R-532-NYC/HUD
MARCH 1971

RESPONSE AREAS FOR TWO EMERGENCY UNITS

Grace M. Carter, Jan M. Chaiken and Edward Ignall

THE
NEW YORK CITY
RAND
INSTITUTE

545 MADISON AVENUE NEW YORK NEW YORK 10022 (212) 758-2244

Rand

The New York City-Rand Institute is a nonprofit research institution founded, as its Articles of Incorporation state, "... primarily to conduct programs of scientific research and study, and provide reports and recommendations, relevant to the operations, planning or administration of the City of New York." The Institute was established in 1969 by the City of New York and The Rand Corporation as a center for the continuing application of scientific and analytic techniques to problems of urban life and local government. It is governed by a Board of Trustees appointed jointly by the City and Rand.

R-532-NYC/HUD
MARCH 1971

RESPONSE AREAS FOR TWO EMERGENCY UNITS

Grace M. Carter, Jan M. Chaiken and Edward Ignall

This study was sponsored by the City of New York and by the U. S. Department of Housing and Urban Development. Its contents, however, do not purport to represent the official views or policy of its sponsors.

THE
NEW YORK CITY
RAND
INSTITUTE

545 MADISON AVENUE NEW YORK NEW YORK 10022 (212) 758-2244

Rand

Published by The Rand Corporation

PREFACE

Problems of allocating men and equipment to respond to emergency calls for service have been studied at the New York City-Rand Institute as part of its work for the Fire Department of the City of New York and the U. S. Department of Housing and Urban Development. Among the questions to be answered in allocating emergency units are:

(1) How many units should be on duty?

(2) Where should the units be located?

(3) How many units should be dispatched to each call?

(4) Which particular unit(s) should be dispatched?

One approach to answering the fourth question is analyzed in this report for the case where there are only two units, and rigorous mathematical results are presented. This work led to recommendations for the design of response areas in realistic cases where there are many units. The suggestions were tested by simulation and were adopted by the Fire Department in certain parts of New York City which have high alarm rates.

A companion report by R. Larson and K. Stevenson, R-533, discusses the facility-location problem for the model presented here as well as for other models.

ABSTRACT

For a model in which two units cooperate in serving a region, the average response time to calls and the workload of each unit are calculated as functions of the boundary which separates their response areas. The boundaries which minimize average response time and the ones which equalize workload are determined. Some boundaries can be dominated, in the sense that another boundary improves both workload balance and response time. The set of undominated boundaries is found.

ACKNOWLEDGMENTS

We are grateful for the support of the Fire Department of New York City and the U. S. Department of Housing and Urban Development. Edward H. Blum and Richard C. Larson suggested several useful techniques for applying queueing theory to this problem.

CONTENTS

PREFACE iii

ABSTRACT v

ACKNOWLEDGMENTS vii

Section
I. INTRODUCTION 1

II. ILLUSTRATIVE RESULTS 4

III. QUEUEING ASPECTS 17

IV. AVERAGE RESPONSE TIME 22

V. RESPONSE AREAS WITH MINIMUM AVERAGE RESPONSE TIME 26
Theorem 1 26
Theorem 2 29
Theorem 3 31

VI. DOMINANCE 33
Theorem 4 38
Corollary 41
Theorem 5 42

VII. SENSITIVITY OF FINDINGS TO MODEL ASSUMPTIONS ... 45

REFERENCES 47

I. INTRODUCTION

In recent years several researchers have begun to apply analytical techniques to the problem of improving the allocation policies of urban emergency services [4, 11, 13]. Our efforts have concentrated on a type of emergency service which is supplied by dispatching one or more vehicles (units) from fixed locations to each call for service (alarm). Typically, fire departments, ambulance services, emergency repair services, and even certain special police services operate in this fashion.

In this paper we describe a portion of our work directed toward determining the particular units to dispatch to each alarm, given the locations of the units and the number to be dispatched to an alarm. Our results are general enough to apply to any service system of the fixed-location type, but they by no means encompass all of the important allocation problems of such a system. Other work by ourselves and our colleagues is designed to determine the number of units which ought to be on duty, the number to send to each alarm, the best locations for the units, and the circumstances under which units should be relocated from one place to another [1, 2, 10]. For services which utilize mobile units, such as police patrol cars, a different class of models is being developed [6, 8, 9].

The results presented here concern the case where the region under consideration is essentially served by only two units, and exactly one unit is dispatched to each alarm. This oversimplified case is interesting because it leads to qualitative conclusions which remain true for more complicated systems. Moreover, it is possible to present the results for the two-unit case without elaborate assumptions and notation. Our findings for the n-unit case will be presented elsewhere; they have already been applied

by the New York City Fire Department to modify response patterns in certain parts of the City during periods with a high alarm rate.

Our approach is to select a geographical district, or response area, for each unit to serve. The unit responds to all calls for service inside its response area unless it happens to be busy servicing another call. In the latter case, the other unit will respond unless it is also busy. Thus we explicitly allow units to cross district boundaries.

Under assumptions described later, we obtain the following results:

- Formulas for the workload of each unit and the average response time to all incidents, as functions of the way in which district boundaries are drawn.
- A determination of the district boundary which minimizes average response time. This frequently turns out to be different from the commonly used boundary - the line equidistant from the home locations of the two units.
- Conditions under which the commonly used boundary is dominated by other boundaries, in the sense that another way of drawing the line yields both reduced average response time and a more equal distribution of work between the units.

Most of the results are derived under very general assumptions about the geographical distribution of the calls for service and about the travel speeds, so the discussion is fairly abstract. Therefore we have prefaced the rigorous theory in Sections III to VII by an informal presentation in Section II of the major results for some important examples.

Our work differs from the typical approach to urban location-allocation problems [5, 7, 14, 16] in two important respects: (1) A unit serves its own district only when it is available. Thus, the selection of a particular response area determines not

only the geographical arrangement of the points to be served from each location but also the probability of each unit's being busy. Both of these enter into the formulas for the objective functions. (2) We consider two types of objective functions at once. One is a measure of the quality of the service, for example the response time, and the other is a measure of the strains placed on the components of the system, for example the workloads of the units.

Thus we are more concerned with the form of the objective functions and the shapes of candidate response areas than we are with techniques for utilizing a computer to find optimal response areas. A first approach to the problem of devising computer routines for locating facilities utilizing our findings is contained in recent work by Larson and Stevenson [10].

II. ILLUSTRATIVE RESULTS

We summarize our results concerning the response areas for two units by considering some illustrative examples.

Consider first the example shown in Figure 1. Here we have a rectangular region B on which we have superimposed a coordinate system whose origin is at the center of B and whose axes are parallel to the sides of B. The two units are located on the x-axis, symmetrically placed a distance d from the y-axis.

We suppose that the streets in the region run parallel to the axes, so that each unit's route to an incident occurring in B would consist of segments parallel to the axes. A possible route from the location of unit 1 to an incident at (x,y) is shown in the figure. (In this paper, we always assume that each response by a unit begins at the unit's "home" location. In the case of a service where the vehicle must return to its home after a call before it can service another one, this is a correct description. Some ambulances operate in this fashion; for other services our model is an approximation.)

We suppose further that the units travel at a constant speed v_1 on any one of the streets parallel to the x-axis, and at a constant speed v_2 on any of the streets in the y-direction. If each unit, when dispatched to an incident, selects a route having the smallest possible travel time, then the total time from location 1 to (x,y) is

$$(2.1) \qquad t_1(x,y) = \frac{|x+d|}{v_1} + \frac{|y|}{v_2} ,$$

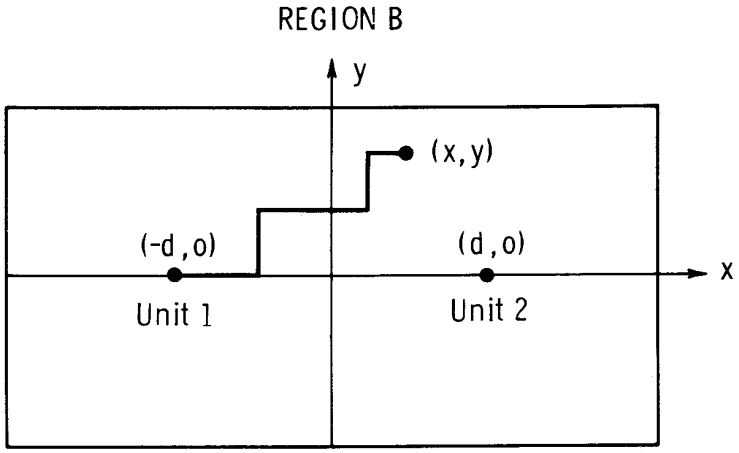
and the total time from location 2 to (x,y) is

$$(2.2) \qquad t_2(x,y) = \frac{|x-d|}{v_1} + \frac{|y|}{v_2} .$$

These travel times can be visualized by imagining that the units first travel as far as necessary along the x-axis and then turn in the

Figure 1

UNITS SYMMETRICALLY LOCATED IN A RECTANGLE



y-direction, whether or not such a route actually exists. Thus we refer to them as right-angle travel times. See Figure 2.

