IIE Transactions

Allocation and reallocation of ambulances to casualty clusters in a disaster relief operation

Qiang Gong [a] & Rajan Batta [b]

[a] United Airlines, 1200 East Algonquin, Elk Grove Township, IL, 60007, USA

[b] 438 Bell Hall, Department of Industrial and Systems Engineering and Center for Multisource Information Fusion, University at Buffalo (SUNY), Buffalo, NY, 14260, USA
Published online: 07 Apr 2011.

PLEASE SCROLL DOWN FOR ARTICLE

Allocation and reallocation of ambulances to casualty clusters in a disaster relief operation

QIANG GONG[1] and RAJAN BATTA[2,*]

[1] *United Airlines, 1200 East Algonquin, Elk Grove Township, IL 60007, USA*
[2] *438 Bell Hall, Department of Industrial and Systems Engineering and Center for Multisource Information Fusion,*
University at Buffalo (SUNY), Buffalo, NY 14260, USA
E-mail: batta@acsu.buffalo.edu

In this paper, we consider ambulance allocation and reallocation models for a post-disaster relief operation. The initial focus is on allocating the correct number of ambulances to each cluster at the beginning of the rescue process. We formulate a deterministic model which depicts how a cluster grows after a disaster strikes. Based on the model and given a number of ambulances, we develop methods to calculate critical time measures, e.g., the completion time for each cluster. Then we present two iterative procedures to optimize the makespan and the weighted total flow time, respectively. The second problem analyzes the ambulance reallocation problem on the basis of a discrete time policy. The benefits of redistribution include providing service to new clusters and fully utilizing ambulances. We consider the objective of minimizing the makespan. A complication is that the distance between clusters needs to be factored in when making an ambulance reallocation decision. Our model permits consideration of the travel distance between clusters. Results of our allocation method are illustrated via a case study, which is based on an earthquake scenario in Northridge, CA.

Keywords: Ambulance allocation, ambulance reallocation, cluster, makespan, weighted total flow time

1. Introduction

Our research is motivated by a disaster relief situation which deals with rescuing casualties after the occurrence of a natural or man-made disaster. Examples of such events include land earthquakes and tsunamis (see (Anon, 2006) for a detailed discussion on tsunamis).

Information on casualties can be obtained from satellite images, sensor systems embedded in the infrastructure (e.g., cameras), police reports, property owners, civilians and other individuals. A large number of casualties can easily overwhelm the ambulance system because the number of ambulances is usually determined by reference to a major but not disasterous event. Management of resources in such an environment becomes critical. In this paper we provide efficient methods to improve the performance of rescue work, specifically the allocation of ambulances in an appropriate manner to the casualty clusters and then to reallocate ambulances between clusters as the disaster evolves.

The rest of this paper is organized as follows. Section 2 contains a literature review of related modeling areas from which we have drawn ideas for our model development/analysis. Section 3 provides a brief description of the general setting of our problem. Section 4 focuses on initial ambulance allocation. We first calculate time measures associated with a casualty cluster once a given number of ambulances have been assigned to it at the initial time point. Then these time measures are used to develop algorithms for the minimization of makespan and the minimization of weighted total flow time. Section 5 addresses the ambulance reallocation problem. We consider a discrete time policy which allows redistribution to occur at predetermined instants. Section 6 applies the algorithms of Section 4 to a case study based on an earthquake simulation in the Northridge, CA area. Finally, Section 7 contains a discussion along with future work directions.

2. Literature review

There are several modeling areas that relate to resource allocation in a disaster relief setting. One such area is forest-fire management. Another is the management of enforcement efforts in illicit drug markets. A third is the interplay between data fusion and dispatch/routing of ambulances. We briefly review relevant papers in each of these areas. We also point out which modeling aspects we have utilized in the development of our ambulance allocation model.

*Corresponding author

Our study of relevant research begins with an interesting problem which relates to forest-fire management. Parks (1964) presented a deterministic model to study the initial attack on wildland fires. His model is also focused on the economic aspect of the problem and the objective is to determine manpower requirements such that the total cost is minimized. The optimal resource allocation balances all the costs including the operating cost of the organization, transportation and logistic cost, emergency cost, damage cost, and cost associated with the length of time for suppression and the size of suppressing force. Islam (1998) considered a daily airtanker deployment problem for forest-fire management, which describes how many airtankers are required per day and where they should be deployed dynamically throughout the day. The growth pattern of a forest fire is used in our problem to describe how a cluster of casualties grows. A cluster is composed of a sufficient number of casualties whose locations are close to one another.

Another related problem can be found in the management of illicit drug markets. To aid the understanding of how drug market management is closely related to our problem, we first identify the similarity between these two problems. The drug dealers and the police enforcement resources can be regarded as casualties and relief resources, respectively. The operation of cracking down on the drug dealers may be thought of as the operation of rescuing casualties. Becker (1976) was one the pioneers who pointed out that utility maximization models are useful in the context of drug dealing. Caulkins (1990) presented an economic model to quantify the rate of growth or decay of a drug market under crackdown enforcement. In his model, the rate of change in the number of dealers is proportional to the difference between the profit available to dealers active in the market and the discouraging utility (including the risk from crackdown enforcement and the reservation wage). Based on the framework of Caulkins' model, Baveja (1993) studied the problem of finding the best rate to crack down on a drug market. He considered both discrete and continuous-time policies. The solutions to both cases gave the same conclusion that the best way is to allocate the maximum possible effort from the earliest possible time. Naik *et al.* (1996) provided an analytical approach to schedule crackdown enforcement for a series of drug markets. They also extended the problem to incorporate the dealer displacement effect (dealers might respond to a crackdown by relocating from one market to another). Behrens *et al.* (2000) studied the drug treatment and prevention problem in the framework of dynamic optimal control under different assumptions. Tragler *et al.* (2001) considered whether to allocate resources to treatment rather than enforcement. They formulated this as an optimal control problem and provided recommendations for a variety of market situations.

Data fusion first appeared in the scientific literature in the late 1960s and found use in multiple disciplines in the 1970s and 1980s (Gros, 1997). Scott and Rogova (2004) gave a general introduction of how to conduct disaster relief management in a data fusion synthetic task environment. Gong *et al.* (2004) studied the casualty pickup problem and casualty delivery problem based on data fusion methods. Gong and Batta (2004) and Gong (2005) also studied the problem of managing casualties with different priorities. Their work requires the consideration of a methodology to address the problem of identification of casualty clusters. A dynamic method is presented under the assumption that every ambulance can be dispatched to every cluster. In this paper, we study the problem from another aspect: namely the case in which a set of ambulances serve only one cluster until the cluster no longer exists. This requires an efficient method to allocate the available resources to clusters. We later study the problem of reallocating ambulances between clusters as the disaster evolves.

3. Problem description

After a disaster occurs, hundreds or thousands of spatially distributed casualties need to be treated. As we mention in Section 1, information on casualties is assumed to be reported by sensors, which could include calls from injured people, passers by, law enforcement officers, ambulance drivers, etc. The large number and different types of these sources leads to imprecise data, making it difficult to determine the precise situation, i.e., how many distinct casualties there are and where these casualties are located. Instead of exact spatial coordinates for each casualty, we have an estimate of a confidence region, typically using data fusion concepts. Data fusion can be defined as the synergistic use of information from multiple sources in order to assist in the overall understanding of a phenomenon. In our case information flowing from multiple sources has a highly variable character (e.g., human intelligence, signal intelligence, etc.). It is necessary to align the data and develop a comprehensive picture rapidly and accurately in order to take full advantage of it in our emergency relief services. The typical output of a data fusion algorithm is that a casualty is located in a specific region with a probability no smaller than a given value p $(0 < p \leq 1)$. If the number of casualties that are likely to be in a small area exceeds a threshold, say N, then these casualties are regarded as a cluster; otherwise, these casualties are treated as individuals. The choice of N should be made so that the total number of clusters does not dilute the available resources for assignment to clusters to an extent that makes the relief operation ineffective.

We assume that the emergency services only respond to casualties in a cluster. An ambulance dispatched to an isolated casualty may take a considerable time to find the patient. Also, in a disaster situation it is more likely that three to four casualties are loaded on to the ambulance prior to a trip back to the hospital. To minimize casualty search time and to simultaneously maximize ambulance utilization, it is more efficient, in general, only to respond to casualties in a cluster.

