A Combined Vehicle Routing and Inventory Allocation Problem

AWI FEDERGRUEN and PAUL ZIPKIN

Columbia University, New York, New York

(Received September 1980; accepted June 1983)

We address the combined problem of allocating a scarce resource among several locations, and planning deliveries using a fleet of vehicles. Demands are random, and holding and shortage costs must be considered in the decision along with transportation costs. We show how to extend some of the available methods for the deterministic vehicle routing problem to this case. Computational results using one such adaptation show that the algorithm is fast enough for practical work, and that substantial cost savings can be achieved with this approach.

THIS PAPER addresses the combined problem of allocating a scarce resource available at some central depot among several locations (or "customers"), each experiencing a random demand pattern, while deciding which deliveries are to be made by each of a set of vehicles and in what order.

This problem is an extension of the standard vehicle routing problem (VRP). There, one must design a set of vehicle routes of minimal total cost, leaving from and eventually returning to the depot, while satisfying capacity constraints and meeting customer requirements. These requirements almost universally include exogenously determined, deterministically known delivery sizes.

Most existing resource allocation models, on the other hand, assume a cost function that is smooth as well as separable and additive in the activities. As a consequence, their application to physical distribution problems is confined to situations where all delivery points receive individual deliveries rather than being served in routes.

There are many situations where the vehicle schedules and the delivery sizes are (or should be) determined simultaneously. Such is often the case, for example, when at each location the demand for the resource is random. Here, the deliveries serve to replenish the inventories to levels that appropriately balance inventory carrying and shortage costs, but thereby incur transportation costs as well. This type of problem is the major focus of the paper.

Subject classification: 331 inventory allocation/VRP, 831 vehicle routing/inventory.

Operations Research
Vol. 32, No. 5, September–October 1984

0030-364X/84/3205-1019 $01.25

We specify a mathematical programming model of the problem. While quite complex, the model can be manipulated to induce a convenient partial separation into subproblems. Using this separation property, we show that well-known interchange heuristics for the deterministic VRP can be modified to handle the current problem. Experience with one such algorithm suggests that this approach is effective, both in producing good solutions and in requiring a modest amount of computation.

This separation property can also be exploited to adapt quite different approaches to the VRP. To illustrate this fact, we derive an exact algorithm for the problem using generalized Benders' decomposition. (While it is interesting to know that an exact algorithm exists, we have no evidence concerning its computational performance.)

The decomposition algorithm can be interpreted (in the conventional manner) as using separate calculations for the allocation and routing decisions, but coordinating them appropriately. The somewhat novel approach here is to show that a comparable coordinating mechanism can be built into an efficient heuristic.

The basic VRP described above is sometimes complicated by other factors. Some can be handled with straightforward extensions of our approach, as described in Section 3.

For the sake of precision, we now describe the scenario envisioned in somewhat greater detail: First, the initial inventory (perhaps supply remaining from the previous day) for each location is reported to the depot. This information is used to determine for the following day the allocation of the available product among the locations. The assignment of locations to vehicles and the routes are set at the same time. After the deliveries are made (say at the end of the day) the demands occur, and inventory-carrying and shortage costs are incurred at each location, proportional to the end-of-the-day inventory levels. Thus, our model is a one-period, "myopic" version of the problem. Note, it is possible to choose not to visit some of the locations.

This scenario applies mainly to *internal* distribution problems, since all decisions are made centrally. An example is Magnanti and Golden's [1978] description of deliveries of fuel oil to automotive service stations. Some *external* distribution processes satisfy our assumptions as well. In fact, we were motivated by applications in the industrial gas industry, where the gas producers themselves install tanks at customer locations and determine the replenishment frequency and delivery sizes. Another potential application is the allocation of a perishable product such as blood, where the supply to various locations in a particular region is coordinated by a regional center. The multilocation allocation model of Prastacos [1978], as applied in Brodheim and Prastacos [1979], treats a problem of this type.

The importance of the interrelationships between inventory and transportation planning is easily recognized from the above examples and has recently been discussed by Herron [1979]. To our knowledge, ours is the first attempt to integrate the allocation and routing problems in a single model. More recently, Assad et al. [1982] reported on a simulation study for a deterministic inventory/routing system. In a broader context, our model may be viewed as a contribution to recent efforts to integrate related areas of physical distribution management, which until now have only been treated separately. Such efforts include Laporte and Norbert [1981] and Federgruen and Lageweg [1980] in the context of network design; cf. also Schrage [1981].

Random demands have been treated in the VRP literature, but in a manner quite different from ours. In Golden and Stewart [1978] and Golden and Yee [1979] a *primary error* occurs when any one vehicle is unable to satisfy the demands of the customers on its route. These authors suggest procedures to search for minimal-cost routes, subject to some upper limit on the probability of a primary error, cf. also Stewart and Golden [1982]. Tillman [1969] treats a simplified version of our problem, where in effect demands are realized before deliveries are made, using a very different computational approach. Cook and Russell [1978] have undertaken simulation studies of VRPs with random delivery sizes.

The scenario treated here is perhaps the simplest possible one which accounts for uncertain demands and control over delivery sizes. Having shown that inventory and routing models are not entirely incompatible, we have reason to hope that further research will lead to systematic treatments of more fully dynamic scenarios. For example, the demand at a location may be revealed when a vehicle reaches it. Alternatively, starting inventories may be random at the beginning of the planning period, and actual inventory at each location may be discovered only when a vehicle visits it. In these scenarios, costs may be reduced by a dynamic determination of the allocations, or even the sequence in which locations are visited. In the terminology of stochastic programming, our model can serve as a "here-and-now" approximation to many such systems.

To summarize the remainder of the paper, in Section 1 we state the problem, introduce notation, and discuss the separation property. Section 2 discusses an inventory allocation problem (IA) which must be solved repeatedly during the algorithm. In Section 3 the interchange heuristics are discussed, and in Section 4 we present the generalized Benders' decomposition approach. Section 5 presents numerical results. An Appendix explores some continuity properties of the model.

1. STATEMENT OF THE PROBLEM AND NOTATION

Much of our notation follows that of Fisher and Jaikumar [1978].

