Chapter 12

Inventory Routing in Practice

Ann M. Campbell
Lloyd W. Clarke
Martin W.P. Savelsbergh

12.1 Introduction

PRAXAIR (www.praxair.com) is a large industrial gases company with about 60 production facilities and more than 10,000 customers across North America. PRAXAIR recently negotiated a policy with its customers in which PRAXAIR is in charge of managing its customers' inventories. Customers will no longer call PRAXAIR to request a delivery. Instead, PRAXAIR will determine who receives a delivery each day and what the size that delivery will be. PRAXAIR will use gauge readings received from remote telemetry units as well as regular customer phone calls to monitor and forecast product inventories. The distribution planning problems associated with such vendor-managed resupply policies are known as *Inventory Routing Problems* (IRPs).

IRPs are very different from VRPs. VRPs occur when customers place orders and the delivery company, on any given day, assigns the orders for that day to routes for trucks. In inventory routing problems, the delivery company, not the customer, decides how much to deliver to which customers each day. There are no customer orders. Instead, the delivery company operates under the restriction that its customers are not allowed to run out of product. Another difference is the planning horizon. VRPs typically deal with a single day, and the only requirement is that all orders have to be delivered by the end of the day. Inventory routing problems deal with a longer horizon. Each day the delivery company makes decisions about which customers to visit and how much to deliver to each of them, while keeping in mind that decisions made today impact what has to be done in the future. The objective is to minimize the total cost over the planning horizon while making sure no customers run out of product. The flexibility to decide when customers receive a delivery

and how large these deliveries will be may significantly reduce distribution costs. However, this flexibility also makes it very difficult to determine a good, much less an optimal, cost-effective distribution plan. When the choice becomes which of the customers to serve each day and how much to deliver to them, the choices become virtually endless.

Vendor-managed resupply policies can be used in many situations. In some instances, the use of such a policy is natural, such as when the "customers" are really part of the same company. In others, the use of a vendor-managed resupply policy is often the result of lengthy negotiations with customers who have for years followed a policy in which they call in their orders. Examples of industries where vendor-managed resupply policies are being used or considered are the petrochemical industry (gas stations), the grocery industry (supermarkets), the soft drink industry (vending machines), and the automotive industry (parts distribution). The number of industries using vendor-managed resupply policies is increasing rapidly. An important reason for this is technology. For a variety of industries and products, the monitoring technology that existed several years ago was not sophisticated enough to make a vendor-managed resupply system possible. The only way to check a customer's inventory for many types of products was for the vendor to call the customer and for the customer to go look at the meter on the tank, to count the number of items in the vending machine, and so forth. Now the use of remote telemetry units, scanners, computers, and modems allows monitoring of inventory levels directly by the vendor, opening up new opportunities for vendor-managed resupply policies.

In section 12.2, we formally introduce the IRP, and in section 12.3, we give a brief literature review. In section 12.4, we discuss the two-phase approach we have chosen to solve instances of the IRP. In section 12.5, we present the results of some computational experiments on real-world instances from PRAXAIR.

12.2 Problem Definition

The IRP is concerned with the repeated distribution of a single product from a single facility to a set N of customers over a planning horizon of length T (expressed in days), possibly infinity. Customer i consumes the product at a rate u_i (volume per day) and has the capability to maintain a local inventory of the product up to a maximum of C_i. The inventory at customer i is I_i^0 at time 0. A fleet M of homogeneous vehicles, with capacity Q, is available for the distribution of the product. The objective is to minimize the average daily distribution cost during the planning period without causing stockouts at any of the customers. Vehicles are allowed to make multiple trips per day. Three decisions have to be made:

- When to serve a customer?
- How much to deliver to a customer when served?
- Which delivery routes to use?

Real-life inventory routing problems are obviously stochastic. No customer will use product the same way every single day. In many situations, however, usage is relatively predictable and customers generally use about the same amount each day if we look at their

total usage for several days in a row. Therefore, solution approaches developed for the IRP as defined above provide useful planning tools.

12.3 Literature Review

Although the IRP is a long-term problem, almost all proposed solution approaches solve only a short-term version of the problem to make it easier. In early work, short-term was often just a single day, but in later work this was expanded to several days. Besides the number of days modeled, key features that distinguish different solution approaches include how the long-term effects of short-term decisions are modeled, how it is determined which customers are included in the short-term problem, and whether demand at the customers is treated as deterministic or stochastic. Summaries of various approaches were made by Ball [3], Dror, Ball, and Golden [14], Nori [27], and Campbell et al. [11]. In the remainder of this section, we discuss a number of proposed approaches in more detail. This discussion is not meant to provide a complete overview of work done in this area, but is an introduction to the types of approaches that have been taken.

Those following a single-day approach include Federgruen and Zipkin [17], Golden, Assad, and Dahl [22], and Chien, Balakrishnan, and Wong [13]. Federgruen and Zipkin [17] in their single-day approach capitalized on many of the ideas from vehicle routing. Their model, which is a nonlinear integer program, decomposes into a routing portion and inventory portion. They construct an initial feasible solution to the routing part of the problem and iteratively improve the solution by exchanging customers between routes and then resolving the inventory part of the problem. Golden, Assad, and Dahl [22] developed a heuristic based on a measure of the urgency of each customer, which is defined as the ratio of tank inventory level to tank size. All customers with an urgency smaller than a certain threshold are excluded. Customers are iteratively selected to receive a delivery according to the highest ratio of urgency to extra time required to visit this customer. Chien, Balakrishnan, and Wong [13] also developed a single-day approach, but it does not treat each day as a completely separate entity. By passing some information from one day to the next, the system simulates a multiple-day planning model.

The work of Fisher et al. [19, 8] was motivated, as is our work, by an application in the industrial gases industry. They took profit maximization from product distribution over several days as their objective. Demand is given by upper and lower bounds on the amount to be delivered to each customer for every period in the planning horizon. An integer program is formulated that captures delivery volumes, assignment of customers to routes, assignments of vehicles to routes, and assignment of start times for routes. It is solved using a Lagrangian dual-ascent approach.

The first serious effort to develop an approach that considers what happens beyond the next few days was made by Dror, Ball, and Golden [14] and Dror and Ball [16]. They considered demand to be stochastic and used the probability that a customer will run out on a specific day in the planning period, the average cost to deliver to the customer, and the anticipated cost of a stockout to find the optimal replenishment day t^* for each customer. If t^* falls within the short-term planning period of the next few days, the customer will be visited, and a value c_t is computed for each of the days in the planning period that reflects the expected increase in future cost if the delivery is made on day t instead of on t^*. An

integer program is then solved that assigns customers to a vehicle and a day, or just a day, that minimizes the sum of these costs plus the transportation costs. Delivery amounts are considered to be dictated by the day of the week on which the delivery is made and thus are not a decision to be made by the integer program.

Some of the ideas of Dror and Ball were extended and improved by Trudeau and Dror [29]. Dror and Levy [15] used a similar analysis to yield a weekly schedule but applied node and arc exchanges to reduce costs in the planning period. Bard et al. [5, 4, 23] discussed another extension of this idea. They took a rolling-horizon approach to the problem by determining a schedule for 2 weeks but implementing only the first week. An analysis similar to Dror and Ball's is done to determine an optimal replenishment day for each customer, and incremental costs are computed that represent the cost for changing the next visit to a customer to a different day but keeping the optimal schedule in the future. These costs are used in an assignment problem formulation that assigns each customer to a day in the 2-week planning horizon.

