ALLOCATION OF EMERGENCY UNITS: RESPONSE AREAS

Jan M. Chaiken

December 1971

P-4745

ALLOCATION OF EMERGENCY UNITS: RESPONSE AREAS

Jan M. Chaiken*

The New York City-Rand Institute
545 Madison Avenue
New York, N. Y. 10022

ABSTRACT

The average travel time for emergency units such as fire engines, ambulances, and police patrol cars, which respond to spatially distributed incidents, is not necessarily minimized by always dispatching the closest available unit(s) to each incident. Methods are described for changing response areas so as to reduce average travel time and also reduce the imbalance of workload among units.

---

*Any views expressed in this paper are those of the author. They should not be interpreted as reflecting the views of The New York City-Rand Institute or the official opinion or policy of the City of New York. Papers are reproduced by The Rand Corporation as a courtesy to members of its staff.

This paper was prepared for presentation at the Fourth Colloquium on Optimization Techniques of the International Federation for Information Processing, Los Angeles, October, 1971.

In recent years optimization techniques have begun to be applied to problems of urban emergency service systems such as fire departments, police patrol systems, and ambulance services. This paper describes an example of such an application which has led to new principles for designing the response areas of emergency units.

To set the stage, I will give a brief description of the dispatch operations of a typical emergency service. A member of the public ordinarily reports an emergency by activating an alarm box located on the street or in a building, or by calling an emergency telephone number. Each alarm box has a code number which indicates its location, and when the box is activated the code number registers at the appropriate dispatch center. (Some of these boxes permit voice communication between the caller and the dispatcher, so that the dispatcher can obtain additional information about the nature of the emergency.) Once the box number is known, the dispatcher consults a file giving, in the case of mobile units such as police patrol cars, the identification of the unit in whose patrol area the box is located, or, in the case of units dispatched from fixed facilities such as fire houses, the identity of the closest facility or a sequenced list of facilities in order of their distance from the box.

If the call is reported by telephone rather than by alarm box, the dispatcher determines the address or approximate location of the incident and then consults a file which identifies the patrol area or alarm box closest to the address. Once the dispatcher knows the alarm box number, he can find the list of facilities as if he had received a box alarm.

The selection of units which will be dispatched to the incident is essentially determined by the information obtained from the file by the dispatcher. In the case of mobile units, if the patrol area found in the file contains an available unit, that unit will be dispatched to the incident. (There is, of course, no guarantee that this unit is actually closer to the incident than any other unit.) If the emergency is a fire, ordinarily several units of different types will be dispatched. When n_i units of type i are supposed to be dispatched, the

first n_i units on the sequenced list will be dispatched, if they are available for service. If one or more of the first n_i are unavailable, the dispatcher may in some cases dispatch those units which are available, while in other cases he may search further down the list for additional available units to dispatch.

In any event, the units dispatched to the incident will always be the ones which the dispatcher believes are closer to the incident than any other available units of the same type. The objective underlying this dispatching protocol is to minimize each component of the vector which gives the travel times of units responding to incidents. (It is useful to note that the dispatching method described here allows us to define a response area for each unit; it consists of all locations to which the unit will be dispatched if all units are available.)

The possibility of changing the standard dispatching strategy arose as part of our work for the New York City Fire Department. This analysis was undertaken by Grace Carter, Edward Ignall, and myself. The fire department faced the problem that some parts of the city had so many fire alarms that the units in those areas might fight as many as 20 fires in a busy night, leading to a condition in which the firefighters felt overworked. However, nearby areas, sometimes not more than a mile or two away, might have a much lower level of activity.

A natural idea for helping to relieve the excessive workloads of some units is to contract the response areas for the busiest units and expand the response areas for the least busy ones. This would tend to distribute the workload more evenly among the units. But we felt we should find out what would happen to travel times if such a change were made. Presumably travel times would increase, and the problem was to estimate the magnitude of the increase. Thus, we wished to calculate expected travel times and workloads of units as functions of the response areas.

Because the travel times and workloads depend on which units are dispatched to each alarm, and the choice of units to dispatch depends on the availability of units, in general one finds that the quantitites

of interest are functions of the arrival rates for alarms, the service times of units, and the dispatching policy. The arrival process can be modeled by assuming that there are several types of alarms, with type m alarms arriving in any region A according to a Poisson process with rate $\lambda_m(A)$. Then each λ_m is a measure on the region under consideration. (Poissonicity is an excellent approximation, but in practice the arrival rates are not actually constant over time. The alarm "types" may distinguish building fires from brush fires, or telephone alarms from box alarms, or other appropriate characteristics of the incidents.)

