The metered inventory routing problem, an integrative heuristic algorithm

Yale T. Herer*, Roberto Levy

Faculty of Industrial Engineering and Management, Technion – Israel Institute of Technology, Haifa 32000, Israel

Abstract

The Metered Inventory Routing Problem (MIRP) involves a central warehouse, a fleet of trucks with a finite capacity, and a set of customers, for each of whom there is an estimated consumption rate, and a known storage capacity. The objective is to determine when to service each customer, as well as the route to be performed by each truck, in order to minimize the total discounted costs. The problem is solved on a rolling horizon basis, taking into consideration holding, transportation, fixed ordering, and stockout costs. The algorithm we develop uses the concept of 'temporal distances': in short, the temporal distance between two customers is the cost of moving these customers to a common period. A simulation study is performed to demonstrate the effectiveness of our procedure.

Keywords: Inventory routing; Vehicle routing; Temporal heuristic

1. Introduction

Today's challenge for management is to remain competitive. To do this, logistic systems must be significantly improved. Usually, the delivery step (moving goods to customers or retailers) has been seen as the most costly of the distribution process. Consequently, making the planning and execution of the transportation activities more efficient will help provide management with a competitive edge.

The importance of the distribution stage is evident from the magnitude of the associated distribution costs. Some surveys (see e.g. Bodin et al., 1983) show that physical distribution costs account for about 16% of the value of an item. A recent study provided by Data Resources Inc. (see Anily, 1986) specifies that transportation costs alone may, in certain sectors of the economy, amount to a fifth (lumber and wood products) or even a quarter (petroleum, stone, clay and glass products) of the average value of sales. Consequently, a small percentage of savings in transportation expenses could result in substantial overall savings.

In order to optimize the system, numerous problems must be solved at different levels of the business hierarchy. At the strategic level, the main decisions concern the location of facilities (plants, warehouses and customers). At the tactical level, we must determine the fleet size and mix (vehicles characteristics, e.g. capacity), as well as customer storage conditions (e.g. the tank sizes for each customer in the case of industrial gases and fuel oil). Finally, at the operational level, we must determine the routing and scheduling...
of the vehicles in order to service the customers and we must determine the corresponding quantities to be delivered.

In the present work we are concerned with the last problem, namely, determining the delivery policy that meets the periodic demands of each customer while at the same time minimizing the total discounted costs.

Our problem, which can be termed the Metered Inventory Routing Problem (MIRP), can be briefly stated as follows: A firm is faced with supplying a set of geographically dispersed customers who are characterized by cost parameters (fixed ordering, holding, and stockout), storage capacity, and a stochastic demand process. The deliveries are to be performed at the start of each period using trucks which are characterized by their fixed rental cost, per mile usage cost, and capacity. We are to determine how to route the trucks so as to minimize the total discounted costs. A more in-depth and mathematical definition of our problem can be found in Section 3.2.

At this point, we wish to highlight the main difference between our model and other similar models that appear in the literature under the name “the Inventory Routing Problem (IRP)”. Our model, as the name implies, is essentially the IRP with the addition of a meter at each customer. In the standard IRP the customer pays for the delivery (in full) when it is made (note that the timing of the delivery is determined by the supplier and not by the customer). In contrast, under the MIRP formulation the customer pays for the inventory he uses, as he uses it. Thus, under the MIRP formulation, the supplier, not the customer, pays for inventory held at the customer.

Consider the billing practice used for certain industrial gases. The customer pays as he uses the gas (measured using a meter) from a tank (sometimes shared among multiple customers) located at the customer. The MIRP formulation is also in line with today’s reengineering trend. Customers no longer see a reason to handle their own raw material inventories (Hammer and Champy, 1993). They allot warehouse space to the supplier and basically say “we will pay for what we use. All you have to do is keep us stocked”.

As presented, the MIRP model makes a strict separation between the tactical and operational levels. This should not, in general, be done because the influence of decisions between levels is very strong. For example, the ideal capacity at each customer, although a tactical decision, can only be truly optimized taking into account the routing strategy to be implemented (an operational decision). This interaction between decision levels will be explored in Section 6.

This research contributes a model formulation of the MIRP, as well a solution procedure for this problem. The solution procedure is unique as it implements the temporal distance concept developed in Herer (1996). This paper also contributes a heuristic method for determining the amount of capacity to be installed at each customer. In the next section, we quickly review the relevant literature. In Section 3, we present a detailed definition of our problem. In Section 4, we present our solution procedure, which we test in Section 5. In Section 6, we use our model to develop a method for determining capacities to be used by each customer. Finally, in Section 7, we present our conclusions.

2. Literature review

The IRP in its original form and the metered version discussed here are both closely related to the Vehicle Routing Problem (VRP).

2.1. The Vehicle Routing Problem (VRP)

The basic VRP is easily stated. Given a set of nodes each with known demand, and a set of trucks each with known capacity, what are the delivery routes from the central warehouse that minimize the total distance traveled? When there is only one truck (of sufficient capacity) the VRP is equivalent to the Traveling Salesman Problem (TSP).

Many approaches have been taken to solve the VRP (see Bodin et al., 1983; Golden and Assad, 1988): exact methods including branch and bound, dynamic programming and cutting plane algorithms have been explored. Heuristics based on mathematical programming formulations have also been attempted. Interactive optimization techniques designed to exploit the knowledge of the experienced dispatcher have also been tried.