Constants

K = number of vehicles
n = number of locations, indexed from 1 to n; index 0 denotes the central depot
b_k = capacity (weight or volume) of vehicle k
c_{ij} = cost of direct travel from location i to location j
$F_i(\cdot)$ = cumulative distribution function of the one period demand in location i, which is assumed to be strictly increasing
h_i^+ = inventory carrying cost (or disposal cost minus salvage value) per unit in location i
h_i^- = shortage cost per unit in location i
β_i = initial inventory at location i
A = total amount of product available at the central depot.

Variables

We define a dummy route $k = 0$ consisting of those locations to which nothing is to be shipped ($b_0 = 0$).

$$y_{ik} = \begin{cases} 1 \text{ if delivery point } i \text{ is assigned to route } k \\ 0 \text{ otherwise} \end{cases}$$

$$x_{ijk} = \begin{cases} 1 \text{ if vehicle } k \text{ travels directly from location } i \text{ to location } j; \\ 0 \text{ otherwise} \end{cases}$$

w_i = amount delivered to location i.

We shall use the notation y^k, y, w, and so forth to denote vectors of variables taken over the suppressed subscripts.

The inventory cost function $q_i(\cdot)$ and its derivative $q_i'(\cdot)$ are given by

$$q_i(w_i) = \int_{\beta_i+w_i}^{\infty} h_i^-(\xi - \beta_i - w_i)dF_i(\xi) + \int_0^{\beta_i+w_i} h_i^+(\beta_i + w_i - \xi)dF_i(\xi),$$

$$q_i'(w_i) = (h_i^+ + h_i^-)F_i(\beta_i + w_i) - h_i^-, \qquad i = 1, \cdots, n.$$

It is straightforward to verify that the $q_i(\cdot)$ are strictly convex and C^1.

With this notation, the problem can be stated as follows:

$$\text{(P): } \min \textstyle\sum_{i,j,k} c_{ij}x_{ijk} + \sum_i q_i(w_i) \tag{1}$$

$$\text{subject to } \textstyle\sum_i w_i y_{ik} \le b_k, \qquad k = 0, \cdots, K \tag{2}$$

$$\textstyle\sum_i w_i \le A; \quad w_i \ge 0, \qquad i = 1, \cdots, n \tag{3}$$

$$\left.\begin{aligned} &\textstyle\sum_{k=1}^K y_{0k} = K \\ &\textstyle\sum_{k=0}^K y_{ik} = 1, \qquad i = 1, \cdots, n \end{aligned}\right\} \tag{4}$$

$$y_{ik} = 0 \text{ or } 1, \qquad j = 0, \cdots, n; \quad k = 0, \cdots, K \tag{5}$$

$$\textstyle\sum_i x_{ijk} = y_{jk}, \qquad j = 0, \cdots, n; \quad k = 1, \cdots, K \tag{6}$$

$$\textstyle\sum_j x_{ijk} = y_{ik}, \qquad i = 0, \cdots, n; \quad k = 1, \cdots, K \tag{7}$$

$$\textstyle\sum_{(i,j)\in S\times S} x_{ijk} \le |S| - 1, \qquad S \subseteq \{1, \cdots, n\}, \quad 2 \le |S| \le n - 1; \quad k = 1, \cdots, K \tag{8}$$

$$x_{ijk} = 0 \text{ or } 1, \qquad i = 0, \cdots, n; \quad j = 0, \cdots, n; \quad k = 0, \cdots, K. \tag{9}$$

Constraint (2) (which is nonlinear) ensures that the load assigned to each vehicle is within its capacity, and constraint (3) guarantees that the total amount shipped is available at the depot. The remaining constraints appear in the deterministic VRP. Constraints (4) and (5) ensure that every delivery point is assigned to a single route (possibly the dummy route 0). Constraints (6)–(9) define a traveling salesman problem (TSP) over the customers assigned to vehicle k.

Observe, (1)–(9) do not include storage capacity limits at the customer locations. The (slight) modifications required to handle this case are described briefly at the end of Section 2.

Unlike the deterministic VRP, our problem is feasible for any vector y satisfying (4) and (5). In this sense, the model has a less intricate combinatorial structure than the VRP, resulting in some simplification of our algorithms, which partly compensates for the added complexities introduced. The ability to adapt delivery sizes and to eliminate a few "inconvenient" locations (or locations with relatively large inventories) enables substantial cost savings, as exhibited in Section 5. Also (in contrast to the deterministic model and to most mixed integer programs), the minimal cost of (P) is continuous in all cost and capacity parameters. (See the Appendix for a proof of this statement. If the F_i are continuous in one or more parameters, e.g. normals, Weibulls or gammas, then this continuity result extends to these parameters as well.) The continuity property is important in planning studies where some of the data may be uncertain. It excludes the possibility that a small change in the data may induce a sudden change in the optimal cost, cf. also the continuity analysis on p. 834 in Geoffrion and Graves [1974] as well as Williams [1973]. (In the deterministic VRP, a slight increase in the delivery size of a customer may require reassignment of several customers or even an additional truck.)

Observe that with y fixed, the problem decomposes into simpler subproblems, namely, an inventory allocation problem (discussed in the next section) and K TSPs, one for each vehicle. This fundamental separation property is the basis of the computational approach of Section 3.

Our algorithm for (1)–(9) requires an initial set of routes (i.e., feasible values for x and y). An obvious procedure would be first to solve the inventory allocation problem obtained by relaxing (2):

$$\min \textstyle\sum_i q_i(w_i) \quad \text{s.t.} \quad \sum_i w_i \le A; \quad w_i \ge 0, \qquad i = 1, \cdots, n. \quad (10)$$

(Zipkin [1980] summarizes solution methods for this problem. Alternatively, instead of solving (10), one could minimize each q_i separately.) Let $\{w_i^*\ i = 1, \cdots, n\}$ denote an optimal solution. Then, use one of the available initialization procedures for the deterministic VRP with the delivery size at location i fixed at w_i^* ($i = 1, \cdots, n$). (The adaptability of the delivery sizes allows for certain simplifications in these initialization procedures, cf. Section 5).