The service times may be modeled by assuming that if a group G of units is dispatched to an alarm at location $\underline{x}$, then the service times are given by a matrix $F_m(\underline{x},G)$ whose ij-component is the service-time distribution of the j^{th} unit of type i. The dispatching policy can then be thought of as a function $H_n(\underline{x},\lambda,F,\underline{a})$ which specifies the group of companies to dispatch to a type n alarm at $\underline{x}$ if the λ_m's are the arrival rates, the F_m's are the service times, and the current availabilities of units are given by the vector $\underline{a}$ (whose components are either zero or an unexpired service time, depending on whether the corresponding unit is available or not). Typical dispatching strategies actually in use, as described above, are independent of λ and F and consider all units as being either available or unavailable, ignoring unexpired service times.

If F and H are sufficiently simple, it is possible to calculate the workloads of the units and the expected travel time for the kth-arriving unit at an incident. To take a simple example, let us suppose that there are only two units in a region B, that they are located at fixed facilities, and that exactly one of them will be dispatched to each alarm unless both are unavailable (in which case the alarm is served by some other unit which does not concern us for the moment). Then the "group" G to be dispatched is either {unit 1} or {unit 2}. We assume further that there is only one type of alarm,

with arrival rate measure λ, and that the service-time distribution $F(\underline{x},G)$ does not depend on $\underline{x}$ or G and has finite mean $1/\mu$.

The dispatching policy in this example is as follows: We select a response area A for unit 1. If an alarm arrives in A when both units are available, unit 1 is dispatched. If the alarm arrives in the complement B-A when both are available, unit 2 is dispatched. If an alarm arrives when one unit is available, that unit is dispatched.

Given these assumptions, one can calculate the steady-state probabilities P_{ij} of the states

00 = both units available
10 = unit 1 busy, unit 2 available
01 = unit 2 busy, unit 1 available
11 = both units busy.

These depend on the service-time distribution only through its mean $1/\mu$, as is proved in Reference 2.

The workload W_j of unit j can be defined as the steady-state probability that unit j is busy, and thus $W_1 = P_{10} + P_{11}$, $W_2 = P_{01} + P_{11}$. The workload difference $|W_1 - W_2|$ can be calculated to be $P_{00}|\lambda(A) - \lambda(B-A)| \div (\lambda(B) + \mu)$, which is minimized when A is selected so that half of the alarms arrive in A, and half in B-A.

The expected travel time can be calculated as follows: Of those alarms which arrive when the state is 00, a fraction $\lambda(A)/\lambda(B)$ will be served by unit 1 and have average response time $T_1(A)$ = average time to travel from the location of unit 1 into A. Similarly, a fraction $\lambda(B-A)/\lambda(B)$ will be served by unit 2 and have average response time $T_2(B-A)$. Those alarms which arrive when the state is P_{01} (resp. P_{10}) will be served from location 1 (resp. 2) and have average response time $T_1(B)$ (resp. $T_2(B)$). Thus, the expected travel time, given A is the response area for unit 1, is

$$\overline{T}(A) = P_{00}(T_1(A)\lambda(A) + T_2(B-A)\lambda(B-A))/\lambda(B) + P_{01}T_1(B) + P_{10}T_2(B) + P_{11}\tau,$$

where τ is the average travel time for calls served by units outside B.

After performing a calculation which is given in detail in Reference 1, one finds that

$$\bar{T}(A) = \frac{P_{00}}{\lambda(B)} \int_A (t_1 - t_2 - s_0) d\lambda + \alpha$$

where $t_j(\underline{x})$ is the travel time from location j to point $\underline{x}$, $s_0 = \lambda(B)(T_1(B) - T_2(B))/(\lambda(B) + \mu)$, and α is a constant (i.e., independent of A).

It is a property of the integral on the right that it attains its minimum when A is the set

$$X = \{\underline{x}: t_1(\underline{x}) - t_2(\underline{x}) < s_0\}$$

Unless $s_0 = 0$, this means that average travel time is not minimized by always dispatching the closest available unit to each alarm. In fact, if the policy of dispatching the closest available unit, i.e., of selecting A to be the set $\{\underline{x}: t_1(\underline{x}) < t_2(\underline{x})\}$, produces a situation where unit 2 has a greater workload than unit 1, it will commonly happen that $s_0 > 0$. Thus, the response area for unit 1 can be expanded in such a way that not only is the workload of unit 2 reduced but also the expected travel time decreases. The intuitive notion that one must pay a penalty in travel time in order to improve the balance in workloads is not correct.