Other heuristics can generally be divided into three categories. The first two treat the VRP as two separate sub-problems, (I) finding the optimal assignment of customers to trucks (clustering) and (II) finding
the best route. (Category 1) The heuristics in this category, termed “cluster first – route second”, as the name implies, first solve sub-problem (I) and then sub-problem (II). (Category 2) The second heuristics, termed “Route first – cluster second”, as the name again implies, first solve sub-problem (II) and then sub-problem (I). (Category 3) The heuristics in the third category use a more global approach and are termed “insertion methods”. These methods attempt to solve sub-problems (I) and (II) together, which is in general a better approach than stepwise optimization. The classic example is the heuristic developed by Clarke and Wright (1964). We will use this heuristic in Section 4.7.

2.2. Delivery models integrated with inventory concerns

The IRP can be interpreted as an enrichment of the VRP to include inventory concerns (see Ball, 1988). It is clear that in order to optimize a system that includes both routing and inventory costs, routing and stocking decisions must be integrated. Artificial constraints separating these decisions are often imposed by the organizational structure causing the complex interaction between routing and stocking policies to be ignored.

Most inventory models propagate this artificial situation by using a simple structure for the replenishment (routing) costs, usually assuming separability across locations. In papers that relax this assumption, the cost structures are usually not rich enough to include true routing costs. We briefly mention some of the papers that integrate routing and inventory costs, most of which assume that demand is deterministic.

Federgruen and Zipkin (1984) consider a one-warehouse multiple retailer single period problem with a scarce resource. Federgruen et al. (1986) consider a model for perishable items. Burns et al. (1985) and Blumenfeld et al. (1985) are the first to consider explicitly the problem of integrating inventory and routing in an infinite-horizon model. Their model, however, uses information on the spatial density of the customers rather than their exact locations. The ideas from this approach resulted in an optimization tool called TRANSPART 2 that was implemented at General Motors (see Blumenfeld et al., 1987). Bell et al. (1983) developed a computerized planning system for a multi-period problem which has been implemented at Air Products and Chemicals, Inc. (industrial gases). Anily and Federgruen (1990) present a more complex approach to the same distribution system using the actual location of the customers. They show their heuristic to be asymptotically optimal within a given class of policies. Gallego and Simchi-Levi (1990) use a similar model; they propose a direct shipping strategy whose cost is no more than 1.061 times the cost of an optimal policy if the economic order quantity for each retailer is at least 71% of a truck load.

Herer and Roundy (1997) investigate the one warehouse multiple retailer distribution problem with traveling salesman tour vehicle routing costs in the framework of a more general production/distribution network with arbitrary near submodular joint order costs. Herer (1990) proposes good submodular approximations to TSP tour lengths, thus allowing the model of Herer and Roundy to include the routing element in the order costs.

2.3. The Inventory Routing Problem (IRP)

The model that motivated our interest was proposed by Dror et al. (1986) and Dror and Ball (1987), under the name “the Inventory Routing Problem (IRP)”. The IRP involves a central warehouse and a set of customers, each of whom faces a stochastic demand process. As stated above, the IRP is closely related to the VRP, whereas the VRP, is solved for a single period with known demands, the IRP is solved over several periods with stochastic demands.

The model can be described as follows: Each customer possesses a known storage capacity (e.g., the tank size in the case of heating oil or industrial gas). The objective is to “minimize the annual delivery costs while attempting to ensure that no customer runs out of the commodity at any time” (Dror and Ball, 1987). No attempt is made to explicitly bring carrying costs into the model.

The replenishment policy adopted is to always fill the customer to capacity when servicing. Since the customer’s expected consumption rate is known, the expected amount to be delivered can be determined before making a delivery, yet the exact amount is only determined once the truck arrives at the customer.
To solve the long-range problem, Dror et al. (1986) and Dror and Ball (1987) reduce the horizon to the short range and introduce penalties into the objective function by defining two short-range costs that reflect long-term delivery costs. The first represents the increase in future delivery costs if a delivery is made earlier in the planning horizon. The second represents the decrease in future delivery costs if an unplanned-for customer is serviced during the current period. Another feature of their work is the way they consider the uncertainty in the consumption rate. A comparison is made between stockout and delivery costs in order to obtain a 'safety stock' level for each customer. Using this safety stock level, Dror et al. (1986) and Dror and Ball (1987) treat consumption as deterministic.

A more general objective function. The basic IRP objective function was restricted to delivery costs. We feel that a complete integrated model should include fixed ordering, holding, and stockout costs. We include all of these costs in our objective function.

- **Holding costs** – The original IRP model assumes that holding costs are irrelevant. We assume that an inventory holding cost per unit per period is incurred at each customer.
- **Fixed ordering costs** – A cost component not examined by the original IRP formulation is the fixed ordering cost incurred by a customer. From the moment the order is placed until the delivery is made, certain costs are incurred irrespective of the route used to make the delivery and the quantity delivered. This cost would include the overhead expenses of placing an order at the warehouse, loading/unloading the truck, checking the quantity delivered, etc.
- **Shortage costs** – Instead of defining a safety stock level and then treating the demand as deterministic, we incorporate the randomness of the demand into the model by estimating the stockout costs for a particular period as a function of the probability of stocking out.

**Flexible fleet.** The original IRP assumes a constant fleet of vehicles each with a known capacity. For our model, we consider instead a fixed cost for using a truck (e.g., renting from a sub-contractor) each time a route of deliveries is made; this cost would include the charge for the vehicle. Additionally, we account for a higher rental cost for special or exclusive deliveries, needed whenever a customer stocks out before his planned delivery period.

3. Detailed problem definition (MIRP)

3.1. **MIRP – Refinements on the basic model**

The main objective in this work is to examine and solve a more encompassing version of problem than the IRP as defined in the previous section. Despite the positive characteristic of joining the inventory and delivery problems in its definition, it lacks an appropriate treatment of inventory carrying costs. Though some systems do not have this component, many do. In order to begin our analysis of the MIRP, we first describe our main changes to the IRP formulation.