Anily and Federgruen [1, 2] looked at minimizing long-run average transportation and inventory costs by determining long-term routing patterns for a set of customers with deterministic demand. The routing patterns are determined using a modified circular partitioning scheme. After the customers are partitioned, customers within a partition are divided into regions to make the demand of each region roughly equal to a truck load. A customer may appear in more than one region, but then a certain percent of the customer's demand is allocated to each region. When one customer in a region gets a visit, all customers in the region are visited. They also determine a lower bound for the long-run average cost to be able to evaluate how good their routing patterns are. Using ideas similar to those of Anily and Federgruen, Gallego and Simchi-Levi [21] evaluated the long-run effectiveness of direct shipping (separate loads to each customer). They concluded that direct shipping is at least 94% effective over all inventory routing strategies whenever minimal economic lot size is at least 71% of truck capacity. This shows that direct shipping becomes a bad policy when many customers require significantly less than a truck load, making more complicated routing policies the appropriate choice.

Another adaptation of these ideas was made by Bramel and Simchi-Levi [10]. They considered the variant of the IRP in which customers can hold an unlimited amount of inventory. To obtain a solution, they transform the problem to a capacitated concentrator location problem (CCLP), solve the CCLP, and transform the solution back into a solution to the IRP. The solution to the CCLP will partition the customers into disjoint sets, which in the inventory routing problem will become the fixed partitions. These partitions are then served in a way similar to the regions of Anily and Federgruen.

In the last few years, several researchers started to investigate a stochastic version of the problem, in which it is assumed that a probability distribution is known for customer usage. This adds more realism, since in practice customer usage is never deterministic, but obtaining probability distributions of customer usage in practice is extremely complex. Kleywegt, Nori, and Savelsbergh [25] formulated the inventory routing problem as a Markov decision process and proposed approximation methods to find good solutions with reasonable computational effort. Computational results are presented for the inventory routing problem with direct deliveries. Other work in this direction includes Minkoff [26], Bassok and Ernst [7], Barnes-Schuster and Bassok [6], Berman and Larson [9], Cetinkaya and Lee [12], and Fumero and Vercellis [20].

12.4 Solution Approach

A short-term approach has the tendency to defer as many deliveries as possible to the next planning period, which may lead to an undesirable situation in the next planning period. Therefore, the proper projection of a long-term objective into a short-term planning problem is essential. It needs to capture the costs and benefits of delivering to a customer earlier than necessary. Our focus has been on developing a flexible system capable of handling large instances that properly balances short-term and long-term goals and that considers all the key factors, i.e., geography, inventory, capacity, and usage rate. We wanted also to create a system that would consider routing customers together on a day where none of them are at the point of run-out but where they combine to make a good, full-truckload delivery route. We found that most systems reduce the problem by starting with only the "emergency" customers, never putting together certain combinations that make sense with regard to location and delivery size. The basis for our system is a two-phase solution approach. In the first phase, we determine which customers receive a delivery on each day of the planning period and decide on the size of the deliveries. In the second phase, we determine the actual delivery routes and schedules for each of the days.

As mentioned, real-life inventory routing problems are stochastic. Therefore, any distribution plan covering more than a couple of days will never be executed completely as planned. Actual volumes delivered differ from planned volumes because usage rates deviate from their forecasts, planned driving time is off due to traffic congestion, and so forth. Therefore, any planning system needs to be flexible. It needs to take advantage of the latest changes in the data. Given this, our approach is to embed our two-phase solution approach in a rolling-horizon framework. We always construct a distribution plan for a month to reflect the long-term nature of the planning problem, but we expect to implement only the first few days. We repeat this as often as necessary using the latest information available.

12.4.1 Phase I: Integer Programming Model

At the heart of the first phase is an integer program. Central to the model are two quantities: $L_i^t = \max(0, tu_i - I_i^0)$, a lower bound on the total volume that has to be delivered to customer i by the end of day t, and $U_i^t = tu_i + C_i - I_i^0$, an upper bound on the total volume that can be delivered to customer i by the end of day t. Let d_i^t represent the delivery volume to customer i on day t; then to ensure that no stockout occurs at customer i and to ensure that we do not exceed the inventory capacity at customer i, we need to have that

$$L_i^t \leq \sum_{1 \leq s \leq t} d_i^s \leq U_i^t, \qquad i \in N,\ t = 1, \ldots, T.$$

To model the resource constraints with some degree of accuracy and to have a meaningful objective function, we found it necessary to explicitly use delivery routes. We added another dimension to the d variable, changing it from d_i^t to d_{ir}^t. However, when we refer to a "route," we are really referring to a set of customers without enforcing a specific ordering among the customers in the set. We estimate the distance required to visit the customers in the set by the length of the optimal traveling salesman tour through all the customers. Now, let R be the set of delivery routes, let T_r denote the duration of route r (as a fraction

of a day), and let c_r be the cost of executing route r. Furthermore, let x_r^t be a 0-1 variable indicating if route r is used on day t ($x_r^t = 1$) or not ($x_r^t = 0$). The total volume that can be delivered on a single day is limited by a combination of capacity and time constraints. Since vehicles are allowed to make multiple trips per day, we cannot simply limit the total volume delivered on a given day to be the sum of the vehicle capacities. To be more precise, the resource constraints can be modeled by

$$\sum_{i \in r} d_{ir}^t \leq Q x_r^t \qquad \forall\, r \in R,\ t = 1, \ldots, T,$$

and

$$\sum_{r \in R} T_r x_r^t \leq |M| \qquad \forall\, t = 1, \ldots, T.$$

These constraints ensure that we do not exceed the vehicle capacity on any of the selected routes and that the time required to execute the selected routes does not exceed the time available.

The basic Phase I integer programming model is given by

$$\min \sum_{t=1}^{T} \sum_{r \in R} c_r x_r^t$$

subject to

$$\begin{aligned}
& L_i^t \leq \sum_{1 \leq s \leq t} \sum_{r \in R} d_{ir}^s \leq U_i^t \qquad \forall\, i \in N,\ t = 1, \ldots, T, \\
& \sum_{i \in r} d_{ir}^t \leq Q x_r^t \qquad \forall\, r \in R,\ t = 1, \ldots, T, \\
& \sum_{r \in R} T_r x_r^t \leq |M| \qquad \forall\, t = 1, \ldots, T, \\
& x_r^t \in \{0, 1\} \qquad \forall\, r \in R,\ t = 1, \ldots, T, \\
& d_{ir}^t \geq 0 \qquad \forall\, i \in N,\ t = 1, \ldots, T.
\end{aligned}$$

The first variation of the basic model handles fixed and variable stop times at the customers as well as a vehicle reloading time at the facility. The duration T_r of route r can be modified to include not only the estimated time to drive the distance between the customers on the route but also a fixed stop time for each customer and an initial fill time for the vehicle required before the route can start. Dispense time at a customer clearly cannot be included in T_r a priori because it depends on the size of the delivery. Therefore, we must alter the resource constraint as follows, where F is the percentage of the day required to dispense each unit of product:

$$\sum_{r \in R} \left(T_r x_r^t + \sum_{i \in r} F d_{ir}^t \right) \leq |M| \qquad t = 1, \ldots, T.$$

The second variation handles operating modes of customers. Operating mode refers to the start and end times of customer usage on each day of the week. Earlier, we assumed that

each customer i uses product 24 hours per day every day. Operating modes are important. When a customer does not use product on the weekend, for example, this has a big impact on properly timing the deliveries. Operating modes can be handled easily by appropriately modifying the lower- and upper-bound parameters. The value for the upper bound and lower bound on day t now depend on where in the week days 1 through t fall.