For the purpose of our mathematical model, we assume that there are m clusters that need to be responded to and n ambulances which are ready to be distributed to the clusters (initial information on the clusters is received through sensors prior to the allocation time point). We further assume that all the ambulances are identical in service rate and capacity, and are initially located at hospitals. After receiving a dispatching order, an ambulance is sent to the pickup location and then returns to the hospital. The service time is defined as the period between the time at which the ambulance is dispatched and the time at which it returns to the hospital.

Several strategic problems arise, including those of deciding the sequence in which clusters are to be attacked, determining the optimal time to quit a cluster, and the coordination of casualty deliveries with available hospital capacity. Our objective in this paper is to address the problems of initial allocation and subsequent reallocation of ambulances among casualty clusters.

4. Initial ambulance allocation

4.1. *Time measures of a cluster*

We first calculate the finish time of a cluster. The finish time, denoted as T, of a cluster is defined as the time at which the number of casualties in that cluster is reduced to N. When the number of casualties in a cluster is less than N the remaining casualties can no longer be thought of as a cluster. At this point the assigned ambulances quit the cluster and switch to serve one of the other clusters.

We use a continuous function to represent the number of casualties in a cluster at any time t. Figure 1 displays how casualties are likely to be discovered, as a function of time, in a cluster. At time 0, there are N_0 ($N_0 \geq N$) casualties in the cluster. If no action is taken (i.e., no ambulances are assigned to pick up casualties) the number of casualties accumulate continuously. From time 0, the arrival rate escalates until it reaches its peak at time t_m. We call this time period phase I. After that, the arrival rate diminishes gradually up to time t_f when it becomes zero. We call this

Fig. 1. Model for the growth of a cluster.

period phase II. The cluster stops growing when the total number of casualties is equal to N_{t_f}. We assume that both t_m and t_f are known, and write:

$$N(t) = \begin{cases} N_1(t) & \text{if } 0 \leq t \leq t_m, \\ N_2(t) & \text{if } t_m \leq t \leq t_f, \end{cases} \quad (1)$$

where $N_1(t)$ corresponds to phase I in which the arrival rate keeps increasing, and $N_2(t)$ corresponds to phase II in which the arrival rate keeps decreasing.

Since the arrival rate is increasing during the time interval $[0, t_m]$, we can define it as follows:

$$\lambda_1(t) = kt + \lambda_0, \quad 0 \leq t \leq t_m, \quad (2)$$

where k is the growth acceleration parameter, $k > 0$, and λ_0 is the observed arrival rate at time 0.

The arrival rate peaks at time t_m and starts to decrease with a deceleration parameter k', which yields:

$$\lambda_2(t) = k'(t - t_m) + kt_m + \lambda_0. \quad (3)$$

Applying the condition that $\lambda_2(t_f) = 0$, we have:

$$\lambda_2(t) = -\left(\frac{kt_m + \lambda_0}{t_f - t_m}\right)(t - t_m) + kt_m + \lambda_0. \quad (4)$$

Since $\lambda_1(t) = dN_1(t)/dt$, we integrate both sides and use the initial condition $N_1(0) = N_0$, which yields:

$$N_1(t) = \frac{1}{2}kt^2 + \lambda_0 t + N_0. \quad (5)$$

Similarly, we can obtain an expression for $N_2(t)$ as follows:

$$N_2(t) = -\frac{1}{2}\left(\frac{kt_m + \lambda_0}{t_f - t_m}\right)t^2 + \left[\left(\frac{kt_f + \lambda_0}{t_f - t_m}\right)t_m + \lambda_0\right]t - \frac{1}{2}\left(\frac{kt_f + \lambda_0}{t_f - t_m}\right)t_m^2 + N_0. \quad (6)$$

By setting $t = t_f$ in Equation (6), the total number of casualties in a cluster is given by:

$$N_{t_f} = \frac{1}{2}kt_m t_f + \frac{1}{2}\lambda_0(t_m + t_f) + N_0. \quad (7)$$

We assume that a cluster is served at time 0 with an initial service rate μ. This rate is directly related to the number of ambulances allocated to the cluster. If the service rate of each ambulance is $\tilde{\mu}$ and a total of a ambulances serve this cluster, the service rate for this cluster can be calculated as:

$$\mu = a\tilde{\mu}. \quad (8)$$

Since the finish time T is defined as the time at which the number of casualties in a cluster is reduced to N, we should consider the time, denoted as $\tilde{T}$, at which the number of casualties in a cluster reaches $N_{t_f} - N$. Clearly, the finish time T cannot be earlier than the time $\tilde{T}$. We need to consider the following three cases to determine the finish time.
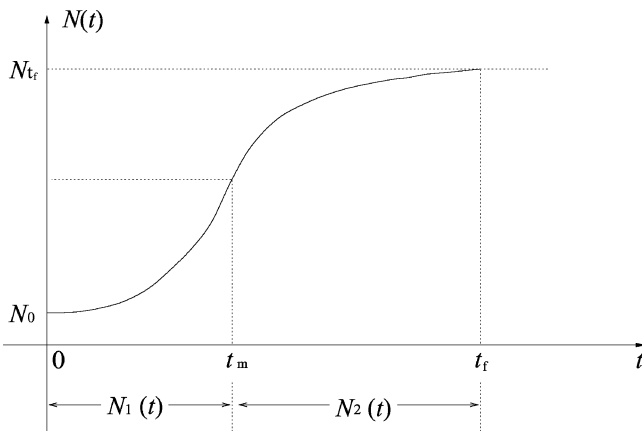
Case 1: $N_{t_f} - N < N_0$.

In this case, the number of casualties reaches the threshold $N_{t_f} - N$ before the ambulances start to serve the cluster. The finish time is therefore:

$$T = \frac{N_{t_f} - N}{\mu}. \tag{9}$$

Case 2: $N_0 \leq N_{t_f} - N < (1/2) k t_m^2 + \lambda_0 t_m + N_0.$

In this case, the time at which the number of casualties reaches $N_{t_f} - N$ occurs during phase I and it is given by the solution of the following equation:

$$N_1(t) = N_{t_f} - N. \tag{10}$$

The left-hand side of Equation (10) represents the total number of casualties discovered by time t. Substituting $N_1(t)$ as given by Equation (5) into Equation (10), we obtain:

$$\frac{1}{2} k t^2 + \lambda_0 t + N_0 - N_{t_f} + N = 0. \tag{11}$$

$\tilde{T}$ is given by the positive root as follows:

$$\tilde{T} = \frac{-\lambda_0 + \sqrt{\lambda_0^2 + 2k(N_{t_f} - N - N_0)}}{k}. \tag{12}$$

After the ambulances begin serving the cluster, casualties will flow out of the cluster. However, at the same time, new casualties will flow into the cluster. Here, we need to consider two subcases.

Case 2.1: $\mu \tilde{T} \leq N_{t_f} - N.$

Even though all the ambulances are busy from time 0 to time $\tilde{T}$, the number of casualties sent to hospital is no more than $N_{t_f} - N$. Therefore, the finish time can be calculated as:

$$T = \frac{N_{t_f} - N}{\mu}. \tag{13}$$

As more ambulances are allocated to this cluster, the finish time keeps reducing. Then we may have the following subcase:

Case 2.2: $\mu \tilde{T} > N_{t_f} - N.$

The first time, denoted as $\hat{T}$, at which the number of casualties is reduced to zero occurs before time $\tilde{T}$. The time $\hat{T}_i$ is given by the solution of the following equation:

$$N_1(t) = \mu t. \tag{14}$$

Substituting $N_1(t)$ as given by Equation (5) into Equation (14) yields:

$$\frac{1}{2} k t^2 + (\lambda_0 - \mu) t + N_0 = 0. \tag{15}$$

The roots of Equation (15) are:

$$t_1 = \frac{\mu - \lambda_0 - \sqrt{(\mu - \lambda_0)^2 - 2k N_0}}{k}, \tag{16}$$

and

$$t_2 = \frac{\mu - \lambda_0 + \sqrt{(\mu - \lambda_0)^2 - 2k N_0}}{k}. \tag{17}$$

It can be verified that $(\mu - \lambda_0)^2 - 2k N_0 > 0$ (see Gong (2005) for details). Since both roots are positive, we just take the one with the smallest value and get:

$$\hat{T} = \frac{\mu - \lambda_0 - \sqrt{(\mu - \lambda_0)^2 - 2k N_0}}{k}. \tag{18}$$

After the number of casualties is reduced to zero at time $\hat{T}$, the ambulances no longer have full loads. This means that the actual service rate during the time interval $[\hat{T}, \tilde{T}]$ is less than μ. In other words, the arrival rate during this period is less than μ. As the number of ambulances increases, the service rate μ also goes up. Once the service rate exceeds the threshold $(N_{t_f} - N)/\tilde{T}$, the finish time remains constant with a value equal to $\tilde{T}$. However, the time $\hat{T}$ still decreases steadily as μ rises. Based on the above analysis, it is worth pointing out that a policy of allocating an excessive number of ambulances has its own strengths and weaknesses. The disadvantage is that the efficiency of the ambulances decreases after time $\hat{T}$. But the benefit is that the average waiting time experienced by a casualty can be reduced.