- **A more general objective function.** The basic IRP objective function was restricted to delivery costs. We feel that a complete integrated model should include fixed ordering, holding, and stockout costs. We include all of these costs in our objective function.

- **Holding costs** – The original IRP model assumes that holding costs are irrelevant. We assume that an inventory holding cost per unit per period is incurred at each customer.

- **Fixed ordering costs** – A cost component not examined by the original IRP formulation is the fixed ordering cost incurred by a customer. From the moment the order is placed until the delivery is made, certain costs are incurred irrespective of the route used to make the delivery and the quantity delivered. This cost would include the overhead expenses of placing an order at the warehouse, loading/unloading the truck, checking the quantity delivered, etc.

- **Shortage costs** – Instead of defining a safety stock level and then treating the demand as deterministic, we incorporate the randomness of the demand into the model by estimating the stockout costs for a particular period as a function of the probability of stocking out.

- **Flexible fleet.** The original IRP assumes a constant fleet of vehicles each with a known capacity. For our model, we consider instead a fixed cost for using a truck (e.g., renting from a sub-contractor) each time a route of deliveries is made; this cost would include the charge for the vehicle. Additionally, we account for a higher rental cost for special or exclusive deliveries, needed whenever a customer stocks out before his planned delivery period.
each customer. The only information used from these distributions is the mean consumption of \( \lambda_i \) units at customer \( i \) per period and the probability \( p_i(t) \) that customer \( i \) will stock out before period \( t \). In addition, customer \( i \) has an identifiable holding cost of \( h_i \) dollars per unit stored per period, and a fixed ordering cost of \( A_i \) dollars for inclusion in a route.

**Distribution system:** As a matter of policy we fill customer stocks to capacity whenever we service a customer. Furthermore, if a customer stocks out, then we immediately make an emergency delivery which fills his stock to capacity. Both these policies are reasonable from a management point of view. The former guarantees that we will not visit any particular customer too often; the latter enables us to maintain good relationships with our customers.

Deliveries from the central warehouse to the customers are made by a fleet of trucks characterized by a finite capacity of \( K \) units, a fixed usage cost of \( V \) dollars per vehicle, and a cost of \( m \) dollars per mile traveled by each truck. We are also given the cost of an emergency service (usually much higher than a regular one). The cost of this service includes two components \( \hat{V}_i \) and \( \hat{m} \) which correspond to \( V \) and \( m \). These costs take into account the cost of paralyzing the business from the time the stockout occurs until the delivery is completed, the general overhead costs, the fixed cost for renting a special truck, etc.

**Objective:** To define the service policy (the period each customer is to be serviced and the routes to be performed by each vehicle) which minimizes the total discounted costs (fixed ordering, holding, stockout, and delivery costs). Since we are using discounted costs, we assume a periodic interest rate of \( r \). We solve the problem on a rolling horizon basis, using a planning horizon of \( T \) periods. We assume that the planning horizon is short enough that a customer being serviced during the planning horizon will not stock out before the end of the planning horizon. If such high demand customers do exist then they should, in any case, be treated separately. We have chosen a rolling horizon approach (and not an approach based on a periodic schedule; see e.g. Herer and Roundy, 1997) because our routing structure dynamically changes from planning horizon to planning horizon. This is, in turn, due to the stochastic demand which implies that customer stock levels are themselves unlikely to be periodic.

### 4. Solution procedure

The MIRP is a "hard" problem. In order to evaluate the complexity of the MIRP, consider an instance of the problem in which daily consumption at each customer is deterministic and equal to its capacity. This instance of our problem (with \( \hat{V}_i = \infty \)) is the VRP over all customers for each period. From the NP-completeness of the VRP (Bodin et al., 1983) we know that our problem is NP-hard. In this section we describe our heuristic solution procedure which requires less than 1 min to run on a PC. We expect our procedure to be run typically once a week (the planning horizon). With this in mind the running time can be considered instantaneous.

The first step in our algorithm is to estimate the effects of present decisions on future costs. We consider demand to be stable in the long range, and therefore, take the policy of making deliveries when inventory reaches zero as the starting point of our analysis. Using this starting point, effects of short-term decisions on future costs are estimable, excepting transportation costs. Since there is no periodic pattern of deliveries (as we explained in Section 3.2), we do not attempt to predict the effects of short-term decisions on long-range routing costs. In Section 4.2 we present the ideal situation, i.e., the one corresponding to stable demand and delivery when the inventory reaches zero. In Sections 4.3, 4.4 and 4.5 we examine the effects of short-term decisions on fixed ordering, holding, and stockout costs, respectively. We first give an overview of our algorithm.

#### 4.1. Overview of our algorithm

Our solution procedure begins with the calculation of the "best period of replenishment", \( T_i^{\text{opt}} \) for each customer \( i \). The value of \( T_i^{\text{opt}} \) is based on the total long-run discounted fixed ordering, holding, and stockout costs. For service level reasons we impose the constraint of never planning to service a customer after period \( L_i/\lambda_i \), i.e., the period he is supposed to stock out. Even if it would be optimal to assume the risk of these customers stocking out, we feel the added loss of good will would be too great. A customer may be willing to accept an occasional stockout when his demand is above average, but he would never be willing to
accept that the planned delivery period was after he was to stock out. After finding $T_{i}^{opt}$ for each customer we are able to determine the temporal distances between customers. This is a crucial concept for it allows us to service customers in a single route, even when they are initially assigned to different periods. The concept of temporal distances was introduced in Herer (1996).

Several of the VRP heuristics are based on the fact that customers who are spatially close (i.e., physically near each other) tend to be on the same route in the optimal solution. The use of temporal distances extends this idea to include time. Customers who are spatially close, but whose individual optimal delivery periods are different, will tend to be on the same route in the optimal solution if they are temporally close (i.e., the individual optimal delivery periods are not too far apart).