The third variant handles time windows at customers. An operating mode restricts when a customer uses product. A time window restricts when a customer can receive a delivery. Time windows may be day dependent as well. To handle time windows, the lower- and upper-bound parameters need to be modified again, but in a slightly different way. Now the lower bound L_i^t needs to be defined as the total volume that has to be delivered to customer i by the closing of the time window on day t to allow customer i to last until the opening of the time window on day $t+1$ (or the opening of the time window on the first available day for the next delivery if no deliveries can be made on day $t+1$). The upper bound U_i^t is now defined as the largest volume that customer i can receive by the close of the delivery window on day t.

12.4.2 Phase I: Solving the Integer Programming Model

The integer programming model presented above is not very practical for two reasons: the huge number of possible delivery routes and, although to a lesser extent, the length of the planning horizon. To make the integer program computationally tractable we consider a small (but good) set of delivery routes and aggregate periods toward the end of the planning horizon.

12.4.2.1 Clusters

Our approach to reduce the number of routes is based on allowing customers to be on a route together only if they are in the same *cluster*. A cluster is a group of customers that can be served cost effectively by a single vehicle for a long period. The cost of a cluster is an approximation of the distribution cost for serving the customers in the cluster for a month. The cost of serving a cluster depends on not only the geographic locations of the customers in the cluster but also on whether the customers in the cluster have compatible inventory capacities and usage rates. Therefore, to evaluate the cost of a cluster, we need a model that considers all these factors.

The following approach is used to identify a good set of disjoint clusters covering all customers:

1. Generate a large set of possible clusters.
2. Estimate the cost of serving each cluster.
3. Solve a set-partitioning problem to select clusters.

Observe that the selection of clusters has to be done only once as a preprocessing step before the actual planning starts. It does not have to be rerun before every execution of the Phase I integer program. In practice it makes sense to recluster when new customers have been added or there have been significant changes to the data.

Since we generate a large number of clusters to choose from, we need a costing procedure that is fast but able to provide an accurate estimate of the cost of serving the cluster. We decided to use a simple integer program with key features represented.

12.4.2.2 Aggregation and Relaxation

Because our two-phase solution approach will be embedded in a rolling-horizon framework, the emphasis should be on the quality and detail of the decisions concerning the first few days of the plan. This provides us with an excellent opportunity to reduce the size of the integer program by aggregating days toward the end of the planning period.

For the first k days, we will still have route selection variables for each day, but for the days after that, we will have route selection variables covering periods of several days. Instead of making a decision on whether to execute each route on days 8 to 14 individually, for example, we now decide how many times each of the routes will be executed during the whole week. Several aggregation schemes were tested. We found that considering weeks rather than days toward the end of the planning horizon still does a good job of preserving the costs associated with the effect of short-term decisions on the future and yields a significant reduction in CPU time. Therefore, the daily variables associated with these later days are replaced by weekly variables. Upper and lower bounds are altered accordingly as well.

A further simplification is obtained by relaxing the integrality restrictions on the variables representing the weekly decisions. Therefore, the only binary variables appearing in the integer program will be those representing route selections for the first k days.

12.4.3 Phase II: Scheduling

A solution to the integer program of Phase I specifies the volumes to deliver to each customer for the next k days. It does not specify departure times and customer sequences for the different vehicles. Therefore, we still need to construct vehicle routes and schedules.

Since the delivery volumes specified by the solution to the integer program may not fit before a specific time of the day and may need to be received before a certain later time to prevent run-out, these deliveries have self-imposed time windows. Therefore, to convert the information provided by the solution to the integer program to daily vehicle routes and schedules, we can solve a sequence of VRPs with time windows.

However, such an approach does not capitalize on the flexibility inherent in the IRP. The delivery volumes specified by the solution to the integer program are good from a long-term perspective; they may not be good from a short-term perspective. Therefore, we treat the delivery volumes and timing specified by the solution to the integer programs as suggestions. We try to follow these suggestions as closely as possible, since this helps to achieve our long-term goals, but we allow small deviations when it helps to construct better short-term plans. To be more precise, we construct vehicle routes and schedules for two consecutive days, where we force the total volume delivered to a customer over the 2 days to be greater than or equal to the total delivery volume specified by the solution to the integer program for these 2 days, but we do not enforce specific delivery volumes on individual days. In this way, we stay close to the delivery volumes suggested by the integer program, which is good from a long-term perspective, but we introduce some flexibility in the daily routing and scheduling, which is good from a short-term perspective. Deliveries

can be split into smaller pieces, delivering one part on the first day and the second part on the second day if this works out to be better, for example, when resources are very tight on one of the days. This flexibility is even more important when we consider that, in practice, a few customers may not follow a vendor-managed resupply policy and may call in orders that need to be added to the daily routing and scheduling problem. With new orders and new accurate up-to-date information on customer inventory levels, it may make sense to shift around some of the deliveries over the next couple of days.

Because of customer usage and customer inventory capacities, there may be customers that require a delivery on both days or even multiple times a day. Consequently, in our 2-day routing and scheduling problem, we can distinguish two types of customer: customers that require multiple deliveries over the 2 days and customers that require only one.

We have developed and implemented an insertion heuristic for this 2-day routing and scheduling problem. The heuristic is a logical progression of commonly used techniques in insertion heuristics for the vehicle routing problem with time windows; see, for example, Solomon [28] and Kindervater and Savelsbergh [24].

In the description of the heuristic, we assume, for ease of presentation, that there are no operating modes and no time windows restricting when deliveries can take place. Both complications can easily be handled. We also do not discuss explicitly the use of fixed stop times and unloading times, though both can be included in the travel-time value used here.

The flexibility to change delivery volumes makes checking the feasibility of insertions much more complex than in the VRP. For example, the insertion of a customer on a route can affect the delivery volume of another customer on an earlier or later route for the same vehicle, which can affect the size and timing of other deliveries for the customers on that route and so forth.

To be able to evaluate the feasibility of an insertion, we maintain several quantities related to deliveries to customers already scheduled. Consider a delivery to customer i on route r. The predecessor on the route is denoted by $p(i)$ and the successor on the route is denoted by $s(i)$. The total volume to be delivered to customer i over the 2 days prescribed by the solution to the Phase I integer program is d_i. We consider a day as ranging from time 0 to 1 for convenience. There is a slight difference for customers that need multiple deliveries over the 2 days, but the basic quantities we maintain are the following:

- The minimum delivery volume, $q_{ri}^{\min}$,

$$q_{ri}^{\min} = d_i.$$

- The earliest time a delivery can be made, t_{ri}^{early},

$$t_{ri}^{early} = \max\left(t_{rp(i)}^{early} + tt_{p(i),i}, (q_{ri}^{\min} - C_i + I_i)/u_i\right),$$

  where $t_{r0}^{early} = t_r^{earlystart}$, the earliest time the route can start, and $tt_{j,k}$ is the travel time from customer j to k. The first term of the maximum represents the time to get to customer i from $p(i)$. The second term represents the time that the minimum delivery volume can fit at customer i.