The equations for t_1 and t_2 show that all points to the left of the y-axis (set B_1 in Figure 2) are "closer" (in the sense of travel time) to the location of unit 1 than they are to the location of unit 2, and the reverse is true for points to the right of the y-axis (set B_2). Points on the y-axis are equally "distant" from the two units. Hence it would be natural to divide B into response areas by assigning to unit 1 all alarms which arrive in a region consisting of B_1 plus some points on the y-axis and assigning calls which arrive in the remainder of B to unit 2. This is usually referred to as the "closest-unit" division, and many fire departments and ambulance services actually establish their response areas in this fashion.

Now, however, we introduce the complication of cross-district dispatching. Suppose we have somehow selected a response area A for unit 1. (For example, the set B_1 discussed above is one possible choice for A.) If a call arrives in A at a time when unit 1 is busy servicing a previous alarm, we will not enter the call into a queue. Instead we will dispatch unit 2 or, if necessary, another unit. To be precise, the system operates in the following way:

(1) A call arriving when both unit 1 and unit 2 are available is served by unit 1 if it is in A and is served by unit 2 otherwise.

(2) A call arriving when exactly one of the two units is available is served by the available one, no matter where in B the incident is located.

(3) A call arriving when both unit 1 and unit 2 are busy is served by a unit from another location which does not concern us for the moment.

(4) Units 1 and 2 never respond to calls outside B.

The asymmetry between units outside B, which sometimes respond to calls in B, and units 1 and 2, which never respond outside B, is discussed in Section VII.

Under these circumstances it may happen that our earlier choice for the response area of unit 1 does not in fact give the minimum average response time. Intuitively, one can understand this phenomenon by considering Figure 3. Suppose the rate at which calls arrive in B_2

Figure 2

CLOSEST-UNIT RESPONSE AREAS

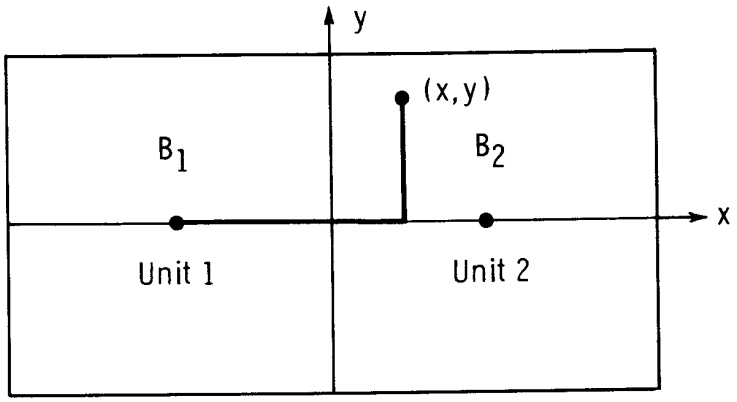
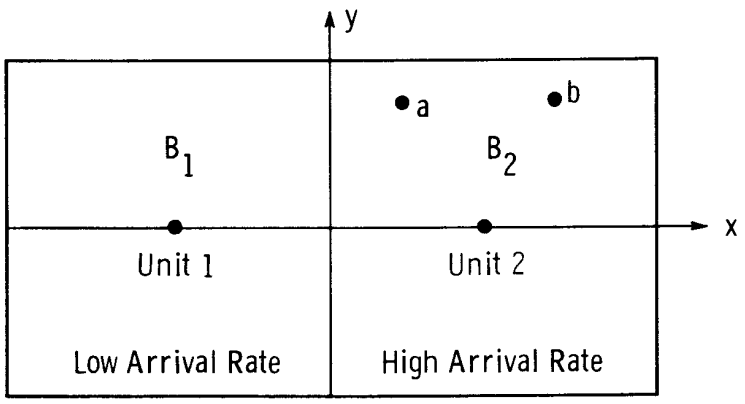


Figure 3



is much higher than the rate in B_1. If an incident occurs at the point a, and unit 2 is dispatched, there may be a good chance of another call arriving while unit 2 is busy.

If such a call does arrive, the chances are good that it will be in B_2 rather than in B_1. Suppose it occurs at point b. Then unit 1 will travel a large distance to b to service the call. On the other hand, if we had originally sent unit 1 to point a, then unit 2 would have been available to go to b, and it is clear that the average travel time for these 2 dispatches is lower than for the original pair of dispatches. Hence it may have been better to have point a in the response area for unit 1.

In Section V we shall give a rigorous argument which shows* that the dividing line between response areas which minimizes average response time is in fact a vertical line somewhat to the right of the y-axis, in the case just described. We also give a formula for the location of the line. In the example being considered here, it is the line

$$(2.3) \qquad x = \frac{v_1}{2} \frac{\lambda}{\lambda+\mu} (T_1 - T_2)$$

where λ is the arrival rate of calls in B, μ is the reciprocal of the average service time, and T_i is the expected response time if every arriving call were to be serviced from the location of unit i. (The precise definition of T_i is given in Section IV.) Since we have assumed that the alarm rate is higher in B_2 than in B_1, we will find that $T_1 > T_2$, so that the line is to the right of the y-axis, as we indicated.

If we had not located our coordinate system so that the origin falls midway between the two units, then unit 1 could be located at $(x_1, 0)$ and unit 2 at $(x_2, 0)$. The dividing line which minimizes average response time would then be

$$x = \frac{1}{2}(x_1 + x_2) + \frac{1}{2} v_1 \frac{\lambda}{\lambda+\mu}(T_1 - T_2).$$

It should be noted that response time to certain locations in B has been increased by moving the dividing line, but the average over all locations has decreased.

---

*under some additional assumptions about the system

Notice that the position of the optimal dividing line moves increasingly to the right as the total alarm rate in B increases, even if we keep the ratio of the rate in B_2 to that in B_1 the same. (Of course, it never moves to the right of the location of unit 2, although this may not be apparent from the formula.)

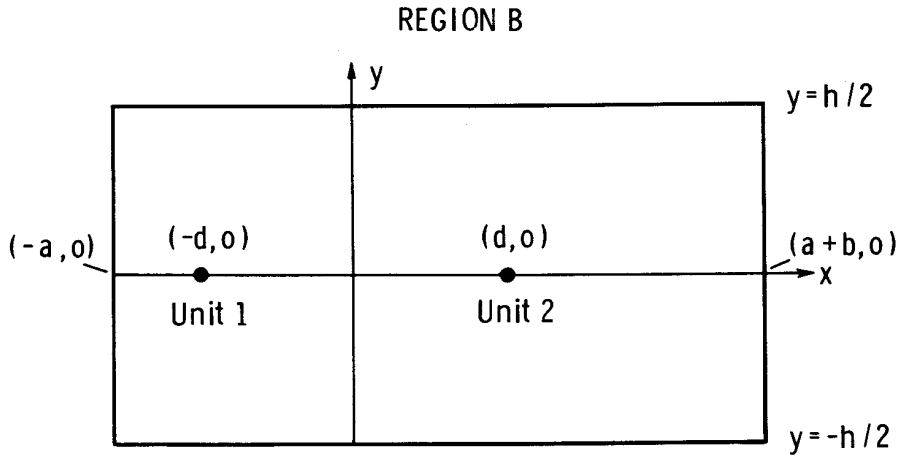
We shall also show that as the dividing line is moved to the right from the y-axis toward the optimal dividing line, the average response time decreases toward the minimum. This is of interest because, by moving the dividing line to the right, we reduce the number of calls serviced by unit 2. Since initially unit 2 was responding to more alarms than unit 1, for some of these dividing lines the average response time is reduced compared to the closest-unit division and the balance of workload is improved. We say that such dividing lines dominate the closest unit division. In more complicated examples than the one being considered here, we find that the closest unit division can be dominated if $T_1 > T_2$ and more than half of the alarms arrive in B_2. In Section VI we discuss dominance in greater detail.

Similar considerations apply even if the alarm rate is uniform across B, but the units are not symmetrically located. Consider the example illustrated in Figure 4. In this case the y-axis again consists of all points which are equally "distant" from the two units, but there is more area in B to the right of the y-axis than there is to the left.

