

Optimizing Urgent Material Delivery by Maximizing Inventory Slack

Adam Montjoy
Jeffrey Herrmann

The
Institute for
Systems
Research



A. JAMES CLARK
SCHOOL OF ENGINEERING

ISR develops, applies and teaches advanced methodologies of design and analysis to solve complex, hierarchical, heterogeneous and dynamic problems of engineering technology and systems for industry and government.

ISR is a permanent institute of the University of Maryland, within the A. James Clark School of Engineering. It is a graduated National Science Foundation Engineering Research Center.

www.isr.umd.edu

Optimizing Urgent Material Delivery by Maximizing Inventory Slack

Adam Montjoy Jeffrey W. Herrmann

June 12, 2012

Abstract

Motivated by the need to create plans for delivering medication quickly during a public health emergency, this paper formulates the Inventory Slack Routing Problem. The planning goal is to deliver material as early as possible to demand sites from a central depot at which material arrives over time. The objective function is to maximize the minimum slack of the deliveries so that all demand sites are treated equitably. This paper presents and analyzes the problem, discusses solution techniques, and discusses the results of computational experiments used to compare the solution techniques. Although motivated by planning for public health emergencies, this work is also applicable to other settings in which material must be delivered quickly to multiple facilities that rely upon the material to operate.

1 Background

This paper presents the Inventory Slack Routing Problem (ISRP). Our work on this problem has been motivated by the need to create plans for delivering medication quickly during a public health emergency, but the work is also applicable to other settings in which material is required urgently at multiple facilities that rely upon the material to operate, so stockouts is critically important.

Events in the last ten years have highlighted the increased need for emergency preparedness by government officials. Events such as the terrorist attacks on September 11, 2001, Hurricane Katrina, and the 2008 earthquake

in Chengdu, China, have provided real world examples of ill-preparation for major disasters [4]. Thus, it is important for government officials to anticipate disasters and plan accordingly. Mathematical models and decision support tools can be used to support planning activities.

Some scenarios could require the quick and efficient distribution of medication to a large number of people. For instance, the widespread release of anthrax in a metropolitan area could result in casualties equivalent to that of a small nuclear explosion [11]. In this scenario (and others involving mass vaccination against communicable diseases such as smallpox and influenza), it is logical to create Points of Dispensing (PODs) such that large populations can be given medication without having to travel to one location. PODs may be setup in schools, recreation centers, churches, and other non-medical facilities. The medication to be distributed at these PODs must be delivered quickly from a central depot as soon as it arrives.

The proposed research is motivated by work with public health officials in the state of Maryland who must plan the logistics for distributing medication to the PODs from a central location. We consider the problem at the state and local levels (not the national level). After the decision for mass dispensing is made, local health departments will begin preparing to open multiple PODs simultaneously at a designated time. The state will request medication from the federal government, who will deliver an initial but limited supply of medication to a state receipt, storage, and stage (RSS) facility (which we call the “depot”). Contractors will deliver more medication to the depot, but the state will begin shipping medication from the depot to the PODs before everything arrives from the contractors. The deliveries to the depot arrive in batches that we call “waves.”

Poor medication distribution plans will delay the time that some PODs receive medication. This can delay the opening of these PODs, or cause PODs to stop dispensing when they have no medication, and some residents may not get their medication in a timely manner, which increases their risk of death or illness. Clearly, there are many uncertainties in medication distribution, including the timing of shipments to the depot, the time needed to load and unload vehicles, travel times, and the demand for medication at each POD. For this reason, planners need a robust plan. In particular, it is better if the plan calls for delivering medication to PODs much earlier than it is needed. This improves the likelihood that the PODs will open on-time, will not run out of medication during operations, and will dispense medication to the largest number of people in a timely manner.

This problem, which we call the Inventory Slack Routing Problem (ISRP), has some features of the inventory routing problem but also has some unique assumptions, constraints, and objectives. In this case, a set of sites are served by a set of vehicles that deliver material during a short time span. Thus, the objective is not to minimize the cost or maximize the profit. Instead the objective is to increase the earliness of the deliveries, the interval between each delivery and the time at which the site would its inventory if the delivery were delayed. This value will be known as the slack.

1.1 Related Work

Much research has been done to develop models to improve emergency preparedness planning. Hupert et al. [11] presented a model to predict the hospital surge after a large-scale anthrax attack and emphasized the importance of timely antibiotic distribution, making logistics of delivery equally important. Similarly, researchers have created simulation methods and planning tools for PODs in makeshift locations such as school gymnasiums [1, 2, 12].

The operations of firefighters, emergency medical services, and police departments have motivated research into location models [3, 8, 9] and dynamic vehicle routing models [18]. However, these models are not relevant to the medication distribution problem, which is more closely related to the inventory routing problem, which will be discussed later in this section.

Planning humanitarian logistics is related to the Vehicle Routing Problem (VRP) and Inventory Routing Problem (IRP). These problems have been applied to a variety of commercial, military, and government applications. The following description of the VRP is by Toth and Vigo [17].

The VRP details the delivery of a set of goods to a set of customers by a set of vehicles. These goods are stored at a depot, or a set of depots, and are delivered by a road network. This road network is usually detailed using a graph with arcs representing roads and vertices as the sites and depots. The solution to the VRP specifies a route for each vehicle that begins and ends at the depot. Typical VRP problems have the following characteristics: customer locations, demands for the customers, time windows for the customers, loading and unloading times, and a set of available vehicles that can be used.

In many cases, it may not be possible to fully satisfy all of the customer demand, and priorities or penalty functions must be employed. With this, it is possible to formulate various objective functions to obtain a solution,

including minimization of global transportation cost, minimization of vehicles used, balancing routes for travel times and load, and minimization of penalties. The VRP is a well-researched technique, and many heuristics, mathematical programming, and search techniques are available.

The VRP has many variations including the Inventory Routing Problem (IRP) [5, 6, 13]. The following description of the IRP is by Campbell et al. [6]. The IRP differs from the VRP because the delivery company decides when and how much to deliver to customers. The objective is to minimize total cost over the planning horizon while ensuring that customers do not run out of product. A single product is delivered from a single depot to a set of n customers over a specified time period. These customers are served by a set of V identical vehicles, each of which has a capacity of Q . A problem solution answers three questions: when to serve a customer, how much to deliver, and which routes to follow?