2. THE INVENTORY ALLOCATION PROBLEM

Any specific value of y determines a partition $\{Y_k: k = 0, \cdots, K\}$ of indices $\{1, \cdots, n\}$, where $Y_k = \{i: y_{ik} = 1\}$. Thus Y_k is the set of locations to be serviced by vehicle k according to the assignment y, $k = 0, \cdots, K$. The inventory allocation problem, then, can be written as follows:

$$\text{(IA): } \min \textstyle\sum_{i=1}^{n} q_i(w_i) \qquad (11)$$

$$\text{subject to } \textstyle\sum_{i=1}^{n} w_i \le A \qquad (12)$$

$$\left.\begin{aligned} \textstyle\sum_{i\in Y_k} w_i &\le b_k, \qquad k = 0, \cdots, K \\ w_i &\ge 0, \qquad i = 1, \cdots, n. \end{aligned}\right. \qquad (13)$$

(For actual solution of this problem, of course, the variables w_i for $i \in Y_0$ and the constraint for $k = 0$ can be ignored.)

Let us project this problem onto new variables $W_k = \sum_{i\in Y_k} w_i$. That is, for each $k = 1, \cdots, K$, define

$$\begin{aligned} Q_k(W_k) &= \min \textstyle\sum_{i\in Y_k} q_i(w_i) \\ &\text{subject to } \textstyle\sum_{i\in Y_k} w_i = W_k \\ &\qquad w_i \ge 0, \qquad i \in Y_k. \end{aligned} \qquad (14)$$

The problem (IA) is equivalent in an obvious sense to the following model:

$$\begin{aligned} &\min \textstyle\sum_{k=1}^{K} Q_k(W_k) \\ &\text{subject to } \textstyle\sum_{k=1}^{K} W_k \le A \qquad 0 \le W_k \le b_k, \qquad k = 1, \cdots, K. \end{aligned} \qquad (15)$$

Since each q_i is C^1 and strictly convex, the methods developed in Zipkin for problems of form (14) can be applied. Zipkin also shows that each Q_k is C^1 and strictly convex. Similar methods, therefore, can be used to solve (15) itself, although additional complications arise from the upper bounds on the W-variables and the fact that the Q_k functions are defined implicitly. Federgruen and Zipkin [1983] present efficient methods to solve (15), and hence (IA). For now, suffice it to say that

(a) The procedure is finite;
(b) Problem (14) must be solved explicitly for $k = 1, \cdots, K$ at $W_k = b_k$, but never for any other value of W_k;
(c) The method is well-suited for recovery of optimality when y is changed, especially when it changes slightly. The latter case will require very few major iterations (and often none);
(d) The procedure yields optimal dual variables (Lagrange multipliers) for problem (IA).

We now show how to estimate the change in optimal cost in problem (IA) resulting from switches of locations between subsets. Such estimates are used in the algorithm of Section 3.

Suppose (IA) is solved, yielding a (unique) optimal solution $\bar{w}$ and a (not necessarily unique) optimal dual solution $(\bar{\rho}, \bar{\nu})$ where $\bar{\nu} = (\bar{\nu}_k)_{k=0}^{K}$. For some k_1 and k_2, suppose we revise y so that some elements of Y_{k_1} and Y_{k_2} are traded. Specifically, we desire an estimate of the change in total inventory costs when switching the set of locations $J_1 \subseteq Y_{k_1}$ from k_1 to k_2, and the set of locations $J_2 \subseteq Y_{k_2}$ from k_2 to k_1. For $i \in J_1$, let $\bar{w}_i'$ be the solution to $q_i'(w_i) = \bar{\rho} + \bar{\nu}_{k_2}$ if $q_i'(0) < \bar{\rho} + \bar{\nu}_{k_2}$, and let it equal zero otherwise. Similarly, let $\bar{w}_i'$ solve $q_i'(w_i) = \bar{\rho} + \bar{\nu}_{k_1}$ or $\bar{w}_i' = 0$ for $i \in J_2$. Then $(\bar{\rho}, \bar{\nu})$ is the optimal dual solution to the new problem (IA) with modified righthand sides

$$A - \textstyle\sum_{J_1} \bar{w}_i - \sum_{J_2} \bar{w}_i + \sum_{J_1} \bar{w}_i' + \sum_{J_2} \bar{w}_i'$$

for the constraint (12),

$$b_{k_1} - \textstyle\sum_{J_1} \bar{w}_i + \sum_{J_2} \bar{w}_i' \quad \text{and} \quad b_{k_2} - \sum_{J_2} \bar{w}_i + \sum_{J_1} \bar{w}_i'$$

for the k_1-th and k_2-th constraint in (13), respectively, and all other righthand sides unchanged. The change in cost from the prior (IA) to this modified new one is just

$$\textstyle\sum_{J_1 \cup J_2} [q_i(\bar{w}_i') - q_i(\bar{w}_i)]. \tag{16}$$

Since $(\bar{\rho}, \bar{\nu})$ is a subgradient of the minimal cost of (IA) as a function of its righthand side at these modified values, moreover, the change in cost from the modified (IA) to the actual new (IA) is bounded below by

$$(\bar{\rho} + \bar{\nu}_{k_1}) \left(\textstyle\sum_{J_1} \bar{w}_i - \sum_{J_2} \bar{w}_i'\right) + (\bar{\rho} + \bar{\nu}_{k_2}) \left(\sum_{J_2} \bar{w}_i - \sum_{J_1} \bar{w}_i'\right). \tag{17}$$

We may sum (16) and (17) to obtain a (lower) estimate $\Delta\mathrm{IA}(k_1, k_2, J_1, J_2)$ for the change in inventory costs of (IA) resulting from switching the locations J_1 and J_2, where

$$\begin{aligned} \Delta\mathrm{IA}(k_1, k_2, J_1, J_2) &= \textstyle\sum_{J_2 \cup J_2} [q_i(\bar{w}_i') - q_i(\bar{w}_i)] \\ &+ (\bar{\rho} + \bar{\nu}_{k_1}) \left(\textstyle\sum_{J_1} \bar{w}_i - \sum_{J_2} \bar{w}_i'\right) + (\bar{\rho} + \bar{\nu}_{k_2}) \left(\sum_{J_2} \bar{w}_i - \sum_{J_1} \bar{w}_i'\right). \end{aligned} \tag{18}$$

It is worth noting that this estimate requires $|J_1| + |J_2|$ inversions (or function evaluations, if the inverse of q_i' is available in explicit form), plus the same number of function evaluations to obtain the $q_i(\bar{w}_i')$, plus $4(|J_1| + |J_2|)$ additions and two multiplications.