We implement this concept by taking the traditional Clarke and Wright (1964) algorithm and modifying it by adding the temporal distance into the savings calculation, when deciding whether or not to join two routes. We thus define the temporal distance as the increase in cost when two customers, initially assigned to be replenished in different periods, are joined together in a single route.

When dealing with routes containing multiple customers the same logic applies. In this case, however, the temporal costs are given by the shift of a group of customers from their joint best period. The effect of the temporal distance, therefore, tends to become more significant when dealing with large routes.

The importance of temporal distances is that they give flexibility to the solution procedure and reflect the non-separability of transportation and inventory decisions.

4.2. Ideal situation

Consider the policy of servicing customer $i$ when his inventory is supposed to reach zero. We identify this policy as policy $A$ (see the solid line in Fig. 1). We call the first point in time that the customer's inventory reaches zero, $T_{i}^{max} = L_{i}/\lambda_{i}$, we continue indefinitely with this strategy (i.e., filling a customer when his inventory reaches zero). The resulting policy resembles the saw-tooth curve of the EOQ policy.

In this section we present our method of computing the change in the total long-range discounted fixed ordering cost from the cost of the ideal situation associated with a one time deviation from policy $A$. Note that policy $A$ assures the minimal total discounted fixed ordering costs.

We define an alternative policy $B$ in which we decide, for the current planning horizon, to service customer $i$ on period $t$ before $T_{i}^{max}$. After this early service we continue servicing customer $i$ only when his inventory reaches zero (this policy is termed a zero-inventory ordering policy), as in policy $A$. In Fig. 1 we can see an illustration of both policies under the assumption of deterministic demand. This assumption is reasonable, if, in the long range, the average consumption is a good approximation for real consumption. Policy $A$ is shown by the solid line, and policy $B$ by the dashed one. We assume in our computations that an "ideal policy" (i.e. zero-inventory ordering) will be followed in the long range even though a decision differing from the "ideal policy" may be made in the short range.

In Fig. 1 we see that the curve associated with policy $B$ is shifted to the left of the curve associated with policy $A$ by $(T_{i}^{max} - t)$; this has consequences in terms of fixed ordering costs. In particular, each fixed ordering cost is incurred $(T_{i}^{max} - t)$ periods earlier in policy $B$ when compared to policy $A$. If we were dealing with a single fixed ordering cost payment shifted by $(T_{i}^{max} - t)$ periods, the cost increase would...
be insignificant. However, we are dealing with an infinite series of shifts. The shifts propagate over time, and the difference between the two policies, brought to their value at time \( t \) (the time when the decision to deviate from the "ideal situation" is to be made), can take on a significant value.

The difference between policies \( A \) and \( B \) in terms of their holding costs valued at time \( t \) for the initial shift in the timing of the order is

\[
A_i - \frac{A_i}{(1+r)^{T_{\text{max}}-t}} = A_i \left(1 - \frac{1}{(1+r)^{T_{\text{max}}-t}}\right).
\]

As we see from the graph the shift repeats itself every \( T_{\text{cyc}} = K_i/\lambda_i \) time units, where \( T_{\text{cyc}} \) can be interpreted as the average time needed for customer \( i \) to consume its capacity. Applying the same analysis for each shift and discounting this infinite series of differences, we obtain the increase in total discounted fixed ordering cost for servicing customer \( i \) on period \( t \) as

\[
O_i(t) = \frac{A_i}{1 - 1/(1+r)^{T_{\text{cyc}}}} \left(1 - \frac{1}{(1+r)^{T_{\text{max}}-t}}\right).
\]

4.4. Effects of short-term decisions on holding costs

We perform a similar analysis for holding costs. The difference in holding costs incurred by policies \( A \) and \( B \) is \( h_i \) times the difference in their stock levels. Between time \( t \) and \( T_{\text{max}} \), the stock level following policy \( B \) is greater than that following policy \( A \) by \( K_i - \lambda_i(T_{\text{max}} - t) \) units. On the other hand, between time \( T_{\text{max}} \) and \( (t + T_{\text{cyc}}) \), the stock level following policy \( A \) is greater than that following policy \( B \) by \( \lambda_i(T_{\text{max}} - t) \) units. As with the fixed ordering costs this pattern is repeated every \( T_{\text{cyc}} \) time units.

Examining the interval \([t, T_{\text{max}}]\) we see that the difference in the holding costs (valued at time \( t \)) is

\[
H_i^1(t) = h_i[K_i - \lambda_i(T_{\text{max}} - t)]((1+r)^{T_{\text{max}}-t} - 1).
\]

Performing the same analysis for the interval \([T_{\text{max}}, t + T_{\text{cyc}}]\), we see that the difference in the holding costs (valued at time \( t \)) is

\[
H_i^2(t) = \frac{h_i\lambda_i(T_{\text{max}} - t)((1+r)^{T_{\text{cyc}}-t} - 1)}{r(1+r)^{T_{\text{max}}-t}}.
\]

Note that because \( H_i^1(t) - H_i^2(t) \) represents the holding of stock earlier in time, it is always non-negative. Accounting for the infinite series of shifts, we obtain the total increase in discounted holding costs for servicing customer \( i \) on period \( t \) as

\[
H_i(t) = \frac{H_i^1(t) - H_i^2(t)}{1 - 1/(1+r)^{T_{\text{cyc}}}.
\]

There is, however, a problem with the above approach. Consider, in opposition to the "ideal policy", an alternative policy \( C \) which services the customer when his inventory level is almost at capacity (see Fig. 2, where policy \( A \) is again represented by a solid line).