Most solutions detailed in literature focus on short-term scenarios solved by mathematical programming techniques. There is a lack of basic heuristics for solving IRPs. The Inventory Slack Routing Problem (ISRP) that we present is similar to an IRP but has some unique assumptions. The main concern is to supply material (medication) as quickly as possible, not to minimize cost. As Hupert et al. [11] emphasize, delaying the start of POD operations will significantly increase the number of people hospitalized. In addition, the limited availability of material (medication) at the depot adds an additional constraint to the problem. Finally, because there is uncertainty in the loading times, unloading times, travel times, and demand, it is necessary to have slack as a hedge against these uncertainties, and more slack is better. The objective of the ISRP is to maximize the minimum slack in order to develop a more robust plan.

1.2 Overview of Paper

The ISRP considers how to deliver inventory from a central depot to the demand sites as early as possible given the availability of inventory at the depot over time. A key constraint is that a solution must specify sufficient deliveries to every site so that it receives the total amount required. We develop a measure known as “slack” to denote how early a delivery reaches its destination before it is needed. Slack is the difference in time between the expected time when all previously delivered inventory will be exhausted and the time that the delivery occurs. Solving the problem with a standard total

travel time or total cost objective function would not adequately address the goal of delivering as early as possible.

A slack value can be calculated for each delivery. Maximizing the sum of these values could yield an inequitable solution in which some deliveries have a large slack (and the sites have excess inventory) and other deliveries have little (or negative) slack (and the sites have little or no inventory). Thus, we maximize the minimum slack, which achieves the primary objective of ensuring that inventory arrives in a timely matter while also equitably allocating inventory to the sites.

Section 2 introduces the problem formulation and presents a simple example. Section 3 describes the solution approach, which separates the problem, and the techniques used to solve the subproblems. Section 4 describes the results of computational tests used to compare the performance of the solution techniques. Section 5 concludes the paper.

2 Problem Formulation

We are given a set of vehicles to deliver material from a depot to a set of sites that will consume this material. In general, only some of the material is available at the depot at the beginning of the time horizon, and more material will become available in “waves,” which are batch deliveries to the depot throughout the time horizon. The schedule of these deliveries to the waves is given. The sites will start operating at a designated time. While operating, each site consumes material at a given rate (for instance, at a POD, the consumption rate equals the dispensing rate, which depends upon the number of personnel at the site), and this demand may vary from site to site. The vehicles must deliver enough material from the depot to the sites to satisfy the total demand over the time horizon. The following section details the notation to be used.

Although vehicles could follow different routes each time they leave the depot and sites could be served by multiple vehicles, this makes supervising and performing the deliveries more complex in practice. We therefore assume that each and every site is assigned to exactly one vehicle, and each vehicle always follows the same route to visit the sites assigned to it.

2.1 Notation

We first describe an instance of ISRP. We denote time as t , where $t = 0$ refers to the first instant that the depot has inventory. There are n sites, and k is the index for the sites. The depot is denoted as site $n + 1$. The cumulative amount of material delivered to the depot between time 0 and t is denoted by $I(t)$ (if the material arrives in batches, this will be a step function that increases at the times that batches arrive). Site k will start consuming material at $t = 0$ and will stop at time e_k . Each site has a non-negative, possibly non-stationary, demand rate of $L_k(t)$ for $0 \leq t \leq e_k$. Let $D_k = \int_0^{e_k} L_k(t) dt$ be the total material requirements at site k . The depot will eventually receive enough material to satisfy all of the material requirements at all of the sites. That is, there is a time t such that $I(t) = \sum_{k=1}^n D_k$.

There are V vehicles that will deliver material to the sites, and v is the index for the vehicles. Vehicle v has a capacity of C_v , which is an upper bound on the amount of material that can be loaded into the vehicle at the depot. At site k , the time required to unload a delivery is p_k ; at the depot, p_{n+1} is time required to load a vehicle. The travel time from site i to j is denoted as c_{ij} .

The decision variables for this problem specify the route for each vehicle to take on each trip and how much to deliver to each site on each trip. Since we are considering a short time period, we assume that a vehicle is assigned a subset of the sites and will visit these sites in the same sequence on each trip (that is, a trip starts at the depot, visits one or more sites, and then returns to the depot). This is done in order to reduce confusion.

A feasible solution specifies, for each vehicle, a route, the number of trips that it makes, the time to start each trip, and the quantity to deliver to each site on each trip. Let r_v be the number of trips that vehicle v makes. The sequence of sites that vehicle v will visit on each trip is denoted by σ_v . Each trip j of vehicle v starts at time t_{vj} by loading at the depot, continues by visiting the sites in σ_v , and ends when the vehicle returns to the depot. The quantity q_{vjk} is delivered to each site $k \in \sigma_v$ on trip j .

The time to complete one trip by vehicle v is denoted by y_v . If vehicle v visits site k , then the the site's delivery will be finished by a vehicle in time w_{vk} (after the vehicle leaves the depot). These quantities can be calculated as follows, where $[i]$ refers to the i -th site in σ_v :

$$\begin{aligned}
w_{v[1]} &= p_{n+1} + c_{n+1,[1]} + p[1] \\
w_{v[i]} &= w_{v[i-1]} + c_{[i-1],[i]} + p[i] \quad i = 2, \dots, r_v \\
y_v &= w_{v[r_v]} + c_{[r_v],n+1}
\end{aligned}$$

We also define $R_k(Q)$ as the time at which site k will finish consuming Q units of material and will need more material to continue operating. We define $R_k(Q) = \max\{T : \int_0^T L_k(t)dt \leq Q\}$ for all $Q < D_k$.

The slack of a delivery measures how early that delivery is. It equals the difference between the time that the delivery is completed (the material is unloaded) and the time that the site would consume the last of all material that was delivered in previous deliveries. If this delivery were delayed by an amount of time greater than its slack, the site would be unable to operate because it would have no material. The objective s is the minimum of all of the slacks of the deliveries.