Suppose now that there are limits on storage capacity at the customer locations. These can be expressed as constraints $w_i \leq u_i$ for $i = 1, \cdots, n$, appended to (1)–(9). The separation property mentioned in Section 1 still holds, so (IA) now includes the same constraints. The methods of Federgruen and Zipkin can easily be extended to this case. Moreover, a lower estimate ΔIA can be calculated as above, with $\bar{w}_i'$ redefined to ensure feasibility; that is, $\bar{w}_i'$ is defined as above, except $\bar{w}_i' = u_i$ if $q_i'(u_i) \leq \bar{\rho} + \bar{\nu}_{k_2}$, $i \in J_1$, or if $q_i'(u_i) \leq \bar{\rho} + \bar{\nu}_{k_1}$ for $i \in J_2$.

3. MODIFIED INTERCHANGE HEURISTICS

A number of the successful approaches to the deterministic TSP and VRP can be described as interchange heuristics. Each such method starts with a given tour (or set of routes) and improves it by a sequence of small changes. Many potential changes are evaluated before any one is implemented. This general description covers the "r-opt" methods of Croes [1958], Lin [1965] and Lin and Kernhigan [1973] for the TSP, and extensions to the VRP by Russell [1977] and Christofides and Eilon [1969], as well as the procedures of Wren and Holliday [1972] and Cassidy and Bennet [1972], among others.

We now show that methods of this kind can be adapted easily to accommodate inventory allocation. We concentrate on r-opt methods (specifically 2-opt for ease of presentation), but extension to other interchange heuristics should be straightforward. (For example, an earlier version of this paper adapts the method of Lin and Kernhigan.)

Figures 1 and 2 illustrate the basic idea of 2-opt as applied to the VRP. Figure 1 shows an initial set of routes, while Figure 2 shows a potential interchange. The interchange consists of dropping two links from the current routes (the dashed lines) and adding two new ones (the double lines). In the process, the partition of customers among vehicles changes; the singleton subset J_1 moves to set k_2 and J_2 to k_1. (Other switches, of course, might change only the sequence in which a vehicle visits its customers, not the partition.)

Evaluation of such a switch in a deterministic problem clearly requires adding the costs of the new links and subtracting the costs of the old ones, a very cheap computation. In our stochastic-demand problem the same evaluation suffices when the partition does not change. When it does change, problem (IA) changes, so we must also assess the changes in expected penalty and holding costs. While we could resolve (IA) for each potential switch, a more attractive approach is to use the approxi-

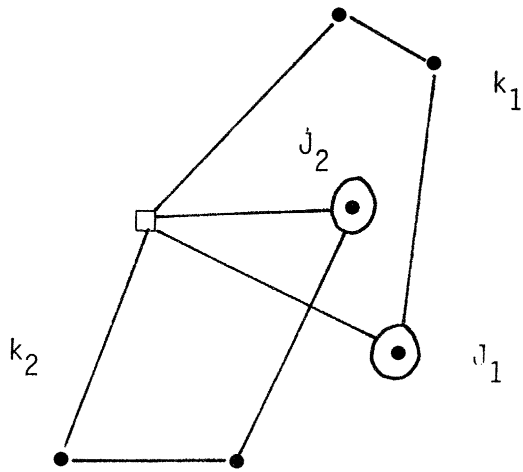

Figure 1. Initial routes.

mation ΔIA, as defined in (18), and to resolve (IA) only when considering implementing a switch. (The exact cost change should then be computed to check that an overall improvement is truly achieved and, if so, to update the multipliers (ρ, ν).) Calculation of ΔIA certainly adds to the work of evaluation, but not unreasonably much. (Note, since ΔIA is a lower bound, if there exists an advantageous 2-opt switch, the algorithm will identify one.)

Prior to evaluation, moreover, a potential switch must be checked for feasibility in a deterministic VRP. This step can be skipped in our problem, since vehicle capacities do not restrict the feasible partitions.

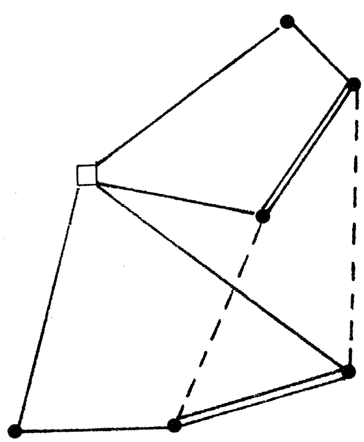

Figure 2. Potential interchange.

We remark that 2-opt indeed changes at most two subsets at a time, so when (IA) must be resolved, special techniques can be applied; cf. the Appendix in Federgruen and Zipkin. These techniques can be used also in the method of Lin and Kernhigan.

Section 5 reports our experience with a hybrid heuristic including 3-opt with the modifications above.

The VRP is sometimes complicated by restrictions on feasible routes involving, for example, the duration of routes and/or the timing of deliveries (cf., e.g., Fisher and Jaikumar [1978]). Given an interchange heuristic that *distinguishes* feasible routes in a deterministic VRP with such restrictions, clearly, the revised method of *evaluating* routes described above can be used.

Problems with multiple depots (cf., e.g., Gillett and Johnson [1976]) can be handled as well. Consider the case where the assignment of vehicles to depots is fixed. Formulate the analogue to (1)–(9) and fix the variables analogous to y. The problem then separates into some TSPs and one problem of the form of (IA) for each depot. An interchange code for the multiple-depot VRP can thus be adapted as in the single-depot case. (A slight modification is required in ΔIA to account for the fact that the sets of variables in the several allocation problems may change between iterations.)

4. AN EXACT ALGORITHM USING GENERALIZED BENDERS' DECOMPOSITION

This section shows how to modify a very different approach to the VRP. The result is an algorithm that solves (1)–(9) exactly, and that provides a lower bound on the true optimal cost in each iteration. Our purpose here is to demonstrate the flexibility of computational techniques permitted by the separation property.

