TRANSPORTATION SCIENCE
Vol. 38, No. 2, May 2004, pp. 210–223
ISSN 0041-1655 | EISSN 1526-5447 | 04 | 3802 | 0210

informs®
DOI 10.1287/trsc.1030.0042

Delivery Volume Optimization

Ann Melissa Campbell
Tippie College of Business, Management Sciences Department, University of Iowa, Iowa City, Iowa 52242,
ann-campbell@uiowa.edu

Martin W. P. Savelsbergh
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332-0205,
martin.savelsbergh@isye.gatech.edu

This work is motivated by the need to solve the inventory routing problem when implementing a business practice called vendor managed inventory replenishment. With vendor managed inventory replenishment, vendors monitor their customers' inventories, and decide when and how much inventory should be replenished at each customer. The inventory routing problem attempts to coordinate inventory replenishment and transportation in such a way that the cost is minimized over the long run. In this paper, we develop a linear time algorithm for determining a delivery schedule for a route, i.e., a given sequence of customer visits, that maximizes the total amount of product that is delivered on the route. This problem is not as easy as it may seem at first glance because of delivery windows at customers and the two dueling effects of increased inventory holding capacity at customers as time progresses and increased delivery times as more product is delivered at customers. Efficiently constructing such delivery schedules is important because it has to be done numerous times in insertion heuristics and local search procedures employed in solution approaches for the inventory routing problem.

Key words: logistics; inventory routing; variable delivery times; variable delivery quantitites
History: Received: June 2001; revisions received: June 2002, December 2002; accepted: December 2002.

Recently the business practice called *vendor managed inventory replenishment* (VMI) has been adopted by many companies. VMI refers to the situation in which a vendor monitors the inventory levels at its customers and decides when and how much inventory to replenish at each customer. This contrasts with conventional inventory management, in which customers monitor their own inventory levels and place orders when they think that it is the appropriate time to reorder. VMI has several advantages over conventional inventory management. Vendors can usually obtain a more uniform utilization of production resources, which leads to reduced production and inventory holding costs. Similarly, vendors can often obtain a more uniform utilization of transportation resources, which in turn leads to reduced transportation costs. Furthermore, additional savings in transportation costs may be obtained by increasing the use of low-cost full-truckload shipments and decreasing the use of high-cost less-than-truckload shipments, and by using more efficient routes by coordinating the replenishment at customers close to each other.

VMI also has advantages for customers. Service levels may increase, measured in terms of reliability of product availability, because vendors can use the information that they collect on the inventory levels at the customers to better anticipate future demand and to proactively smooth peaks in the demand. Also, customers do not have to devote as many resources to monitoring their inventory levels and placing orders, as long as the vendor is successful in earning and maintaining the trust of the customers.

Our work on this problem was motivated by our collaboration with a producer and distributor of air products. The company operates plants worldwide and produces a variety of air products, such as liquid nitrogen, oxygen, and argon. The company's bulk customers have their own storage tanks at their sites, which are replenished by tanker trucks under the supplier's control.

Under VMI, the vendor typically manages a fleet of vehicles to transport the product to the customers. The objective of the vendor is to coordinate the inventory replenishment and transportation in such a way that the total cost is minimized over the long run. The problem of optimal coordination of inventory replenishment and transportation is called the *inventory routing problem* (IRP). Most of the methodologies developed for solving the IRP are deterministic and proceed in two phases. In the first phase, it is decided which customers are visited in the next few days, and a target amount of product to be delivered to these customers is set. In the second phase, vehicle routes are determined taking into account vehicle capacities, customer delivery windows, driver restrictions, etc. Surveys of algorithms used for the IRP can be found in Nori (1999), Campbell et al. (1998), and Kleywegt et al. (2002).

In any delivery schedule, there is usually some flexibility in the sense that routes may be started earlier or later without affecting feasibility, which can be exploited to increase the amount of product delivered. Note that even though there is a target amount of product to be delivered, in vendor-managed inventory resupply environments, we have the flexibility to deliver more than the target amounts. Therefore, if we consider the target amount as a minimum delivery quantity to each customer rather than a fixed delivery quantity, we may be able to use the flexibility in the schedule to increase the amount of product delivered on each route. It is clearly desirable from the vendor's perspective to deliver a larger percentage of the vehicle capacity, and it is desirable from a customer's perspective because it creates better insurance against running out of product.

Because the maximum quantity that can be delivered at a customer is dependent on the time of delivery, the selection of actual delivery times (between the earliest and latest delivery times) will affect the total volume deliverable on a trip. Because the customer consumes product over time, the later the truck arrives, the more inventory holding capacity there is and the more product can be delivered. On the other hand, delivery is not instantaneous but is dependent on the size of the delivery. Furthermore, the later the truck arrives at a customer, the less time may be available for making a delivery because of the delivery time restrictions of customers to be visited later on the trip. These two dueling effects make determining an optimal delivery schedule for given a sequence of customer visits on a trip not as easy as it may seem at first glance. In this paper, we develop an algorithm for this problem, which we call the *delivery volume optimization* (DVO) problem, which runs in linear time. Efficiently constructing such delivery schedules is important because it has to be done numerous times in insertion heuristics and local search procedures employed in solution approaches for the second-phase problem mentioned above.

The remainder of the paper is organized as follows. We start by introducing the concept of delivery volume profiles, which help in visualizing and analyzing the problem. Next, we introduce the simplest version of the problem and describe the basic algorithm for optimizing the delivery volume. In later sections, we introduce various extensions based on complexities encountered in practical applications and discuss how the basic algorithm must adjust to handle these variations. Finally, we present computational experiments that demonstrate the value of embedding delivery volume optimization in algorithms for the inventory routing problem.

1. Delivery Volume Profile

As indicated earlier, customer usage rates and volume dependent delivery times can have a conflicting relationship when it comes to determining the maximum quantity deliverable at a customer. A customer's delivery volume profile is the graph representing the maximum delivery quantity at the customer as a function of the time of delivery. These delivery volume profiles play a critical role in understanding and analyzing the problem as well as the algorithms.

Because we are working with a given route, i.e., a fixed sequence of customer visits, we will refer to a customer by its position on a route. Thus, the ith customer on the route will consume product at a constant rate of u_i per time unit. This customer will have a capacity of C_i and an inventory of I_i^0 at the start of the planning period.

Additionally, we will use Q to represent the vehicle's capacity, and R to represent the vehicle's delivery rate, which is the amount of product that can delivered per time unit. This implies that, for now, we assume that the delivery rate is the same for all customers. (Note that we must have $u_i \leq R$, because otherwise we cannot deliver fast enough to keep up with usage.) Finally, we assume that customer i has a delivery window specifying that a delivery can begin anytime between e_i and l_i and that the target delivery quantity, i.e., minimum delivery quantity, is $q_i^{\min}$.

We will call the points where the slope of the delivery volume profile changes the *breakpoints*. For these points, we will use the notation t_i^k and q_i^k to denote the associated delivery times and maximum delivery quantities, where $k = 1, 2, \ldots$ is the index of the breakpoint.

Before proceeding, there is one other assumption that should be stated. We assume that the delivery windows of the customers are such that

$$l_i \leq l_{i+1} - \frac{q_i^{\min}}{R} - d_{i,i+1}, \qquad (1)$$

$$e_{i-1} + \frac{q_{i-1}^{\min}}{R} + d_{i-1,i} \leq e_i, \qquad (2)$$

where $d_{j,k}$ represents the travel time from j to k. This guarantees that there will be at least one feasible delivery schedule. Starting delivery within the window, delivering the minimum quantity, and traveling directly to the next customer will yield a feasible solution.