The analysis is as before: We assume that the decision to order at time \( t \) is a one time deviation from the "ideal policy", and then we resume the zero-inventory ordering policy for the long range. In this case, the difference in the total holding cost is low. Therefore, the above deviation is unlikely to be a one time occurrence and, thus, the zero-inventory ordering part of policy \( C \) is unrealistic. The reason that this deviation would not be unique is that if, in this period we are led to believe that servicing when the customer's capacity is almost full will not adversely effect the long-term policy, we will be led to the same decision in every new planning horizon. Thus, the stock level will be constantly high. We therefore decided to consider servicing a customer only when his stockout probability exceeds a minimal threshold value. In this way we guarantee that a customer's inventory will not be constantly high.
4.5. Effects of short-term decisions on stockout costs

Since stockouts are rare events, their future occurrences are unaffected by the immediate delivery decision. Thus, the same analysis used for the fixed ordering and holding costs cannot be used when considering stockout costs. We consider instead expected stockout costs. The pivotal value that we must consider is the probability of stocking out before the planned delivery period, $P_i(t)$.

When a customer stocks out he receives a special delivery. This delivery is processed immediately due to the importance of keeping the customer supplied. The cost of this special delivery is $g + 2d_i\bar{m}$.

As we saw before, fixed ordering and holding costs tend to postpone deliveries as much as possible. Stockout costs, on the other hand, are responsible for bringing deliveries forward in time (the sooner the delivery, the smaller the probability of stocking out). To this end, we define a cost function which reflects the expected value of the stockout cost when servicing customer $i$ on period $t$:

$$S_i(t) = P_i(t)[V_i + 2d_i\bar{m}]$$

Note that $S_i(t)$ is never really incurred. In most cases, where there are no stockouts, it adds an artificial cost which motivates us not to risk a stockout. In the cases where a stockout does occur, the cost incurred is higher than that considered by the model for that period (because the real cost is not reduced by the probability factor). In this way, $S_i(t)$ can be used as an estimator of the stockout cost, because on average it equals the true stockout cost.

4.6. Temporal distance

After obtaining the $H_i(t), O_i(t)$ and $S_i(t)$ for every period from 0 to $T_{\text{max}}(i)$, we can determine the "best period of replenishment", $T_{i,\text{opt}}$, for each customer. $T_{i,\text{opt}}$ is the period which minimizes the sum of the costs considered above. If we let $C_i(t) = O_i(t) + H_i(t) + S_i(t)$, then

$$T_{i,\text{opt}} = \arg \min_t C_i(t).$$

Our methodology uses $T_{i,\text{opt}}$ as the initial delivery period for each customer. When joining two customers initially assigned to different periods we incur the 'temporal cost' of changing the service period for one or both of the customers. We thus define the temporal distance between two customers as the minimal cost incurred (in terms of fixed ordering, holding, and stockout costs), in bringing these two customers to a common delivery period. Obviously, this distance will be zero when the customers are already assigned to the same period. Thus, we define the temporal distance to be

$$d_{ij} = \min_{0 \leq t \leq \min(T_{i,\text{max}}, T_{j,\text{max}})} \{C_i(t) + C_j(t) - C_i(T_{i,\text{opt}}) - C_j(T_{j,\text{opt}})\}. \quad (1)$$

The meaning of Eq. (1) is that the temporal distance between two customers is the cost of servicing them in the same period (terms 1 and 2) and not in their individual optimal periods (terms 3 and 4). The new temporal location is simply the period in which Eq. (1) obtains its minimum, or more simply

$$\arg \min_{0 \leq t \leq \min(T_{i,\text{max}}, T_{j,\text{max}})} \{C_i(t) + C_j(t)\}.$$
4.7. Our algorithm

Our algorithm can simply be interpreted as the Clark and Wright (1964) algorithm for the VRP adapted to consider both spatial and temporal distances. We present here a detailed description of our algorithm, as well the time complexity of each step as a function of \( N \), the number of customers considered in the problem, and \( T \) the length of the planning horizon. The overall algorithm has time complexity of \( O(\max(N^3, N^2T)) \).

**STEP 1: \( O(NT) \)**

(a) Determine \( O_i(t), H_i(t), \) and \( S_i(t) \) for each customer for each period from the first period until period \( \min(4T, T_{i}^{\text{max}}) \). We stop calculating costs at period \( 4T \) because we feel that any customer whose best delivery period is beyond four planning horizons will have a negligible influence on the routes over the next \( T \) periods (i.e., the present planning horizon).

(b) Determine for each customer \( i \), \( t_i^{\text{opt}} \).

(c) Create a route between each customer and the central warehouse and let \( R_i = \{i\} \) for all \( i \).

**STEP 2: \( O(N^2T) \)**. Calculate the temporal distance between each pair of customers.

**STEP 3: \( O(\max(N^3, N^2T)) \)**. Repeat Steps 3.1 and 3.2 until no further gains are possible.

**STEP 3.1: \( O(N^2) \)**. Calculate the following gain function for each pair of customers, \( i, j \), connected to the central warehouse:

\[
V + k(d_{ij0} + d_{ij}^2 - d_{ij}^1) - d_{ij}^0.
\]

The gain function represents how much we would save by joining the two routes into a single route. We would save the cost of a truck rental \( V \), we would have to travel less spatial distance \( k(d_{ij0} + d_{ij}^2 - d_{ij}^1) \), but we may have to incur some temporal costs \( d_{ij}^0 \). Of course, we only calculate the gain function if the truck capacity is large enough to handle the delivery needs of the two routes.

**STEP 3.2: \( O(NT) \)**. Join the routes corresponding to the customer pair having the largest positive gain and update the temporal distances.