The method of Fisher and Jaikumar [1978] for deterministic VRPs relies on the fact that a TSP can be viewed as a linear program, whose feasible set is defined implicitly as the convex hull of all feasible solutions to the TSP (the so-called traveling salesman polytope). From this standpoint, the VRP is a mixed-integer linear program, where the integer variables are the y_{ik}, since with y fixed the VRP reduces to K TSPs. The VRP may thus be solved exactly with Benders' decomposition procedure, cf. Benders [1962].

Benders' decomposition requires dual solutions for the linear subproblem(s) in order to generate cuts. Cutting-plane algorithms for the TSP have enjoyed a resurgence recently (e.g., Grötschel [1980], Miliotis [1976, 1978], Padberg and Hong [1980]), and produce a dual solution as a by-product. We remark that the method of Fisher and Jaikumar [1978] does not require that cutting planes be used to solve each TSP, only that the (relatively few) binding cutting planes be generated and priced once the solution is found.

If y is fixed in problems (1)–(9), we obtain the same K TSPs plus the inventory allocation problem. The latter is a nonlinear program, so Benders' decomposition cannot be used. Generalized Benders' decomposition (Geoffrion [1972]), however, can be adapted nicely to the current problem.

This method can be summarized as follows: A master problem in the y variables, equivalent to the original, is derived by projection and dualization. A sequence of relaxed master problems is solved. Each such problem yields a tentative solution y, which defines the subproblems. The subproblems are solved or determined to be infeasible. Dual solutions or extreme rays then define one or more constraints ("cuts") of the master problem. These cuts are appended to the previous relaxed master problem, and the process continues.

The success of generalized Benders' decomposition for a particular problem depends on the resolution of several issues. The subproblems must be relatively easy to solve; algorithms for them must produce optimal multipliers (or extreme rays, where appropriate); and they must not have duality gaps (if the master problem is to be equivalent to the original version of the problem). The TSPs regarded as linear programs satisfy these conditions. For the allocation problem, the methods developed in Federgruen and Zipkin solve it easily and yield optimal multipliers; as a convex program, it has no duality gap.

Also, what Geoffrion calls "Property P" must hold: The constraints of the master problem are expressed in terms of optimization problems, and these must be "easy" to evaluate. We demonstrate below that this criterion is well-satisfied.

Our discussion assumes all subproblems and the relaxed master problems are solved exactly, and that nonbinding constraints are never dropped. The effects of relaxing these assumptions here are the same as in the general case, cf. Geoffrion.

A lower bound on the true optimal cost is produced in each iteration of generalized Benders' decomposition (cf. Geoffrion). In one iteration, therefore, the suboptimality of any starting solution (e.g., one computed by a heuristic) could be checked.

Master Problem

Recall that $x^k = (x_{ijk})_{ij}$ and $y^k = (y_{ik})_i$ are vectors of variables. Let us rewrite (1)–(9) as follows:

$$\min \textstyle\sum_{i,j,k} c_{ij} x_{ijk} + \sum_i q_i(w_i)$$

$$\text{subject to } G^k(x^k, y^k) \geq 0$$

$$x^k \in X^k, \qquad k = 1, \cdots, K, \qquad y \in Y, \quad \text{and} \quad (2), (3).$$

In this version, Y represents (4) and (5), $X^k = \{x^k\colon 0 \le x_{ijk} \le 1,\ i, j = 0, \cdots, n\}$, and G^k represents all the other linear inequalities defining the kth TSP polytope.

Following Fisher and Jaikumar [1978], we now project this problem onto y. The allocation problem and the TSPs are feasible for any $y \in Y$. For $y \in Y$, let $v(y)$ be the minimal objective value with y fixed. Then, as in Geoffrion,

$$\begin{aligned} v(y) = \textstyle\sum_{k=1}^{K} \sup_{\pi^k \ge 0} \min_{x^k \in X^k} \left\{\sum_{i,j=0}^{n} c_{ij} x_{ijk} - \pi^k G^k(x^k, y^k)\right\} \\ + \sup_{\rho,\nu \le 0} \min_{w \ge 0} \left\{\textstyle\sum_{i=1}^{n} q_i(w_i) + \rho\left(A - \sum_{i=1}^{n} w_i\right)\right. \\ \left. + \textstyle\sum_{k=0}^{K} \nu_k \left(b_k - \sum_{i=1}^{n} w_i y_{ik}\right)\right\}. \end{aligned}$$

The following master problem is thus equivalent to (1)–(9):

$$\begin{aligned} &\min Z \\ &\text{subject to } Z \ge v(y) \\ &\qquad\qquad y \in Y. \end{aligned}$$

Cuts for the Relaxed Master Problem

Cutting-plane methods produce a dual-optimal $\bar{\pi}^k$ for each k. (Although $\bar{\pi}^k$ is a vector of very large dimension, nearly all its components will usually be zero.) Solution of the allocation problem yields optimal multipliers $(\bar{\rho}, \bar{\nu})$. These values generate a cut helping to approximate $v(y)$, of the following form:

$$\begin{aligned} Z \ge \textstyle\sum_{k=1}^{K} \left\{\min_{x^k \in X^k} \sum_{i,j=0}^{n} c_{ij} x_{ijk} - \bar{\pi}^k G^k(x^k, y^k)\right\} + \bar{\rho} A + \sum_{k=1}^{K} \bar{\nu}_k b_k \\ + \min_{w \ge 0} \left\{\textstyle\sum_{i=1}^{n} \left[q_i(w_i) - \left(\bar{\rho} + \sum_{k=0}^{K} \bar{\nu}_k y_{ik}\right) w_i\right]\right\}. \end{aligned}$$

Now we discuss how to simplify this expression.