2.2 General Formulation

The ISRP can be formulated as follows.

$$\begin{aligned} \max s & \tag{1} \\ t_{vj} - t_{v,j-1} &\geq y_v \quad v = 1, \dots, V; j = 2, \dots, r_v \tag{2} \\ \sum_{(a,b):t_{ab} \leq t_{vj}} \sum_{k \in \sigma_a} q_{abk} &\leq I(t_{vj}) \quad v = 1, \dots, V; j = 1, \dots, r_v \tag{3} \\ \sum_{k \in \sigma_v} q_{vjk} &\leq C_v \quad v = 1, \dots, V; j = 1, \dots, r_v \tag{4} \\ \sum_{j=1}^{r_v} q_{vjk} &= D_k \quad v = 1, \dots, V; k \in \sigma_v \tag{5} \\ s + (t_{vj} + w_{vk}) - R_k &(\sum_{i=1}^{j-1} q_{vik}) \leq 0 \quad v = 1, \dots, V; k \in \sigma_v; j = 1, \dots, r_v \tag{6} \\ \sigma_v \cap \sigma_w &= \emptyset \quad v = 1, \dots, V-1; w = v+1, \dots, V \tag{7} \\ \cup_{v=1}^V \sigma_v &= \{1, \dots, n\} \tag{8} \\ t_{vj} &\geq 0 \quad v = 1, \dots, V; j = 1, \dots, r_v \tag{9} \\ q_{vjk} &\geq 0 \quad v = 1, \dots, V; j = 1, \dots, r_v; k \in \sigma_v \tag{10} \end{aligned}$$

Equation 1 is the objective value which denotes maximization of the minimum slack where slack can be defined for each delivery to each site. This value is defined as the difference in time between when a shipment is made and when it is needed. Equation 2 ensures that no vehicle can start a new trip before it has completed its previous trip. Equation 3 limits the amount of material that is available to be loaded onto a vehicle at the start of a trip. Equation 4 is the vehicle capacity constraint. Equation 5 ensures that every site receives the amount of material required. The minimum slack is bounded by the slack of each delivery, as described by Equation 6.

Equation 7 ensures that no two vehicles service the same sites, and Equation 8 ensures that all sites are serviced. The nonnegativity constraints for the decision variables are given by Equations 9 and 10.

If all of the required material were available at $t = 0$, the vehicle capacity constraint (Equation 4) did not exist, and the total rate at which the sites consume material were greater than the rate at which the vehicles can deliver material, then the slack of each delivery would be less than the slack of the previous delivery to that site. In that case, this problem would be equivalent to minimizing the maximum delivery time. A problem similar to this was

studied by Campbell et al. [7] in an application for humanitarian logistics.

This paper will focus on a specific version of the ISRP. We assume that all vehicles have the same capacity (all $C_v = C$); that all sites will begin operating at the same time T_1 and end operating at the same time T_2 (that is, all $e_k = T_2$), with $0 < T_1 < T_2$; and that each site k has a constant demand rate L_k so that $L_k(t) = 0$ for $0 \leq t < T_1$ and $L_k(t) = L_k$ for $T_1 \leq t \leq T_2$. Thus, $D_k = L_k(T_2 - T_1)$.

Note that the slack of the first delivery to each site does not depend on previously delivered material because none has been previously delivered. Thus, in some situations the minimum slack will occur on the first wave, and the delivery quantities do not influence the minimum slack. It is possible to show that this happens when, for every delivery, the time needed to consume all inventory previously delivered is greater than the difference between that delivery time and the first delivery time (i.e., $\sum_{i=1}^{j-1} q_{vjk} / L_k \geq t_{vj} - t_{v1}$).

2.3 Combining Deliveries

Our approach for solving the ISRP determines the trip start times using a deliver-when-possible algorithm that begins a new trip whenever there is material available at the depot (see Section 3.2). If more material is due to arrive at the depot soon, however, waiting for that material and then starting a trip may increase slack. When vehicle capacity is sufficient, it is always better to make a delivery of available inventory if the vehicle will return before the next wave to the depot. However, when the vehicle can not return before the next wave, it may be better to wait for the next wave.

Consider a feasible solution in which vehicle v visits sites $k \in \sigma_v$ and makes trips j and $j + 1$ (which start at times t_{vj} and $t_{vj} + y_v$) but has the option of waiting and combining them into one trip that would start at t_{vj}^* (and this is feasible with respect to capacity). Let s_{vjk} and $s_{v(j+1)k}$ denote the slacks at site k for the two deliveries, and let s_{vjk}^* denote the slack at site k for the combined delivery. Let Q_{vjk} be the cumulative amount of material delivered to the site prior to t_{vj} , and let q_{vjk} denote the amount of the first delivery to site k .

$$\begin{aligned} s_{vjk} &= T_1 + \frac{Q_{vjk}}{L_k} - (t_{vj} + w_{vk}) \\ s_{v(j+1)k} &= T_1 + \frac{Q_{vjk} + q_{vjk}}{L_k} - (t_{vj} + y_v + w_{vk}) \\ s_{vjk}^* &= T_1 + \frac{Q_{vjk}}{L_k} - (t_{vj}^* + w_{vk}) \end{aligned}$$

If $\frac{q_{vjk}}{L_1} \leq t_{vj} + y_v - t_{vj}^*$, then $s_{v(j+1)k} \leq s_{vjk}^* < s_{vjk}$. That is, because the delivery quantity is sufficiently small, the minimum slack of the two deliveries to this site is $s_{v(j+1)k}$ and this is not greater than s_{vjk}^* , so the combined delivery does not decrease the minimum slack at this site. If this condition holds for all $k \in \sigma_v$, then the combined delivery does not decrease the minimum slack of this solution, and the solution formed by combining the two deliveries is not worse than the original.

This discussion shows that, if a wave brings to the depot a relatively small amount of material just before another wave arrives, it is reasonable to delay vehicle trips until the second wave arrives. (The delivery quantities of any trips that did start would also be relatively small.) Thus, the first wave does not need to be considered separately when scheduling deliveries (it is added to the second wave). Thus, we do not consider cases with such small waves.