5. Computational study

5.1. Overview

In this section we describe our computational study of our proposed algorithm. As the MIRP is a real life problem, our aim is to find a solution that fits the needs of industry. This means an algorithm which produces a good solution with a short run time. In this context, the search for an exact policy (which would undoubtedly require long run times) would have very little use for us. Thus, we compare our method to a different heuristic approach. Both algorithms (coded in Pascal) had nearly identical run times, about half a minute for processing one planning horizon. As discussed above, run lengths of this magnitude should be considered instantaneous. In the following subsection we present the alternative algorithm; this is followed by the description of the experimental design, and finally the results and analysis.

5.2. The alternative algorithm

Since no other algorithm exists for the MIRP, our goal here is to develop a simple intuitive heuristic. This procedure, as opposed to our method of evaluating temporal and spatial costs, tackles the problem in a hierarchical fashion, assigning customers to periods and then solving a VRP for each period. This type of hierarchical approach has been used for the IRP (see e.g. Dror et al., 1986; Dror and Ball, 1987). We denote this alternative procedure \( \mathcal{H} \) (for \( \mathcal{H} \)-hierarchical) and our procedure by \( \mathcal{T} \) (for \( \mathcal{T} \)-temporal).

In the first phase of the procedure, assigning customers to periods, we implemented assignment rules based on Golden et al. (1984). For each period we defined two sets of customers. The customers whose consumption (as a fraction of the present inventory level) is expected to be above a given parameter (which we name \( x_1 \)) are considered potential candidates to be serviced (i.e. \( \frac{t_i - L_i}{L_i} > x_1 \)). Those customers, for whom this fraction is above a second parameter \( x_2 (x_2 > x_1) \), are tagged as obligatory customers (the delivery date will not exceed the period in question). The parameters \( x_1 \) and \( x_2 \) can be set by trying different values and adopting those which give the best results. For reasons identical to those found in Section 4.4, we decided that \( x_1 \) should be at least 0.5, meaning that we
only test values for \( z_1 \) which consider service to customers who would, supposedly, consume at least half of their inventory. Clearly, the larger the value of \( z_1 \) and \( z_2 \), the greater are the risks of stocking out.

In the second and final phase of procedure \( H \), we process the periods of the planning horizon one by one, solving for each period a traditional VRP using the Clarke and Wright (1964) method. If a route in the final configuration does not contain any obligatory customers (i.e., contains only potential customers), it is eliminated from the solution.

### 5.3. Experimental design

Simulation runs to evaluate the performance of the heuristics over time were performed for 16 different scenarios (see Table 1). Each problem set consisted of 200 customers uniformly distributed over a square centered at the warehouse. Initially, we tried several different configurations of the customers around the warehouse and found that the customers' configuration had a negligible effect on the performance of the algorithms.

For each scenario, we used a five period planning horizon (one week) and simulated the system for 200 weeks (about 4 years). The system was designed to service customers at the beginning of a period, before the beginning of their activity. In the case of a stock-out, the customer is immediately serviced by a special delivery, after which he continues with his activities. A graphical analysis of the system performance led us to believe that the system reached 'steady-state' well before 40 weeks; thus, we eliminated the first 40 weeks from each run. Therefore, we evaluated the system's performance from week 41 until week 200.

The daily consumption for each customer was obtained by generating a truncated normal distribution (consumption cannot be negative), with a fixed coefficient of variance of \((\sigma_i/\mu_i)\) of a quarter. In order to reduce variances when comparing two different methods, the same random number seeds were used for every scenario (see Law and Kelton, 1991).

We collected weekly values for the total cost (Table 1) as well as for each cost component (holding, transportation, fixed ordering, and stockout) (Table 2) in order to facilitate analysis of the heuristics. Before examining the results of the simulation experiment, we first note a few points relating to all the scenarios.

- For the hierarchical heuristic, \( H \), \( x_1 \) and \( x_2 \) were chosen as 0.75 and 0.90, respectively. These values were obtained through a two-dimensional search.
- Cost per mile traveled and truck rental were viewed as representing the same measure – transportation costs. Thus, they were always changed together in the same direction.
- Since we wished to focus on the algorithms and not the system design, identical customers were chosen.
- Truck capacity was defined in terms of customer capacity (essentially, the number of customers we can service under a zero-inventory ordering policy).

In Table 1, we can see the 16 scenarios which resulted from our two-level four factor full factorial experiment. The factors were the transportation costs \((m \text{ and } V)\), holding costs \((h_i)\), truck capacity \((K/K_i)\) and average demand as a function of tank size \((K_i/\lambda_i)\).

### 5.4. Experimental analysis

In Table 1 we present the weekly averages for our temporal heuristic \( \mathcal{T} \) and the hierarchical alternative \( \mathcal{H} \). In the final two columns we report the amount by which our model outperforms the alternative (DIF), as well as the standard deviation of this difference \( \sigma_{\text{DIF}} \).