The minimum over x^k is independent of y^k, and reduces to the expression

$$\textstyle\sum_{k=1}^{K} \left(d_k + \sum_{i=0}^{n} d_{ik} y_{ik}\right) = \sum_{k=1}^{K} \left(d_k + d_{0k} + \sum_{i=1}^{n} d_{ik} y_{ik}\right),$$

where all d's represent constants. In the minimum over w, note that by constraints (3) exactly one y_{ik} is 1 for each i. Thus, the minimum equals

$$\textstyle\sum_{k=0}^{K} \sum_{i=1}^{n} y_{ik} \min_{w_i \ge 0} \{q_i(w_i) - (\bar{\rho} + \bar{\nu}_k) w_i\}.$$

For each i, k, the inner minimization here is just a newsboy problem. Call its cost f_{ik}. (This simplification is what satisfies Property P.) Thus, the form of the cut is

$$\begin{aligned} Z \ge \bar{\rho} A + \textstyle\sum_{k=1}^{K} (\bar{\nu}_k b_k + d_k + d_{0k}) + \sum_{i=1}^{n} f_{i0} y_{i0} \\ + \textstyle\sum_{k=1}^{K} \sum_{i=1}^{n} (d_{ik} + f_{ik}) y_{ik}. \end{aligned} \tag{19}$$

The relaxed master problem is thus min Z, subject to cuts of the form (19), as well as constraints (4) and (5).

In general, since Y is finite, only a finite number of subproblems can be generated, and the algorithm converges to the true optimum in a finite number of iterations. Using the lower bounds, we can terminate the procedure prior to optimality when any given error tolerance is achieved.

5. COMPUTATIONAL RESULTS

In this section we report our computational experience with a modified interchange heuristic (Section 3), using problems adapted from the literature. We discuss both computation times and cost comparisons between our problems and deterministic versions of them.

Description of the Code

With unlimited supply and vehicle capacities, each location would be given an amount w_i^*, the (unique) minimum of the function $q_i(\cdot)$. (w_i^* is often referred to as the "newsboy solution.") An initial assignment of locations to vehicles (as well as an initial set of routes) is obtained by applying a sweep heuristic, as in Gillett and Miller [1974] or Gillett and Johnson to the deterministic VRP with w_i^* as the (fixed) delivery size of location i. (A new vehicle is used as soon as the cumulative load on the current truck exceeds $1/K \sum_{i \in Y} w_i^*$.) For this set of assignments we next solve the inventory allocation problem (IA) *exactly*, and find 3-opt solutions to the K TSPs. The solution procedure for the problem (IA) follows the overall approach of Federgruen and Zipkin.

The improvement part of the procedure iteratively constructs a set of routes that cannot be improved by serving more or fewer locations or by a 3-opt switch involving at most two routes. (A 3-opt interchange could involve three routes; such switches were ignored.) Potentially beneficial switches are identified by evaluation of the ΔIA function; for these the exact cost change is computed to check whether an improvement is truly achieved. (The optimal solution of (IA) is recovered using the methods in Federgruen and Zipkin.) Whenever the assignment of one or more locations is changed, we recover a 3-opt solution of all (at most two) routes involved.

The improvement routine has two phases. In Phase I, we consider switches only between pairs of routes that are adjacent in the initial solution. Phase II considers all pairs of routes.

The Problem Set

A 50-location and a 75-location problem introduced by Christofides and Eilon were used as the basis for our computational experiments. (These correspond to problems 8 and 9 in Chapter 9 of Eilon et al.

[1971].) The vehicle capacities in these problems are 160 and 140, respectively. The supply in the depot was fixed at 160 and 1000 units, respectively. All $F_i(\cdot)$ for $i = 1, \cdots, n$ were taken to be normal with coefficients of variation equal to one. We generated starting inventories β_i for $i = 1, \cdots, N$ from a uniform distribution between 0 and 15; and used identical penalty and holding cost rates in all locations. To construct a set of demand distributions corresponding roughly to the deterministic problems, we proceeded as follows. The costs h^+ and h^- were temporarily fixed at 0.5 and 5. Mean demands (and hence standard deviations) were chosen to equate the "ideal" delivery size w_i^* for $i = 1, \cdots, n$ with the fixed delivery sizes used in Christofides and Eilon. Next, fixing the demand parameters, we varied h^+ and h^-. We also varied the number of vehicles K.

Computational Results

Table I summarizes the results of 18 runs on an IBM 4341. In reporting computational times, we distinguish between the times spent in inventory allocation subroutines ("*alloc*"), and subroutines administering potential switches ("*switch*"); a third category ("*other*") includes all remaining procedures (such as the sweep-procedure resulting in an initial solution). (Since the clock routine often consumed more than 50% of the total time, we ran all problems twice, once with and once without the clock routine. Half of the time consumed by the clock routine is attributed to "alloc" and half to "switch," since the number of clock routine calls are almost identical in both sets of subroutines.) Very roughly speaking, we can interpret the "switch" plus "other" time as the time a deterministic VRP heuristic would require, and the "alloc" time as the additional computation required to handle stochastic demands.

The following observations can be made. The two-phase variant is far more time consuming than the Phase I version. Improvements in the second phase are rare, and in any case of limited size. Our discussion below thus reflects the Phase I results only.

Computational times vary between 3–7 and 7–16 CPU seconds for the two problem sizes. The time spent for allocation is certainly a substantial fraction of the total, but in every case the total time is well within the same order of magnitude as the "switch" plus "other" time. Thus, while the combined routing/allocation problem requires more effort than the VRP (as one would expect), the overall computational demands of the combined approach are reasonable for many applications.

As expected, when h^+ and h^- increase, a different set of routes is chosen, usually with larger routing costs, however, enabling a less than proportional increase in inventory carrying and/or shortage costs.