Figure 2 depicts how the time T and the time $\hat{T}$ change as the service rate μ varies. In this example, we consider four different service rates with $\mu_1 < \mu_2 < (N_{t_f} - N)/\tilde{T} < \mu_3 < \mu_4$. Line i $(i = 1, 2, 3, 4)$ corresponds to rate μ_i, respectively. Both the finish time T_1 related to μ_1 and the finish time T_2 related to μ_2 are longer than $\tilde{T}$. More specifically, we have $T_1 > T_2$ due to $\mu_1 < \mu_2$. As we can see, the finish time moves to the left as the service rate increases. In this period, we also have $\hat{T} = T$. As μ exceeds $(N_{t_f} - N)/\tilde{T}$ and increases further, the finish time T stops at the value of $\tilde{T}$ but $\hat{T}$ keeps decreasing (see $\hat{T}_3$ and $\hat{T}_4$).

Case 3: $(1/2) k t_m^2 + \lambda_0 t_m + N_0 \leq N_{t_f} - N.$

In this case, the time at which the number of casualties reaches $N_{t_f} - N$ occurs during phase II and it is given by the solution of the following equation:
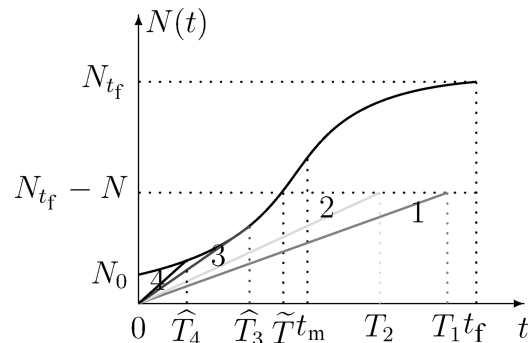
$$N_2(t) = N_{t_f} - N. \tag{19}$$

Fig. 2. Finish time of a cluster under different service rates (case 2).

$$t_{1,2} = \frac{[(kt_{\mathrm{m}} + \lambda_0)t_{\mathrm{f}} - \mu(t_{\mathrm{f}} - t_{\mathrm{m}})] \pm \sqrt{[(kt_{\mathrm{m}} + \lambda_0)t_{\mathrm{f}} - \mu(t_{\mathrm{f}} - t_{\mathrm{m}})]^2 - (kt_{\mathrm{m}} + \lambda_0)\left[(kt_{\mathrm{f}} + \lambda_0)t_{\mathrm{m}}^2 - 2N_0(t_{\mathrm{f}} - t_{\mathrm{m}})\right]}}{kt_{\mathrm{m}} + \lambda_0}. \quad (26)$$

Substituting $N_2(t)$ as given by Equation (6) into Equation (19), we obtain:

$$\frac{1}{2}\left(\frac{kt_{\mathrm{m}} + \lambda_0}{t_{\mathrm{f}} - t_{\mathrm{m}}}\right)t^2 - \left[\left(\frac{kt_{\mathrm{f}} + \lambda_0}{t_{\mathrm{f}} - t_{\mathrm{m}}}\right)t_{\mathrm{m}} + \lambda_0\right]t$$
$$+ \frac{1}{2}\left(\frac{kt_{\mathrm{f}} + \lambda_0}{t_{\mathrm{f}} - t_{\mathrm{m}}}\right)t_{\mathrm{m}}^2 + N_{t_{\mathrm{f}}} - N_0 - N = 0. \quad (20)$$

The roots of Equation (20) are:

$$t = \frac{(kt_{\mathrm{m}} + \lambda_0)t_{\mathrm{f}} \pm \sqrt{(kt_{\mathrm{m}} + \lambda_0)^2 t_{\mathrm{f}}^2 - (kt_{\mathrm{m}} + \lambda_0)\left[(kt_{\mathrm{f}} + \lambda_0)t_{\mathrm{m}}^2 + 2(N_{t_{\mathrm{f}}} - N_0 - N)(t_{\mathrm{f}} - t_{\mathrm{m}})\right]}}{kt_{\mathrm{m}} + \lambda_0}. \quad (21)$$

Substituting $N_{t_{\mathrm{f}}}$ as given by Equation (7) into Equation (21), we have:

$$t_1 = t_{\mathrm{f}} - \sqrt{\frac{2N(t_{\mathrm{f}} - t_{\mathrm{m}})}{kt_{\mathrm{m}} + \lambda_0}} \quad \text{and} \quad t_2 = t_{\mathrm{f}} + \sqrt{\frac{2N(t_{\mathrm{f}} - t_{\mathrm{m}})}{kt_{\mathrm{m}} + \lambda_0}}. \quad (22)$$

Since we know that $t_{\mathrm{m}} \le \tilde{T} < t_{\mathrm{f}}$, it follows that:

$$\tilde{T} = t_{\mathrm{f}} - \sqrt{\frac{2N(t_{\mathrm{f}} - t_{\mathrm{m}})}{kt_{\mathrm{m}} + \lambda_0}}. \quad (23)$$

Again, we need to consider two subcases.

Case 3.1: $\mu\tilde{T} \le N_{t_{\mathrm{f}}} - N$.

In terms of the same analysis of case 2.1, we know that the finish time is given by Equation (13).

Case 3.2: $\mu\tilde{T} > N_{t_{\mathrm{f}}} - N$.

From case 2.2, we know that only the time $\hat{T}$ decreases under this condition. In order to calculate $\hat{T}$, we need further consider two cases.

Case 3.2.1: $\mu t_{\mathrm{m}} \le N_1(t_{\mathrm{m}})$.

In this case, $\hat{T}$ takes place during phase II and it is given by the solution of the following equation:

$$N_2(t) = \mu t. \quad (24)$$

Substituting $N_2(t)$ as given by Equation (6) into Equation (24), we obtain:

$$-\frac{1}{2}\left(\frac{kt_{\mathrm{m}} + \lambda_0}{t_{\mathrm{f}} - t_{\mathrm{m}}}\right)t^2 + \left[\left(\frac{kt_{\mathrm{f}} + \lambda_0}{t_{\mathrm{f}} - t_{\mathrm{m}}}\right)t_{\mathrm{m}} + \lambda_0 - \mu\right]t$$
$$-\frac{1}{2}\left(\frac{kt_{\mathrm{f}} + \lambda_0}{t_{\mathrm{f}} - t_{\mathrm{m}}}\right)t_{\mathrm{m}}^2 + N_0 = 0. \quad (25)$$

The roots of Equation (25) are:

It is not easy to identify which root gives the answer to $\hat{T}$, so we just calculate it by:

$$\hat{T} = \min\{t_i : t_i > 0, i = 1, 2\}. \quad (27)$$

The time $\hat{T}$ keeps shortening as the service rate goes up, so another case may happen.

Case 3.2.2: $\mu t_{\mathrm{m}} > N_1(t_{\mathrm{m}})$.

In this case, $\hat{T}$ occurs in phase I. According to case 2.2, it is given by Equation (18).