- The latest time a delivery can be made, t_{ri}^{late},

$$t_{ri}^{late} = \min\left(t_{rs(i)}^{late} - tt_{i,s(i)},\, I_i/u_i\right),$$

  where $t_{r(n+1)}^{late} = t_r^{lateend}$, the latest time route r can end. The first term of the minimum represents the latest departure time from i to be able to reach $s(i)$ by the latest time for its delivery. The second term represents the time when customer i runs out of product.

- The maximum delivery volume, $q_{ri}^{\max}$,

$$q_{ri}^{\max} = \min\left(Q - \sum_{j \neq i \in r} q_{rj}^{\min},\, C_i,\, d_i,\, C_i - I_i + u_i t_{ri}^{late}\right).$$

  The first term of the minimum is the capacity remaining in the vehicle if we assume all other customers on the route will receive their minimum delivery volumes, the second and third terms are obvious, and the fourth term represents the volume that will fit at the latest time a delivery can be made.

Because vehicles can drive multiple routes per day, we also maintain several quantities for each route:

- the earliest time the route can start, $t_r^{earlystart}$,
- the latest time the route can start, $t_r^{lateststart}$,
- the earliest time the route can end, $t_r^{earlyend}$, and
- the latest time a route can end, $t_r^{lateend}$.

Given these quantities, the feasibility of an insertion is checked as follows. First, we check whether the minimum delivery volume fits in the vehicle given the other planned deliveries. Next, we compute the earliest time and the latest time a delivery can take place. If the earliest delivery time is greater than the latest delivery time, the insertion is infeasible. Using the latest delivery time, we compute the maximum delivery size. If it is smaller than the minimum delivery size, the insertion is infeasible. If the insertion passes both of these tests, it is feasible.

If an insertion is feasible, the cost of the insertion is evaluated. The cost of an insertion is a weighted sum of several components. The first component is the increase in distance and the second component is an approximation of the minimum increase in waiting time if the insertion is carried out. The third component is a charge for making routes inflexible. In the final 2-day plan, we like to have near-capacity routes. Therefore, we want to discourage the construction of routes with a small difference $t_r^{lateststart} - t_r^{earlystart}$ and a large difference $Q - \sum_{j \in r} q_{rj}^{\max}$, since it is unlikely that such routes can be extended to near-capacity routes. A charge is incurred if the insertion forces a route to have a gap between earliest and latest starting time that is less than x minutes and a total maximum delivery volume that is less than $y\%$ of capacity. The charge is inversely related to the size of the gap.

For each delivery to a customer, we maintain the cheapest feasible insertion and the second cheapest feasible insertion, if it exists. Since we can always construct a feasible route with just a delivery to a single customer, there exists at least one feasible insertion.

All that remains to complete the description of the insertion heuristic is to specify how we select the deliveries to be inserted in each iteration. Note that we select deliveries rather than customers, because customers may require multiple deliveries over the 2 days. We use the following selection rule:

1. If there are deliveries that cannot be inserted into any existing route, then among those deliveries select the one with the most expensive route for itself.

2. If all deliveries can be inserted into at least one existing route, then select the one with the largest difference between the cost of its cheapest and second cheapest insertion.

The first part of the rule captures the idea that if there are deliveries that cannot be inserted in the current set of routes, we know that we have to create at least one more route, so we may as well do it now. The second part of the rule captures the idea of trying to insert a delivery well and before all its good potential insertion points become infeasible.

These rules are first applied to the deliveries to customers that require multiple deliveries over the 2 days. The idea is that these deliveries will be the most difficult to schedule feasibly, so we need to handle these first. When all of these are scheduled, these same insertion rules are then applied to the remaining deliveries.

After a feasible schedule is created, we run one more heuristic, the delivery-amount optimization routine, which finalizes the schedule. It reviews the current schedule, decides which of the customers should have their delivery amounts set above the minimum and, if so, the new amount, and decides where in the final feasible time ranges the delivery times should be set.

The insertion heuristic described above is embedded into a greedy randomized adaptive search procedure (GRASP; see Feo and Resende [18]). A GRASP combines a greedy heuristic with randomization. Whenever the heuristic selects the next delivery to be inserted, it will pick randomly from the q best choices, where q is prespecified. This allows the algorithm to make choices that do not seem to be the best at the time but may provide better opportunities later. In a GRASP framework, the heuristic is executed many times and the best plan obtained is picked.

12.5 Computational Experience

In this section, we present the results of various computational experiments that demonstrate the viability and value of the approach presented in section 12.4 and illustrate many of the complexities of inventory routing problems.

12.5.1 Instances

For our computational experiments we used actual data from two of PRAXAIR's production facilities. We chose these two production facilities because the characteristics of the set of customers they serve are quite different in terms of geography, tank capacities, and usage rates.

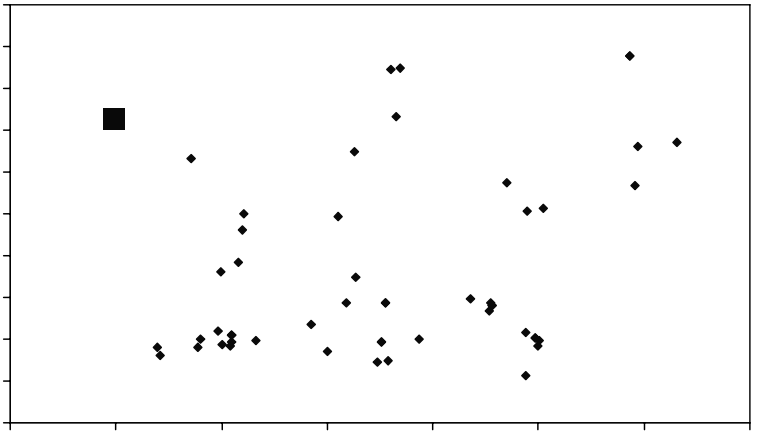

Figure 12.1. *Map of plant A and its customers.*

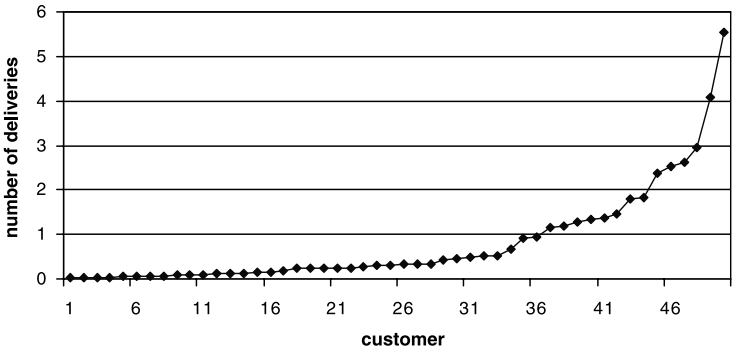

Figure 12.2. *Deliveries per week for plant A customers.*

Production facility A serves 50 customers that are fairly spread out, covering a mostly rural area with some small clusters of customers near cities. The facility is located in the northwestern corner of the state, not in the center, and is represented graphically by the large square (see Figure 12.1). Customers are between 4 minutes and 4.5 hours driving time from the facility, with an average of 3 hours. The average driving time between two customers is 2 hours and 10 minutes. Of the 50 customers, 72% require less than one delivery per week, 16% require between one and two, 8% require between two and three, and 4% require between three and six (see Figure 12.2). With respect to tank capacities, 22% of the customers can receive a delivery of more than a truckload, but 58% cannot receive even half a truckload (see Figure 12.3). In the graph, the heavy line indicates truck capacity.