2.4 Example Instance

To illustrate the importance of scheduling quantities, this section presents a small example. This example is taken from the instances used in the computational results section. This instance has three vehicles and five sites. Three waves of material arrive at the depot and are available at $t = 0$, $t = 180$, and $t = 360$. The material available in each wave is 48,000, 98,000, and 73,000. The first vehicle visits sites 5 and 4; the travel times are $w_{15} = 45$, $w_{14} = 73$, and $y_1 = 90$. The second vehicle visits only site 3; the travel times are $w_{23} = 46$ and $y_2 = 76$. The third vehicle visits sites 2 and 1; the travel times are $w_{32} = 45$, $w_{31} = 74$, and $y_3 = 91$. Each vehicle begins a trip at each wave and returns before the next wave, so the start times $t_{v1} = 0$, $t_{v2} = 180$, and $t_{v3} = 360$. The consumption start and end times are $T_1 = 600$ and $T_2 = 1200$, and the site demand rates are $L_1 = 50$, $L_2 = 75$, $L_3 = 100$, $L_4 = 60$, and $L_5 = 80$.

If the delivery quantities to each site for each trip were decided by allocating the material in each wave so that the quantities are proportional the site demand rates, then we would have the solution displayed in Table 1. The slack calculations for this solution are displayed in Table 2. The minimum slack of 478 minutes occurs on the second delivery to site 1 by vehicle 3.

Given these routes and trip start times, this solution is not the best allocation of inventory. It is possible to find a feasible solution that, in the earlier trips, delivers more material to sites that are visited later in each route. The revised schedule in Table 3 was found using the linear programming

Table 1: Schedule for example.

Vehicle v	Site k	Delivery quantity		
		q_{v1k}	q_{v2k}	q_{v3k}
1	5	10,521	21,479	16,000
	4	7,890	16,110	12,000
2	3	13,151	26,849	20,000
3	2	9,863	20,137	15,000
	1	6,575	13,425	10,000

Table 2: Slack calculations for example.

Vehicle v	Site k	w_{jk}	Slack		
			s_{v1k}	s_{v2k}	s_{v3k}
1	5	45	555	507	595
	4	73	527	479	567
2	3	46	540	492	580
3	2	45	555	507	595
	1	74	526	478	566

formulation presented in Section 3.3. The slack calculations for this solution can be seen in Table 4. The new minimum slack is 494, and the slack of every delivery in the second wave equals this value. Note that sites 1, 3, and 4 received more material in the first wave, and sites 2 and 5 received less. Also note that first wave slacks did not change.

3 Solution Approaches

The ISRP can be separated into three sets of decisions (subproblems): (“routing”) assigning sites to each vehicle and sequencing them to create routes, (“scheduling”) determining how many trips each vehicle will make and when each trip will start, and (“allocation”) determining the amount of material to deliver to each site on each trip. We developed the following algorithms for these subproblems. For the routing problem, we developed a route-first, cluster-second heuristic (described in Section 3.1). For the scheduling problem, we developed a deliver-when-possible algorithm (described in Section

Table 3: Revised schedule for example.

Vehicle v	Site k	Delivery quantity		
		q_{v1k}	q_{v2k}	q_{v3k}
1	5	9,506	20,789	17,705
	4	8,809	16,756	10,435
2	3	13,382	23,949	22,669
3	2	8,912	19,819	16,269
	1	7,391	14,611	7,998

Table 4: Slack calculations for revised schedule.

Vehicle v	Site k	w_{jk}	Slack		
			s_{v1k}	s_{v2k}	s_{v3k}
1	5	45	555	494	574
	4	73	527	494	593
2	3	46	540	494	553
3	2	45	555	494	578
	1	74	526	494	606

3.2). For the allocation problem, we developed a linear program (described in Section 3.3).

From these components we developed three solution approaches for the ISRP: (1) the Routing-and-Scheduling (R&S) approach, which solves the sub-problems in order with the route-first, cluster-second heuristic, the deliver-when-possible algorithm, and the linear program; (2) an adaptive large neighborhood search (ALNS), which first searches over the routes, using the deliver-when-possible algorithm to schedule deliveries and a heuristic that can generate good material allocations, and then uses the linear program after the search ends (this is described in Section 3.4); and (3) a branch-and-bound approach that enumerates the routes, uses the deliver-when-possible algorithm to schedule deliveries, and uses the linear program both to create bounds for incomplete solutions and to evaluate complete solutions (this is described in Section 3.5).

3.1 Routing Vehicles to Sites

Our route-first, cluster-second heuristic approach is based on the procedure presented in Toth and Vigo [17]. First, a “big route” is constructed with the depot and all of the sites. Second, the big route is divided into clusters, one for each vehicle. Because the vehicles are identical, we used procedures that try to keep the clusters the same “size” with respect to travel time or total demand rate. The sequence from the big route is used to sequence the sites in each vehicle’s route. This method was first presented by Montjoy et al. [14].

To construct the “big route,” we used two different methods that do not use X-Y coordinates. In many real world situations, the X-Y coordinates are not as important as the travel times between sites, and, in some situations, the X-Y coordinates may be unavailable. The first method started at the depot and used the nearest neighbor heuristic to select the next site in the route until all sites were visited. The route ended back at the depot. Given this route, the second method used a 2-opt exchange to improve the route by reducing the total travel time.

Once a big route was obtained, it was necessary to divide (“cluster”) the sites among all of the vehicles available. The sites are first divided between vehicles as equally as possible. Each vehicle route includes a subsequence of the big route and is completed by adding the trip from the depot to the first site and the trip from the last site back to the depot. Because this division of the big route ignores both the demand at the sites and the travel times, the durations (and demands) of the routes may vary widely, which can reduce the minimum slack of any solution constructed from these routes. Thus, we used improvement algorithms to reduce the variation; one considered the travel time, and one considered the total demand.

The first improvement algorithm method sought to make the route durations as similar as possible by minimizing the range of route durations. This method begins by calculating each vehicle’s route duration. In each iteration, the algorithm examines the vehicles with maximum and minimum travel times and considers moving sites at the beginning (or end) of one route to the previous (or next) vehicle’s route. If the potential move decreases the range of route durations, then the routes are updated to reflect this change. This continues until no further improvement can be made.

The second improvement algorithm is similar but sought to make the total demand of the routes as similar as possible. For route σ_v , the total demand

of the route $D_v = (T_2 - T_1) \sum_{k \in \sigma_v} L_k$. At each iteration, the algorithm moves a site from one route to another to decrease the range of total demand.