The beginning of a customer's delivery window will be the first breakpoint of the delivery volume profile, i.e., $t_i^1 = e_i$. The maximum delivery quantity at this point, q_i^1, is the minimum of three values:

- the volume that can fit at customer i at the time of the breakpoint, i.e.,

$$C_i - I_i^0 + t_i^1 u_i; \tag{3}$$

- the volume that can feasibly be delivered at customer i given the time available for delivery, i.e.,

$$(l_{i+1} - (t_i^1 + d_{i,\,i+1}))R; \tag{4}$$

- the amount of product that is available in the truck for delivery at customer i, i.e.,

$$Q - \sum_{j\in r,\, j\neq i} q_j^{\min}. \tag{5}$$

If the maximum delivery quantity q_i^1 is determined by the first of these three values, then the volume deliverable to i will increase linearly with time due to the usage rate until a limit is reached (the next breakpoint). This limit will occur at a time that is the earliest of the following three times:

- the time t^α when the maximum amount of product in the vehicle that can be delivered to customer i will fit into customer i's storage facility

$$t^\alpha = t_i^1 + \frac{Q - \sum_{j\in r,\, j\neq i} q_j^{\min} - q_i^1}{u_i}; \tag{6}$$

- the time t^β when the volume deliverable to the customer starts to decrease because there is not enough time to deliver more and make it to the next customer by the latest time delivery can start. The change in slope is caused due to the restrictions imposed by the next customer:

$$t^\beta + \frac{C_i - I_i^0 + t^\beta u_i}{R} + d_{i,\,i+1} = l_{i+1}, \tag{7}$$

$$t^\beta + \frac{t^\beta u_i}{R} = l_{i+1} - \frac{C_i - I_i^0}{R} - d_{i,\,i+1}, \tag{8}$$

$$t^\beta = \frac{l_{i+1} - (C_i - I_i^0)/R - d_{i,\,i+1}}{1 + u_i/R}. \tag{9}$$

- l_i is reached. We assume that l_i is at or before the time the customer's inventory reaches zero, so in finding breakpoints we do not have to consider where the volume deliverable to the customer equals C_i.

The earliest of these three times will be t_i^2, and the corresponding delivery quantity will be q_i^2. Note that this delivery quantity will be the maximum volume deliverable for customer i.

If the last breakpoint occurred when the maximum amount of product deliverable is reached, as in Equation (5) or (6), the delivery volume will remain at this value until either there is not enough time left to deliver this volume and reach the next destination by the end of its delivery window or when l_i is reached.

If the last breakpoint was determined by the delivery time available, as dictated by Equation (4) or (7), then this forces the maximum delivery quantity to decline from then on, because there is less time available to deliver product at i and reach the next customer by l_{i+1}. The rate of decline is determined by the delivery rate R, and the graph will slope downward until l_i.

The last breakpoint will always occur at l_i, by definition, and the quantity deliverable at this time will again be based on the volume that can fit at the customer, the remaining truck capacity, and the time available to make the delivery. Note, though, that if the volume deliverable at the prior breakpoint was determined by time available, this one will be as well.

Therefore, the maximum number of breakpoints for a single customer will be four: the start of delivery window (1), the earliest point in time where the maximum quantity deliverable can be delivered (2), the earliest point in time where the maximum quantity deliverable will start to decline because limited time is available for delivery (3), and the end of the delivery window (4). An example of a delivery volume profile is given in Figure 1.

It is key to realize that if any of the profiles for the customers on a route has all four of these breakpoints, the delivery volume optimization problem is trivial. The reason is that if customer k, for example, has four breakpoints in its profile, then this means that at one point in time the volume deliverable to customer k is equal to the truck capacity minus the minimum delivery quantities required at the remaining customers. All we need to do, in that case, is to set the delivery time and quantity for customer k equal to t_i^2 and q_i^2, respectively, and deliver the minimum to the remaining customers. This would yield a total delivery volume equal to Q, the truck's capacity, which is clearly optimal.

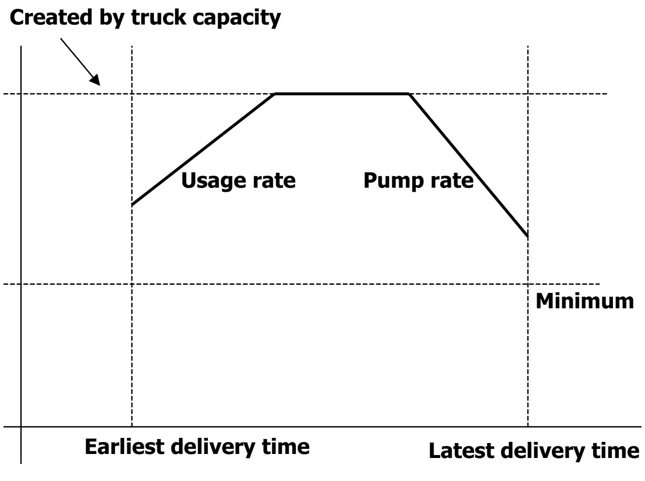

Figure 1 General Customer Delivery Volume Profile

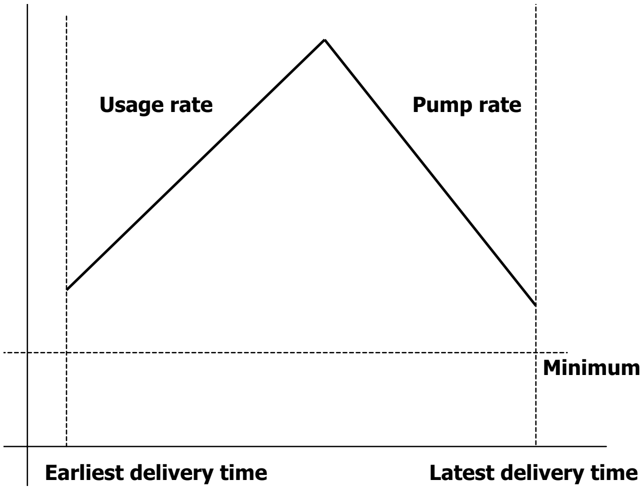

Figure 2 Customer Delivery Volume Profile with Rise and Decline

Therefore, we switch our focus to situations where all customer profiles have fewer than four breakpoints. In that case, all customers have profiles that resemble the ones depicted in Figures 2, 3, and 4.

In Figure 2, the volume deliverable increases because of customer usage and then declines because of delivery time available. Truck capacity does not affect the volume deliverable.

In Figure 3 the delivery volume never declines because there is always enough time available for delivery. This happens, for example, when there is significant waiting time between customers i and $i+1$.

In Figure 4 the maximum volume deliverable to a customer is at the start of its delivery window. Delivery time available restricts the volume deliverable over the entire window, usually indicative of a narrow delivery window for customer $i+1$ or large inventory holding capacity at customer i.

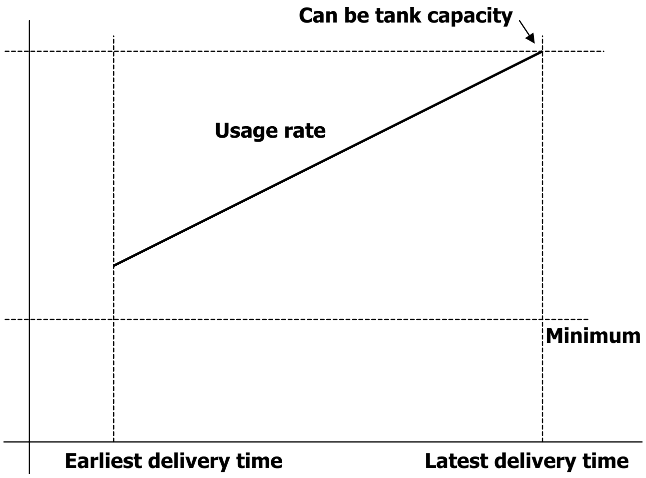

Figure 3 Customer Delivery Volume Profile—Incline Only

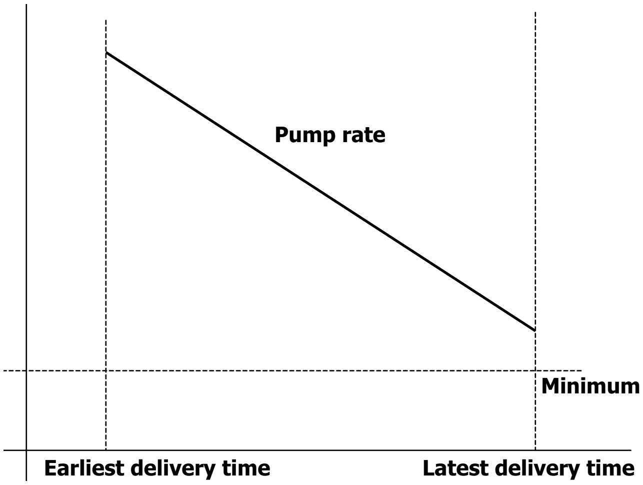

Figure 4 Customer Delivery Volume Profile—Decline Only

2. Developing Algorithms

We use the delivery volume profiles for the customers on a route to illustrate ideas for an optimization algorithm.