As in the previous example, we can decrease the average response time and improve the balance of workload by moving the dividing line to the right. Equation (2.3) still gives the dividing line which produces the minimum average response time, and for this example we can illustrate how to calculate $T_1 - T_2$. Since we have assumed the alarm rates are uniform, the probability of an arriving alarm lying in an area element with dimensions dx and dy is $dxdy/\ell h$, where ℓ $(=2a+b)$ is the length of the rectangle, and h is its height. Thus

Figure 4

UNITS NOT SYMMETRICALLY LOCATED



$$T_i = \int_{y=-h/2}^{h/2} \int_{x=-a}^{a+b} t_i(x,y)\, dxdy/\ell h.$$

Using equations (2.1) and (2.2) we see that $t_1(x,y)-t_2(x,y)$ is independent of y, and so

$$T_1 - T_2 = \frac{1}{v_1} \int_{-a}^{a+b} (|x+d| - |x-d|)\, dx/\ell$$

$$= \frac{1}{v_1} \int_{a}^{a+b} 2d\, dx/\ell$$

$$= 2bd/v_1\ell.$$

Entering this in equation (2.3) we see that the optimal dividing line is

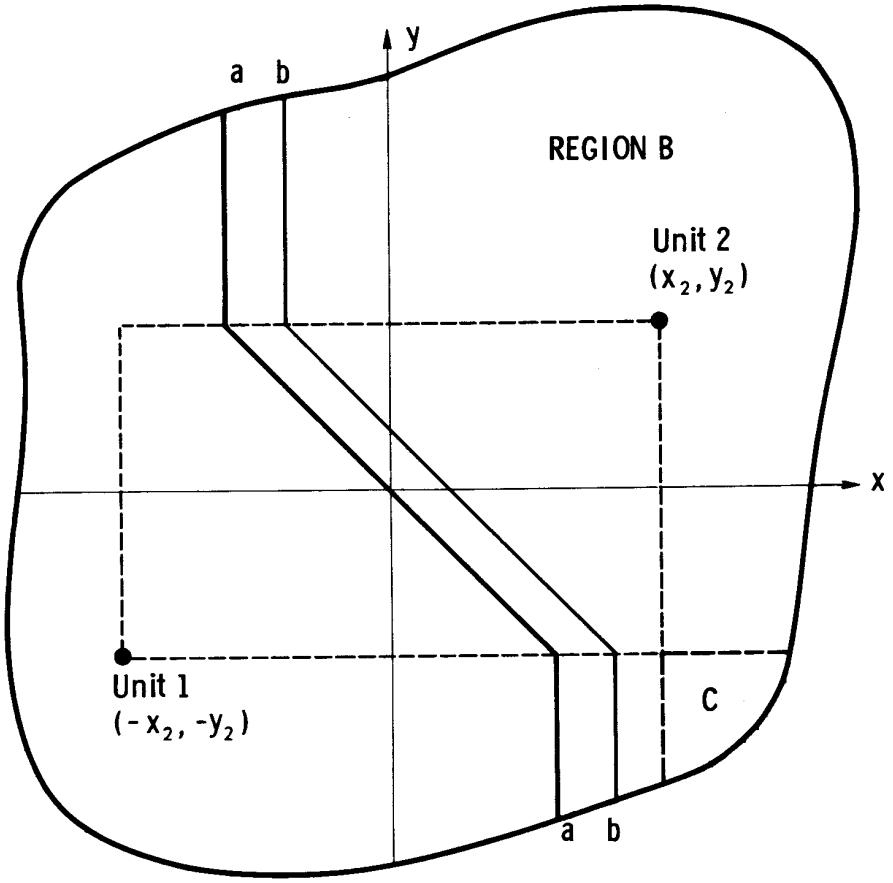
$$x = d \cdot \frac{\lambda}{\lambda+\mu} \cdot \frac{b}{\ell}.$$

This lies to the right of the y-axis, and to the left of the location of unit 2.

Finally, we consider one additional example which illustrates some of the complications which arise in the more general problem which we discuss in the following sections. The configuration for this example is shown in Figure 5. The region B is no longer a rectangle, and the line joining the locations of the two units no longer runs parallel to the streets. But we are free to choose a coordinate system parallel to the streets in such a way that the units are symmetrically located with respect to the origin. Assuming, for simplicity, that the x-direction speed v_1 equals the y-direction speed v_2, we find that the line consisting of points "equidistant" from the locations of the two units is no longer straight, but in fact is the line **aa** shown in Figure 5.

Figure 5

UNITS NOT ON THE SAME STREET



Now when we consider the possibility of changing the dividing line to reduce average response time, we see that we can modify the shape of the line as well as translate it. (Of course, this possibility also existed in the previous examples, but it seemed "natural" to look only at vertical dividing lines.) In Section V, we prove a theorem which says, roughly speaking, that for this example the minimum average response time will be obtained by selecting the dividing line between the response areas for the two vehicles to lie somewhere in the set

$$D = \left\{\underline{x}\in B:\ t_1(\underline{x}) = t_2(\underline{x}) + \frac{\lambda}{\lambda+\mu}\,(T_1 - T_2)\right\},$$

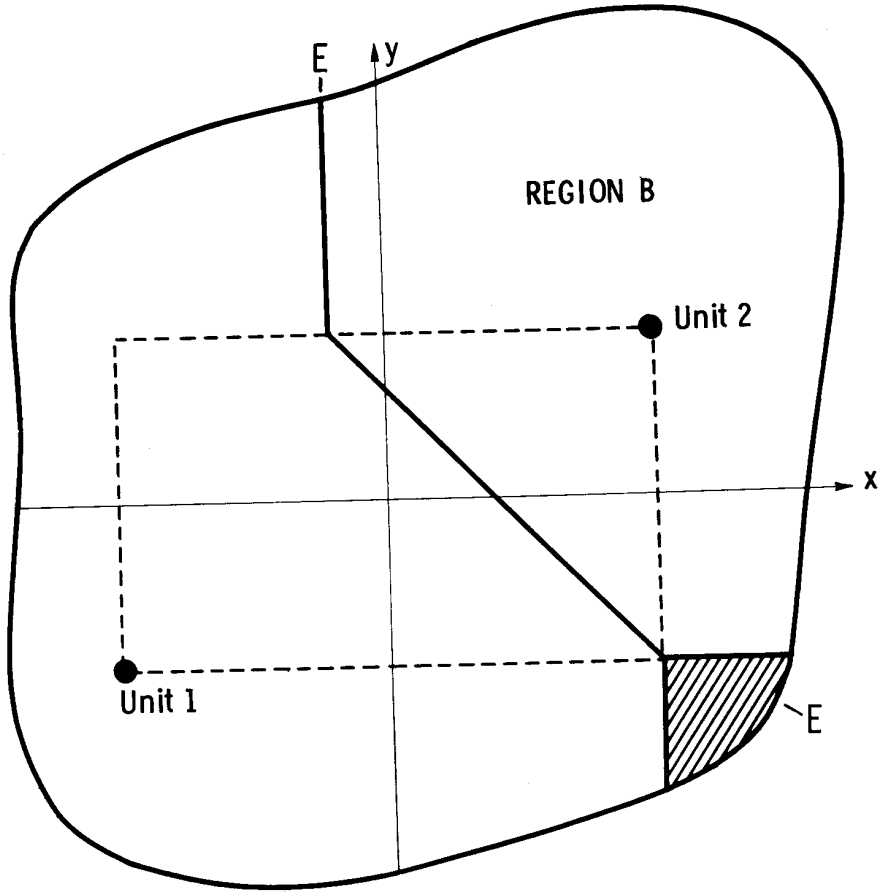
where t_1 and t_2 are the response-time functions of equations (2.1) and (2.2) and T_i is as defined after equation (2.3). Thus, if time is measured in minutes, D consists of all points which are $\lambda(T_1-T_2)/(\lambda+\mu)$ minutes "closer" to unit 2 than to unit 1.

Usually D is a dividing line itself. For instance, if $T_1 = T_2$, then D is the line aa in Figure 5. If T_1-T_2 is a small positive number, then D is the line bb, which is actually a translate of aa by a distance $\frac{1}{2}v\ \lambda(T_1-T_2)/(\lambda+\mu)$, where v is the common speed in the x- and y-directions. In these cases there is no ambiguity about where to place the dividing line.

However, for some values of T_1-T_2, strange things can happen. Notice, for instance, that all the points $\underline{x}$ in the region C of Figure 5 have the property that $t_1(\underline{x}) = t_2(\underline{x}) + 2(x_2-y_2)/v$. Thus, if $\lambda(T_1-T_2)/(\lambda+\mu) = 2(x_2-y_2)/v$, the set D coincides with the region E indicated by heavy lines and shading in Figure 6, and the dividing line can lie anywhere in it. Our Theorem 1, in Section V, shows that all such dividing lines will yield the same minimum average response time. (This result is not at all obvious, since if the set D is not the region E on Figure 6, for instance if it is the line bb on Figure 5, then different lines through E would have different average response times.) This

Figure 6

AMBIGUOUS LOCATION OF DIVIDING LINE



flexibility in selecting the dividing line allows us to consider other factors, such as convenience or workload balance between the units, in making the final selection. For values of T_1-T_2 larger than the one just considered, the dividing line will once again be unique, but it will not be a translate of aa.

Our treatment of the two-unit problem in the sections which follow will allow even more general geographical arrangements than those already described. For instance, we shall consider travel times other than the right-angle times of equations (2.1) and (2.2), and we shall permit the alarm rates to be distributed over B in a fairly complicated way. But most of the analytical problems which arise are relevant even for the simple examples in this section.

III. QUEUEING ASPECTS

We now begin our discussion of the general problem of designing response areas for two units serving a region B, with exactly one unit sent to each alarm. The region B is permitted to be any subset* of the plane. If A is any subset of B, we assume that calls arrive in A according to a Poisson process with parameter $\lambda(A)$, which means that the probability of n calls arriving in the region A during a time period of length T is $(\lambda(A)T)^n \exp(-\lambda(A)T)/n!$. We also assume that arrivals in any two disjoint subsets are independent.

Since the sum of any number of independent Poisson processes is a Poisson process with the sum rate, λ is a measure on B. The total alarm rate $\lambda(B)$ is assumed finite (and not zero), but otherwise λ can be arbitrary for present purposes. For instance, it might be conceptually desirable to think of the calls as occurring only on the streets, in which case λ might be concentrated on a number of intersecting straight lines in B.