We use Fig. 3 to illustrate how T and $\hat{T}$ behave as the service rate changes. In this example, we consider three different rates with $\mu_1 < (N_{t_{\mathrm{f}}} - N)/\tilde{T} < \mu_2 < N_1(t_{\mathrm{m}})/t_{\mathrm{m}} < \mu_3$. The explanation of Fig. 3 is similar to that of Fig. 2.

4.2. *System performance measures*

4.2.1. *Minimization of makespan*

The makespan problem considers minimizing the maximal finish time of the clusters. Generally speaking, this problem requires a good method that does not simply concentrate on a portion of the clusters in order to avoid the case in which the finish time of some clusters is short but the finish time of other clusters is quite long. With this in mind we start by allocating ambulances as evenly as possible to the
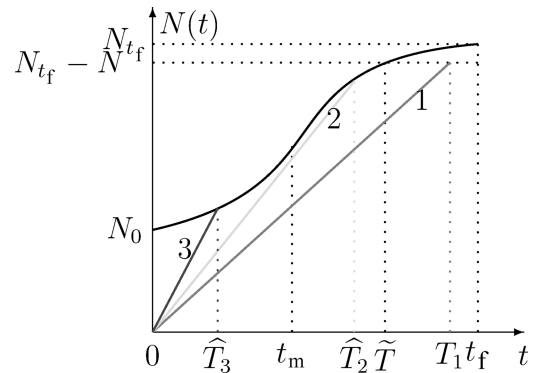

Fig. 3. Finish time of a cluster under different service rates (case 3).

clusters. The initial finish time of all clusters are calculated and maintained in a list.

We use an iterative method. In each iteration, we take one ambulance from the cluster with the shortest finish time (e.g., cluster i) and give it to the cluster with the longest finish time (e.g., cluster j). Since all other clusters are untouched, we just recalculate the finish time for these two clusters. If the new finish time of cluster j is equal to the old one, we know that the finish time of cluster j cannot be shortened further. We therefore stop at the current solution.

However, if the longest finish time is reduced by adding one ambulance to it, we have to study how the shortest time changes under this situation. One possible case is that the new finish time of cluster i is longer than the old finish time of cluster j. The result is given by the following property.

Property 1. *In any iteration, if the new makespan is greater than the old one, this iteration is invalid and the cluster with the shortest finish time is deleted from the list.*

Proof. We need to prove this property from two aspects: (i) this cluster cannot help the others; and (ii) the other clusters are also unable to help this cluster.

Assume that we take one ambulance from cluster i and assign it to cluster j. As we know, the new makespan is equal to the finish time of cluster i and is even longer than the previous one. In the next iteration, cluster i has the longest finish time. An ambulance from the cluster (e.g., cluster k) with the shortest finish time is allocated to it. If that cluster happens to be cluster j, then this procedure falls in a dead loop. If not, we can see that the process of allocation is $k \rightarrow i \rightarrow j$ which is equivalent to $k \rightarrow j$. The process $i \rightarrow j$ can be skipped. Therefore, we know that cluster i does not help reduce the makespan.

Next, we are going to show that cluster i will not receive any ambulances from the other clusters. It is important to point out that in each iteration the shortest finish time is the lower bound for the makespan. So there is no benefit to allocating more ambulances to it.

We can conclude in this case that the cluster with the shortest finish time can be deleted from the list. ∎

Another possible case is that the new makespan is shorter than the old one, which means that the makespan has been improved. What we need to do in this case is to repeat the iterative procedure.

4.2.2. *Minimization of the weighted total flow time*

The total flow time is defined as the summation of the finish time of all clusters. Depending upon the information about the number and severity of casualties in each cluster, it would be natural to associate different weights with them. There are many methods to determine these weights. A typical and simple way is given by:

$$w_i = \frac{N_{i,t_{\mathrm{f}}} - N}{\sum_{i=1}^{m}(N_{i,t_{\mathrm{f}}} - N)}, \quad i = 1, \ldots, m. \quad (28)$$

The basis for this weight assignment is the number of casualties by which a cluster exceeds the base number N which allows it to be classified as a cluster.

For any cluster, if there are no ambulances serving it, its finish time is infinity. Therefore, initially we assign one ambulance to each cluster. Remember that the finish time of a cluster is first strictly decreased and then remains constant. After the initial allocation, we take one ambulance at a time from the pool of remaining ambulances. We consider the addition of an additional ambulance to each cluster. We calculate the new finish time T^{new} for all the clusters and compare them with the corresponding current finish time T_i. If we have $T_i = T_i^{\mathrm{new}}$ for $\forall i$, we stop and quit the procedure. Otherwise, the ambulance is given to the cluster which satisfies:

$$i \in \arg\max_i \ w_i\big(T_i - T_i^{\mathrm{new}}\big), \quad i = 1, \ldots, m. \quad (29)$$

We repeat the same procedure until all the ambulances are allocated.

5. Subsequent ambulance reallocation

The emergence of new clusters as the disaster evolves creates new management issues. The analysis in this section assumes that the reallocation of ambulances can be made only at predetermined discrete points in time, e.g., every hour. The length of the time unit is critical to the efficiency of the policy. If the length is too short, frequent reallocations will cause a considerable amount of time to be wasted on traveling between clusters. If the length is too long, quite a few of the newly emerging clusters can accumulate and have to wait for service for a long time.

The technical difficulties in studying the reallocation problem are: (i) in determining the status of a cluster at a reallocation time epoch; and (ii) in calculating the updated time measures of a cluster given that reallocation occurs. Once these time measures are calculated the iterative scheme of Section 4.2.1 can be readily modified to minimize the makespan. The rest of this section is devoted to addressing the two aforementioned technical problems.

5.1. *Cluster status at a reallocation time epoch*

Consider a cluster which is first reported at time t_0. The initial number of casualties and arrival rate are given by N_0 and λ_0, respectively. We assume that the reallocation decision is made at time $t_0^{\mathrm{new}}(t_0^{\mathrm{new}} > t_0)$. We need to consider two cases: in the first case the reallocation time occurs in phase I, whereas in the second case the reallocation time happens in phase II.

Case 5: $t_0^{\mathrm{new}} \leq t_0 + t_{\mathrm{m}}$.

We start by noting that $t_{\mathrm{m}}^{\mathrm{new}}$ and $t_{\mathrm{f}}^{\mathrm{new}}$ are updated to $t_0 + t_{\mathrm{m}} - t_0^{\mathrm{new}}$ and $t_0 + t_{\mathrm{f}} - t_0^{\mathrm{new}}$, respectively. The values N_0^{new}

and λ_0^{new} at time t_0^{new} also differ from the values N_0 and λ_0 at time t_0. In order to calculate N_0^{new} and λ_0^{new}, we need to further consider two subcases based on whether or not the cluster has been serviced.

Case 5.1: New cluster.

A *new* cluster means that it has not yet been served. This cluster occurs during the period between the new reallocation time and the last reallocation time. In this case, according to Equation (5) the number of casualties at time t_0^{new} (denoted as N_0^{new}) can be calculated as:

$$N_0^{\text{new}} = \frac{1}{2}k\left(t_0^{\text{new}} - t_0\right)^2 + \lambda_0\left(t_0^{\text{new}} - t_0\right) + N_0. \quad (30)$$

Similarly, from Equation (2) the arrival rate at time t_0^{new} (denoted as λ_0^{new}) can be given as follows:

$$\lambda_0^{\text{new}} = k\left(t_0^{\text{new}} - t_0\right) + \lambda_0. \quad (31)$$

Since this cluster has not been serviced, every casualty that appears before time t_0^{new} remains in the cluster and thus the updated total number of casualties (denoted as $N_{t_f}^{\text{new}}$) is still given by Equation (7).

The relationship between the information on a cluster at time t_0 and the updated one at time t_0^{new} is shown in Fig. 4. The dashed curve represents the original growth pattern of the cluster, whereas the solid curve depicts the updated information on the cluster. We can see that the dashed curve and the solid curve overlap one another, which points out the fact that the cluster is untouched.

Case 5.2: Old cluster.