When we consider a single customer route, visiting customer i, the total delivery volume on the route is maximized by starting delivery to i at the breakpoint corresponding with the highest q_i value. This quantity will be referred to as the peak, and we will refer to the time it occurs as t_i^{peak}. When there are more customers in a route, though, it is not as clear how to maximize the total delivery volume. Consider the example in Figure 5.

Figure 5 shows three graphs that represent the delivery volume profiles for consecutive customers on a three customer route. The left profile is for the first customer on the route, the center profile is for the second, and the right profile is for the third. The arrows at the top of the figure and the accompanying numbers represent the travel times. All delivery volume profiles displayed in this paper, unless otherwise specified, will assume a delivery rate of 1. Thus, a delivery of 12.5 to customer 1 starting at time 12.5 dictates a completed delivery at customer 1 at time $12.5+12.5=25$ and arrival at customer 2 at $25+5=30$.

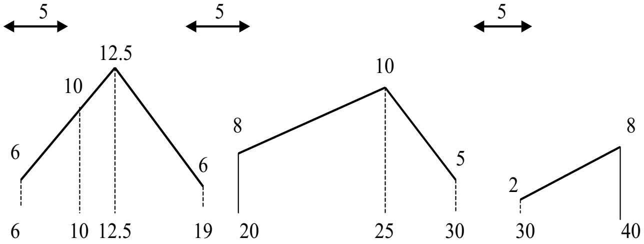

Figure 5 Three-Customer Delivery Volume Optimization

Let us examine various initial ideas to construct a delivery schedule that maximizes the total volume delivered on the route.

- Can we deliver the peak volume to all of the customers on the route?

From Figure 5, it is clear that it is not always possible to deliver the peak volume to all customers on a route. In Figure 5, if we try to deliver at the peak for customer 1 (12.5) this implies an arrival at customer 2 at time 30, which corresponds to a maximum delivery volume of 5 rather than 10. Because the peak for the last customer is always at the end of its window, the total delivery quantity will be $12.5+5+8=25.5$.

In fact, because customer 1's profile declines because of time available, delivering at its peak corresponds to arriving at the next customer at its latest time. If this latest time is not the peak for the second customer, delivering at the peak for both is clearly not possible.

- Because customer usage dictates that the longer we wait, the more inventory holding capacity will be available at the customers, should we try to deliver at the latest time possible at all customers?

Waiting to make deliveries as late as possible is popular in vehicle routing literature because it minimizes waiting time, but it is not the best idea here because of the delivery time associated with making deliveries. Waiting until the latest time does indeed create a situation where there is the largest amount of capacity available at the customers on a route, but it may not allow as much time for the actual delivery as if we started delivering earlier. Delivering at the latest time allows for a delivery quantity of only 6 at customer 1, 5 at customer 2, but still 8 at customer 3, yielding a total delivery volume of only 19.

- Because an extra unit of time allows the largest increase in volume deliverable at the customer with the largest usage rate, does it make sense to construct a delivery schedule by looking first at the customer with the highest usage rate? What if we set the delivery volume/time to be at the peak for this customer and work from there?

For the example given in Figure 5, this results in a maximum total delivery volume of 25.5. Note that the usage rate for customer 1 is $(12.5-6)/(12.5-6)=1$, the usage rate for customer 2 is $(10-8)/(25-20)=2/5$, and the usage rate for customer 3 is $(8-2)/(40-30)=3/5$.

On the other hand, if we deliver 10 units at time 10 to customer 1, 10 units at time 25 for customer 2, and 8 units for customer 3 at time 40, the total delivery volume is 28, which in fact is optimal. In the next section, we will discuss the optimal policy that resulted in this delivery schedule.

3. Optimal Policy

Theorem 1. *It is optimal to deliver the peak quantity at the peak time for the last customer, and then to deliver the maximum quantity possible at preceding customers subject to the restriction that it is possible to complete the delivery and travel to the succeeding customer in time to start the delivery at its specified time.* (*A more precise description is given in Algorithm* 1.)

Algorithm 1: Optimal Policy.

```
$\bar{t}_n = l_n$
$\bar{q}_n = \min\{(C_n - I_n^0) + \bar{t}_n u_n,\ Q - \sum_{k=1,\dots,n-1} q_k^{\min}\}$
for $j = n-1, \dots, 1$ do
    $\bar{t}_j = \min\left\{l_j,\ \frac{\bar{t}_{j+1} - d_{j,j+1} - ((C_j - I_j^0)/R)}{1 + u_j/R}\right\}$
    $\bar{q}_j = \min\{(C_j - I_j^0) + \bar{t}_j u_j,\ Q - \sum_{k=1,\dots,j-1} q_k^{\min} - \sum_{k=j+1,\dots,n} \bar{q}_k\}$
end for
```

Proof. For ease of presentation, we will first discuss the case where customer capacities sum up to less than or equal to the truck's capacity. This guarantees all profiles will have no more than three breakpoints and no delivery quantities will be truncated because of product availability.

Case 1. $\sum_{i=1..n} C_i \leq Q$.

The policy may produce a delivery schedule that includes waiting time at customers. Waiting time at a customer j occurs when the start of delivery at $j-1$ plus the delivery time at $j-1$ plus the travel time from $j-1$ to j is less than the start of delivery at j. We will refer to a maximal set of consecutive deliveries without intermediate waiting time as a section. We first analyze the effectiveness of the policy on a section. After that, we will consider the effectiveness of the policy on the route as a whole.

Note that the delivery volume profile of the last customer in a section will look like the one in Figure 3. The maximum delivery quantity will increase until the end of the delivery window. The end of the delivery window is determined by either some customer-specific restriction or by the fact that the customer will run out of product at that point in time. Waiting time between segments indicates that there is no restriction on the time available for making a delivery at the last customer of a segment by its successor.

We prove that we deliver the maximum delivery volume to a section by induction.

When there is only one customer in a section, the proof of optimality is trivial because the policy delivers the peak quantity at the peak time, which is clearly optimal.

When there are two customers in a section, the proof of optimality is not much more difficult. Because the peak quantity at the last (second) customer is deliverable at the end of its delivery window, the last customer does not really limit the options at its predecessor (the first customer). The delivery volume profile for the first customer already ensures that it is possible to reach the last customer at or before the end of its delivery window. Thus, we can deliver the peak volume at the peak time for the first customer as well on a two customer route. We clearly cannot do better than this.

Now, assume that the policy leads to the optimal delivery volume for the last k customers in a section, say $j+1,\ldots,n$, and consider customer j. If we begin delivery at $j+1$ at time $\bar{t}_{j+1}$, the policy dictates that the start of a delivery at j must be such that it is possible to complete the delivery and to travel to $j+1$ and arrive at $j+1$ at or before $\bar{t}_{j+1}$. To determine the start of a delivery at j that gets us to $j+1$ at exactly $\bar{t}_{j+1}$, we have to solve

$$\bar{t}_j + \frac{(C_j - I_j^0) + \bar{t}_j u_j}{R} + d_{j,\,j+1} = \bar{t}_{j+1}.$$

Because there is no waiting time at any customer in the section, $\bar{t}_j$ will correspond to a point on the delivery volume profile before t_j^{peak}, and it enables a delivery quantity of $\bar{q}_j = (C_j - I_j^0) + \bar{t}_j u_j$, which is the most possible at customer j given a start time of $\bar{t}_j$.