TABLE I
COMPUTATIONAL RESULTS FOR 18 RUNS

Problem Parameters				Phase I							Phase II	
				Solution Cost			Times (CPU sec)					
n	K	h^-	h^+	Inventory	Routing	Total	Alloc.	Switch	Other	Total	Cost (% of Phase I)	Total time
75	6	5.0	0.5	946.7	648.2	1594.9	6.920	1.516	0.875	9.311	100.0	13.105
75	7	5.0	0.5	868.9	669.9	1538.7	6.827	1.512	0.916	9.255	100.0	14.140
75	8	5.0	0.5	856.9	697.1	1554.0	13.802	1.381	0.895	16.078	99.9	30.022
75	9	5.0	0.5	858.8	734.8	1593.6	11.516	0.911	0.913	13.340	97.2	56.138
75	6	10.0	1.0	1868.9	679.2	2548.1	7.086	1.712	0.880	9.678	100.0	14.138
75	7	10.0	1.0	1673.6	729.4	2402.9	5.242	1.233	0.886	7.361	99.8	22.574
75	8	10.0	1.0	1649.5	753.5	2403.0	10.997	1.513	0.885	13.395	99.9	24.721
75	9	10.0	1.0	1656.3	759.4	2415.7	10.327	1.045	0.926	12.298	99.3	33.999
75	6	20.0	1.0	3052.0	681.1	3733.1	12.159	2.262	0.899	15.320	100.0	21.016
75	7	20.0	1.0	2551.1	724.2	3275.2	7.385	1.340	0.904	9.629	100.0	16.036
75	8	20.0	1.0	2486.7	753.5	3240.2	12.928	1.360	0.924	15.212	99.9	28.420
75	9	20.0	1.0	2486.1	801.6	3287.6	12.659	0.973	0.919	14.551	98.7	118.328
50	3	5.0	0.5	574.9	414.7	989.6	2.909	1.219	0.532	4.660	100.0	4.681
50	4	5.0	0.5	515.3	430.9	946.2	1.817	0.651	0.552	3.020	100.0	3.641
50	3	10.0	1.0	1110.6	467.7	1578.3	1.723	0.659	0.540	2.922	100.0	2.942
50	4	10.0	1.0	971.1	476.1	1447.2	3.549	1.227	0.558	5.334	100.0	6.222
50	3	20.0	1.0	1770.6	484.2	2254.7	4.146	1.367	0.556	6.069	100.0	6.091
50	4	20.0	1.0	1398.6	493.8	1892.4	4.919	1.192	0.592	6.703	100.0	7.742

Comparisons between the Combined Problem and the VRP

To enable a meaningful comparison between the combined inventory allocation/routing problem and the VRP, we reran the 50- and 75-location problem with depot supplies of 800 and 1500 units, respectively. (These suffice to cover all the fixed deliveries in the corresponding VRPs.) Thus, the known VRP solutions reported in Table 9.2 of Eilon et al. are feasible solutions in the combined model as well. Table II exhibits to what extent these solutions can be improved by using the combined approach. The results show that substantial savings (6–7%) can be achieved in operating costs while reducing the number of required trucks

TABLE II
($h^+ = 0.5$; $h^- = 5$)

n		k	Customers Visited	Total Quantity Delivered	Inventory Cost	Routing Cost	Total Cost
75	Stochastic	7	70	980.0	868.9	669.9	1538.7
75	Stochastic	8	71	1120.0	797.4	720.8	1518.1
75	Stochastic	9	73	1259.0	749.5	799.9	1549.3
75	Stochastic	10	74	1345.0	742.1	842.3	1584.4
75	Determin.	10	75	1364.0	736.4	876.0	1612.4
50	Stochastic	3	44	480.0	574.9	414.7	989.6
50	Stochastic	4	46	640.0	496.0	440.2	936.2
50	Stochastic	5	44	729.0	483.8	454.1	937.8
50	Determin.	5	50	777.0	450.5	556.0	1006.5

by no less than 20%! (These savings are achieved by eliminating only four locations.) The usual caution applies when extrapolating these results to other problems.

APPENDIX

Let α represent the vector of model parameters, cf. Section 1. Let $z(\alpha)$ be the value of (P) for parameter vector α. Let $z(\alpha; x, y)$ be the value of (P) for fixed vector α, and fixed vectors x, y and let $(z(\alpha) = \min_{x,y} \{z(\alpha; x, y)$, subject to (4)–(9)$\}$.

LEMMA. *The value of (P) is continuous in A, $b_k (k = 1, \cdots, K)$, β_i, h_i^+, h_i^-, and c_{ij} $(i, j = 1, \cdots, n)$.*

Proof. Let $\{\alpha_n\}_{n=1}^{\infty} \rightarrow \alpha_0$. Let (x_n, y_n) be part of the optimal solution for (P) with $\alpha = \alpha_n$; $n \geq 0$. Note that for $n \geq 1$, $z(\alpha_n; x^n, y^n) - z(\alpha_0; x_n, y_n) \leq z(\alpha_n) - z(\alpha_0) \leq z(\alpha_n; x^0, y^0) - z(\alpha_0; x^0, y^0)$. Since there are only finitely many (x, y) vectors satisfying (4)–(9), it suffices to show that for any such pair (x^*, y^*)

$$z(\alpha;\, x^*,\, y^*) = \textstyle\sum_{\{(i,j,k):x^*_{ijk}=1\}} c_{ij} + \bar{z}(\alpha;\, y^*) \tag{A1}$$

is continuous in α where

$$\bar{z}(\alpha;\, y^*) = \min \textstyle\sum_i q_i(w_i) \quad \text{subject to(12), (13),} \quad \text{with} \quad Y_k = \{i\colon y^*_{ik} = 1\}, \qquad k = 0, \cdots, K. \tag{A2}$$

Continuity of the first term in (A1) is immediate. To verify continuity of $\bar{z}$, let w_n for $n \geq 1$ solve $\bar{z}(\alpha_n;\, y^*)$. Since $\{w_n\}_{n=1}^{\infty}$ is contained in a compact set, it has a convergent subsequence. Restricted to this subsequence, $w_n \rightarrow w_0$ (say). Limiting arguments show that w_0 is feasible for $\alpha = \alpha_0$. Hence $\lim_{n\to\infty} \bar{z}(\alpha_n;\, y^*) = \lim_{n\to\infty} \sum_i q_i(\alpha_n;\, w_n) = \sum_i q_i(\alpha_0;\, w_0) \geq \bar{z}(\alpha_0;\, y)$. To prove the converse inequality, let $\bar{w}$ solve $\bar{z}(\alpha_0;\, y^*)$ and construct $\{w_n{}'\} \rightarrow \bar{w}$ with $w_n{}'$ feasible for $\alpha = \alpha_n$. Then $\lim_{n\to\infty} \bar{z}(\alpha_n;\, y^*) \geq \lim_{n\to\infty} \sum_i q_i(\alpha_n;\, w_n{}') = \sum_i q_i(\alpha_0;\, \bar{w}) = \bar{z}(\alpha_0, y)$.