The system consisting of two units located in B and responding as described in Section II can be reformulated in queueing terms. Suppose that some region A (a subset of B) has been chosen as the response area for unit 1. We can think of calls arriving from A as "Type 1 customers" for our queue, and calls arriving from the remainder of B, namely B-A, as "Type 2 customers". Arrivals of Type 1 customers constitute a Poisson process with rate $\lambda_1 = \lambda(A)$, and arrivals of Type 2 customers are a Poisson process with rate $\lambda_2 = \lambda(B)-\lambda(A)$.

We then have a two-server queue with two kinds of customers, and the service discipline is as follows:

(a) If a Type j customer arrives when both servers are available, it is served by unit j, j=1,2.

---

*For mathematical precision, we assume that all sets (or "areas" or "regions") mentioned in this paper are Borel sets, all functions are Borel-measurable, and all measures are Borel measures.

(b) If a Type j customer arrives when exactly one server is available, it is served by the available unit, j=1,2.

(c) If a Type j customer arrives when neither server is available, the customer is lost.*

The above description is simply a rephrasing of conditions (1), (2), and (3) of Section II. In order to determine the state probabilities for the system, we make the following additional assumptions:

(d) The service times for all customers are identically distributed with a finite average $1/\mu$, independent of the history or the state of the system at arrival, the type of the customer, and the identity of the serving unit.

(e) The system is in steady state.

Because "servicing" a call involves responding to the location of the incident, performing some work at the scene, and returning home, assumption (d) is realistic only if the difference in travel times for the two units is a negligible fraction of the total service time, and if the length of time it takes to serve an incident does not depend on how long it takes for the serving unit to arrive. For the vast majority of minor fires and police emergencies this is approximately the case, but we cannot pretend that this model is adequate for all purposes. In Section VII we discuss certain features of a model in which assumption (d) is relaxed.

One can classify the states of this system in a number of different ways. For the purposes at hand (locating the optimal response areas and measuring workloads) we need only consider the states

00 = both units available

01 = unit 1 available, unit 2 busy servicing a call

10 = unit 1 busy servicing a call, unit 2 available

11 = both units busy.

Later we shall consider alternative collections of states for the system.

If the service times happen to be exponentially distributed, our system is a continuous-time Markov process with states as shown in

*In reality, we have in mind that the customer is served from another location, outside the system under consideration.

Figure 7. The numbers on the arrows are transition rates. Thus, the equations of detailed balance* for P_{ij}, the steady-state probability of state ij, are as follows:

$$\lambda P_{00} = \mu(P_{10} + P_{01})$$

$$(\lambda+\mu)P_{10} = \lambda_1 P_{00} + \mu P_{11}$$

$$(\lambda+\mu)P_{01} = \lambda_2 P_{00} + \mu P_{11}$$

$$2\mu\, P_{11} = \lambda(P_{10} + P_{01}).$$

The solution which has total probability 1 is

$$(3.1) \qquad P_{00} = 1/(1+\rho+\rho^2/2)$$

$$P_{01} = P_{00}\,(\rho_2+\rho^2/2)/(1+\rho)$$

$$P_{10} = P_{00}\,(\rho_1+\rho^2/2)/(1+\rho)$$

$$P_{11} = P_{00}\,\rho^2/2,$$

where $\rho=\lambda(B)/\mu$, $\rho_1=\lambda_1/\mu$, and $\rho_2=\lambda_2/\mu$.

These values for the steady-state probabilities are actually valid for an arbitrary service-time distribution with mean $1/\mu$, as we show in Reference 3.

We define the workload of unit j to be the steady-state probability that unit j is servicing a call, and we denote it by W_j. Thus $W_1 = P_{10} + P_{11}$, and $W_2 = P_{01} + P_{11}$, so we have

$$(3.2) \qquad W_j = P_{00}\,(\rho_j + \rho^2 + \rho^3/2)/(1+\rho),\ j=1,2.$$

The difference in workload is thus

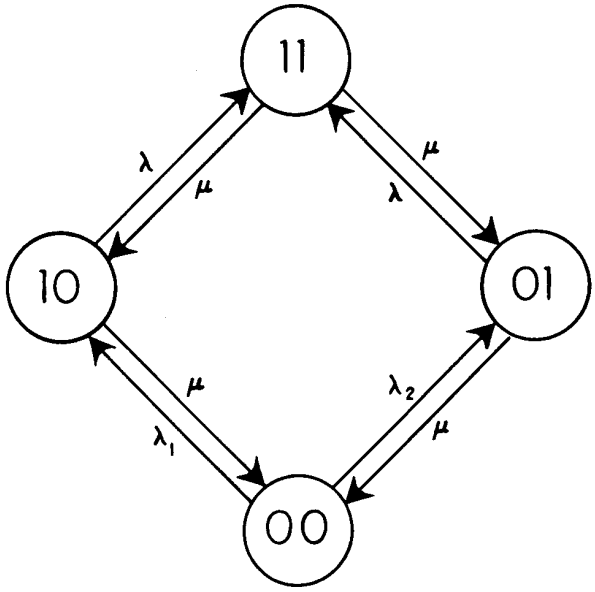
$$(3.3) \qquad \Delta W = \left|W_1-W_2\right| = P_{00}\,\left|\rho_1-\rho_2\right|/(1+\rho).$$

---

*See, for example, Reference 12.

Figure 7

STATES AND TRANSITION RATES



The quantity ΔW is intended as a measure of the extent to which one unit works harder than the other. Under the assumptions we are using, W_j is proportional to the fraction of all calls served by unit j, and is thus also proportional to the total number of hours which unit j spends serving calls during a given time period. Thus, many reasonable measures of workload imbalance are proportional to the one we are using. In more complicated cases (for instance if different types of calls are assumed to have different service times) a different definition of workload imbalance may be desirable.

IV. AVERAGE RESPONSE TIME

The state probabilities calculated in the preceding section can be used to determine the average response time to calls for service. Suppose we denote by $T_j(C)$ the average time it takes to respond from the location of unit j to a call for service arising in a region C. Then, for example, $T_j(B)$ is the average response time if every arriving call is answered from location j.

Since unit 1 responds to all calls if unit 2 is unavailable, $T_1(B)$ represents the average response time for calls arriving when the state is 01. Similarly $T_2(B)$ is the average response time for calls arriving when the state is 10.

If a region A is selected as the response area of unit 1, then for alarms arriving when the state is 00, a fraction $\lambda(A)/\lambda(B)$ will occur in A and have average response time $T_1(A)$, while a fraction $\lambda(B-A)/\lambda(B)$ will occur in B-A and have average response time $T_2(B-A)$. Thus the average response time for calls arriving when the state is 00 will be

$$T_1(A)\cdot\lambda(A)/\lambda(B) + T_2(B-A)\cdot\lambda(B-A)/\lambda(B).$$

Putting together the observations in the previous two paragraphs, we see that the average response time to all alarms, given that A is the response area of unit 1, is*

$$(4.1)\qquad \overline{T}(A) = P_{00}\left[T_1(A)\cdot\lambda(A) + T_2(B-A)\cdot\lambda(B-A)\right]\Big/\lambda(B) + P_{01}\cdot T_1(B) + P_{10}\cdot T_2(B) + P_{11}\cdot\tau,$$

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

---

*Here we use the fact that a fraction P_{ij} of arriving alarms find the system in state ij. See Strauch, [15].

In order to use this formula for $\overline{T}(A)$, we have to be able to calculate $T_j(A)$ and $T_j(B)$, $j=1,2$. To do this, we introduce functions t_1 and t_2 which are analogous to the right-angle travel-time functions of Section II.

Conceptually, $t_j(x,y)$ is the mean time it takes for unit j to travel to the point (x,y) in B from its home location, using the fastest possible route.* Mathematically, t_j can be any bounded non-negative function on B which is integrable with respect to the measure λ.

Although it may seem unnecessary to allow such general travel-time functions, it does no harm. Moreover one can easily imagine realistic examples where the travel-time functions are not continuous. This happens, for instance, if the region B has one-way streets or obstructions such as railroad crossings.

It is not even necessary to define the travel time to places where incidents cannot occur. This may be desirable, for example, if there is a lake in the region, or if we assume that the units can only respond to places on the streets, and the alarm-rate measure λ is concentrated on the streets.