Contrary to a *new* cluster, an *old* cluster represents one that has been serviced before the current reallocation time. Since the arrival rate is an intrinsic characteristic of a cluster, its pattern cannot be influenced by whether or not the cluster has been previously serviced. Thus, in this case the arrival rate at time t_0^{new} is also given by Equation (31).
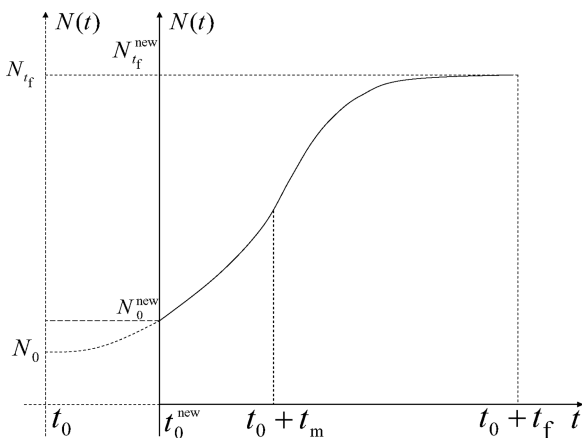

Fig. 4. Information about a cluster at the reallocation time (case 5.1).

However, the number of casualties at time t_0^{new} is influenced by the service. It is calculated by subtracting the number of casualties sent to hospital from the total number of casualties who have appeared in the cluster so far. As in the calculation of the arrival rate, the total number of casualties that have occurred before time t_0^{new} is not affected by the service and is given by Equation (30).

The calculation of the number of casualties who have been sent to hospital is not straightforward. Consider a situation in which the ambulances allocated to a cluster come from different locations. In all likelihood these ambulances arrive at the cluster at different time points, which causes the service rate to change with time. Suppose that the cluster is first reported to the emergency center at time t_0 and the initial service rate is μ_0. At time t_1, the first batch of ambulances arrives (or leaves) and the service rate is changed to μ_1. The process continues in this manner. The service rate changes whenever a batch of ambulances arrives (or leaves). If there are a total of k batches of ambulances that arrive (or leave) before the current reallocation time, the number of casualties sent to hospital by time t_0^{new} can be calculated as:

$$\sum_{i=0}^{k} \int_{t_i}^{t_{i+1}} \mu_i \, \mathrm{d}t, \quad \text{where } t_{k+1} = t_0^{\text{new}}. \quad (32)$$

Then the number of casualties at time t_0^{new} in this cluster is given by:

$$N_0^{\text{new}} = \frac{1}{2}k\left(t_0^{\text{new}} - t_0\right)^2 + \lambda_0\left(t_0^{\text{new}} - t_0\right)$$
$$+ N_0 - \sum_{i=0}^{k} \int_{t_i}^{t_{i+1}} \mu_i \, \mathrm{d}t. \quad (33)$$

Similarly, the updated total number of casualties is computed by subtracting the number of casualties sent to hospital from the total number of casualties in the cluster. As in the arrival rate, the total number of casualties is another intrinsic characteristic of a cluster and is always given by Equation (7). Thus, we can calculate $N_{t_f}^{\text{new}}$ as follows:

$$N_{t_f}^{\text{new}} = \frac{1}{2}k t_m t_f + \frac{1}{2}\lambda_0(t_m + t_f) + N_0 - \sum_{i=0}^{k} \int_{t_i}^{t_{i+1}} \mu_i \, \mathrm{d}t. \quad (34)$$

Again, the relationship between the information about a cluster at time t_0 and the updated one at time t_0^{new} is shown in Fig. 5. We can see that this time the solid curve no longer overlaps the dashed curve. The fact that the solid curve is located below the dashed curve indicates that a given number of casualties has been delivered to hospital. This number is equal to the difference between these two curves.

Case 6: $t_0^{\text{new}} > t_0 + t_m$.

We now turn to consideration of the second case in which the decision time point occurs during the second phase of
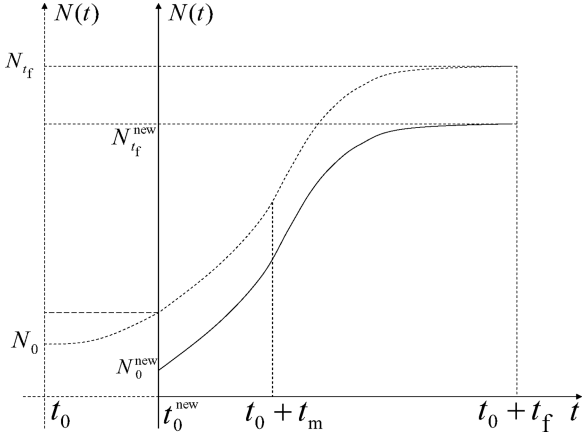
Fig. 5. Information about a cluster at the reallocation time (case 5.2).

a cluster. The time t_f^{new} is given by $t_0 + t_f - t_0^{new}$. This situation also happens in the following two kinds of clusters.

Case 6.1: New cluster.

The cluster has not been serviced but the first phase of it is quite short, specifically, $t_m < t_0^{new} - t_0$. In this case, from Equation (6) the number of casualties at time t_0^{new} is given by:

$$N_0^{new} = -\frac{1}{2}\left(\frac{kt_m + \lambda_0}{t_f - t_m}\right)\left(t_0^{new} - t_0\right)^2$$
$$+ \left[\left(\frac{kt_f + \lambda_0}{t_f - t_m}\right)t_m + \lambda_0\right]\left(t_0^{new} - t_0\right) \quad (35)$$
$$-\frac{1}{2}\left(\frac{kt_f + \lambda_0}{t_f - t_m}\right)t_m^2 + N_0.$$

Equation (4) yields the arrival rate of the cluster at time t_0^{new} as:

$$\lambda_0^{new} = -\left(\frac{kt_m + \lambda_0}{t_f - t_m}\right)\left(t_0^{new} - t_0 - t_m\right) + kt_m + \lambda_0. \quad (36)$$

The total number of casualties $N_{t_f}^{new}$ is still given by Equation (7).

Case 6.2: Old cluster.

The cluster has been serviced and the reallocation time t_0^{new} occurs in the second phase of the cluster. By a similar argument to case 5.2, the number of casualties at time t_0^{new} in this cluster can be calculated as:

$$N_0^{new} = -\frac{1}{2}\left(\frac{kt_m + \lambda_0}{t_f - t_m}\right)\left(t_0^{new} - t_0\right)^2$$
$$+ \left[\left(\frac{kt_f + \lambda_0}{t_f - t_m}\right)t_m + \lambda_0\right]\left(t_0^{new} - t_0\right) \quad (37)$$
$$-\frac{1}{2}\left(\frac{kt_f + \lambda_0}{t_f - t_m}\right)t_m^2 + N_0 - \sum_{i=0}^{k}\int_{t_i}^{t_{i+1}}\mu_i\,dt.$$

The arrival rate and the updated total number of casualties at time t_0^{new} are given by Equations (36) and (34), respectively.

5.2. *Updated time measures of a cluster*

As mentioned in Section 4.1, the finish time of a cluster should be calculated first because it is the basis of analyzing many system performance measures, e.g., makespan. Based on the updated information on clusters at the reallocation time point, only those clusters which satisfy $N_{t_f}^{new} > N$ (minimum requirement of a cluster) are qualified to be serviced.

Referring to Section 5.1, a qualified cluster is classified as either an *old* or a *new* cluster. Thus, the reallocation of ambulances may happen in the following two cases: (i) reallocation of ambulances between an *old* cluster and another *old* cluster; and (ii) reallocation of ambulances from an *old* cluster to a *new* cluster. Recall the iterative procedure for minimizing the makespan described in Section 4.2.1. The core part of this procedure requires, in each iteration, extracting one ambulance from the cluster with the shortest finish time and then allocating it to the cluster with the longest finish time. We call the cluster that releases an ambulance a *granting* cluster, and we name the cluster that receives the ambulance a *receiving* cluster. The way to compute the updated finish time of a *granting* cluster is different from the method to calculate the updated finish time of a *receiving* cluster. In the following we will describe the processes to obtain the finish time for both types of clusters.

Case 7: Granting cluster.

As discussed above, one ambulance that is serving an existing cluster is released and is sent to another one. The consequence of the release only affects the service rate of that cluster. The service rate changes immediately after the ambulance leaves. Based on the time that the reallocation occurs, we consider the following two cases.