Given two methods for constructing the big route and two algorithms for improving the routes after dividing the big route, we have four versions of this heuristic.

3.2 Scheduling Deliveries

Given routes for the vehicles, our deliver-when-possible algorithm was used to find feasible start times for each trip. This procedure begins by temporarily allocating to each vehicle a amount of material from each wave that is proportional to the demand rates (and material requirements) of the sites that the vehicle visits. Let $J_k(t) = I(t)L_k / \sum_{i=1}^n L_i$. Then, let $J_v(t) = \sum_{k \in \sigma_v} J_k(t)$ be cumulative material available for vehicle v .

Every vehicle starts its first trip at $t = 0$. All of the material available for vehicle v (which equals $J_v(0)$) will be delivered on the first trip if the vehicle capacity is sufficient (that is, if $J_v(0) \leq C$). If not, then, when the vehicle returns to the depot, the vehicle will pick up the undelivered material, which equals $J_v(y_v) - C$, and start another trip. This will continue until no more material is available for that vehicle.

Whenever no material is available for a vehicle, it will wait at the depot until the next wave, when more material becomes available (and $J_v(t) > 0$), and will start a trip then.

3.3 Material Allocation

Given the routes for the vehicles and the delivery start times, it is possible to allocate material optimally by solving a linear program (LP). In this LP, the decision variables are a_{vj} , which is the amount of material delivered by vehicle v for trip j , and q_{vjk} , which is the amount of material delivered to site k by vehicle v on trip j . We know that vehicle v takes r_v trips.

The notation from the original formulation remains as follows: C is the capacity of a vehicle, t_{vj} is the time that vehicle v begins trip j , σ_v is the route that vehicle v follows, w_{vk} is the delay from the beginning of trip j until the delivery at site k by vehicle v , $I(t)$ is the cumulative amount of material that has arrived at the depot by time t , and Q_{vjk} is the cumulative amount of material delivered to site k by vehicle j before trip j .

In addition, let $n_v(t)$ be the number of trips taken by vehicle v up to, and including, time t . Let $T_v = \{t_{vj}, j = 1, \dots, r_v\}$ be the set of all trip start times for vehicle v . Let T be the union of all trip start times for the V vehicles: $T = T_1 \cup \dots \cup T_V$.

The LP can be described as follows:

$$\max s \quad (11)$$

$$\sum_{v=1}^V \sum_{j=1}^{n_v(t)} a_{vj} \leq I(t) \quad \forall t \in T \quad (12)$$

$$a_{vj} \leq C \quad v = 1, \dots, V; j = 1, \dots, r_v \quad (13)$$

$$\sum_{k \in \sigma_v} q_{vjk} = a_{vj} \quad v = 1, \dots, V; j = 1, \dots, r_v \quad (14)$$

$$\sum_{j=1}^{r_v} q_{vjk} = L_k(T_2 - T_1) \quad v = 1, \dots, V; k \in \sigma_v \quad (15)$$

$$Q_{v1k} = 0 \quad v = 1, \dots, V; k \in \sigma_v \quad (16)$$

$$Q_{vjk} = \sum_{m=1}^{j-1} q_{vmk} \quad v = 1, \dots, V; j = 2, \dots, r_v; k \in \sigma_v \quad (17)$$

$$s + t_{vj} + w_{vk} \leq T_1 + \frac{Q_{vjk}}{L_k} \quad v = 1, \dots, V; j = 1, \dots, r_v; k \in \sigma_v \quad (18)$$

$$a_{vj} \geq 0 \quad v = 1, \dots, V; j = 1, \dots, r_v \quad (19)$$

$$q_{vjk} \geq 0 \quad v = 1, \dots, V; j = 1, \dots, r_v; k \in \sigma_v \quad (20)$$

Equation 12 is the objective function, which seeks to maximize the minimum slack. Equation 13 limits the amount of material that the vehicles can deliver to the amount available at the depot. Equation 14 is the vehicle capacity constraint. Equation 15 ensures that a vehicle delivers all of the material that it takes from the depot every trip. Equation 16 ensures that every site receives the amount of material required there. Equations 16 and 17 define the cumulative quantity amounts. Equation 18 is the upper bound on the slack and ensures that the objective function value is the minimum slack. Equations 19 and 20 are the nonnegativity constraints for the decision variables.

3.4 Adaptive Large Neighborhood Search

Our search approach was a version of the Adaptive Large Neighborhood Search (ALNS) [16]. One advantage of this method is that each iteration of the search uses simple heuristics (and is thus quick), but the diversity of these heuristics reduces the likelihood of the search being trapped at a locally optimal solution. The search begins with an initial feasible solution created by the heuristic technique. Then each iteration of the search partially destroys and rebuilds the solution.

The general framework for this search can be summarized as follows:

1. Start with initial route and calculate objective value.
2. Choose removal heuristic randomly based on past performance.
3. Choose insertion heuristic randomly based on past performance.
4. Calculate objective function value and accept or reject based on simulated annealing framework.
5. Update heuristic performances and cooling parameters.
6. Return to Step 2 (unless the number of iterations has reached the desired limit).
7. Output best solution found.

Our ALNS used four removal heuristics that are appropriate for the ISRP and have the ability to diversify the search. These removal heuristics take a complete sequence of sites for each vehicle, relax a specified number of sites, and output a partial solution. The search also used four insertion heuristics, which take a partial solution and insert the removed sites to form a complete solution. When a partial solution is presented to an insertion heuristic, the first step is to ensure that each vehicle has at least one site to service. Montjoy and Herrmann [15] described these heuristics and their parameters in more detail.

Like the procedure of Pisinger and Ropke [16], our ALNS selects a removal heuristic and an insertion heuristic each iteration. A heuristics selection probability is proportional to its weight. Both selected heuristics are rewarded in three cases: (1) a new global best solution is found, (2) the new

solution is better than the previous one and has not been accepted before, and (3) the new solution is not better than the previous one and it has not been accepted before. If rewarded, the heuristics observed weights are increased by 5 (in case 1), 3 (in case 2), or 1 (in case 3). The search process is divided into segments of 50 iterations. At the beginning of a segment, every heuristic has an observed weight of zero. At the end of a segment, the ALNS calculates new weights based on the weights from the previous segment and the observed weights. Our ALNS used a simulated annealing procedure to determine if a new solution is accepted. The starting temperature of the simulated annealing procedure was selected by observing the objective value of an initial solution and choosing by a desired probability for a relatively lesser value. The cooling rate was then tuned to have reasonable acceptance probabilities towards the end of the search.