Adding customer j does not impact the delivery times (and delivery quantities) of customers $j+1,\ldots,n$. For the policy not to be optimal, it has to be possible to deliver more to $j, j+1,\ldots,n$ by delivering more to j. The only way to deliver more to j is to start the delivery to customer j later. Delaying the start of delivery at j by Δ, where Δ is small, allows the delivery quantity to increase by Δu_j. Delivering more increases the actual delivery time, which in turn delays arrival at $j+1$ by $\Delta + u_j\Delta/R$. This inevitably leads to a loss in total delivery time for the rest of the section of $\Delta + u_j\Delta/R$, which equates to a loss of total delivery volume of $(\Delta + u_j\Delta/R)R$ or, rather, $\Delta R +$ *earlier increase*. This loss occurs because all the time up until the end of the window for the last customer of the section is currently being used for traveling and delivering product. We cannot arrive at the last customer any later than in the current solution, so if we delay an earlier delivery, this leads to a reduction in the time available for delivering product. A delay of Δ at j yields a net reduction of ΔR in the volume deliverable. Thus, delivering more at the first customer than dictated by the policy does not yield an overall improvement.

This demonstrates that the policy is optimal for the sections of the route. It follows relatively easily that if each of the sections themselves is indeed optimal, in terms of delivering the maximum total volume to the customers in the section, then the whole route is optimal as well. If there is waiting time between sections, this implies that the optimal decisions for one section do not restrict the decisions made for the other sections.

Case 2. $\sum_{i=1..n} C_i \geq Q$.

When the sum of customer capacities is more than the truck capacity, the peak of a customer's delivery profile may be the result of available truck capacity. If any of the deliveries on the route are limited by available truck capacity, then (as we discussed earlier) we know we can deliver a full truckload to the customers on the route, and the problem is trivial. On the other hand, the fact that the sum of customer capacities is more than the truck capacity does not imply that one of the customer delivery profiles will reach its peak because of available truck capacity. It is even possible that the maximum volume deliverable on the route may be strictly less than the truck capacity. Furthermore, if the sum of the customer capacities exceeds the truck capacity, but all delivery profiles have at most three breakpoints, then to be able to deliver a full truckload on the route, more than one customer must receive a delivery above its minimum.

In any case, when the sum of customer capacities is more than the truck capacity, we may find that when we are applying the policy we are not able to deliver as much as the profile indicates to some customer j because of product availability. We will need to always determine which amount is smaller:

- the quantity dictated by the profile (based on available capacity and time available),
- the product availability $Q - \sum_{k=1,\ldots,j-1} q_k^{\min} - \sum_{j=j+1,\ldots,n} \bar{q}_k$.

If product availability is smaller than the volume dictated by the profile, we will deliver this smaller volume. All customers preceding j will receive their minimum delivery quantity. Furthermore, if this situation is encountered, it implies that a full truckload will be delivered on the route, which is clearly optimal. On the other hand, if this situation never occurs, then we are in the exact same situation as in Case 1, which we have already shown to be optimal. □

For each customer, we compute a maximum of three breakpoints, and each breakpoint requires no more than three computations. This yields a total of at most $9n$ computations for the breakpoints on the route. Given the profiles, we can determine the best time to begin delivery at each customer. This is either the peak or the point on the delivery volume profile corresponding to arrival at the next customer at its delivery time. Each check requires one computation. There is an exception for the last customer who always begins delivery at its l_n value. This yields a total of $9n + n = 10n$ computations, which is linear in n, to optimize the delivery quantity over a route.

4. Route Duration Constraints

As mentioned in the introduction, our work on DVO and more generally on IRP, was motivated by our collaboration with a producer and distributor of air products. One of the key practical complexities that had to be dealt with in that context was route duration constraints. Route duration constraints come about because drivers must obey Department of Transportation (DOT) restrictions governing how long they can drive and how long they can be on duty.

DVO capitalizes on the flexibility within a delivery schedule to maximize total delivery volume on a route. However, the "optimized" delivery schedule may have an increased duration, which could lead to infeasibility. In this section, we discuss the modifications necessary to properly handle route duration constraints.

We will let D represent the maximum duration of the route, and e_r^{start}, e_r^{end}, l_r^{start}, and l_r^{end} denote the earliest feasible start time, the earliest feasible end time, the latest feasible start time, and the latest feasible end time of route r, respectively. These values follow from the delivery windows of the first and last customers on the route, the travel times to the depot, and the minimum delivery quantities. We assume the following conditions hold for feasibility:

$$e_r^{start} \geq e_r^{end} - D, \tag{10}$$

$$l_r^{end} \leq l_r^{start} + D. \tag{11}$$

To adapt the optimal policy from the previous section to obtain a feasible solution with regard to route duration constraints, we must consider two additional values as we set the time and quantities from customers n to 1. The values are:

- How much time is spent on deliveries to customers succeeding the customer currently under consideration, e.g., waiting time, travel time, and delivery time. Note that this value can be computed easily as we apply our policy.
- How much time is still needed for making minimum deliveries to customers preceding the customer currently under consideration. Note that this value can be computed easily in advance because it involves travel times and time required to deliver the minimum quantity.

If the policy described in the previous section prescribes a start time of a delivery at i that would lead to a violation of the maximum duration constraint, as determined by the delivery time at i combined with the delivery time at preceeding and succeeding customers, we can adjust the start of delivery at i and the associated delivery quantity appropriately. Delivering a smaller quantity to i at a later time ensures that all preceding customers can still be visited within their delivery window and receive their minimum delivery quantity. Depending on the characteristics of the delivery schedule produced, we may need to re-run the algorithm with an earlier delivery time for the last customer in order to deliver the maximum volume possible on the route.

Example. The graphs in Figure 6 represent delivery volume profiles for three customers on a route with a maximum duration of 44. The latest end time for the route, given the delivery window [1, 10] for customer 1 and the maximum duration of 44, is time 49 (start route at time 5, travel to customer 1, and start delivery at time 10 (the latest delivery start time possible) plus the remaining, as of yet unused, portion of the maximum duration, i.e., 39). Assuming a delivery rate of 1, we can deliver the peak quantity at customer 3, which is 6, at time 41. Note that because of the latest end time of the route, the delivery volume profile of the last customer no longer has the form of Figure 3. Consequently, we need to leave customer 2 by $41 - 5 = 36$. Because starting delivery of the peak quantity of 5 at customer 2 at the peak time of 17 leads to a completed delivery at time 22, there will be waiting time before the start of the delivery at customer 3. If we start the delivery at customer 2 at 17, then we must leave customer 1 by time 12. Because of the

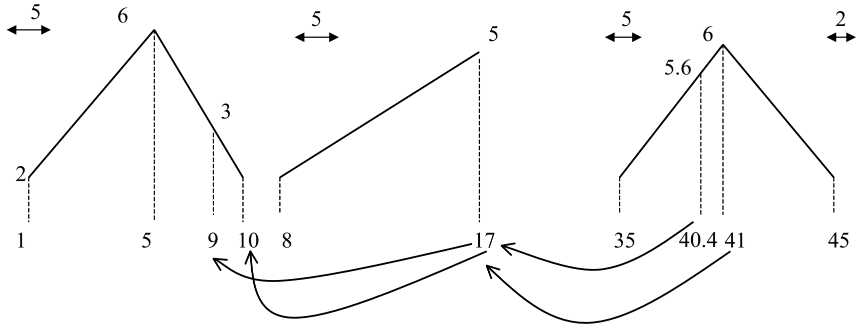

Figure 6 **Example of Instance with Wait**

maximum duration of 44, the remaining time available at customer 1 is $44-(49-12)$ (the time already allocated) $-\,5$ (travel time from depot) $=2$. Thus, we only have time to deliver two units of product at customer 1. The total delivery volume is $2+5+6=13$.