Production facility B serves 87 customers spread over a large geographic area in the northern United States. The customers are concentrated heavily in the middle of the area, where the facility is located, and become less concentrated as the distance from the center

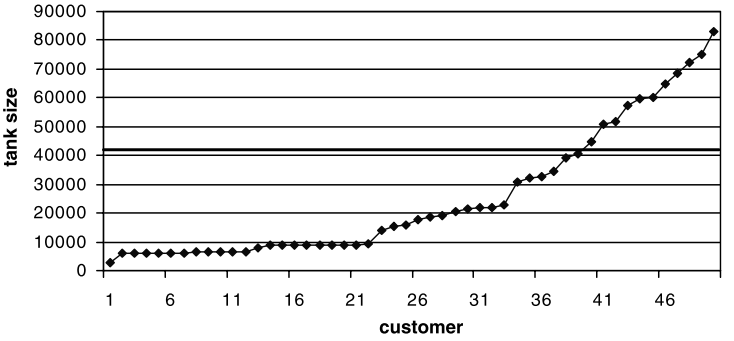

Figure 12.3. *Plant A tank capacity.*

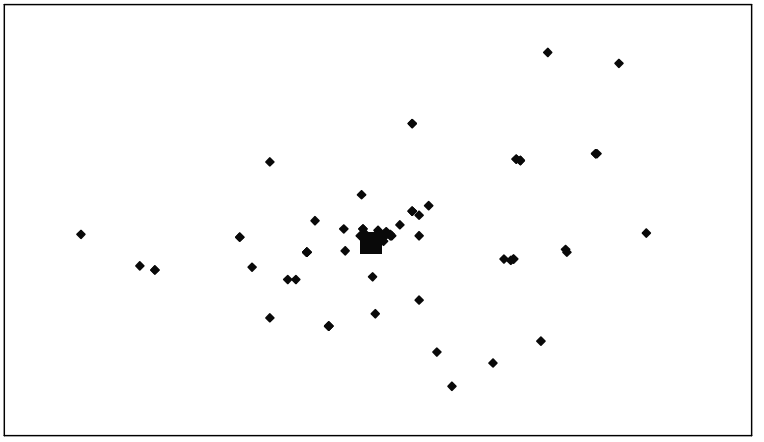

Figure 12.4. *Map of plant B and its customers.*

increases (see Figure 12.4). Customers are between 6 minutes and 10 hours driving time from the facility, with an average of 2 hours and 20 minutes. The average driving time between two points is 3 hours and 40 minutes. In terms of usage, 90% need less than one delivery per week, 8% need between one and two, and only 2% require more than one delivery per week (see Figure 12.5). Furthermore, 21% can receive a delivery of a truckload and 41% cannot receive half of a truckload (see Figure 12.6).

Roughly 75% of the customers at both plants use product 24 hours a day, 7 days a week. Of the customers that are not constant users, many change how they use product depending on the day of the week. Most use product roughly the same way Monday through Friday, but often only 8 to 10 hours per day. The usage pattern usually changes on the weekend, with many of these customers not using product at all on Sundays and less than half of a weekday amount on Saturdays.

Other relevant information used in our computational experiments is that the time of a delivery is calculated as 0.5 + (vehicle pump rate)·(quantity delivered), that it takes 1 hour

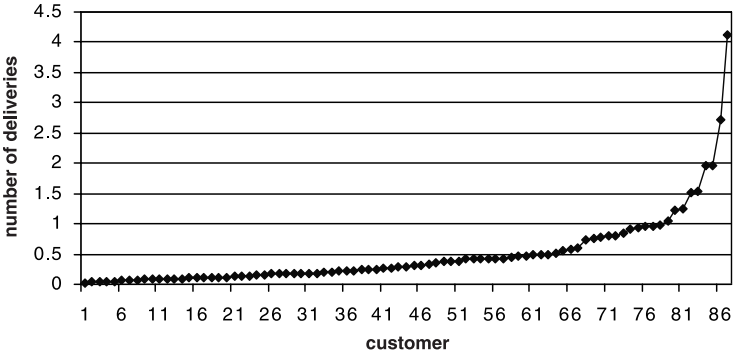

Figure 12.5. *Deliveries per week for plant B customers.*

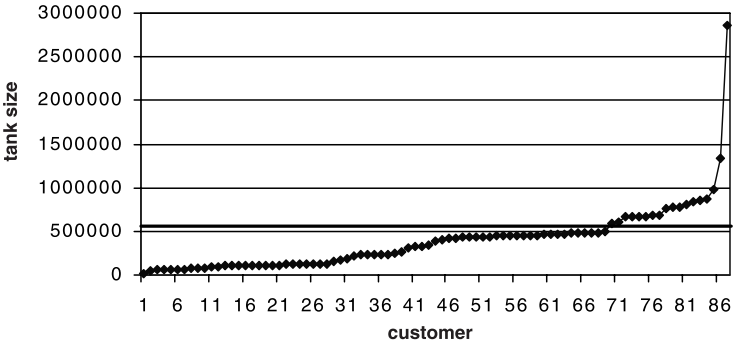

Figure 12.6. *Plant B tank capacity.*

to reload a vehicle at the facility before it can depart again on another route, that all vehicles drive at a speed of 45 mph, and that deliveries can be made 24 hours a day.

Finally, the initial inventory for all customers was chosen randomly, with the restriction that the inventory level should be sufficient to last the customer until the first time a vehicle would be able to arrive at the customer to refill its tank.

12.5.2 Solution Quality

A solution to the IRP for a given planning period specifies which vehicles are visiting which customers on each day of the planning period, in what order the deliveries are being made, and how much is delivered to each customer. However, even with all this information it is still nontrivial to evaluate the quality of the solution. Since the IRP is really an infinite-horizon problem, we have only specified the first part of a solution. For example, if we consider a planning period of 2 weeks, as we do in our computational experiments, it is not obvious how to compare two solutions and claim that one is better than the other. If the total distance traveled in one solution is less than in the other solution, this represents a smaller driving cost. However, if in the solution with a higher total distance traveled,

only full truckload deliveries are made, how can we say this solution is worse? It utilizes the trucks extremely well and may end in a state that is a much better starting point for the deliveries that have to be made in the following weeks.

Therefore, in addition to looking at the obvious statistics, such as the number of trucks (indicated as T), number of routes (R), number of stops (S), percent utilization of the vehicles (U), total volume delivered (V), the total distance traveled (Mile), we look at several other statistics to evaluate the quality of a solution for a 2-week planning period. Some of these statistics are used by PRAXAIR to evaluate their own performance, others are proposed in the literature, and some we just found to be interesting.