Case 7.1: $t_0^{new} \leq t_0 + t_m$.

In this case, the finish time can be calculated by the method discussed in Section 4.1. We just need to use t_0^{new}, t_m^{new}, t_f^{new}, λ_0^{new}, N_0^{new}, and $N_{t_f}^{new}$ to replace the corresponding parameters.

Case 7.2: $t_0^{new} > t_0 + t_m$.

The reallocation time occurs in phase II of the cluster. Two cases are further developed.

Case 7.2.1: $N_{t_f}^{new} - N \leq N_0^{new}$.

Since the threshold has been reached, all ambulances simply need to work until the required number of casualties have been delivered to hospital. We obtain:

$$T = t_0^{new} + \frac{N_{t_f}^{new} - N}{\mu}. \quad (38)$$

Case 7.2.2: $N_{t_f}^{\text{new}} - N > N_0^{\text{new}}$.

The arrival rate at any time t is given by:

$$\lambda_2(t) = k''(t - t_0^{\text{new}}) + \lambda_0^{\text{new}}. \tag{39}$$

By setting $\lambda_2(t_0^{\text{new}} + t_f^{\text{new}}) = 0$, we get:

$$\lambda_2(t) = -\frac{\lambda_0^{\text{new}}}{t_f^{\text{new}}}t + \frac{(t_0^{\text{new}} + t_f^{\text{new}})\lambda_0^{\text{new}}}{t_f^{\text{new}}}. \tag{40}$$

Since $\lambda_2(t) = \mathrm{d}N_2(t)/\mathrm{d}t$, we integrate both sides of Equation (40) and apply the initial condition $N_2(t_0^{\text{new}}) = N_0^{\text{new}}$, which yields:

$$N_2(t) = -\frac{1}{2}\left(\frac{\lambda_0^{\text{new}}}{t_f^{\text{new}}}\right)\left[t^2 - (t_0^{\text{new}})^2\right]$$
$$+ \left[\frac{(t_0^{\text{new}} + t_f^{\text{new}})\lambda_0^{\text{new}}}{t_f^{\text{new}}}\right](t - t_0^{\text{new}}) + N_0^{\text{new}}. \tag{41}$$

The time $\tilde{T}$ is given by solution of the following equation:

$$N_2(t) = N_{t_f}^{\text{new}} - N. \tag{42}$$

The roots of Equation (42) are:

$$t_{1,2} = t_0^{\text{new}} + t_f^{\text{new}} \pm \frac{\sqrt{(t_0^{\text{new}} + t_f^{\text{new}})^2(\lambda_0^{\text{new}})^2 + 2\lambda_0^{\text{new}}C}}{\lambda_0^{\text{new}}}, \tag{43}$$

where $C = (1/2)\lambda_0^{\text{new}}(t_0^{\text{new}})^2 - (t_0^{\text{new}} + t_f^{\text{new}})\lambda_0^{\text{new}}t_0^{\text{new}} + (N_0^{\text{new}} + N - N_{t_f}^{\text{new}})t_f^{\text{new}}$.
It is easy to see that $\tilde{T}$ is given by:

$$\tilde{T} = t_0^{\text{new}} + t_f^{\text{new}} - \frac{\sqrt{(t_0^{\text{new}} + t_f^{\text{new}})^2(\lambda_0^{\text{new}})^2 + 2\lambda_0^{\text{new}}C}}{\lambda_0^{\text{new}}}. \tag{44}$$

After obtaining the value of $\tilde{T}$, we further consider two cases.

Case 7.2.2.1: $\mu(\tilde{T} - t_0^{\text{new}}) \le N_{t_f}^{\text{new}} - N$.

Even though all ambulances are busy in the time interval $[t_0^{\text{new}}, \tilde{T}]$, the finish time happens after the time points $\tilde{T}$. Thus the finish time is calculated using Equation (38).

Case 7.2.2.2: $\mu(\tilde{T} - t_0^{\text{new}}) > N_{t_f}^{\text{new}} - N$.

The time $\hat{T}$ occurs before time $\tilde{T}$. This means that the ambulances are not fully loaded during $[\hat{T}, \tilde{T}]$. Then we have:

$$T = \tilde{T}. \tag{45}$$

Case 8: Receiving cluster.

Since clusters are spatially distributed, in this case we need to consider the travel time of an ambulance which moves between clusters. The ambulances that are allocated to a
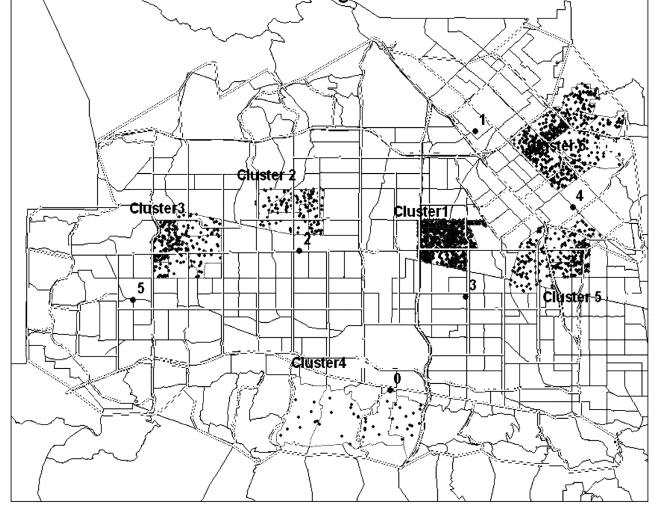

Fig. 6. Case study based on a simulated earthquake in Northridge, CA.

cluster may come from several clusters, which causes ambulances to reach the cluster sequentially in time. Consequently, the service rate changes as a function of time. Consider a typical cluster which receives q ambulances. When we calculate the finish time, this factor should be taken into account. We consider the following two cases.

Case 8.1: $t_0^{\text{new}} \le t_0 + t_m$.

The reallocation time occurs in phase I of the cluster. In light of the analysis presented in Section 4.1, we further consider three cases.

Case 8.1.1: $N_{t_f}^{\text{new}} - N < N_0^{\text{new}}$.

The number of casualties exceeds the threshold $N_{t_f}^{\text{new}} - N$ at the reallocation time. The initial time begins with $\tilde{t}_0$, where $\tilde{t}_0 = t_0^{\text{new}}$. The initial service rate is given by μ_0. After that the ith ($i = 1, \ldots, q$) ambulance arrives at time $\tilde{t}_i$ and the service rate is changed to μ_i. Assume that the finish time happens in the time interval $[\tilde{t}_{q_1}, \tilde{t}_{q_1+1}]$ where $0 \le q_1 \le q$ and $\tilde{t}_{q+1} = \infty$. (If $\sum_{i=0}^{q_1-1}\int_{\tilde{t}_i}^{\tilde{t}_{i+1}}\mu_i\mathrm{d}t \le N_{t_f}^{\text{new}} - N \le \sum_{i=0}^{q_1}\int_{\tilde{t}_i}^{\tilde{t}_{i+1}}\mu_i\mathrm{d}t$, we can say that the finish time falls in the interval $[\tilde{t}_{q_1}, \tilde{t}_{q_1+1}]$.) The finish

Table 1. Detailed information on each casualty cluster in the case study

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
λ_0	56	45	37	43	54	50
N_0	165	141	112	105	116	130
k	48.29	99.32	32.55	34.66	32.51	65.73
t_m	3.7	2	3.2	2.5	4.2	3
t_f	5.5	4.4	4.8	4.2	6	5
N_{t_f}	914	722	510	431	801	823