Now consider completing the route at 48, i.e., one time unit earlier. The delivery start time at customer 3 that maximizes the quantity deliverable while completing the route by time 48 is 40.4 with a delivery quantity of 5.6. Starting delivery at 40.4 at customer 3 still allows delivery of five units at the peak time at customer 2. However, at customer 1, we now have time to deliver three units of product $(44-(48-12)-5=3)$. Therefore, we start delivery at customer 1 at time 9, resulting in a total delivery volume of $3+5+5.6=13.6$, which is an improvement over 13!

The example illustrates that when the route contains waiting time, it is possible to trade waiting time for delivery time. In the above example, we have traded 0.6 units of waiting at customer 3 for 0.6 units of delivery time (one more unit of product delivery at customer 1 and 0.4 fewer units of product delivery at customer 3).

Note that the duration of a route is taken up by travel, waiting, and product delivery. Therefore, minimizing the waiting time is equivalent to maximizing the volume deliverable. Note that it may not be possible to eliminate all waiting time due to the delivery windows.

In the example above, completing the route δ units earlier (δ small) can be accomplished by starting delivery at customer n, the last customer on the route, $(1+u_n/R)^{-1}\delta$ units earlier. To see this, realize that starting $(1+u_n/R)^{-1}\delta$ units earlier reduces the available storage capacity by $(1+u_n/R)^{-1}u_n\delta$, which in turn reduces the product delivery time by $(1+u_n/R)^{-1}(u_n/R)\delta$. Therefore, the completion of the route is reduced by

$$\frac{1}{1+u_n/R}\delta+\frac{1}{1+u_n/R}\frac{u_n}{R}\delta=\delta,$$

i.e., start time reduction plus product delivery time reduction. Consequently, we will deliver $(1+u_n/R)^{-1}\delta u_n$ less product at customer n. On the other hand, we can directly convert the δ units of time to product delivery of $R\delta$ for customers in the first segment. Because $R\delta>(1+u_n/R)^{-1}\delta u_n$, the total delivery volume of the route will increase.

If there are several customers in the last segment, then completing the route δ units earlier causes a reduction in the volume deliverable as well as a change in delivery start time for all of the customers in the last segment. The total reduction in the volume deliverable, though, is always less than $R\delta$, because $u_i\le R$ for all i. For example, if there are two deliveries after the wait, then the reduction in volume is

$$\frac{\delta u_n}{1+u_n/R}+\frac{\delta u_{n-1}}{(1+u_n/R)(1+u_{n-1}/R)}\le\frac{\delta R}{2}+\frac{\delta R}{4}\le\delta R.$$

Next, we consider computing the optimum value of δ, i.e., the value of δ that results in the maximum increase in volume deliverable on the route. Three bounds on the maximum value of δ have to be computed:

- We have to make sure that the schedule remains feasible. It suffices to ensure that we start delivery at customer n no earlier than e_n. Therefore, we have

$$\bar{t}_r^{end}-e_r^{end} \tag{12}$$

We also have to recognize the situations in which increasing δ cannot lead to additional improvements. There are two cases to consider:

- The maximum amount of waiting time is removed from the route. As mentioned earlier, it may not be possible, due to delivery windows at customers, to remove all waiting time. We can compute the waiting time that is unavoidable in any schedule as follows. Start delivery at the last customer at e_n. Working backward, deliver the maximum quantity possible at each customer, always making sure that you arrive at the next customer in time to make its planned delivery. (Note that delivering less than the maximum quantity possible means starting delivery earlier and spending less time on delivering product, which leads to an earlier arrival at the next customer, which can only increase waiting time.) Let w be the total waiting time in this computed schedule, then we have

$$\delta\le w.$$

It is important to realize that there may exist schedules in which delivery at the last customer starts later than e_n but the schedule contains only unavoidable waiting time.

- The maximum route duration will no longer be restrictive for selecting the delivery start time for the customers of the first segment of the route. Note that when our policy creates a delivery schedule that includes waiting time, time available never restricts the selection of the delivery quantity and the delivery start time for customers not in the first segment. Working backward from the last customer in the first segment, selecting optimal delivery quantities and delivery start times ignoring any time restrictions, we can determine a desired delivery start time of the route, say $\hat{t}_r^{start}$. There is no reason to increase δ further if we know it is possible to commence delivery at the first customer at the desired delivery time. Therefore, we have

$$\delta\le\bar{t}_r^{start}-\hat{t}_r^{start}. \tag{13}$$

If the value of δ is determined by the first or second of the three values listed above, then we have increased the total volume deliverable on the route until it is no longer feasible to reduce waiting on the route any further. The total delivery volume is optimal because all time other than travel and waiting time is spent on delivering product.

If the value of δ is determined by the third of the three values listed above, then the maximum duration no longer restricts our policy, and it will produce the maximum total delivery volume for the route.

5. Customer-Specific Delivery Rates

In our study of the delivery of industrial gases, it is not always the case that each vehicle has pumping equipment. Some vehicles do not and thus can make deliveries only to customers that have pumps at their tanks. This introduces differences in the delivery rates at the customers. Customer-specific delivery rates are not limited to industrial gas delivery. In many industries, the delivery rate is influenced more by the customer than by the vendor. Inventory check-in procedures, distances from the truck to the delivery points, and quality inspections are all ways a customer affects the rate at which the product is delivered. In this section, we discuss the necessary modifications to the basic policy to handle customer-specific delivery rates.

First, consider how customer-specific delivery rates impact the delivery profiles. When delivery profiles decline because of time available, it will now be with a wide range of slopes. Also, because of the change in delivery rate, the point in time the delivery profile will begin to decline will change, too. As the delivery rate increases, the peak will be later, because less time is needed to deliver the same amount of product. As the delivery rate decreases, the peak will be earlier because more time is needed to deliver the same amount of product.

A high delivery rate at an early customer can now make it advantageous to reduce delivery time, and thus delivery quantity, at later customers in exchange for the much larger quantity that can now be delivered earlier. With varying delivery rates, the decisions at later customers are now impacted by the earlier customers, making the optimal delivery quantities and delivery times harder to determine.

To better understand the impact of customer specific delivery rates, consider any two consecutive customers $i-1$ and i on an n customer route. In applying our basic policy, a delivery at i at its peak dictates a delivery to $i-1$ at or before its peak. Recall that, when there is no unavoidable wait between these two customers, the delivery at $i-1$ will begin at or before its peak because a delay of Δ at $i-1$ leads to an increase in delivery quantity of Δu_j at $i-1$ but a decrease in delivery quantity of $\Delta R + \textit{earlier increase}$ at i.

Consider how this changes when the delivery rates are customer specific. It is better to deliver more at customer $i-1$ if the increase at $i-1$ is more than the decrease at i:

$$u_{i-1} > R_i\left(1+\frac{u_{i-1}}{R_{i-1}}\right). \tag{14}$$

Rearranging,

$$R_i < \frac{u_{i-1}R_{i-1}}{u_{i-1}+R_{i-1}}. \tag{15}$$

If all delivery rates are the same, this inequality clearly will never hold. However, if the delivery rate at $i-1$ is sufficiently faster than at i, the inequality will hold, and it is better to deliver to $i-1$ at its peak and shift the delivery start at i to l_i. On the other hand, if the inequality does not hold, it is better to deliver to i and $i-1$ as in the original policy.

Let $v_R(i) = (u_iR_i)/(u_i+R_i)$. The above example demonstrates that to decide on the start of delivery at customer k and the amount to deliver at customer k it is necessary to consider preceding customers with $v_R(j) > v_R(k)$. In fact, it is in our best interest to allow a large quantity to be delivered at j even if it means postponing the start of delivery and reducing the quantity delivered at customers after customer j.

We will show that it suffices to consider only the last customer on the route preceding k with $v_R(j) > v_R(k)$ in making the decision of when to begin delivery at k.