Step 4 in the search calculates the objective function value by using the deliver-when-possible algorithm (Section 3.2) to determine the start times and then using simple rules to find quickly a good material allocation (instead of using the LP described in Section 3.3). These rules first allocate material proportionally to the demand rates and then adjust the allocations to improve the minimum slack; these are described in detail in Herrmann et al. [10].

When the search is complete (after 1,500 iterations), we used the LP introduced in Subsection 3.3 to optimize the material allocation given the routes and delivery schedule in the best solution found by the search (this approach was first presented as the “LinProg” version of the ALNS in Yan et al. [19]).

3.5 Branch-and-Bound Approach

We developed a branch-and-bound algorithm to find optimal solutions where possible. The approach implicitly enumerates all feasible routes, uses the deliver-when-possible procedure to schedule deliveries (as described in Section 3.2, and uses the LP to find the optimal material allocations (as described in Section 3.3).

The algorithm proceeds by examining each vehicle in turn. The general scheme can be described as follows. At the initial node, there are no sites assigned to the first vehicle. The branches from this node correspond to placing different sites as the first site in the route of the first vehicle. At any other node, the branches correspond to adding remaining sites to the current vehicle’s route or terminating this route and starting the route for the next

vehicle. For example, the first branch from the initial node assigns site 1 as the first site in the route of the first vehicle. If there are three remaining sites (2, 3, and 4), then the branches from this node will be 1-2, 1-3, 1-4, and 1-T, where 1-T denotes that the route for the first vehicle is terminated (that is, the vehicle will return to the depot after visiting site 1) and the second vehicle is now the current vehicle. The strategy for searching this tree is to start from the “far left” and systematically look at all combinations while ignoring redundant or non-desirable solutions. This can be done by ensuring that the vehicles in a solution are ordered by the first site in their sequence and by using the following rules:

1. Consider only the first $n - V + 1$ sites to be the first site on the first vehicle;
2. If the remaining number of sites equals the remaining number of vehicles, evaluate this solution if the solution can still be increasingly ordered by the first site assigned to each vehicle;
3. Among a set of remaining sites to be the first in the route on a new vehicle when V_r vehicles have not yet been considered, the $V_r - 1$ last remaining sites can not start the next vehicle (for example, if there are three vehicles and 1-2-3-8-9-T is the first vehicle’s route, then there are $V_r = 2$ remaining vehicles, sites 4, 5, 6, and 7 are the remaining sites, and site 7 can not be the first site assigned to the second vehicle);
4. When adding sites to the route of a vehicle that begins with site a , at least V_r sites with indices that are greater than a must remain unassigned, so do not add to this route a site with an index that is greater than a if there are only V_r such sites remaining (for example, in a three-vehicle, nine-site problem, if the first vehicle’s route is currently 5-3-7-8-2, then neither 6 nor 9 can be added to this route).

Recall that the objective function is to maximize the minimum slack. An initial lower bound is found by applying the routing-and-scheduling approach. At each node explored in the branch-and-bound approach, an upper bound can be found by assuming each and every site not yet assigned to any route is assigned to an extra vehicle that has infinite capacity, visits only that site, and starts each delivery when each new wave arrives. After combining these hypothetical vehicles and routes with the actual routes being considered, the

start times were determined using the deliver-when-possible procedure, and the optimal material allocations were determined using the LP.

At any node, if the upper bound is below the current lower bound, the node is not explored.

4 Computational Testing

To compare the quality of the solutions generated and the computational effort of each solution technique, instances based on real-world settings and traditional VRP test instances were generated. For datasets that had street addresses, Toursolver and Google Maps were used to calculate travel times between the sites. We invented demand and wave delivery information to be similar to real world mass dispensing plans from Maryland. All of the instances had loading times of 15 minutes. Montjoy et al. [14] described the process of generating the instances. (These instances are available upon request.)

A subset of 21 instances were used for this work. There were seven baseline instances (which specify the site locations and other details), but the number of vehicles was varied to create three instances from each baseline instance.

On each instance, we ran the routing-and-scheduling (R&S) approach four times (once with each version of the route-first-cluster-second heuristic), ran the ALNS search five times (with 1,500 iterations per run), and conducted a branch-and-bound search. The branch-and-bound search was terminated if it did not complete within 24 hours.

4.1 Solution Quality

The quality of each solution is its minimum slack. We kept the best solution of those generated from the four versions of the routing-and-scheduling (R&S) heuristic and averaged the minimum slack of the solutions generated by the five runs of the ALNS.

Table 5 lists the minimum slack of these solutions and the upper bound, which presumes that every site has its own vehicle (cf. Section 3.5) and does not depend upon the number of vehicles in the instance. For those instances on which the B&B approach terminated without finding an optimal solution, no value is listed for the B&B. The instances with an asterisk are those

in which the solution’s minimum slack occurs on the first wave. In these instances, the minimum slack is determined by the first delivery to a site. The later deliveries have more slack.

The vehicle capacity constraint was not tight on any deliveries, and none of the instances had small waves that arrived soon before larger ones. Thus, no solution could be improved by combining any two deliveries (as discussed in Section 2.3). Thus, if the branch-and-bound approach generated a solution, it is an optimal solution.

The three solution techniques generate solutions that have similar quality. The most notable exceptions are the 20-site instances, on which the ALNS found solutions that were approximately 20 minutes better than the solutions generated by the R&S heuristics. A representation of the routes from the best heuristic solution and ALNS can be seen in Figure 4.1 for the 6 vehicle variant. (Note that the instance comes from a real-world road network and this is an estimation of the points in a Euclidean space.) The slack occurs on the first delivery, so the route that minimizes the latest delivery time will yield the best solution. The ALNS is able to find a better solution because the heuristic focuses on servicing sites that are closer together by the same vehicle.

We also noted that adding more vehicles increases the minimum slack, which is expected because deliveries can be completed sooner. The size of the increase is limited, however, because the minimum slack approaches the instance upper bound.