#### Table 1

<table>
<thead>
<tr>
<th>Scenario</th>
<th>( m )</th>
<th>( V )</th>
<th>( h_i )</th>
<th>( K/K_i )</th>
<th>( K_i/\lambda_i )</th>
<th>( \mathcal{T} )</th>
<th>( \mathcal{H} )</th>
<th>DIF</th>
<th>( \sigma_{\text{DIF}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0.5</td>
<td>3</td>
<td>20</td>
<td>301</td>
<td>304</td>
<td>32</td>
<td>250</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>50</td>
<td>0.5</td>
<td>3</td>
<td>20</td>
<td>312</td>
<td>324</td>
<td>32</td>
<td>479</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>20</td>
<td>1.0</td>
<td>3</td>
<td>20</td>
<td>311</td>
<td>312</td>
<td>3</td>
<td>303</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>50</td>
<td>1.0</td>
<td>3</td>
<td>20</td>
<td>322</td>
<td>323</td>
<td>1</td>
<td>303</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>20</td>
<td>0.5</td>
<td>10</td>
<td>20</td>
<td>311</td>
<td>322</td>
<td>3</td>
<td>269</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>50</td>
<td>0.5</td>
<td>10</td>
<td>20</td>
<td>311</td>
<td>322</td>
<td>3</td>
<td>269</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>20</td>
<td>1.0</td>
<td>10</td>
<td>20</td>
<td>311</td>
<td>322</td>
<td>3</td>
<td>269</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>50</td>
<td>1.0</td>
<td>10</td>
<td>20</td>
<td>311</td>
<td>322</td>
<td>3</td>
<td>269</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>20</td>
<td>0.5</td>
<td>3</td>
<td>10</td>
<td>311</td>
<td>322</td>
<td>3</td>
<td>269</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>50</td>
<td>0.5</td>
<td>3</td>
<td>10</td>
<td>311</td>
<td>322</td>
<td>3</td>
<td>269</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>20</td>
<td>1.0</td>
<td>3</td>
<td>10</td>
<td>311</td>
<td>322</td>
<td>3</td>
<td>269</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>50</td>
<td>1.0</td>
<td>3</td>
<td>10</td>
<td>311</td>
<td>322</td>
<td>3</td>
<td>269</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>20</td>
<td>0.5</td>
<td>10</td>
<td>20</td>
<td>311</td>
<td>322</td>
<td>3</td>
<td>269</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>50</td>
<td>0.5</td>
<td>10</td>
<td>20</td>
<td>311</td>
<td>322</td>
<td>3</td>
<td>269</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>20</td>
<td>1.0</td>
<td>10</td>
<td>20</td>
<td>311</td>
<td>322</td>
<td>3</td>
<td>269</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>50</td>
<td>1.0</td>
<td>10</td>
<td>20</td>
<td>311</td>
<td>322</td>
<td>3</td>
<td>269</td>
</tr>
</tbody>
</table>
Table 2
Cost component comparisons (dollars per week)

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Holding DIF</th>
<th>σ_DIF</th>
<th>Transportation DIF</th>
<th>σ_DIF</th>
<th>Fixed ordering DIF</th>
<th>σ_DIF</th>
<th>Stockout DIF</th>
<th>σ_DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>529</td>
<td>169</td>
<td>124</td>
<td>51</td>
<td>41</td>
<td>28</td>
<td>1184</td>
<td>155</td>
</tr>
<tr>
<td>2</td>
<td>753</td>
<td>167</td>
<td>41</td>
<td>118</td>
<td>47</td>
<td>32</td>
<td>2404</td>
<td>399</td>
</tr>
<tr>
<td>3</td>
<td>616</td>
<td>261</td>
<td>202</td>
<td>38</td>
<td>22</td>
<td>24</td>
<td>1274</td>
<td>134</td>
</tr>
<tr>
<td>4</td>
<td>1006</td>
<td>401</td>
<td>202</td>
<td>116</td>
<td>32</td>
<td>32</td>
<td>2410</td>
<td>475</td>
</tr>
<tr>
<td>5</td>
<td>487</td>
<td>125</td>
<td>159</td>
<td>30</td>
<td>34</td>
<td>30</td>
<td>1501</td>
<td>190</td>
</tr>
<tr>
<td>6</td>
<td>671</td>
<td>130</td>
<td>5</td>
<td>73</td>
<td>43</td>
<td>29</td>
<td>2716</td>
<td>373</td>
</tr>
<tr>
<td>7</td>
<td>481</td>
<td>282</td>
<td>283</td>
<td>36</td>
<td>26</td>
<td>30</td>
<td>1872</td>
<td>188</td>
</tr>
<tr>
<td>8</td>
<td>921</td>
<td>220</td>
<td>309</td>
<td>70</td>
<td>36</td>
<td>26</td>
<td>3221</td>
<td>425</td>
</tr>
<tr>
<td>9</td>
<td>1080</td>
<td>90</td>
<td>436</td>
<td>50</td>
<td>157</td>
<td>31</td>
<td>4335</td>
<td>323</td>
</tr>
<tr>
<td>10</td>
<td>1490</td>
<td>97</td>
<td>757</td>
<td>109</td>
<td>204</td>
<td>35</td>
<td>10255</td>
<td>754</td>
</tr>
<tr>
<td>11</td>
<td>1642</td>
<td>174</td>
<td>496</td>
<td>64</td>
<td>112</td>
<td>38</td>
<td>4046</td>
<td>334</td>
</tr>
<tr>
<td>12</td>
<td>2172</td>
<td>185</td>
<td>838</td>
<td>109</td>
<td>145</td>
<td>33</td>
<td>6527</td>
<td>634</td>
</tr>
<tr>
<td>13</td>
<td>778</td>
<td>182</td>
<td>371</td>
<td>42</td>
<td>108</td>
<td>65</td>
<td>4537</td>
<td>370</td>
</tr>
<tr>
<td>14</td>
<td>1277</td>
<td>75</td>
<td>517</td>
<td>64</td>
<td>171</td>
<td>37</td>
<td>9693</td>
<td>678</td>
</tr>
<tr>
<td>15</td>
<td>1151</td>
<td>246</td>
<td>484</td>
<td>37</td>
<td>88</td>
<td>51</td>
<td>4921</td>
<td>452</td>
</tr>
<tr>
<td>16</td>
<td>1699</td>
<td>248</td>
<td>781</td>
<td>65</td>
<td>121</td>
<td>42</td>
<td>8699</td>
<td>779</td>
</tr>
</tbody>
</table>

In Table 2 we compare the results of each of the scenarios by examining the cost components: holding costs, transportation costs (truck rental costs and delivery costs), fixed ordering costs, and stockout costs (fixed stockout costs and special delivery costs). We again report the amount by which our heuristic outperforms the alternative, as well as the standard deviation of this difference.