Now, to determine the average response time from location j in a region C, we simply have to weight the travel time $t_j(\underline{x})$ by the conditional probability that an alarm in C occurs at $\underline{x}$ and integrate over all values of $\underline{x}$. Thus

$$T_j(C) = \int_C t_j(\underline{x})\, d\lambda(\underline{x})/\lambda(C).$$

Entering this in equation (4.1) we find

$$\overline{T}(A) = \frac{P_{00}}{\lambda(B)} \left(\int_A t_1(\underline{x})\, d\lambda(\underline{x}) + \int_{B-A} t_2(\underline{x})\, d\lambda(\underline{x}) \right) \tag{4.2}$$

$$+ \frac{P_{01}}{\lambda(B)} \cdot \int_B t_1(\underline{x})\, d\lambda(\underline{x}) + \frac{P_{10}}{\lambda(B)} \int_B t_2(\underline{x})\, d\lambda(\underline{x})$$

$$+ P_{11}\tau \,,$$

*The travel time from location j to (x,y) is a random variable, which we assume to be independent of the history or current state of our system; $t_j(x,y)$ is the expected value of this random variable.

or, simplifying,

$$(4.3) \qquad \bar{T}(A) = \frac{P_{00}}{\lambda(B)} \int_A (t_1 - t_2)\, d\lambda + P_{01}\, T_1 + (P_{10} + P_{00})\, T_2 + P_{11}\tau,$$

where

$$(4.4) \qquad T_j = T_j(B) = \int_B t_j(\underline{x})\, d\lambda(\underline{x}) \Big/ \lambda(B).$$

Either (4.2) or (4.3) can be used to calculate $\bar{T}(A)$ for particular choices of A. However, for the purpose of finding the response areas which minimize average response time, it will be convenient to rewrite $\bar{T}(A)$. Using the equations (3.1) for the P_{ij}, and the fact that $\rho_1 = \lambda(A)/\mu$, $\rho_2 = \rho - \rho_1$, we find

$$(4.5) \qquad \bar{T}(A) = \frac{P_{00}}{\lambda(B)} \left(\int_A (t_1 - t_2)\, d\lambda - \frac{\rho}{\rho+1} \lambda(A)(T_1 - T_2) \right) + \alpha,$$

where

$$(4.6) \qquad \alpha = P_{00} \left(\frac{\rho+\rho^2/2}{1+\rho} T_1 + \frac{1+\rho+\rho^2/2}{1+\rho} T_2 + \rho^2\tau/2 \right),$$

and α is independent of A.

Setting

$$(4.7) \qquad s_0 = \frac{\rho}{\rho+1} (T_1 - T_2)$$

we can rewrite (4.5) as

$$(4.8) \qquad \bar{T}(A) = \frac{P_{00}}{\lambda(B)} \int_A (t_1 - t_2 - s_0)\, d\lambda + \alpha.$$

This form for the average response time, given A as the response area for unit 1, is fundamental for finding the optimal response areas. The form of equation (4.8) suggests that points $\underline{x}$ such that

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

will play an important role in the sequel.

It should be noted that since the functions t_j are arbitrary, they do not have to be interpreted as response times. Thus $t_j(\underline{x})$ can be thought of as the utility of responding from location j to $\underline{x}$. For example, the importance of responding rapidly to certain points may be high, whereas for others it is low; in this case $t_j(\underline{x})$ could be a weighted response time. The function $\overline{T}(A)$ then becomes an average utility function. However, to simplify the terminology, we shall continue to call $\overline{T}(A)$ the "average response time".

V. RESPONSE AREAS WITH MINIMUM AVERAGE RESPONSE TIME

To summarize, we have a system described by a finite alarm rate measure λ, two bounded response-time functions t_1 and t_2, and the queueing discipline (a)-(e) of Section III. In this section we describe the response areas for unit 1 which yield the minimum average response time, and we prove that our description is correct.

In the remainder of this section, we let

$$s_0 = \rho(T_1 - T_2)/(1+\rho) \tag{5.1}$$

where $\rho = \lambda(B)/\mu$, and T_j is defined by equation (4.4).

THEOREM 1

1. Choosing the set

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

as the response area for unit 1 (and B-X as the response area for unit 2) yields the minimum average response time. More precisely, $\overline{T}(A) \geq \overline{T}(X)$ for any subset **A** of B.

2. For any set A lying between the set X and the set

$$Y = \left\{\underline{x}\epsilon B: \ t_1(\underline{x}) - t_2(\underline{x}) \leq s_0\right\}$$

(i.e. $X \subseteq A \subseteq Y$), we have $\overline{T}(A) = \overline{T}(X)$. That is, A also gives minimum average response time.

3. For any A such that $\lambda(X-A)>0$ or $\lambda(A-Y)>0$, we have $\overline{T}(A) > \overline{T}(X)$. That is, if A differs substantively from the sets already described as yielding minimum average response time, then A does not yield the minimum.

Proof: Let A be any subset of B.
By equation (4.8) we have

$$\overline{T}(A) - \overline{T}(X) = \frac{P_{00}}{\lambda(B)} \left(\int_A (t_1-t_2-s_0)\, d\lambda - \int_X (t_1-t_2-s_0)\, d\lambda \right)$$

or

$$\overline{T}(A) - \overline{T}(X) = \frac{P_{00}}{\lambda(B)} \left(\int_{A-X} (t_1-t_2-s_0)\, d\lambda - \int_{X-A} (t_1-t_2-s_0)\, d\lambda \right), \tag{5.2}$$

where A-X denotes all points in A which are not in X.

Now if $\underline{x}$ is in A but not X, then the definition of X tells us that $t_1(\underline{x}) - t_2(\underline{x}) \geq s_0$. Hence,

$$\int_{A-X} (t_1-t_2-s_0)\, d\lambda \geq 0. \tag{5.3}$$

Similarly, if $\underline{x}$ is in X but not A, we have $t_1(\underline{x}) - t_2(\underline{x}) < s_0$, so

$$\int_{X-A} (t_1-t_2-s_0)\, d\lambda \leq 0. \tag{5.4}$$

[The equality can hold if and only if $\lambda(X-A)=0$.]

Entering (5.3) and (5.4) in (5.2) shows that

$$\overline{T}(A) - \overline{T}(X) \geq 0. \tag{5.5}$$

This proves part 1 of the theorem.

To prove part 2, suppose $X \subseteq A \subseteq Y$. Then X-A is empty, and for every point $\underline{x}$ in A-X we have

$$t_1(\underline{x}) - t_2(\underline{x}) = s_0.$$

Hence (5.2) becomes

$$\overline{T}(A) - \overline{T}(X) = \frac{P_{00}}{\lambda(B)} \int_{A-X} (s_0 - s_0)\, d\lambda = 0.$$

This proves part 2.

As for part 3, we have already observed that if $\lambda(X-A)>0$, then strict inequality holds in (5.4), and hence, since (5.3) is still true, strict inequality holds in (5.5). This proves that $\overline{T}(A) > \overline{T}(X)$ if $\lambda(X-A)>0$.

The other possibility in part 3 is that $\lambda(A-Y)>0$. In this case, since any point $\underline{x}$ which is in A but not Y satisfies

$$t_1(\underline{x}) - t_2(\underline{x}) > s_0,$$

the integrand in (5.3) is positive on a set of positive measure, and thus

$$\int_{A-X} (t_1 - t_2 - s_0)\, d\lambda > 0.$$

This also shows, using (5.4) and (5.2), that $\overline{T}(A) > \overline{T}(X)$. This completes the proof of Theorem 1.

To see the relation of the theorem to the examples in Section II, refer back to Figure 6. If the shaded area and heavy lines together constitute the set

$$E = \left\{\underline{x}\epsilon B: \quad t_1(\underline{x}) - t_2(\underline{x}) = s_0\right\} ,$$

then X is the set to the left of E. The theorem says that any optimal response area for unit 1 must contain all the points to the left of the shaded area and heavy lines, and <u>may</u> contain some points in E. Speaking loosely, this means that the dividing line between the response areas of the two units must lie within the set indicated in Figure 6.

In practical cases it may not be easy to calculate s_0, and so the theorem is applied by noting that an optimal response area for unit 1 is one of the sets

$$X(s) = \left\{\underline{x}\epsilon B: \quad t_1(\underline{x}) - t_2(\underline{x}) < s\right\} .$$

It may be possible to describe this whole family of sets and, by trial and error, select the best [10].

Another reason for considering this family of sets is that the response area which gives the minimum average response time may be undesirable for other reasons, such as workload balance. We therefore prove some additional properties of this family, which are of interest in themselves.

<u>THEOREM 2</u>

Let g(s) denote the average response time if X(s) is chosen as the response area of unit 1, i.e.

$$g(s) = \overline{T}(X(s)).$$

Then g(s) increases monotonically as s moves away from s_0 in either direction. More precisely:

1. If $s_2 > s_1 > s_0$ then $g(s_2) \geq g(s_1) \geq g(s_0)$. Moreover, if $\lambda(X(s_2)) > \lambda(X(s_1))$, then $g(s_2) > g(s_1)$.

2. If $s'' < s' < s_0$, then $g(s'') \geq g(s') \geq g(s_0)$. Moreover, if $\lambda(X(s'')) < \lambda(X(s'))$, then $g(s'') > g(s')$.

Proof: We already know from Theorem 1 that $g(s_0)$ is the minimum value for g, so we just have to prove the inequalities about $g(s_1)$ vs. $g(s_2)$ and $g(s')$ vs. $g(s'')$.

Since $s_1 < s_2$, the definition of X(s) shows that $X(s_1) \subseteq X(s_2)$. Thus, using (5.2) with A replaced by $X(s_2)$ and X replaced by $X(s_1)$, we have

$$g(s_2) - g(s_1) = \frac{P_{00}}{\lambda(B)} \int_{X(s_2) - X(s_1)} (t_1 - t_2 - s_0)\, d\lambda.$$

The integrand is positive everywhere on $X(s_2) - X(s_1)$. (In fact, it lies between $s_1 - s_0$ and $s_2 - s_0$.) Thus $g(s_2) - g(s_1) \geq 0$, and if $\lambda(X(s_2) - X(s_1)) > 0$ we have $g(s_2) - g(s_1) > 0$. The condition $\lambda(X(s_2) - X(s_1)) > 0$ is the same as $\lambda(X(s_2)) > \lambda(X(s_1))$, so part 1 is proved. Part 2 is proved similarly.