A popular statistic used in industry is *average volume per mile* (aV/M). This statistic averages the volume per mile of all the trips, where the volume per mile of a trip is what we expect it to be, namely, the total volume delivered on a trip divided by the total distance traveled on the trip. It is easy to see that this number is very sensitive to the distance of customers from the facility and therefore does not seem to provide reliable information in an averaged form. For example, if we consider a trip to a customer 4 miles from the facility where a full truckload is delivered, the volume per mile is equal to truckload divided by 8. If we consider another trip to a customer 40 miles from the facility where we also deliver a full truckload, the volume per mile is equal to truckload divided by 80. The average of these two volumes per mile ($\frac{11}{160}$ truckload) does not provide much information.

A more sensible statistic, especially over a period of several days, is *total volume per mile* (V/M), defined as the total volume delivered to all customers over the period considered divided by the total distance traveled over this period. Since we are looking only at the first piece of a long-term problem, it makes sense, in this first piece, to deliver more product than required to ensure that customers will not run out, if it can be done at a relatively small cost, i.e., a small increase in distance traveled. A large value of total volume per mile indicates that we are successful at doing so.

A third statistic, inspired by Bell et al. [8], is *weighted volume per mile* (wV/M). In [8], the authors discussed computing a "weighted delivery radius," which for a period equals the amount delivered to each tank times the distance of that tank from the depot summed over all tanks and divided by the amount delivered. With a representative from PRAXAIR, we modified this statistic so that it can be computed for an individual route and such that comparisons of this value among different routes can have meaning. The weighted volume per mile for a trip with n customers is computed as

$$\frac{d_1 \cdot tt_{0,1} + d_2 \cdot tt_{0,2} + \cdots + d_n \cdot tt_{0,n}}{\text{total round trip distance}}$$

The intuition behind this statistic is revealed when we look at the values it gives for the example given above. The value it gives for both trips is 0.5 truckload (which is the largest value possible). It says that both trips are equally good, in fact, as good as possible, since the best we can do when serving a customer for a long period is to deliver full truckloads. When a trip contains several stops to deliver a full truckload or when a trip does not deliver a full truckload, the value of this statistic goes down. The other benefit of this statistic is that it still provides relevant information when it is averaged over a number of trips (assuming all vehicle capacities are equal).

Other statistics that are also important to consider include the *average inventory level before delivery* and *average inventory level after delivery* (indicated as Bef and Aft, respec-

tively), both given as percent of capacity. Obviously, higher values are preferred, especially with respect to the average inventory level before delivery, because a high value indicates that we are less likely to experience stockouts due to fluctuations in usage rates. Furthermore, the average vehicle utilization is an interesting statistic. It tracks what percent of the truck's capacity is used in making deliveries to the customers on a route. We would like this value to be high, but not at all costs. We do not want to drive many extra miles just to ensure a high vehicle utilization. (In practice, there is a strong belief that every vehicle should leave the facility fully loaded and return empty. In part, however, this is motivated by the inherent stochasticity that sometimes allows for larger-than-expected deliveries.)

Finally, we may also want to look at the number of vehicles used. However, we do not want to put too much importance on this statistic in our experiments. In the long term, eliminating a vehicle represents significant savings for a company, but in the short term, we cannot really argue that one solution is really better than another just because it uses one less vehicle.

In our tests, we used the number of vehicles used in practice as the maximum number of vehicles available. We operated under the assumption that the number used in practice was necessary (long-term) and that therefore minimizing this number (short-term) does not make sense. If everything else is equal, however, this may be used as criterion for preferring one solution over another.

12.5.3 Alternate Heuristic

To compare the quality of the solutions produced by our proposed approach to current industrial practices, we implemented a solution approach based on the rules-of-thumb idea and ideas most commonly used in practice. After many discussions with the planners at PRAXAIR, we identified the following rules:

- Create trips around customers that must receive a delivery on the day being considered. Fill up that customer to capacity and then add nearby customers to the trip if there is remaining inventory in the vehicle.
- All customers on a trip are filled to capacity except for the last one.
- Discourage a vehicle from returning to the facility without delivering its remaining capacity to some customer.
- Do not create trips involving only customers that do not require a delivery on the day being considered unless there are excess resources that day and it appears that there will be insufficient resources when the first of these customers requires delivery.

We refer to the heuristic that implements these ideas in our computational experiments as IND APP since it represents an approximation of what is being done in industry.

12.5.4 Computational Experiments

The first experiment compares the solutions obtained by our proposed approach to the solutions obtained by the industry approximation approach. The results can be found in

Table 12.1. *Base case versus industry approximation.*

Setting	T	R	S	U	V	Mile	V/M	aV/M	wV/M	Bef	Aft
BASE	3	65	118	95.72	2613027	18841	138.69	410	18856	24.54	81.64
IND APP	4	67	90	89.58	2519989	18988	132.71	309	18357	11.41	81.27
BASE	3	61	106	90.26	30283480	14226	2128.74	8019	215095	19.91	92.49
IND APP	3	61	93	85.41	28656473	15042	1905.10	7016	206276	9.83	88.97

Table 12.1. In all the tables, the results for plant A appear first, and the results for plant B appear after the dividing line.

Our approach clearly outperforms the industry approximation approach. It does better for both facilities on all the important statistics. The difference in the underlying ideas of the two approaches is most clearly observed in the Bef column. The industry approximation approach is driven by customers that are getting close to running out and that have to be visited, which results in a low average inventory before delivery, whereas our approach looks further ahead and attempts to identify good opportunities to visit customers before they are near run-out.

As we indicated above, we believe that the strength of our approach is that it considers "enough" of the future to make the right decisions. In the next two experiments, we investigate the impact of varying the amount of future considered. In our chosen approach, we consider 5 days in full detail plus 4 weeks in aggregated form beyond this. We note that considering 5 days in full detail is already more than many of the solution approaches proposed in the literature. In Table 12.2, we show the results when we vary the amount of future considered in aggregated form.

It is interesting to observe the increase in the number of deliveries when 6 weeks are considered. When 6 weeks are considered, a larger portion of the objective function value represents future costs, and optimizing with this objective apparently allows us to make some unwise and expensive decisions in the part of the planning period that really counts, i.e., the first 5 days. If we do not consider any future beyond the 5 days, we appear to be missing some beneficial opportunities. There is not much difference, however, between considering 1 week or 4 weeks beyond the 5 days.

Next, we decided to investigate the effect of considering fewer days in full detail. By reducing this number from 5, we make the IPs smaller and therefore easier to solve, but it is not clear what the impact will be on the solutions. In Table 12.3, we show the results when we vary the number of days considered in full detail.