ACKNOWLEDGMENTS

We are indebted to Harry Groenevelt for developing the computer code described in Section 5, and for many useful suggestions and improvements. The referees' comments on an earlier draft were most helpful.

REFERENCES

ASSAD, A., B. GOLDEN, R. DAHL AND M. DROR. 1982. Design of an Inventory-Routing System for a Large Propane-Distribution Firm. Working Paper, University of Maryland.

BENDERS, J. 1962. Partitioning Procedures for Solving Mixed-Variables Programming Problems. *Numer. Math.* **4,** 238–252.

BRODHEIM, E., AND G. PRASTACOS. 1979. The Long Island Blood Distribution System as a Prototype for Regional Blood-Management. *Interfaces* **9,** 3–35.

CASSIDY, P., AND H. BENNETT. 1972. TRAMP—A Multi-depot Vehicle Scheduling System. *Opnl. Res. Quart.* **23,** 151–163.

CHRISTOFIDES, N. 1976. The Vehicle Routing Problem. *Rev. Franc. Automat. Informat. Rech. Opnl.* **10,** 55–70.

COOK, T., AND R. RUSSELL. 1978. A Simulation and Statistical Analysis of Stochastic Vehicle Routing with Timing Constraints. *Decision Sci.* **9,** 673–687.

CROES, A. 1958. A Method for Solving Traveling-Salesman Problems. *Opns. Res.* **5,** 791–812.

EILON, S., C. WATSON-GANDY AND N. CHRISTOFIDES. 1971. *Distribution Management: Mathematical Modelling and Practical Analysis.* Hafner Publishing Co., New York.

FEDERGRUEN, A., AND B. LAGEWEG. 1980. Hierarchical Distribution Modelling with Routing Costs, Working Paper No. 314A. Graduate School of Business, Columbia University, New York.

FEDERGRUEN, A., AND P. ZIPKIN. 1983. Solution Techniques for Some Allocation Problems. *Math. Program.* **25,** 13–24.

FISHER, M., AND R. JAIKUMAR. 1978. A Decomposition Algorithm for Large-Scale Vehicle Routing, Working Paper. Dept. of Decision Sciences, University of Pennsylvania, Philadelphia.

GEOFFRION, A. 1972. Generalized Benders' Decomposition. *J. Optim. Theor. Appl.* **10,** 237–260.

GEOFFRION, A., AND G. GRAVES. 1974. Multicommodity Distribution System Design by Benders' Decomposition. *Mgmt. Sci.* **20,** 822–844.

GILLETT, B., AND J. JOHNSON. 1976. Multi-Terminal Vehicle-Dispatch Algorithm. *Omega* **4,** 711–718.

GILLETT, B., AND L. MILLER. 1974. A Heuristic Algorithm for the Vehicle-Dispatch Problem. *Opns. Res.* **22,** 340–349.

GOLDEN, B., AND W. STEWART. 1978. Vehicle Routing with Probabilistic Demands. In *Computer Science and Statistics: Tenth Annual Symposium on the Interface*, pp. 252–259, D. Hoyben and D. Fife (eds.), NBS Special Publication 503. National Bureau of Standards, Washington, D.C.

GOLDEN, B., AND J. YEE. 1979. A Framework for Probabilistic Vehicle Routing. *AIEE Transact.* **11** (2), 109–112.

GRÖTSCHEL, M. 1980. On the Symmetric Traveling Salesman Problem: Solution of a 120-City Problem. *Math. Program. Study* **12,** 61–77.

HERRON, D. 1979. Managing Physical Distribution for Profit. *Harvard Business Review* **79** (May–June), 121–132.

LAPORTE, G., AND Y. NORBERT. 1981. An Exact Algorithm for Summarizing Routing and Operating Costs in Depot Location. *Eur. J. Opns. Res.* **6,** 224–227.

LIN, S. 1965. Computer Solutions of the Traveling-Salesman Problem. *Bell Syst. Tech. J.* **44,** 2245–2269.

LIN, S., AND B. KERNIGHAN. 1973. An Effective Heuristic Algorithm for the Traveling Salesman Problem. *Opns. Res.* **21,** 498–513.

MAGNANTI, T., AND B. GOLDEN. 1978. Transportation Planning. In *Studies in Operations Management*, Chap. 16, A. Hax (ed.). North-Holland, Amsterdam.

MILIOTIS, P. 1976. Integer Programming Approaches to the Traveling Salesman Problem. *Math. Program.* **10,** 367–378.

MILIOTIS, P. 1978. Using Cutting Plans to Solve the Symmetric Traveling Salesman Problem. *Math. Program.* **13,** 177–188.

PADBERG, M., AND S. HONG. 1980. On the Symmetric Traveling Salesman Problem: A Computational Study. *Math. Program. Study* **12,** 78–107.

PRASTACOS, G. 1978. Optimal Myopic Allocation of a Product with Fixed Lifetime. *J. Opns. Res. Soc.* **29,** 905–913.

RUSSELL, R. 1977. An Effective Heuristic for the *M*-Tour Traveling Salesman Problem with Some Side Conditions. *Opns. Res.* **25,** 517–524.

SCHRAGE, L. 1981. Formulation and Structure of More Complex/Realistic Routing and Scheduling Problems. *Networks* **11,** 229–232.

STEWART, W., AND B. GOLDEN. 1982. Stochastic Vehicle Routing: A Comprehensive Approach, Working Paper. University of Maryland, Baltimore.

TILLMAN, F. 1969. The Multiterminal Delivery Problem with Probabilistic Demands. *Trans. Sci.* **3,** 192–204.

WILLIAMS, A. 1973. Sensitivity to Data in LP and MIP; presented at VIII International Symposium on Mathematical Programming, Stanford, Calif.

WREN, A., AND A. HOLLIDAY. 1972. Computer Scheduling of Vehicles from One or More Depots to a Number of Delivery Points. *Opnl. Res. Quart.* **23,** 333–344.

ZIPKIN, P. 1980. Simple Ranking Methods for Allocation of One Resource. *Mgmt. Sci.* **26,** 34–43.