We can always deliver the maximum quantity to customer n, since his maximum always occurs at l_n. This creates t_n^* and q_n^*. Therefore, we effectively start the algorithm at customer $n-1$.

Find the last customer prior to $n-1$ such that $v_R(j) > v_R(n-1)$. First, we observe that it is feasible to deliver the peak quantity at the peak time to customer j (unless truck capacity has been reached, which means we have an optimal solution). We set the delivery time and delivery quantity at j equal to the peak time t_j^* and peak quantity q_j^*. Given t_j^* and q_j^*, we can compute the earliest arrival time at $n-1$, say t'_{n-1}, considering the travel time between each customer in between and the time associated with delivering the minimum (or committed) quantity, i.e.,

$$t'_{j+1} = \max\left(t_j^* + \frac{q_j^*}{R_j} + d_{j,\,j+1},\, e_{j+1}\right)$$
$$\cdots$$
$$t'_{n-1} = \max\left(t'_{n-2} + \frac{q_{n-2}^{\min}}{R_{n-2}} + d_{n-2,\,n-1},\, e_{n-1}\right).$$

Thus, t'_{n-1} represents the earliest possible arrival at $n-1$ after delivering the most at j. If t'_{n-1} occurs prior to the peak delivery time at $n-1$, we can feasibly deliver at the peak time for both j and $n-1$. In that case, we set $t^*_{n-1} = t^{peak}_{n-1}$ and $q^*_{n-1} = q^{peak}_{n-1}$. On the other hand, if t'_{n-1} occurs after the peak delivery time at $n-1$, we cannot feasible deliver at the peak time but have to "settle" for the maximum quantity deliverable at t'_{n-1}. Therefore, we set $t^*_{n-1} = t'_{n-1}$ and q^*_{n-1} accordingly.

Before discussing how to continue, we will argue that it suffices to consider the last customer j prior to $n-1$ with $v_R(j) > v_R(k)$, i.e., that we do not have to be concerned about a customer j' prior to j with $v_R(j') > v_R(j) > v_R(k)$.

Consider customer j. The peak for j occurs at or before l_j. If the peak for j occurs before l_j, then delivering the peak quantity at the peak time at j implies an arrival at $j+1$ at l_{j+1}. Consequently, no decision concerning the start of the delivery and the quantity delivered at a customer prior to j can lead to a later arrival at $j+1$, and therefore at $n-1$. If the peak for j is at l_j, then no decision concerning the start of the delivery and the quantity delivered at a customer prior to j can lead to a later arrival at j, and thus at $n-1$.

It should now be clear that we can repeat the steps outlined above for customers $n-2$ to 1 with only one minor change, because t^* must be feasible with respect to the preceding as well as the succeeding customers. We set t^*_i for customer i as follows:

$$t^*_i = \min\left(\max\left(t'_i, t^{peak}_i\right), t^*_{i+1} - d_{i,i+1} - \frac{q^{\min}_i}{R_i}\right).$$

In this way, we create a schedule of delivery times and quantities that maximizes total delivery quantity given customer delivery rates, but the revised policy no longer runs in linear time.

6. Other Practical Complexities

A driver's shift may involve several routes. In that case, it is natural to optimize the delivery volume over the entire shift. This can easily be accomplished by starting with the last delivery on the last route and working towards the first delivery on the first route. Each time we move to a preceding route, we just "renew" the available truck capacity. It may be possible to deliver more to the customers on an early route than suggested by the schedule produced, but as we have encountered before, this will always be at the expense of a reduction in delivery volume in later routes and a net loss in the delivery volume for the shift.

In practice, usage of a particular commodity rarely occurs at the same rate 24 hours a day, 7 days a week. Many factories shut down overnight and/or on weekends. Further, depending on the production schedule, inventory may be depleted faster earlier in the day than later in the day. This indicates that it is necessary to be able to handle varying usage rates at customers. Fortunately, each change in usage rate at a customer simply amounts to an additional breakpoint in the delivery profile. For example, a delivery profile that looked like Figure 7, for example, may now look like Figure 8.

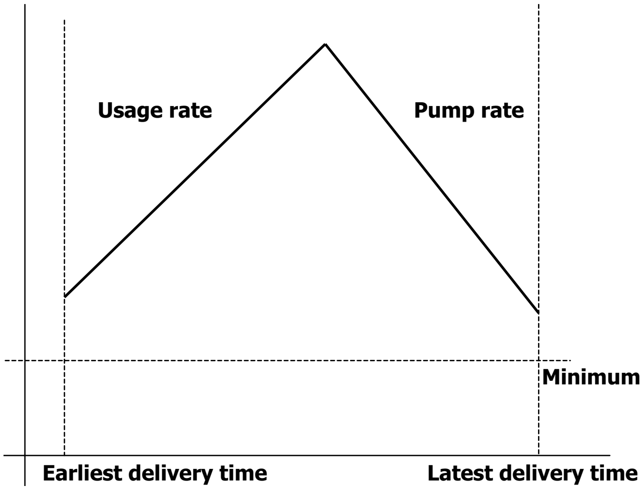

Figure 7 Customer Delivery Profile

The delivery profile will have a steeper slope when the usage rate increases and it will plateau when the usage rate drops to zero. In computing the breakpoints, we set the next breakpoint by determining the minimum between the time dictated by the formulas we described earlier given the current usage rate and the time of the next change in the usage rate.

Because usage rates are always nonnegative, the delivery profiles remain quasi-concave (Bertsekas 1995), i.e., they will not alternately rise and fall. Consequently, there will still be only one time, or block of time, corresponding to the peak delivery quantity for

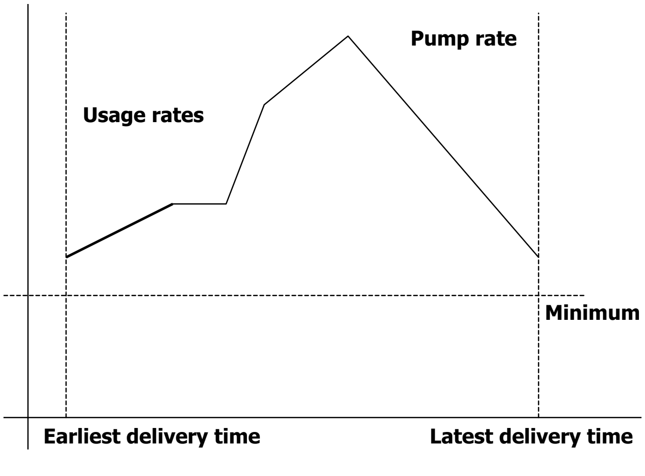

Figure 8 Customer Delivery Profile with Changing Usage Rates

a customer. Therefore, the same basic techniques can be applied.

7. Computational Results

To demonstrate the value of DVO and to determine whether there are environments in which DVO may be especially useful, we have conducted a series of computational experiments.

Because we are the first to investigate DVO, as far as we know, we cannot compare DVO to any existing methods. Therefore, we compare the performance of DVO to a number of other heuristics that can be used to schedule the deliveries on a particular route. These heuristics are described below:

- *Early Method (E)*: deliver to every customer as early as possible, respecting the scheduling decisions at the preceding customers on the route, and at that time deliver the maximum possible.
- *Late Method (L)*: deliver to every customer as late as possible, respecting the scheduling decisions at the preceding customers on the route, and at that time deliver the maximum possible.
- *Greedy Method (G)*: deliver the maximum possible to every customer, respecting the scheduling decisions at the preceding customers on the route.

Note that these three heuristics are easy to implement and traverse the route once from start to finish. The last heuristic we consider is a little harder to implement because it does not traverse the route from start to finish.

- *Maximum Usage Method (U)*: deliver the maximum possible to the customer with the highest usage rate among the as of yet unscheduled customers, respecting the scheduling decisions that have already been made.