Our final theorem in this section states that you can decide in advance what workload balance you would like to have between the units (for instance, equal workload), and the optimal response area will still be very close to one member of the family X(s). The problem is that there may not be any one of the sets X(s) having exactly the desired workload balance. For example, if all the alarms are concentrated at a single point, then each $\lambda(X(s))$ is either 0 or $\lambda(B)$. Hence the result is somewhat more complicated than the previous ones.

Referring to equation (3.2) and noting that $\rho_2 = \rho - \rho_1$, it is apparent that fixing the workload balance is the same as fixing ρ_1. Thus, all potential response areas for unit 1 which have the same workload balance must have the same total alarm rate. This leads us to the following statement of the theorem.

THEOREM 3

Fix the number λ_1, $0 \leq \lambda_1 \leq \lambda(B)$. If it is possible to find a set A_1 having the property that $\lambda(A_1) = \lambda_1$ and $X(s) \subseteq A_1 \subseteq Y(s)$ for some s, then

$$\overline{T}(A_1) \leq \overline{T}(A) \text{ for all A having } \lambda(A) = \lambda_1.$$

Moreover, if $\lambda(A - Y(s)) > 0$ or $\lambda(X(s) - A) > 0$, then $\overline{T}(A_1) < \overline{T}(A)$. Here

$$(5.6) \quad X(s) = \left\{\underline{x}\epsilon B: \ t_1(\underline{x}) - t_2(\underline{x}) < s\right\}$$

$$Y(s) = \left\{\underline{x}\epsilon B: \ t_1(\underline{x}) - t_2(\underline{x}) \leq s\right\}$$

Remark: A condition on λ which guarantees the existence of A_1 is the following:

Non-singularity condition: for any two sets C and C_1 such that $C \subset C_1$ and any number λ' such that $\lambda(C) \leq \lambda' \leq \lambda(C_1)$, it is possible to find a set C' such that $C \subseteq C' \subseteq C_1$ and $\lambda(C') = \lambda'$.

Proof of theorem: Suppose one can find a set A_1 and a number s such that

$$\lambda(A_1) = \lambda_1, \quad X(s) \subseteq A_1 \subseteq Y(s).$$

If A is any other set with $\lambda(A) = \lambda_1$, we have, by (5.2),

$$\overline{T}(A) - \overline{T}(A_1) = \frac{P_{00}}{\lambda(B)} \left\{ \int_{A-A_1} (t_1 - t_2 - s_0)\, d\lambda - \int_{A_1-A} (t_1 - t_2 - s_0)\, d\lambda \right\}.$$

The first integral is $\geq (s-s_0) \cdot \lambda(A-A_1)$, and the second integral is $\leq (s-s_0) \cdot \lambda(A_1-A)$. Since $\lambda(A) = \lambda(A_1)$, we have $\lambda(A-A_1) = \lambda(A_1-A)$. Hence $\overline{T}(A) - \overline{T}(A_1) \geq 0$. This proves the first statement of the theorem.

If $\lambda(A-Y(s)) > 0$, then the first integrand is actually $>(s-s_0)$ on a set of positive measure, and hence the integral is greater than $(s-s_0)\ \lambda(A-A_1)$. Similarly, if $\lambda(X(s)-A) > 0$, the second integral is actually $<(s-s_0)\ \lambda(A_1-A)$. In either case, we find that $\overline{T}(A) - \overline{T}(A_1) > 0$. This proves the second statement.

Proof of Remark: Consider the function $u(s) = \lambda(X(s))$. This function u is monotonically non-decreasing and takes values between 0 and $\lambda(B)$, including the endpoints. It is continuous from the left, since

$$X(s) = \bigcup_{s'<s} X(s').$$

Moreover, $\lim_{\varepsilon\to 0+} u(s+\varepsilon) = u(s+) = \lambda(Y(s))$, since

$$Y(s) = \bigcap_{s'>s} X(s').$$

It follows that, given λ_1 between 0 and $\lambda(B)$, there must be some number s such that $u(s) \leq \lambda_1 \leq u(s+)$. Thus, for this s, we have

$$\lambda(X(s)) \leq \lambda_1 \leq \lambda(Y(s)).$$

The non-singularity condition then guarantees that a set A_1 exists with $\lambda(A_1) = \lambda_1$ and $X(s) \subseteq A_1 \subseteq Y(s)$.

VI. DOMINANCE

We say that a response area A for unit 1 dominates another response area A' if either A yields lower average response time than A', with workload balance at least at good, or A yields better workload balance than A', with average response time no higher. Referring back to equation (3.3) we see that the difference in workload between the two units, if A is the response area for unit 1, can be written

$$\Delta W(A) = P_{00} \left| \frac{\lambda(A)}{\mu} - \frac{\lambda(B) - \lambda(A)}{\mu} \right| \Big/ (1+\rho)$$

or

$$(6.1) \qquad \Delta W(A) = P_{00} \left| 2\lambda(A) - \lambda(B) \right| \Big/ (\mu+\lambda(B)).$$

This depends only on $\lambda(A)$. It decreases as $\lambda(A)$ increases from 0 to $\lambda(B)/2$, and then it increases.

Using this definition of ΔW, and the definition of $\overline{T}$ given in (4.5), we can give a mathematical description of dominance as follows:

> Response area A for unit 1 dominates response area A' if and only if
>
> $$\overline{T}(A) \leq \overline{T}(A')$$
>
> and
>
> $$\Delta W(A) \leq \Delta W(A'),$$
>
> with at least one of the inequalities strict.

Example

We illustrate the concept of dominance by considering an example whose geographical arrangement is the one shown in Figure 1.

For this example, we assume right-angle travel times [equations (2.1) and (2.2)], and the parameters have been set as shown in Table 1.

Table 1

VALUES OF PARAMETERS FOR EXAMPLE (FIGURE 8)

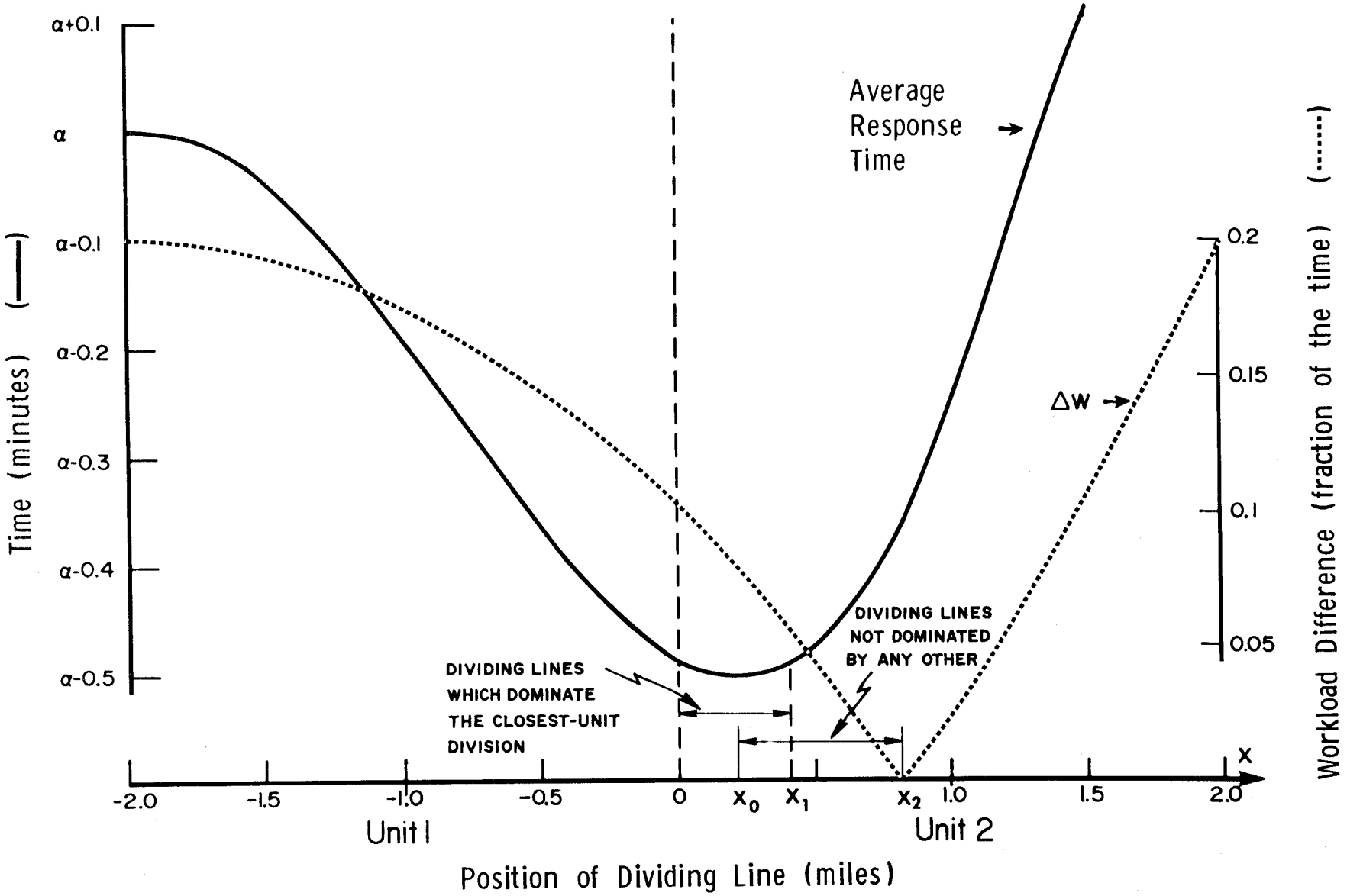
Parameter	Symbol	Value
Total alarm rate	$\lambda(B)$	4 incidents/hour
Average service time	$1/\mu$	15 minutes
Response speeds	$v_1 = v_2$	20 miles/hour
Distance of unit from center	d	1 mile
Length of rectangle	ℓ	4 miles
Height of rectangle	h	arbitrary
Alarm rate in region A	$\lambda(A)$	$\int_A (x+2)\, dxdy/4h$

The alarm rate is assumed to be uniform in the y-direction and to increase linearly with the distance from the left side of the rectangle.

