). Economic ordering decisions with market choice flexibility. *Naval Research Logistics*, *51*(1), 117–136.

Gorman, M. F., & Ahire, S. (2006). A major appliance manufacturer rethinks its inventory policies for service vehicles. *Interfaces*, *36*(5), 407–419.

Grotschel, M. L., Lovasz, L., & Schrijver, A. (1981). The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, *1*, 169–197.

Hadley, G., & Whitin, T. M. (1963). *Analysis of inventory systems*. Prentice-Hall: Englewood Cliffs.

Heyman, D. P., & Sobel, M. J. (1984). *Stochastic models in operations research, vol. 2: stochastic optimization*. New York: McGraw-Hill.

Iwata, S. L., Fleischer, L., & Fujishige, S. (2000). A combinatorial, strongly polynomial-time algorithm for minimizing submodular functions. In *Proceedings of the 32nd annual ACM symposium on theory of computing*, Portland, Oregon (pp. 97–106).

Kellerer, H., Pferschy, U., & Pisinger, D. (2004). *Knapsack problems*. Berlin: Springer.

Kleywegt, A., & Papastavrou, J. D. (2001). The dynamic and stochastic knapsack problem with random sized items. *Operations Research*, *49*(1), 26–41.

Kosuch, S., & Lisser, A. (2010). Upper bounds for the 0-1 stochastic knapsack problem and a B&B algorithm. *Annals of Operation Research*, *176*(1), 77–93.

Merzifonluoğlu, Y., Geunes, J., & Romeijn, H. E. The static stochastic knapsack problem with normally distributed item sizes. *Mathematical Programming* (2011, forthcoming).

Schrijver, A. (2000). A combinatorial algorithm minimizing submodular functions in strongly polynomial time. *Journal of Combinatorial Theory. Series B*, *80*(2), 346–355.

Shen, Z.-J., Coullard, C., & Daskin, M. S. (2003). A joint location-inventory model. *Transportation Science*, *37*(1), 40–55.

Smith, S. A., Chambers, J. C., & Shlifer, E. (1980). Optimal inventories based on job completion rate for repairs requiring multiple items. *Management Science*, *26*(8), 849–852.

Taaffe, K., Geunes, J., & Romeijn, H. E. (2008). Target market selection and marketing effort under uncertainty: the selective newsvendor. *European Journal of Operational Research*, *189*(3), 987–1003.

Teunter, R. H., & Haneveld, W. K. (2002a). Inventory control of service parts in the final phase. *European Journal of Operational Research*, *137*(3), 497–511.

Teunter, R. H., & Haneveld, W. K. (2002b). Inventory control of service parts in the final phase: a central depot and repair kits. *European Journal of Operational Research*, *138*(1), 76–86.

Vasquez, M., & Vimont, Y. (2005). Improved results on the 0-1 multidimensional knapsack problem. *European Journal of Operational Research*, *165*, 70–81.