Next, we turn to the generation of instances. The characteristics of our set of instances are motivated by our experiences with the distribution of industrial gases. The following three guidelines are frequently used during the construction of delivery schedules:

- It is best, i.e., most cost effective, to make a delivery to a customer when the customer's inventory level is close to the safety stock level (as opposed to delivering to a customer when the inventory level is still fairly high). The reason is that over a period of time this reduces the number of visits to the customer, which in turn should reduce the per-unit delivery costs.
- It is best, i.e., most cost effective, when making a delivery to fill up the storage facility at the customer completely (as opposed to filling up the storage facility only partially). Again, the reason is that over a period of time this reduces the number of visits to the customer, which in turn should reduce the per-unit delivery costs.
- It is best, i.e., most cost effective, to deliver an entire truck load on a route (as opposed to returning to the distribution center with product left in the vehicle). This should also reduce the per-unit delivery cost.

Note that it is easy to achieve one or two of these goals, but not as easy to achieve all three simultaneously. For example, it is not difficult to construct a route and delivery schedule for which we can guarantee that it is possible to deliver a full truckload. When the combined storage capacity at the customers at the start of the route is larger than the vehicle capacity, then it will be possible to deliver the entire contents of the vehicle. However, it is also clear that we will not be able to fill up the storage facilities completely at each of the customers. It is in situations where companies try to aggressively pursue achieving all three goals simultaneously that DVO becomes relevant and important, because it is more difficult to evaluate the many trade-offs and make the appropriate choices.

We constructed 20 instances with the following characteristics. Each instance has 10 customers that have to be visited in a prespecified order. The average storage capacity of the customers is 350 and the vehicle capacity is 3,000. The average usage rate of the customers is 20 units of product per hour, and the delivery rate is 600 units of product per hour. The initial inventory at each customer is set in such a way that the customer will run out of product (reach the safety stock level) at some point during the next 9 hours. Given the storage capacities of the customers, the initial inventories and usage rates of the customers, and the vehicle capacity, the instances are such that the total volume deliverable on the route should be reasonably close to a truckload and the quantity delivered to each customer should, in most cases, fill up its storage facility. The average travel time between customers is 50 minutes, and there is a route duration limit of 9 hours. The minimum delivery quantity at all customers is 60 (about 20% of the average customer's capacity). We allow the customer delivery capacity, usage rate, and travel time to vary by 10% above or below the above average values. There are no explicit delivery time windows, but there are, of course, the implicit time windows implied by the earliest time the minimum delivery quantity can be delivered and the time at which the customer will run out of product.

When we present computational results, they will represent, unless clearly stated otherwise, the average over a set of 20 different instances.

In Table 1 we show for each of the different methods the total volume delivered on the route as well as the percentage of vehicle capacity this value represents. The results in Table 1 demonstrate the value of DVO. The average total volume delivered by DVO

Table 1 **Delivery Volume for Different Methods**

DVO	%	E	%	G	%	L	%	U	%
2,709.50	90.32	2,583.40	86.11	2,370.06	79.00	2,105.61	70.19	2,439.39	81.31

Table 2 **Schedules for Various Methods**

	Start	Complete	Quantity	% before	% after
Early (Quantity: 2,366.89, Time: 787.89)					
Depot	0				
1	50.33	58.00	76.71	77.78	100.00
2	106.82	125.50	186.80	49.81	100.00
3	173.29	195.49	221.97	39.93	100.00
4	248.55	268.37	198.18	41.87	100.00
5	316.08	340.07	239.98	30.55	100.00
6	389.85	413.16	233.14	33.86	100.00
7	462.43	491.29	288.62	19.00	100.00
8	540.25	572.20	319.52	13.33	100.00
9	624.17	653.06	288.95	15.40	100.00
10	706.53	737.84	313.02	15.24	100.00
D	787.89				
Late (Quantity: 2,097.12, Time: 859.10)					
Depot	0				
1	107.10	116.56	94.63	72.58	100.00
2	196.45	212.61	161.52	42.42	85.82
3	260.40	268.36	79.59	32.93	54.47
4	321.42	343.67	222.47	34.74	100.00
5	447.02	475.33	283.09	18.07	100.00
6	529.98	547.59	176.09	20.09	70.04
7	596.86	630.27	334.08	6.24	100.00
8	685.82	696.34	105.20	0.51	29.05
9	748.30	777.31	290.06	2.40	87.33
10	830.78	865.82	350.40	5.12	100.00
D	915.87				
Greedy (Quantity: 2,263.09, Time: 859.10)					
Depot	56.77				
1	107.10	116.56	94.63	72.58	100.00
2	191.34	212.61	212.71	42.84	100.00
3	260.40	268.36	79.59	32.93	54.47
4	321.42	343.67	222.47	34.74	100.00
5	447.02	475.33	283.09	18.07	100.00
6	525.10	547.59	224.90	20.57	84.37
7	596.86	630.27	334.08	6.24	100.00
8	679.22	696.34	171.15	1.09	47.52
9	748.30	777.31	290.06	2.40	87.33
10	830.78	865.82	350.40	5.12	100.00
D	915.88				
Maximum Usage (Quantity: 2,314.64, Time: 859.10)					
Depot	56.77				
1	107.10	116.56	94.63	72.58	100.00
2	191.34	212.61	212.71	42.84	100.00
3	260.40	268.36	79.59	32.93	54.47
4	321.42	343.67	222.47	34.74	100.00
5	441.86	470.00	281.39	18.56	100.00
6	519.78	547.59	278.16	21.09	100.00
7	596.86	630.27	334.08	6.24	100.00
8	679.22	692.16	129.37	1.09	36.19
9	744.13	777.31	331.85	2.84	100.00
10	830.78	865.82	350.40	5.12	100.00
D	915.88				

Table 2 **(cont'd.)**

	Start	Complete	Quantity	% before	% after
DVO (Quantity: 2,679.43, Time: 859.10)					
Depot	56.77				
1	107.10	116.56	94.63	72.58	100.00
2	175.48	196.27	207.85	44.15	100.00
3	244.06	268.36	243.00	34.24	100.00
4	321.42	343.67	222.47	34.74	100.00
5	421.24	448.70	274.60	20.53	100.00
6	498.47	525.55	270.78	23.19	100.00
7	574.82	607.48	326.62	8.33	100.00
8	656.44	692.16	357.23	3.10	100.00
9	744.13	777.31	331.85	2.84	100.00
10	830.78	865.82	350.40	5.12	100.00
D	915.88				

exceeds the average total volume delivered by the next best method by more than 4% of the vehicle capacity. Over a period of time, this represents a significant decrease in delivery costs.

To create additional insights in the working and performance of the different methods, we present more detailed results for a single instance. In this instance, the DVO method outperformed the next best method by more than 13.2%. Table 2 present results for the Early, Late, Greedy, Maximum Usage, and DVO methods, respectively.

It is not hard to see what causes the Early Method to perform poorly. Even though the method fills up the storage facility at every customer on the route, it makes deliveries when there is still a reasonable amount of inventory left and therefore less storage capacity available for new product. For example, delivery to the last customer starts at 706.53, and 313.02 units of product are delivered, which fills the customer's inventory to capacity. In comparison, the delivery schedule created by the DVO Method visits the last customer later (830.78), when more product has been consumed. It still fills up the storage facility but is able to deliver a larger total volume (350.4). Delivering to this last customer at the earlier time, as done in the schedule produced by the Early Method, serves no purpose. Because of situations like this, in the schedule produced by the DVO Method more than 300 more units of product are delivered than in the schedule produced by the Early Method.