Table 12.2. *Varying the number of weeks.*

Setting	T	R	S	U	V	Mile	V/M	aV/M	wV/M	Bef	Aft
5 days, 0 wk	4	69	119	92.43	2678599	20284	132.05	408	18114	29.87	87.76
5 days, 1 wk	4	67	117	95.04	2674527	19640	136.18	411	18687	25.26	82.18
5 days, 4 wk	3	65	118	95.72	2613027	18841	138.69	410	18856	24.54	81.64
5 days, 6 wk	4	74	125	88.10	2738189	20836	131.42	392	17514	24.26	80.74
5 days, 0 wk	3	64	105	84.58	29773123	14512	2051.62	7087	204340	20.30	92.33
5 days, 1 wk	3	60	102	89.02	29375272	14050	2090.77	7855	218711	18.03	92.16
5 days, 4 wk	3	61	106	90.26	30283480	14226	2128.74	8019	215095	19.91	92.49
5 days, 6 wk	3	57	111	89.40	28026955	13789	2032.56	7496	214607	19.28	85.22

Table 12.3. *Varying the number of days.*

Setting	T	R	S	U	V	Mile	V/M	aV/M	wV/M	Bef	Aft
2 days, 4 wk	*										
3 days, 4 wk	4	69	103	90.32	2617434	19856	131.82	283	18188	27.99	91.30
5 days, 4 wk	3	65	118	95.72	2613027	18841	138.69	410	18856	24.54	81.64
2 days, 4 wk	3	66	102	86.19	31285911	14027	2230.41	7298	214827	18.40	94.75
3 days, 4 wk	3	65	103	84.53	30220444	13172	2294.29	8357	219112	17.83	92.99
5 days, 4 wk	3	61	106	90.26	30283480	14226	2128.74	8019	215095	19.91	92.49

As expected, the quality of the solutions decreases when we consider fewer days in full detail. In fact, when we consider just 2 days, we are unable to construct a solution in which none of the customers runs out of product during the planning period. In this case, the IP selects delivery amounts for customers that turn out to be impossible to schedule with the routing heuristic, because too many deliveries must occur on a specific day and roughly at the same time. It is interesting to observe that when we consider fewer days, the number of stops decreases significantly. Apparently, when we consider more days in full detail, the IP starts looking for inexpensive opportunities to make deliveries to customers that require a delivery only a few days out, whereas the IP is unable to do that when fewer days are considered in full detail.

Besides the amount of future considered, the quality of the solution also is affected by the parameter settings used in the routing and scheduling heuristic and whether delivery amount optimization is active. When delivery amount optimization is not active, a delivery amount cannot be set above the amount specified by the IP. In Table 12.4 we present the results of our approach with and without delivery optimization. Without delivery optimization (entries IP AMT), we expect the average vehicle utilization and the total volume to be less. On the other hand, we do not want it to be much less because that would suggest that our integer program is not making the right decisions.

Looking at the summary statistics, the delivery amount optimization clearly does improve truck utilization and also leads to a significantly better total volume per mile and weighted volume per mile. The increase in total volume delivered, however, was slightly less than 3%.

Our default settings for the GRASP are to run the routing and scheduling heuristic 25 times and to select from among the three best choices. To investigate the impact of these settings as well as the importance of randomization, we conducted an experiment in which we executed the heuristic without any randomization (pure greedy) and with different settings for the number of replications. The results are presented in Table 12.5.

Table 12.4. *Delivery optimization.*

Setting	T	R	S	U	V	Mile	V/M	aV/M	wV/M	Bef	Aft
IP AMT	3	69	126	87.56	2537554	19929	127.33	330	17077	24.90	78.00
BASE	3	65	118	95.72	2613027	18441	138.69	410	18856	24.54	81.64
IP AMT	3	61	111	85.81	28789623	15274	1884.88	6612	202495	19.12	85.10
BASE	3	61	106	90.26	30283480	14226	2128.74	8019	215095	19.91	92.49

Table 12.5. *Randomization.*

Setting	T	R	S	U	V	Mile	V/M	aV/M	wV/M	Bef	Aft
NO RAND	4	71	121	87.82	2618675	21799	120.13	394	17092	23.54	79.87
5	3	67	116	93.70	2636760	19315	136.51	419	18592	24.76	82.77
25	3	65	118	95.72	2613027	18441	138.69	410	18856	24.54	81.64
50	4	65	121	93.22	2544803	19176	132.71	327	17793	23.74	77.21
NO RAND	3	59	99	86.22	27979401	14378	1945.99	6990	200752	15.28	88.34
5	3	62	111	87.99	30004693	15094	1987.86	6773	201000	19.88	89.13
25	3	61	106	90.26	30283480	14226	2128.74	8019	215095	19.91	92.49
50	3	63	111	87.57	30343074	14605	2077.58	7762	211049	19.88	89.61

Without randomization the solution has a noticeably low average vehicle utilization, total volume per mile, and average weighted volume per mile. On the other hand, going to 50 replications does not seem to improve over 25 replications; in fact, it does slightly worse. This is possible because the replications are for 2 days of the schedule at a time. Which schedule is selected affects what deliveries are made, what the customer inventories are at the end of the 2 days, and thus the input for the next integer program that is solved.

To obtain more insight in the behavior of the GRASP, we kept track of the total distance traveled for all 50 replications at two different points in the 2-week planning period. The results are plotted in Figure 12.7.

The criterion used to pick the best solution out of the 25 produced by the GRASP is total travel distance. However, the results may be quite different if we decide to use average weighted volume per mile as the criterion to pick the best solution. Our last computational experiment relating to the GRASP compares the behavior based on different selection criteria. The results are presented in Table 12.6. (D) indicates schedules selected based on mileage and (WVM) stands for schedules selected based on average weighted volume per mile.

Various other parameters can be set in the routing and scheduling heuristic. Some of these help to construct solutions that reflect company policy. For example, in our default approach, we did not penalize waiting time at customers. In practice, however, waiting time is often strongly discouraged or even not allowed. To show the impact of discouraging waiting time on the quality of the solutions, Table 12.7 presents the solution statistics when we penalize waiting time significantly.

As we expected, when we allow waiting at customers, we get a higher truck utilization, a better total volume per mile, and average weighted volume per mile.

12.6 Conclusion

We presented the IRP and an optimization-based approach for its solution. Extensive computational experiments indicate the value and potential of optimization-based approaches for complex routing and scheduling problems. The IRP is of special interest because it integrates two components of supply chain management: inventory control and vehicle routing. This type of integration is essential to improve overall system performance.