We consider possible response areas for unit 1 which are bounded on the right by a vertical dividing line positioned at x. The results for the average response time [equation (4.5)] and the workload difference [equation (6.1)] are shown in Figure 8 for each value of x. The response time is not fully determined, because we have not specified the height of the rectangle B or the response time τ for incidents served from outside B. Therefore response times are shown relative to the constant α of equation (4.7).

One can calculate, for this example, that $s_0 = 11/480$ hours, so the minimum value of the response time occurs for the value of x equal to $x_0 = 11/48$ miles, as indicated on Figure 8.

FIG. 8: EXAMPLE SHOWING RESPONSE TIME AND WORKLOAD BALANCE
VS
POSITION OF DIVIDING LINE



Let us first find all dividing lines whose response areas for unit 1 dominate the closest-unit division. The closest-unit division corresponds to the line x=0. As the dividing line moves to the right from x=0, both the average response time and the workload difference decrease until we get to x_0. Moving to the right from here, the workload balance continues to decrease, but the response time rises until, at the point $x_1 \doteq 0.444$ miles, the response time returns to the value it has at x=0. Thus any x such that $0 < x < x_1$ gives a division of B which dominates the closest-unit division. No other values of x dominate the closest-unit division, because for them the response time is higher.

However, even some divisions which do not dominate the closest-unit division might be desirable under some circumstances. An interesting class of divisions consists of those which are <u>not dominated by any other division</u>. Of course, the division with minimum average response time has this property, but there are others as well. In fact, noticing that the workload difference becomes zero when $x = x_2 = \sqrt{8} - 2$ miles, we see that any x satisfying $x_0 \leq x \leq x_2$ yields a division which is not dominated by any other division. (In fact, for any such x, if $x' < x$, then the x'-division yields bigger workload difference, and if $x' > x$, then the x'-division yields higher response time. Theorem 4, below, gives further details.)

None of the dividing lines between x_0 and x_2 can be said to be "better" than another one of them unless we specify how important workload balance is, in comparison with average response time. For example, it might appear from Figure 8 that selecting the extreme value $x = x_2$ would not be desirable. But if we look at the numerical values of $\overline{T}$ and ΔW, we see that selecting $x = x_2$ sacrifices less than 10 seconds in average response time, compared to the lowest possible average response time, but reduces the workload of unit 2 nearly 10%. Under

some circumstances, this might be desirable. The fact that it might be acceptable to move the dividing line more than 0.8 miles from the closest-unit line can be attributed to the extreme variation in alarm rate which we selected for this example. In practical cases, all the "admissible" dividing lines are only a few blocks away from the closest-unit dividing line.

General case

We return now to the general case where B is any region, λ is any finite measure on B, and t_1 and t_2 are arbitrary bounded response time functions. To simplify the statements we shall make about dominance, we add the following assumptions in the remainder of this section:

(1) λ is non-singular, in the sense of Theorem 3, Remark.

(2) As s increases, the alarm rate in X(s) also increases. More precisely: if $s < s'$, and $\lambda(X(s)) < \lambda(B)$ and $\lambda(X(s')) > 0$, then $\lambda(X(s)) < \lambda(X(s'))$. (This implies also that $\lambda(Y(s)) < \lambda(X(s'))$.)

Here X(s) and Y(s) are given by (5.6).

These two assumptions will be true for most examples one might construct. The first always holds if the alarm rate is given by a planar density or a linear density along the streets, but it will fail if all the alarms are associated with a finite number of points, for instance fire alarm boxes. The second assumption holds if the units do not pass through regions where no alarms ever occur. For example, using right-angle travel times in Figure 1, the assumption (2) is equivalent to requiring that every vertical strip (of non-zero width) can have alarms in it.

We shall first determine all response areas which are not dominated by any response area. This result will give us a direct method for determining whether the closest-unit division can be dominated by some other choice of response areas.

Before proceeding, we observe that two response areas for unit 1 which differ by a set on which no alarms occur (with probability 1) will have identical average response times and workload. Hence we say that two response areas A and A' are equivalent if and only if $\lambda(A-A') = 0$ and $\lambda(A'-A) = 0$.

THEOREM 4

Assume that the alarm-rate measure λ satisfies conditions (1) and (2) above, s_0 is given by (5.1), and A is a subset of B.

(a) Suppose $\lambda(Y(s_0)) \leq \lambda(B)/2$. A is not dominated by any response area if and only if A is equivalent to a set C satisfying

(6.2) $C \supseteq Y(s_0)$,

$\lambda(C) \leq \lambda(B)/2$,

and $X(s) \subseteq C \subseteq Y(s)$ for some s.

(b) Suppose $\lambda(X(s_0)) \geq \lambda(B)/2$. A is not dominated by any response area if and only if it is equivalent to a C satisfying

(6.3) $C \subseteq X(s_0)$,

$\lambda(C) \geq \lambda(B)/2$,

and $X(s) \subseteq C \subseteq Y(s)$ for some s.

(c) Suppose $\lambda(X(s_0)) < \lambda(B)/2 < \lambda(Y(s_0))$. A is not dominated by any response area if and only if A is equivalent to a set C satisfying $X(s_0) \subseteq C \subseteq Y(s_0)$ and $\lambda(C) = \lambda(B)/2$. (In fact, these sets C dominate any set not equivalent to one of them.)

Remark: In essence, the theorem states that a set has to have two properties in order not to be dominated by any other set: (1) it must lie between some X(s) and the corresponding Y(s), and (2) it must lie between the "minimum response time" area and the "equal workload" area. Complications arise from the fact that X(s) may differ substantially from Y(s), so that two sets which lie between the same X(s) and Y(s) have to be carefully compared to see if one dominates the other.

Proof of theorem: We shall only consider case (a), since the proof for case (b) is similar, and case (c) is almost obvious.

Since equivalent sets have the same $\overline{T}$ and ΔW, we may as well assume that A itself satisfies the conditions $A \supseteq Y(s_0)$, $\lambda(A) \leq \lambda(B)/2$, and $X(s) \subseteq A \subseteq Y(s)$ for some s, and prove from this that no response area dominates A. Note that we must have $s \geq s_0$.

Suppose A' is any region whatsoever. We wish to show that A' does not dominate A. By Theorem 3, there is a set A_1 such that $\lambda(A_1) = \lambda(A')$, $X(s') \subseteq A_1 \subseteq Y(s')$ for some s', and $\overline{T}(A_1) \leq \overline{T}(A')$.

1. If $s' < s$, then

$$\lambda(A') = \lambda(A_1) \leq \lambda(Y(s')) < \lambda(X(s)) \leq \lambda(A) \leq \lambda(B)/2.$$

It follows from (6.1) that $\Delta W(A') > \Delta W(A)$. Hence, in this case, A' does not dominate A.

2. If $s < s'$, it is easy to see that $\overline{T}(A_1) > \overline{T}(A)$, using the fact that $s \geq s_0$ and following the method used in Theorem 3. Thus $\overline{T}(A') \geq \overline{T}(A_1) > \overline{T}(A)$, and so A' does not dominate A.

3. Assume $s = s'$. Then, by (5.2), we have

$$(6.4) \qquad \overline{T}(A_1) - \overline{T}(A) = \frac{P_{00}}{\lambda(B)} (s-s_0) \left[\lambda(A_1-A) - \lambda(A-A_1)\right],$$

$$= \frac{P_{00}}{\lambda(B)} (s-s_0) \left[\lambda(A_1) - \lambda(A)\right],$$

$$= \frac{P_{00}}{\lambda(B)} (s-s_0) \left[\lambda(A') - \lambda(A)\right].$$

The last line follows from the fact that $\lambda(A') = \lambda(A_1)$.

We must have either $\lambda(A') < \lambda(A)$, $\lambda(A') > \lambda(A)$, or $\lambda(A') = \lambda(A)$.

If $\lambda(A') < \lambda(A)$, then $\Delta W(A') > \Delta W(A)$, from (6.1) and the fact that $\lambda(A) \leq \lambda(B)/2$. Hence A' does not dominate A.