Below we present our conclusions which are based on the results shown in Tables 1 and 2.

- We first and most importantly note that our temporal heuristic always outperforms the alternative.
- The difference between the two heuristics when measured in the number and cost of stockouts was large in all the scenarios. Note that these differences could be considerably reduced by using more conservative values for \( x_1 \) and \( x_2 \) in heuristic \( H \). However, increases in the other costs (especially holding costs) from such a change would more than outweigh the decrease in the stockout cost.
- For each cost component, the advantage of the temporal heuristic became more evident when the daily consumption was higher. We attribute this to the fact that no matter which cost component is considered, the timing of the supply decision becomes more crucial when service is more frequent.
- For higher transportation costs, the standard deviation of the difference always increases, even though the difference, itself, sometimes decreases. The same phenomena occurs with holding costs, but with reduced intensity.

6. Optimizing customer capacities

Until now, we have considered customer capacity as determined a priori, exogenously from detailed operational considerations (belonging to what we called the tactical level). In this section we examine the question of what should be the customer capacity when the temporal heuristic is used for distribution planning.

Recall that each time a customer is serviced, he is filled to capacity. Now, for the purpose of determining customer capacities, consider the demand to be deterministic. In this situation, a customer will follow policy \( \mathcal{A} \) of Fig. 1; in particular, the amount delivered (order quantity) will be equal to customer capacity. The decision of what this quantity should be can be made based on classical EOQ-type considerations.
Table 3
Cost comparison (dollars per week) – constant vs. optimized capacity

<table>
<thead>
<tr>
<th>Scenario 2</th>
<th>DIF</th>
<th>σ_DIF</th>
<th>Scenario 7</th>
<th>DIF</th>
<th>σ_DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holding</td>
<td>8814</td>
<td>148</td>
<td>17591</td>
<td>274</td>
<td></td>
</tr>
<tr>
<td>Transportation</td>
<td>(2045)</td>
<td>116</td>
<td>(844)</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>Fixed ordering</td>
<td>(888)</td>
<td>34</td>
<td>(130)</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Stockout</td>
<td>(3107)</td>
<td>281</td>
<td>(3324)</td>
<td>234</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2773</td>
<td>365</td>
<td>13292</td>
<td>366</td>
<td></td>
</tr>
</tbody>
</table>

(Harris, 1915). From the EOQ model we see that the optimal capacity for customer \( i \) is

\[
K_i^* = \sqrt{\frac{2z_i}{h_i}},
\]

where \( z_i \) represents the total order cost incurred when servicing customer \( i \). Note that \( z_i \) is the only unknown in Eq. (2) (\( \lambda_i \) and \( h_i \) are inputs). There are two components of the order cost \( z_i \). The first is the fixed ordering cost \( A_i \), incurred whenever servicing the customer (\( A_i \) is also an input). The second component is the cost associated with transportation. This cost can, in turn, be decomposed into the per mile transportation cost and the truck rental price. This component is much more difficult to evaluate for two reasons:

1. The distribution dynamics, and hence, the routing structure are characterized by high volatility.
2. Transportation costs have to be allocated among the customers.

To overcome the first point we made use of our simulation results. In particular, we used the weekly average total per mile transportation cost and the average number of times each customer was serviced in a week. We allocated the average total per mile transportation cost among customers in proportion to their service frequency and distance from the central warehouse. We did this by dividing the customers into five groups of equal size based on their proximity to the central warehouse. We then proportioned the per mile transportation costs to the customers in each group based on a factor which was larger for groups that were further from the central warehouse.

This method of allocating the per mile transportation costs to the customers is circular in the sense that it uses existing customer capacities in order to run the simulation whose aim is to find new capacities. One could repeat this process iteratively by running new simulations with the updated customer capacities and adjusting the capacities until there is no further gain, but as we see below, one iteration is enough to illustrate the power of this procedure.

For the truck rental cost, we assume that each time a customer is serviced, he is allocated the rental cost according to the proportion of his capacity to the truck capacity (i.e. \( VK_i/K \)).

We chose two scenarios, 2 and 7, which we use to illustrate the effectiveness of our procedure. In Table 3 we present the results of these two scenarios. Each system (with constant customer capacity and improved capacity levels) was simulated using the temporal heuristic. As before, DIF and \( \sigma_{DIF} \) indicate the mean and standard deviations of the improvement in the weekly average costs given by the procedure. As the table shows, the reduction in holding costs was very large, overshadowing the increases in the other costs. The table clearly shows that the use of the improved customer capacity levels led to lower total costs in both scenarios.

7. Conclusions

Our objective was to build and evaluate a heuristic for the Metered Inventory Routing Problem. We present here some of the main conclusions concerning our model.

1. The concept of temporal distances gives the solution method flexibility by allowing customers to be joined in the same route even when assigned to different periods. Thus, effort should be expended in developing efficient routes rather than on allocating customers to ‘good’ days which may eventually be changed.

2. In many situations the MIRP model better reflects the reality than the standard IRP, especially with the growth of reengineering concepts. In these situations, inventory holding, fixed ordering, and stockout costs all represent significant portions of the total cost.

3. The temporal heuristic is superior to the hierarchical approach and thus should be researched further.
Expressing stockout costs using stockout probabilities introduces a stochastic element into the model and was very effective.

Using our model to help determine customer capacity levels shows how an operational model can be used to support tactical decisions.

Viewing this work as an application of the temporal distance concept (Herer, 1996), we immediately see that an outgrowth of this work would be to apply the temporal distance concept to other inventory routing problems. In addition, a more extensive computational test and analysis of our algorithm would be desirable in order to obtain insights into the metered inventory routing problem.

References