4.2 Computational Effort

The computational effort for the ALNS and the B&B grows as the number of sites increases. Table 6 presents the runtime in seconds for the ALNS and the B&B. The run time for the ALNS is the average for the 5 runs. The run times for the R&S heuristic were less than a few seconds.

For the smaller instances, the B&B requires less time than the ALNS because the ALNS runs the same number of iterations for each instance. As the number of sites increases, the run time for the B&B grows quickly.

The run time for the ALNS increases more slowly. The ALNS is able to complete 1,500 iterations in less than 7 minutes for the largest instances.

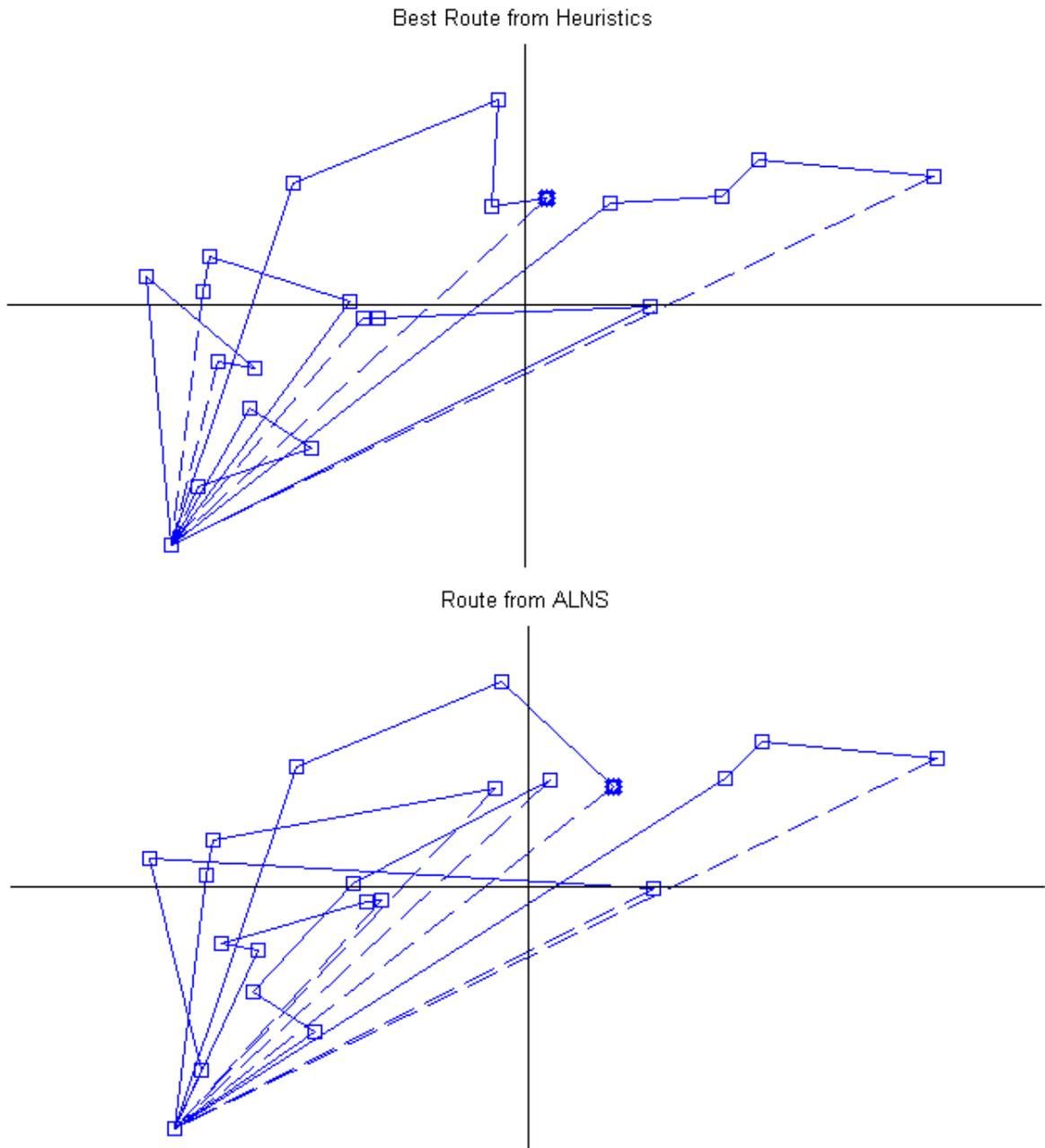


Figure 1: Representation of routes from the heuristic and ALNS for the 20 site, 6 vehicle instance. The dotted line represents the return from the last site to the depot. The site with the minimum slack is denoted in bold.

Table 5: Objective function values (minutes) for the solution approaches. R&S = routing-and-scheduling, ALNS = Adaptive Large Neighborhood Search, B&B = branch-and-bound. The symbol * denotes solutions in which the slack occurs on the first wave for all techniques.

Sites	Vehicles	R&S	ALNS	B&B	Upper Bound
5	2	481.5	481.5	483.7	501.8
	3	493.8	493.8	493.8	
	4	498.1	498.1	498.1	
9	3*	1273.0	1281.0	1281.0	1329.0
	5*	1300.0	1316.0	1316.0	
	7*	1325.0	1328.0	1328.0	
9	3*	458.0	458.0	458.0	496.0
	5*	471.0	482.6	483.0	
	7*	496.0	496.0	496.0	
10	3	1080.8	1083.2	1084.7	1105.5
	5	1095.3	1095.3	1096.8	
	7	1100.0	1100.0	1101.1	
20	6*	124.0	148.8	-	188.0
	10*	167.0	180.4	-	
	14*	167.0	188.0	188.0	
50	15	1273.0	1273.0	-	1302.0
	25	1291.8	1291.8	-	
	35	1297.5	1297.5	-	
189	30*	1244.0	1245.2	-	1351.0
	71*	1318.0	1320.4	-	
	100*	1333.0	1333.0	-	

Table 6: Computational time in seconds for the ALNS and B&B techniques.