Table 2. Results of minimizing makespan and weighted total flow time for the case study

| | Makespan | | Weighted total flow time | | | |
| | | | $w_i = 1$ | | $w_i = \frac{N_{i,t_f} - N}{\sum_i N_{i,t_f} - N}$ | |
	Number of ambulances	Finish time (hours)	Number of ambulances	Finish time (hours)	Number of ambulances	Finish time (hours)
Cluster 1	22	6.167	19	7.140	22	6.167
Cluster 2	17	6.098	17	6.098	17	6.098
Cluster 3	11	6.212	14	4.881	11	6.212
Cluster 4	9	6.130	12	4.597	9	6.130
Cluster 5	19	6.149	18	6.491	19	6.149
Cluster 6	20	6.025	18	6.694	20	6.025
Makespan	—	6.212	—	7.140	—	6.212
Total flow time	—	36.781	—	35.901	—	36.781

time T is given by the following equation:

$$\sum_{i=0}^{q_1-1} \int_{\tilde{t}_i}^{\tilde{t}_{i+1}} \mu_i dt + \mu_{q_1}(T - \tilde{t}_{q_1}) = N_{t_f}^{new} - N. \quad (46)$$

It is easy to get:

$$T = \left(N_{t_f}^{new} - N - \sum_{i=0}^{q_1-1} \int_{\tilde{t}_i}^{\tilde{t}_{i+1}} \mu_i dt + \mu_{q_1} \tilde{t}_{q_1} \right) \Big/ \mu_{q_1}. \quad (47)$$

Case 8.1.2: $N_0^{new} \leq N_{t_f}^{new} - N < (1/2)kt_m^2 + \lambda_0 t_m + N_0.$

The time $\tilde{T}$ is developed from Equation (12) with λ_0, N_0, and N_{t_f} replaced by λ_0^{new}, N_0^{new}, and $N_{t_f}^{new}$, respectively. Then we have that:

$$\tilde{T} = t_0^{new} + \frac{-\lambda_0^{new} + \sqrt{\left(\lambda_0^{new}\right)^2 + 2k\left(N_{t_f}^{new} - N - N_0^{new}\right)}}{k}. \quad (48)$$

Assume that $\tilde{T}$ occurs in the time interval $[\tilde{t}_{q_1}, \tilde{t}_{q_1+1}]$. Then we consider the following two cases.

Case 8.1.2.1: $\sum_{i=0}^{q_1-1} \int_{\tilde{t}_i}^{\tilde{t}_{i+1}} \mu_i dt + \mu_{q_1}(\tilde{T} - \tilde{t}_{q_1}) \leq N_{t_f}^{new} - N.$

The finish time occurs after the time $\tilde{T}$. Assume again that the finish time happens in the time interval $[\tilde{t}_{q_2}, \tilde{t}_{q_2+1}]$. This time interval is later than the one $[\tilde{t}_{q_1}, \tilde{t}_{q_1+1}]$ or these two intervals are the same. In the latter case, the finish time must

fall in the interval $[\tilde{T}, \tilde{t}_{q_1+1}]$. Then we have that:

$$\sum_{i=0}^{q_2-1} \int_{\tilde{t}_i}^{\tilde{t}_{i+1}} \mu_i dt + \mu_{q_2}(T - \tilde{t}_{q_2}) = N_{t_f}^{new} - N, \quad (49)$$

which yields

$$T = \left(N_{t_f}^{new} - N - \sum_{i=0}^{q_2-1} \int_{\tilde{t}_i}^{\tilde{t}_{i+1}} \mu_i dt + \mu_{q_2} \tilde{t}_{q_2} \right) \Big/ \mu_{q_2}. \quad (50)$$

Case 8.1.2.2: $\sum_{i=0}^{q_1-1} \int_{\tilde{t}_i}^{\tilde{t}_{i+1}} \mu_i dt + \mu_{q_1}(\tilde{T} - \tilde{t}_{q_1}) > N_{t_f}^{new} - N.$

The number of casualties is reduced to zero before time $\tilde{T}$. Thus, it is easy to get that:

$$T = \tilde{T}. \quad (51)$$

Case 8.1.3: $(1/2)kt_m^2 + \lambda_0 t_m + N_0 \leq N_{t_f}^{new} - N.$

In this case, the time $\tilde{T}$ occurs in phase II of the cluster. Using Equation (23) it can be shown that:

$$\tilde{T} = t_0^{new} + t_f^{new} - \sqrt{\frac{2N\left(t_f^{new} - t_m^{new}\right)}{kt_m^{new} + \lambda_0^{new}}}. \quad (52)$$

Again, we consider two more cases. These two cases are the same as cases 8.1.2.1 and 8.1.2.2, respectively. The corresponding results are given by Equations (50) and (51).

Case 8.2: $t_0^{new} > t_0 + t_m.$

Table 3. Parameter range for each casualty cluster in the case study

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
λ_0	56	45	37	43	54	50
N_0	165	141	112	105	116	130
t_m	[3.3, 4.0]	[1.8, 2.1]	[3.0, 3.5]	[2.2, 2.6]	[3.8, 4.5]	[2.7, 3.2]
t_f	[5.0, 6.0]	[4.2, 4.5]	[4.5, 5.0]	[4.0, 4.5]	[5.5, 6.5]	[4.8, 5.5]
N_{t_f}	[900, 950]	[710, 750]	[500, 530]	[420, 440]	[780, 800]	[800, 850]

Table 4. Results on minimizing the makespan for 20 randomly generated cases

	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6		Makespan
	Number of ambulances	Finish time	Number of ambulances	Finish time	Number of ambulances	Finish time	Number of ambulances	Finish time	Number of ambulances	Finish time	Number of ambulances	Finish time	
1	22	6.16667	17	6.09804	11	6.21212	9	6.12963	19	6.14912	20	6.025	6.025
2	22	6.25	17	6.14706	11	6.21212	9	6.07407	19	6.01754	20	6.04167	6.01754
3	22	6.37121	17	6.34314	11	6.28788	9	6.09259	19	6.30702	20	6.175	6.09259
4	22	6.43939	17	6.18627	11	6.10606	9	6.07407	19	6.20175	20	6.25	6.07407
5	22	6.15909	17	6.01961	11	6.25758	9	6.11111	20	6.14167	19	6.25439	6.01961
6	22	6.35606	17	6.29412	11	6.15152	9	6.01852	19	6.16667	20	6.15833	6.01852
7	23	6.15217	17	6.12745	11	6.24242	9	6.25926	18	6.37037	20	6.24167	6.12745
8	22	6.18182	17	6.03922	12	5.77778	9	6.09259	19	6.21930	19	6.2807	5.77778
9	22	6.09091	17	6.18627	11	6.34848	9	5.96296	20	6.15833	19	6.17544	5.96296
10	22	6.31061	17	6.34314	11	6.16667	9	6.24074	19	6.18421	20	6.10833	6.10833
11	22	6.10606	18	5.98148	11	6.15152	9	6.20370	19	6.19298	19	6.19298	5.98148
12	22	6.17424	17	6.12745	11	6.25758	9	6.03704	20	5.96667	19	6.27193	5.96667
13	22	6.07576	17	6.17647	11	6.30303	9	6.24074	20	6.03333	19	6.15789	6.03333
14	22	6.26515	17	6.31373	11	6.22727	9	6.18519	19	6.03509	20	6.05833	6.03509
15	22	6.26515	17	6.07843	12	5.95833	9	5.98148	19	6.28947	19	6.37719	5.95833
16	22	6.07576	17	6.15686	12	5.97222	9	5.98148	19	6.11404	19	6.15789	5.97222
17	23	6.07971	17	6.28431	11	6.10606	9	5.94444	18	6.29630	20	6.15833	5.94444
18	22	6.32576	17	6.02941	11	6.24242	9	6.03704	19	6.08772	20	6.12522	6.02941
19	22	6.13636	17	6.21569	11	6.24242	9	6.05556	20	6.00833	19	6.22807	6.00833
20	22	6.07576	17	6.04902	12	5.83333	9	6.29630	19	6.34211	19	6.15789	5.83333

In this case, we only need to study phase II of the cluster. We consider the following two cases.

Case 8.2.1: $N_{t_f}^{\text{new}} - N < N_0^{\text{new}}$.

The analysis of this case is the same as the analysis of case 8.1.1. The result is given by Equation (47).

Case 8.2.2: $N_{t_f}^{\text{new}} - N \geq N_0^{\text{new}}$.

The time $\tilde{T}$ is given by Equation (44). The two cases developed in this case are the same as the ones in case 8.1.2. The results are given by Equations (50) and (51), respectively.