Similarly, it is not hard to see what causes the Late Method to perform poorly. The Late Method spends too much time waiting as opposed to using time to deliver product. This frequently creates situations where a delivery to a customer does not bring the

Table 3 Delivery Volume for Different Average Storage Capacities

Tank	DVO	%	E	%	G	%	L	%	U	%
250	2,177.52	72.58	1,967.42	65.58	1,999.84	66.66	1,830.46	61.02	2,040.91	68.03
300	2,633.69	87.79	2,509.34	83.64	2,320.83	77.36	2,087.13	69.57	2,388.88	79.63
350	2,939.85	98.00	2,924.86	97.50	2,599.88	86.66	2,303.93	76.80	2,682.65	89.42
400	2,982.79	99.43	2,979.35	99.31	2,808.86	93.63	2,483.74	82.79	2,878.30	95.94
450	2,997.96	99.93	2,997.88	99.93	2,925.23	97.51	2,635.06	87.84	2,959.69	98.66

inventory back up to capacity. Consider, for example, customer 2. Because the delivery at customer 1 completes at 116.56 and the travel time between customer 1 and 2 is 48.82, the delivery at customer 2 can begin at 165.38. Instead, in the schedule produced by the Late Method, the delivery is delayed until its latest delivery time of 196.45, when there is only enough time to deliver 161.52 units of product and travel to customer 3 to get there by its latest delivery time of 260.4. The inventory after delivery at customer 2 is only 85.82% of its capacity. In contrast, the schedule produced by DVO dictates a start of delivery at customer 2 at 175.48 when there is time to deliver a larger 207.85.

By examining customers from start to finish and greedily maximizing the volume deliverable at each of them, the Greedy Method can create delays that end up decreasing the volume deliverable in the later portion of the route. Consider customer 2, where the schedule produced by the Greedy Method and the schedule produced by the DVO Method first differ. The Greedy Method chooses to wait much longer with the start of the delivery to increase the quantity that can be delivered (212.71 vs. 207.85 in the schedule produced by the DVO Method), but as a result there is less time available for deliveries at the next customers. In the schedule produced by the Greedy Method, the delivery at customer 2 is completed at 212.61, resulting in an arrival at customer 3 at its latest delivery time with a possible delivery quantity of only 79.59. In the schedule produced by the DVO Method, the delivery at customer 3 starts earlier, allowing a quantity of 243 to be delivered, which is a significant improvement, even considering the smaller delivery quantity at customer 2.

The deficiencies of Maximum Usage are very similar to the Greedy Method, but are not quite as easy to see because of the order in which delivery decisions are made. The Maximum Usage Method sets delivery times and delivery quantities in the order: 9, 6, 7, 5, 4, 1, 8, 2, 3, and 10. To deliver the maximum possible at customer 6 (the second in the series after customer 9), delivery at customer 6 is started at 519.78 with a quantity of 278.16, which is larger than the delivery quantity in the scheduled produced by the DVO Method (270.78). Arrival at customer 7 after completing the delivery at customer 6 is at 574.82, but to optimize delivery quantity at customer 7 (third in the ordering), the start of delivery is delayed until 596.86, creating waiting time between customers 6 and 7. The quantity deliverable at this time is 334.08, which is again larger than delivery quantity of 326.62 in the schedule produced by the DVO method. This delay of approximately 22 minutes allows 7.46 more units to be consumed at customer 7 (and thus to be delivered), but delays arrival at customer 8 by 22 minutes plus the delivery time for the 7.46 units. This delay translates into a decreased quantity deliverable at customer 8 of approximately 228 units, which is why there is a delivery of 129.37 in the schedule produced by the Maximum Usage Method, whereas there is a delivery of 357.23 in the schedule produced by the DVO Method. This difference is much larger than the gains from increased deliveries at customers 6 and 7. This example points out the danger of delaying delivery at a customer to increase the delivery quantity, as it may compromise the possibilities at other customers.

Next, we study the impact of changing the characteristics of the instances. This allows us to compare DVO to other methods in more scenarios and see how certain characteristics affect their relative performance. Table 3 presents the results when we vary the average storage capacity at a customer.

As expected, the total volume delivered on a route decreases when the average storage capacity at a customer decreases. On a route with 10 customers with an average storage capacity of 250 (and thus approximately a combined storage space of 2,500), it will be impossible to deliver an entire truckload of 3,000. On the other hand, with an average storage capacity of 450 (and thus approximately a combined storage space of 4,500), it should be possible to deliver an entire truckload of 3,000 in most cases. The results in the table demonstrate that DVO starts to perform better, compared to the other methods, when it is more difficult to deliver a large volume on a route. It is somewhat surprising to see that the relatively simple Early Method performs reasonably well in most situations.

Similar behavior is observed in Table 4, which presents the results for different vehicle capacities. We see that when the ratio of combined storage capacity to vehicle capacity is relatively large, the opportunity for large individual deliveries, and therefore

Table 4 **Delivery Volume for Different Vehicle Capacities**

Tank	DVO	%	E	%	G	%	L	%	U	%
2,500	2,491.59	99.66	2,467.87	98.71	2,347.97	93.92	2,101.92	84.08	2,394.95	95.80
3,000	2,709.50	90.32	2,583.40	86.11	2,370.06	79.00	2,105.61	70.19	2,439.39	81.31
3,500	2,712.00	77.49	2,583.40	73.81	2,370.06	67.72	2,105.61	60.16	2,439.39	69.70

Table 5 **Delivery Volume for Different Delivery Rates**

R	DVO	%	E	%	G	%	L	%	U	%
300.00	1,761.35	58.71	1,761.29	58.71	1,676.90	55.90	1,499.03	49.97	1,710.62	57.02
400.00	2,317.04	77.23	2,316.39	77.21	2,037.97	67.93	1,818.77	60.63	2,095.09	69.84
600.00	2,716.69	90.56	2,542.84	84.76	2,464.99	82.17	2,235.90	74.53	2,507.12	83.57
1,200.00	2,753.84	91.79	2,372.18	79.07	2,738.17	91.27	2,682.20	89.41	2,742.32	91.41
∞	2,755.28	91.84	2,209.00	73.63	2,755.28	91.84	2,755.28	91.84	2,755.28	91.84

the chances to deliver an entire truckload, are higher. Consequently, DVO seems most useful in situations where this ratio is relatively small.

Table 5 presents the results when we vary the delivery rate. When delivery is instantaneous, it is always best to deliver as late as possible. We see that all methods except for the Early Method do this and therefore produce identical results (row labeled ∞). On the other hand, when the delivery rate is very low and all the time in an optimal schedule is occupied by traveling and delivering product, it is best to deliver as early as possible and avoid waiting. We see that both DVO and the Early Method do so, but the others do not. These results clearly show that the other methods are not robust enough to adjust to changing circumstances. In fact, the results demonstrate that, as expected, the Early Method gets progressively worse when the delivery rate increases, and the Late Method gets progressively worse when the delivery rate decreases.

We find similar results when we vary the average travel time between customers, the average customer usage rates, and the route duration limits.

The computational experiments demonstrate that the truly simple approaches, such as the Early Method and the Late Method, are not robust and may result in very poor performance. The other, somewhat more adaptive methods, e.g., Greedy Method and Maximum Usage Method, appear to be less sensitive but overall do not exhibit the desired performance, either. DVO consistently delivers significantly larger volumes to customers than the other methods across a variety of scenarios with hardly any additional computational cost.

8. Conclusion

We have demonstrated that optimizing the volume deliverable on a route (or a shift comprising a sequence of routes) in situations where vendor managed inventory resupply policies are in place can easily and efficiently be incorporated in existing heuristics. We have also shown that, over a period of time, the use of DVO can result in significant cost savings.

References

Bertsekas, D. P. 1995. *Nonlinear Programming*. Athena Scientific, Belmont, MA.

Campbell, A., L. Clarke, A. Kleywegt, M. Savelsbergh. 1998. Inventory routing. T. Crainic, G. Laporte eds. *Fleet Management and Logistics*. Kluwer Academic Publishers, Boston, MA.

Kleywegt, A., V. Nori, M. Savelsbergh. 2002. The stochastic inventory routing problem with direct deliveries. *Trans. Sci.* **36** 94–118.

Nori, V. 1999. Algorithms for dynamic and stochastic logistics problems. Ph.D. thesis, Georgia Institute of Technology, Department of Industrial and Systems Engineering, Atlanta, GA.