Sites	Vehicles	ALNS	B&B
5	2	12.6	3.3
	3	9.6	1.3
	4	8.2	0.3
9	3	26.1	1762.0
	5	18.2	177.0
	7	15.6	9.7
9	3	39.5	1804.6
	5	26.5	238.1
	7	20.2	1.8
10	3	40.1	46615.4
	5	25.8	7100.0
	7	20.0	330.9
20	6	31.1	-
	10	21.7	-
	14	20.4	68.2
50	15	101.1	-
	25	81.2	-
	35	80.0	-
189	30	380.3	-
	71	221.6	-
	100	216.1	-

For an instance, the run time decreases as the number of vehicles increases. When inserting a site into an existing route, the ALNS tries every location in a vehicle’s route; in an instance with more vehicles, the vehicle routes are shorter and have fewer potential locations that must be considered. Likewise, the B&B has fewer branches to investigate at each point.

5 Conclusions

The inventory slack routing problem is a unique vehicle routing problem that focuses on delivering material as early as possible and emphasizes robustness and equity. This work was motivated by the problem of delivering antibiotics

during the response to an anthrax attack, but the problem can occur in other emergencies or situations in which urgently needed material must be delivered over a short time horizon.

Due to its complexity, we decided to separate the ISRP into three sub-problems: routing, scheduling, and material allocation. For the routing problem, we developed a route-first, cluster-second heuristic; for the scheduling problem, we developed a deliver-when-possible algorithm; and, for the allocation problem, we developed a linear program. From these components we developed three solution approaches for the ISRP: (1) the Routing-and-Scheduling (R&S) approach, (2) an adaptive large neighborhood search (ALNS), and (3) a branch-and-bound (B&B) approach.

A set of instances was used to test the various techniques. The R&S approach performed well in smaller instances but performed much worse than the ALNS and B&B in larger instances. The B&B was unable to find complete solutions within a reasonable time for most instances with more than 10 sites. Thus, the ALNS appears to be a useful technique for solving this problem.

Future work is needed to consider problems in which the demand at each site is uncertain and to develop approaches for updating solutions to the ISRP problem as real-time information about actual demand at the sites becomes available so that the right amount of material can be delivered to the right place.

6 Acknowledgements

The authors were supported by award number 5H75TP000309-02 from the Centers for Disease Control and Prevention (CDC) to National Association of County and City Health Officials (NACCHO) and the Montgomery County, Maryland, Advanced Practice Center for Public Health Emergency Preparedness and Response. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the CDC or NACCHO.

References

- [1] K Aaby, RL Abbey, JW Herrmann, M Treadwell, CS Jordan, and K Wood. Embracing computer modeling to address pandemic influenza in the 21st century. *Journal of Public Health Management and Practice*, pages 365–375, 2006.
- [2] K Aaby, JW Herrmann, CS Jordan, M Treadwell, and K Wood. Montgomery county’s public health service uses operations research to plan emergency mass dispensing and vaccination clinics. *Interfaces*, 36(6):569–579, 2006.
- [3] MO Ball and FL Lin. A reliability model applied to emergency service vehicle location. *Operations Research*, 41:18–36, 1993.
- [4] ML Brandeau, JH McCoy, N Hupert, JE Holty, and DM Bravata. Recommendations for modeling disaster responses in public health and medicine: A position paper of the society for medical decision making. *Medical Decision Making*, pages 1–23, 2009.
- [5] A Campbell, L Clarke, AJ Kleywegt, and MWP Savelsbergh. The inventory routing problem. In TG Crainic and G Laporte, editors, *Fleet Management and Logistics*. Kluwer Academic Publishers, 1998.
- [6] AM Campbell, LW Clarke, and MWP Savelsbergh. Inventory routing in practice. In *The Vehicle Routing Problem*. Society for Industrial and Applied Mathematics, 2001.
- [7] AM Campbell, D Vandenbussche, and W Hermann. Routing for relief efforts. *Transportation Science*, 42(2):127–145, 2008.
- [8] A Ceyhun, H Selim, and I Ozkarahan. A fuzzy multi-objective covering-based vehicle location model for emergency services. *Computers and Operations Research*, 34:705–726, 2007.
- [9] MS Daskin and EH Stern. Hierarchical objective set covering model for emergency medical service vehicle deployment. *Transportation Science*, 15:137–152, 1981.
- [10] Jeffrey W. Herrmann, Sara Lu, and Kristen Schalliol. Delivery volume improvement for planning medication distribution. *Proceedings of the*

2009 IEEE International Conference on Systems, Man, and Cybernetics, San Antonio, TX, October, 2009.

- [11] N Hupert, D Wattson, J Cuomo, E Hollingsworth, and K Neukermans. Predicting hospital surge after a large-scale anthrax attack: A model-based analysis of cdc’s cities readiness initiative prophylaxis recommendations. *Medical Decision Making*, 29:424–437, 2009.
- [12] EK Lee, CH Chen, F Pietza, and B Benecke. Modeling and optimizing the public health infrastructure for emergency response. *Interfaces*, 39(5):476–490, 2009.
- [13] NH Moin and S Salhi. Inventory routing problems: A logistical overview. *Journal of the Operational Research Society*, 58:1185–1194, 2007.
- [14] Adam Montjoy, Stephanie Brown, and Jeffrey W. Herrmann. Solving the inventory slack routing problem for medication distribution planning. *2009-13 ISR Technical Report*, Sept. 2009.
- [15] Adam Montjoy and Jeffrey W. Herrmann. Adaptive large neighborhood search for the inventory slack routing problem. *Proceedings of the 2010 Industrial Engineering Research Conference*, Cancun, MX, June, 2010.
- [16] P Ropke and D Pisinger. An adaptive large neighborhood search heuristic for the pickup and delivery problem with time windows. *Transportation Science*, 40(4):455–472, 2005.
- [17] Paolo Toth and Daniel Vigo. *The Vehicle Routing Problem*. Society for Industrial and Applied Mathematics, New York, 2001.
- [18] A Weintraub, J Aboud, C Fernandez, G Laporte, and E Ramirez. An emergency vehicle dispatching system for an electric utility in chile. *Journal of the Operational Research Society*, 50:690–696, 1999.
- [19] Zijie Yan, Adam Montjoy, and Jeffrey W. Herrmann. Variants of the adaptive large neighborhood search for the inventory slack routing problem. *2011-10 ISR Technical Report*, Sept. 2011.