Note that in this section we do not calculate the time $\hat{T}$. Actually the method shown in Section 4.1 still applies. Similar to the way of calculating the updated finish time, we just need to incorporate the fact that the service rate changes with time.

6. Earthquake disaster case study

The area chosen for this case study is the Los Angeles basin with the simulated earthquake being the one that occurred in Northridge in 1994. The Ground Truth, from which much of the simulated phenomenology is derived, was generated using the HAZUS software (Al-Momani and Harrald, 2003) and includes information such as human casualties.

After running the simulation, we obtained six clusters as shown in Fig. 6. The threshold value to be a cluster was set to 100. Detailed information on each cluster is given in Table 1. There are a total of 19 hospitals in this area. In order to simplify the problem, we only considered the hospitals that are closest to the clusters. We assume that each hospital has enough capacity to treat the casualties in any cluster close to it. There are 98 ambulances belonging to the 19 hospitals. We further assume that all the ambulances are allocated to the six selected hospitals (labeled "0" through "5" in Fig. 6).

Each ambulance is assumed to pick up three casualties and deliver them to the closest hospital. Under a disaster environment, the road conditions are expected to be poor due to debris blocking the roads, road collapses, traffic congestion etc., which directly causes a sharp decrease in the speed of the ambulances. In our case, we assume an average speed of 25 miles per hour. Since the distance between a cluster and the closest hospital is different for each cluster-hospital pairing, the average travel time of the individual ambulances is different. In our case, the travel time for each cluster is roughly equal to 18 minutes. As we mentioned in Section 3, the service time also includes the preparation time and the on-scene service time. The average values for them are assumed to be 4 and 8 minutes, respectively.

Using this data we applied the procedures for minimizing the makespan and for minimizing the weighted total flow time and obtained the results listed in Table 2. As can be readily seen from the table, the allocations for the case of the makespan and when the weights of all clusters are assumed to be the same differ significantly. For equally weighted clusters, we focus on clusters 3 and 4, which have a low arrival rate, so as to generate many casualties with a low flow time, thereby reducing the total flow time considerably. Interestingly, the solution for minimizing the makespan and that for minimizing the weighted total flow time when the weights are given in Equation (28) are the same, which points to the robustness of this particular ambulance allocation.

Since the disaster environment is highly dynamic, we need to capture the randomness of the growth of each cluster. Consider the six parameters listed in Table 1. Here λ_0 and N_0 are the observed initial arrival rate and initial number of casualties, respectively. We treat them as deterministic parameters. However, others are random parameters which are difficult to predict precisely at the allocation time point. An alternative method to address this problem is to determine a range of values rather than fixing a value for each parameter. This range ensures that the actual value of a parameter falls in it with a predetermined confidence α (e.g., $\alpha = 0.9$). Table 3 outlines an example. We note that the value of the parameter k is determined by the values of the other parameters.

Based on the information shown in Table 3, we randomly generated 20 different cases and calculated the makespan to obtain the results shown in Table 4 (details regarding these cases are available in Gong (2005). It is interesting to see that the optimal allocation solutions only change slightly, which demonstrates the robustness of this particular ambulance allocation. This insensitivity could be due to tight bounds on the input data ranges.

7. Conclusions

This paper proposes a deterministic model to depict the process of discovering casualties in a cluster. Casualties are discovered by sensors (e.g., police cars, pedestrians, cars driven by citizens). As more sensors engage in the rescue work, the discovery rate first increases until it reaches its peak and then it decreases gradually to zero. Several aspects of this model warrant further investigation. The research directions include: (i) consideration of ambulances with different capacities; (ii) consideration of clusters with different thresholds; and (iii) development of a dynamic model for the growth of a cluster.

As new clusters are reported to the emergency center, ambulances should be redistributed from time to time, i.e., ambulance reallocation comes into play. In our work, we consider a discrete time policy which allows for the reallocation to occur at only a finite number of discrete points in time. The updated information is calculated for each cluster at any reallocation time point. The drawback of the discrete time policy is that some clusters may wait for a relatively long time before they get serviced. For example, if a cluster is reported to the emergency center immediately

after a reallocation time, it has to wait for the whole period until the next reallocation time occurs. A continuous time policy is a good choice to solve such a problem. We reallocate ambulances whenever a new cluster emerges. The disadvantage of a continuous time policy is that the reallocation may occur too frequently. This would result in ambulances wasting a significant amount of time on traveling between clusters. Creating a balance between long waiting times and frequent reallocations is another problem that should be addressed.

Acknowledgements

References

Al-Momani, N.M. and Harrald, J.R. (2003) Sensitivity of earthquake loss estimation model: how useful are the predictions. *International Journal of Risk Assessment and Management*, **4**(1), 1–19.

Anon (2006) *Tsunami*, Department for Planning and Infrastructure, Government of Western Australia, available at http://www.dpi.wa.gov.au/coastaldata/tidesandwaves/tsunami.html. Accessed 2006.

Baveja, A. (1993) Optimal strategies for cracking down on illicit drug markets. Ph.D. dissertation, University at Buffalo (SUNY), Buffalo, NY.

Becker, G.S. (1976) *The Economic Approach to Human Behavior*, The University of Chicago Press, Chicago, IL.

Behrens, D.A., Caulkins, J.P., Tragler, G. and Feichtinger, G. (2000) Optimal control of drug epidemics: prevent and treat—but not at the same time? *Management Science*, **46**(3), 333–347.

Caulkins, J.P. (1990) The distribution and consumption of illicit drugs: some mathematical models and their policy implications. Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Gong, Q. (2005) Responding to casualties in a disaster relief operation: initial ambulance allocation and reallocation, and switching of casualty priorities. Ph.D. dissertation, The State University of New York at Buffalo, Buffalo, NY.

Gong, Q. and Batta, R. (2004) Methodology to manage priority queues of casualties in a dynamic disaster environment, in *Proceedings of the 13th Industrial Engineering Research Conference*, Institute of Industrial Engineers, Norcross, GA.

Gong, Q., Jotshi, A. and Batta, R. (2004) Dispatching/routing of emergency vehicles in a disaster environment using data fusion concepts, in *Proceedings of the 7th International Conference on Information Fusion*, International Society on Information Fusion, pp. 967–974.

Gros, X.E. (1997) *NDT Data Fusion*, Arnold, London, UK.

Islam, K.M.S. (1998) Spatial dynamic queueing models for the daily deployment of airtankers for forest fire control. Ph.D. dissertation, University of Toronto, Toronto, Canada.

Naik, A.V., Baveja, A., Batta, R. and Caulkins, J.P. (1996) Scheduling crackdowns on illicit drug markets. *European Journal of Operational Research*, **88**, 231–250.

Parks, G.M. (1964) Development and application of a model for suppression of forest fires. *Management Science*, **10**(4), 760–766.

Scott, P.D. (2004) Crisis management in a data fusion synthetic task environment, in *Proceedings of the 7th International Conference on Information Fusion*, International Society on Information Fusion, pp. 330–337.

Tragler, G., Caulkins, J.P. and Feichtinger, G. (2001) Optimal dynamic allocation of treatment and enforcement in illicit drug control. *Operations Research*, **49**(3), 352–362.

Biographies

Qiang Gong is an Operations Analyst at United Airlines. He holds a Ph.D. degree in Operations Research from the Department of Industrial Engineering, University at Buffalo (SUNY). This paper is part of his Ph.D. research which was sponsored by the Air Force Office of Scientific Research.

Rajan Batta is a Professor of Industrial and Systems Engineering, and Associate Dean for Graduate Studies in the School of Engineering and Applied Sciences, University at Buffalo (SUNY). He served as Qiang Gong's dissertation supervisor. His areas of interest include the application of OR to problems in the military and in homeland security. His work has been supported by the National Science Foundation, the National Institute of Justice and by several corporate sponsors, including Lockheed Martin and Boeing. He has co-authored, together with his graduate students and colleagues, 81 journal publications (in journals such as *Operations Research, Transportation Science, Interfaces, IIE Transactions, European Journal of Operational Research, Networks*, etc.). He also has served as the supervisor or co-supervisor of 26 completed doctoral dissertations and 36 completed MS theses. Currently, he is a Departmental Editor for *IIE Transactions* and on the Editorial Advisory Board for *Computers & Operations Research*.