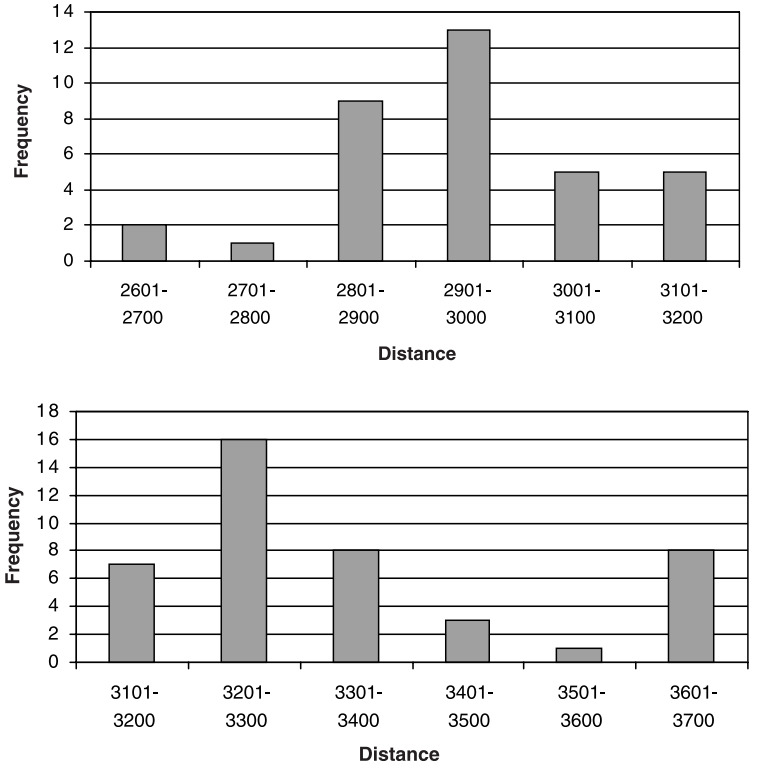

Figure 12.7. *Effects of randomization.*

Table 12.6. *Selection criteria.*

Setting	T	R	S	U	V	Mile	V/M	aV/M	wV/M	Bef	Aft
(D)	3	65	118	95.72	2613027	18441	138.69	410	18856	24.54	81.64
(WVM)	4	65	120	96.83	2643371	19341	136.67	423	19018	24.99	82.64
(D)	3	61	106	90.26	30283480	14226	2128.74	8019	215095	19.91	92.49
(WVM)	3	58	105	91.34	29138802	14235	2046.98	8291	220091	20.50	91.76

Table 12.7. *Waiting time.*

Setting	T	R	S	U	V	Mile	V/M	aV/M	wV/M	Bef	Aft
NO CHG	3	65	118	95.72	2613027	18441	138.69	410	18856	24.54	81.64
WAIT CHG	4	70	119	90.50	2660785	20019	132.91	390	17817	21.26	77.66
NO CHG	3	61	106	90.26	30283480	14226	2128.74	8019	215095	19.91	92.49
WAIT CHG	3	58	103	88.46	28220108	14908	1892.95	6298	197806	18.06	88.17

Bibliography

[1] S. Anily and A. Federgruen. One warehouse multiple retailer systems with vehicle routing costs. *Management Science*, 36:92–114, 1990.

[2] S. Anily and A. Federgruen. Rejoinder to "One warehouse multiple retailer systems with vehicle routing costs." *Management Science*, 37:1497–1499, 1991.

[3] M. Ball. Allocation/routing: Models and algorithms. In B.L. Golden and A.A. Assad, editors, *Vehicle Routing: Methods and Studies*, Elsevier Science, Amsterdam, Netherlands, 1988.

[4] J. Bard, L. Huang, M. Dror, and P. Jaillet. A branch and cut algorithm for the VRP with satellite facilities. *IIE Transactions on Operations Engineering*, 30:821–834, 1998.

[5] J. Bard, L. Huang, P. Jaillet, and M. Dror. A decomposition approach to the inventory routing problem with satellite facilities. *Transportation Science*, 32:189–203, 1998.

[6] D. Barnes-Schuster and Y. Bassok. Direct shipping and the dynamic single-depot/multi-retailer inventory system. *European Journal of Operational Research*, 101:509–518, 1997.

[7] Y. Bassok and R. Ernst. Dynamic allocations for multi-product distribution. *Transportation Science*, 29:256–266, 1995.

[8] W. Bell, L. Dalberto, M.L. Fisher, A. Greenfield, R. Jaikumar, P. Kedia, R. Mack, and P. Prutzman. Improving the distribution of industrial gases with an on-line computerized routing and scheduling optimizer. *Interfaces*, 13:4–23, 1983.

[9] O. Berman and R. Larson. Deliveries in an inventory/routing problem using stochastic dynamic programming. Technical report, Massachusetts Institute of Technology, Cambridge, MA, 1999.

[10] J. Bramel and D. Simchi-Levi. A location based heuristic for general routing problems. *Operations Research*, 43:649–660, 1995.

[11] A. Campbell, L. Clarke, A. Kleywegt, and M. Savelsbergh. Inventory routing. In T. Crainic and G. Laporte, editors, *Fleet Management and Logistics*, Kluwer, Boston, MA, 1998.

[12] S. Cetinkaya and C. Lee. Stock replenishment and shipment scheduling for vendor managed inventory systems. Technical report, Texas A & M University, College Station, TX, 1999.

[13] T. Chien, A. Balakrishnan, and R. Wong. An integrated inventory allocation and vehicle routing problem. *Transportation Science*, 23:67–76, 1989.

[14] M. Dror, M. Ball, and B.L. Golden. Computational comparison of algorithms for the inventory routing problem. *Annals of Operations Research*, 4:3–23, 1985.

[15] M. Dror and L. Levy. Vehicle routing improvement algorithms: Comparison of a "greedy" and a matching implementation for inventory routing. *Computers and Operations Research*, 13:33–45, 1986.

[16] M. Dror and Ball. M. Inventory/routing: Reduction from an annual to a short period problem. *Naval Research Logistic Quarterly*, 34:891–905, 1987.

[17] A. Federgruen and P. Zipkin. A combined vehicle routing and inventory allocation problem. *Operations Research*, 32:1019–1036, 1984.

[18] T.A. Feo and M.G.C. Resende. Greedy randomized adaptive search procedures. *Journal of Global Optimization*, 6:109–133, 1995.

[19] M.L. Fisher, A. Greenfield, R. Jaikumar, and P. Kedia. Real-time scheduling of a bulk delivery fleet: Practical application of Lagrangean relaxation. Technical report, The Wharton School, University of Pennsylvania, 1982.

[20] F. Fumero and C. Vercellis. Synchronized development of production, inventory, and distribution schedules. *Transportation Science*, 33:330–340, 1999.

[21] G. Gallego and D. Simchi-Levi. On the effectiveness of direct shipping strategy for the one-warehouse multi-retailer r-systems. *Management Science*, 36:240–243, 1990.

[22] B.L. Golden, A.A. Assad, and R. Dahl. Analysis of a large scale vehicle routing problem with an inventory component. *Large Scale Systems*, 7:181–190, 1984.

[23] P. Jaillet, L. Huang, J. Bard, and M. Dror. A rolling horizon framework for the inventory routing problem. Working paper, University of Texas, Austin, 1997.

[24] G.A.P. Kindervater and M.W.P. Savelsbergh. Vehicle routing: Handling edge exchanges. In E.H.L. Aarts and J.K. Lenstra, editors, *Local Search in Combinatorial Optimization*, Wiley, Chichester, UK, 1997, pp. 337–360.

[25] A.J. Kleywegt, V.S Nori, and M.W.P. Savelsbergh. The stochastic inventory routing problem with direct deliveries. Technical Report TLI99-01, Georgia Institute of Technology, Atlanta, GA, 1999.

[26] A. Minkoff. A markov decision model and decomposition heuristic for dynamic vehicle dispatching. *Operations Research*, 41:77–90, 1993.

[27] V. Nori. *Algorithms for Dynamic and Stochastic Logistics Problems*. Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA, 1999.

[28] M.M. Solomon. Algorithms for the vehicle routing and scheduling problems with time window constraints. *Operations Research*, 35:254–265, 1987.

[29] P. Trudeau and M. Dror. Stochastic inventory routing: Route design with stockouts and route failures. *Transportation Science*, 26:171–184, 1992.