It should be noted that the dispatch policy which minimizes expected travel time will dispatch unit 1 to $\underline{x}$ if and only if

(a) both units are available and $\underline{x}$ is in the set X defined above, or

(b) only unit 1 is available.

This policy depends on λ and F as well as the availabilities, since s_0 is a function of λ and μ.

The insights derived from this example suggest that in realistic models having more than 2 units it may be desirable to find the response areas which minimize expected travel time, since these may also reduce workload imbalance, compared to the usual response areas. If several units are to be dispatched to each alarm, one may have to substitute some single objective function for the vector of travel times for the kth-arriving unit at a type m alarm, since it is not necessarily possible or desirable to find a dispatching policy which minimizes the expectation of each component of the vector. This objective function would be a function of the travel-time vector, for example a weighted average of the components of the vector, and therefore we call it a generalized travel time. We denote by $t_G(\underline{x},m)$ the generalized travel time if group G is dispatched to a type m alarm at $\underline{x}$, which includes the weight attached to a type m alarm.

Methods have been developed for minimizing the expected generalized travel time in special cases. The example which follows is a case in which linear programming has been applied. This model was developed by Edward Ignall. We assume that there are a total of 2N units, of which N will be dispatched to each alarm. Under these circumstances, the dispatching policy is determined by specifying which group responds to an alarm at $\underline{x}$ if all units are available, since an alarm which arrives when only N units are available will be served by all of them. We denote by $y(\underline{x},m,G)$ a function which is 1 if group G is to be dispatched to a type m alarm at $\underline{x}$ when all units are available, and is 0 otherwise. We continue to assume that the service times $F_m(\underline{x},G)$ are independent of $m,\underline{x}$, and G and have finite mean $1/\mu$; all units in a group complete service at the same time. In addition, we assume that the alarm rate measures λ_m are concentrated at a finite set of points; if $\underline{x}$ is one of these points, we denote $\lambda_m(\{\underline{x}\})$ by $\lambda(\underline{x},m)$. Then λ denotes the total alarm rate in the region.

In analogy with the average response times $T_j(B)$ in the 2-unit sample given above, we introduce

$$T_G = \sum_{\underline{x},m} t_G(\underline{x},m)\lambda(\underline{x},m)/\lambda,$$

which is the expected generalized response time if all alarms are served by group G. One can then show (Reference 3) that with dispatching policy y, the expected generalized travel time is

$$\overline{T} = \sum_{\underline{x},m,G} \frac{\lambda(\underline{x},m)y(\underline{x},m,G)}{\lambda(\lambda+\mu)/\mu}\left[t_G(\underline{x},m) + T_{G'}\beta + \frac{(T_G+T_{G'})\beta^2}{2(1-\beta)}\right] + \alpha$$

where G' denotes the group of all units not in G, $\beta = \lambda/(\lambda + \mu)$, and α is a constant independent of y. A direct method of minimizing $\overline{T}$ is to choose, for each $\underline{x}$ and m, $y(\underline{x},m,G) = 1$ for the group G which minimizes the quantity in brackets on the right. However, one may wish to minimize $\overline{T}$ subject to constraints on the workloads of units.

The steady-state probability that unit j is working can be shown to be

$$W_j = P_{00} \sum_{\underline{x},m,G} \frac{\lambda(\underline{x},m)}{\lambda+\mu}\, y(\underline{x},m,G)\chi(j,G) + C$$

where $\chi(j,G) = 1$ if unit j is in group G, zero otherwise, and C is a constant, independent of y and j. One need only note that the y's enter into the expressions for $\overline{T}$ and W_j linearly to see that a linear programming formulation can be developed to minimize $\overline{T}$ subject to a constraint on the maximum workload difference among units. The optimal values of the policy variables y may turn out to be between 0 and 1, in which case one interprets $y(\underline{x},m,G)$ as specifying the probability of dispatching group G to a type m alarm at $\underline{x}$. Analysis of the results from using such a program is still in progress at this time.

REFERENCES

1. Carter, G., J. Chaiken, and E. Ignall, "Response Areas for Two Emergency Units," Operations Research, to appear, 1972.

2. Chaiken, J., and E. Ignall, "An Extension of Erlang's Formulas which Distinguishes Individual Servers," J. Appl. Prob., to appear, 1972.

3. Ignall, E., "Response Areas For Groups of Fire-Fighting Units, I," unpublished mimeo, The New York City-Rand Institute, 1971.