If $\lambda(A') > \lambda(A)$, then (6.4) shows that $\overline{T}(A') > \overline{T}(A)$ unless $s = s_0$. But it is impossible to have $s = s_0$ and at the same time have $\lambda(A') > \lambda(A)$, because from $A \supseteq Y(s_0)$ we know $\lambda(A) \geq \lambda(Y(s_0))$, and if $s = s' = s_0$ we have $A_1 \subseteq Y(s') = Y(s_0)$ so that $\lambda(A') = \lambda(A_1) \leq \lambda(Y(s_0))$. Hence, whenever $\lambda(A') > \lambda(A)$ we see that A' does not dominate A.

Finally, if $\lambda(A') = \lambda(A)$, then (6.1) and (6.4) together show that A and A' have the same ΔW and $\overline{T}$. Hence A' does not dominate A.

Since we have considered all possible relationships of s' to s, the sufficiency of the three conditions is proved.

To prove the necessity of the three conditions in (6.2), we take as hypothesis the assumption that our set A is not equivalent to any set C satisfying (6.2), and we show that it is possible to find a set which dominates A.

We apply Theorem 3 again to find a number s' and a set A_1 such that $X(s') \subseteq A_1 \subseteq Y(s')$, $\lambda(A_1) = \lambda(A)$, and $\overline{T}(A_1) \leq \overline{T}(A)$. We consider the three cases:

(i) $\lambda(A) < \lambda(Y(s_0))$, (ii) $\lambda(Y(s_0)) \leq \lambda(A) \leq \lambda(B)/2$, and (iii) $\lambda(A) > \lambda(B)/2$.

In case (i), we can show that $Y(s_0)$ dominates A. We already know that $Y(s_0)$ has minimum average response time, so $\overline{T}(Y(s_0)) \leq \overline{T}(A)$. Moreover, since $\lambda(A) < \lambda(Y(s_0)) \leq \lambda(B)/2$, we see, from (6.1), that $\Delta W(A) > \Delta W(Y(s_0))$. Thus $Y(s_0)$ dominates A.

In case (ii), we must have $\lambda(A-Y(s')) > 0$ or $\lambda(X(s')-A) > 0$, or both. (Otherwise A is equivalent to the set $C = A \cup X(s') \cap Y(s')$ which satisfies (6.2). The hypothesis forbids this.) By Theorem 3, $\overline{T}(A) > \overline{T}(A_1)$. Since $\Delta W(A_1) = \Delta W(A)$, we see that A_1 dominates A.

If (iii) holds, we have $\Delta W(A) > 0$. Moreover, by the non-singularity assumption for λ, we can find a set A* and a number s* such that $\lambda(A^*) = \lambda(B)/2$ and $X(s^*) \subseteq A^* \subseteq Y(s^*)$. Since $\Delta W(A^*) = 0$, we just have to show that $\overline{T}(A^*) \leq \overline{T}(A)$ to prove that A* dominates A.

Noting that $A_1 \supseteq A^*$, we obtain, from (5.2),

$$\overline{T}(A_1) - \overline{T}(A^*) = \frac{P_{00}}{\lambda(B)} \int_{A_1-A} (t_1-t_2-s_0)\, d\lambda.$$

Since $\lambda(Y(s_0)) \leq \lambda(B)/2$, we necessarily have $s^* \geq s_0$, which implies that the integrand is positive. Hence $\overline{T}(A_1) \geq \overline{T}(A^*)$. Since we already know $\overline{T}(A) \geq \overline{T}(A_1)$, we have proved that A* dominates A. This completes the proof.

COROLLARY

Under assumptions (1) and (2) above, it is possible to find a response area which dominates all non-equivalent response areas if and only if $\lambda(X(s_0)) = \lambda(B)/2$ or $\lambda(Y(s_0)) = \lambda(B)/2$ (or both). Here s_0 is given by equation (5.1).

Proof: If $\lambda(Y(s_0)) = \lambda(B)/2$, then the only set C which satisfies (6.2) is $Y(s_0)$ itself. Hence, by Theorem 4(a), the only sets which are not dominated by any others are those equivalent to $Y(s_0)$. Thus $Y(s_0)$ dominates all non-equivalent response areas. A similar argument shows that if $\lambda(X(s_0)) = \lambda(B)/2$, then $X(s_0)$ dominates all non-equivalent response areas.

If neither of these equalities holds, then there will be a number of inequivalent sets C satisfying the conditions in Theorem 4, and no set can dominate all of them. We omit the details of the proof of this statement.

Now we can apply the theorem to the case of the closest-unit division. As the example of Figure 6 suggests, the expression "closest-unit division" can be ambiguous, since the set of points which are equidistant (in the sense of travel time) from the two units could possibly form a large region. In general, a set A is called a closest-unit division if it lies between X(0) and Y(0) (i.e. $X(0) \subseteq A \subseteq Y(0)$).

THEOREM 5

Under assumptions (1) and (2) above, there exists a response area which dominates every closest-unit division if and only if

(i) $T_1 > T_2$ and $\lambda(Y(0)) < \lambda(B)/2$

or

(ii) $T_1 < T_2$ and $\lambda(X(0)) > \lambda(B)/2$.

Remarks: 1. This theorem can be roughly paraphrased as follows. Suppose that, with the closest-unit division, one unit works harder than the other and is closer to the alarms (on the average). Then the closest-unit division can be dominated.

2. Since it is possible to construct examples having $T_1 > T_2$ and $\lambda(Y(0)) \geq \lambda(B)/2$ (or $T_1 < T_2$ and $\lambda(X(0)) \leq \lambda(B)/2$), the two conditions in each part of the theorem are independent of each other.

Proof of theorem: One can argue abstractly from Theorem 4 and prove that (i) or (ii) implies the existence of a dominating set. However, it is more interesting to construct the dominating set, so we shall proceed to do so.

Suppose that hypothesis (i) is true. Since $T_1 > T_2$, the number s_0 of equation (5.1) is positive. Thus, if A is any closest-unit division and $0 < s' < s_0$, we have $A \subset X(s')$, and $\lambda(X(s')-A) > 0$ by assumption (2). Hence, by (5.2),

$$\overline{T}(X(s')) - \overline{T}(A) = \frac{P_{00}}{\lambda(B)} \int_{X(s')-A} (t_1-t_2-s_0)\, d\lambda,$$

which is negative, since the integrand lies between $-s_0$ and $-(s_0-s')$ on $X(s')-A$. Thus we see that, for any s' such that $0 < s' < s_0$ the region X(s') has lower average response time than any closest-unit division.

But since $u(s) = \lambda(X(s))$ is an increasing function of s (by assumption (2)), and $u(0+) = \lambda(Y(0)) < \lambda(B)/2$, we can find a value of s' between 0 and s_0 such that

$$\lambda(Y(0)) < u(s') \leq \lambda(B)/2.$$

This gives

$$\lambda(A) < \lambda(X(s')) \leq \lambda(B)/2,$$

for any closest-unit division A.

Using equation (6.1), we see that X(s') has lower ΔW than A has. Since we already know that $\overline{T}(X(s')) < \overline{T}(A)$, we see that X(s') dominates any closest-unit division.

The proof for case (ii) is similar.

Now it remains to show that if neither (i) nor (ii) holds, then there is no response area which dominates all the closest-unit divisions, and for this we apply Theorem 4. Conditions (i) and (ii) can fail in the following ways:

(iii) $T_1 < T_2$ and $\lambda(X(0)) \leq \lambda(B)/2$,

(iv) $T_1 > T_2$ and $\lambda(Y(0)) \geq \lambda(B)/2$,

or (v) $T_1 = T_2$.

If (iii) holds, then $\lambda(Y(s_0)) \leq \lambda(X(0)) \leq \lambda(B)/2$, and Theorem 4(a) implies that no set dominates X(0). If (iv) holds, then $\lambda(X(s_0)) \geq \lambda(Y(0)) \geq \lambda(B)/2$, and Theorem 4(b) implies that no set dominates Y(0). If (v) holds, all the closest-unit divisions have minimum average response time, and any non-equivalent set has higher $\overline{T}$, by Theorem 1, so it cannot dominate a closest-unit division. Hence, in all three cases, it is not possible to find a set which dominates all the closest-unit divisions.

Summary of Findings

For the two-unit response problem, the sets X(s) play an important role. These sets can be determined from the response-time metric without any knowledge of the arrival patterns of alarms, and the sets Y(s) can be determined directly from the X's.

From Theorem 4 we know that the only sets which should be considered as candidates for the response area of unit 1 are those which lie between some X(s) and its corresponding Y(s), because all other response areas are dominated by another choice.

Among those candidate sets which do lie between some X(s) and Y(s), there will usually be a whole family of choices which are "good", in the sense that they cannot be dominated. In order to determine these "good" choices, it is necessary to know the geographical distribution of the alarms. Speaking roughly, Theorem 4 says that this family consists of all candidates which lie between the "minimum response time" area and the "equal alarm rate" area.

Under conditions where the alarm rates vary substantially over small distances, it will often be found that the usual choice of response area - the closest-unit division - is not a good candidate. Theorem 5 shows how to determine whether or not this is the case. To use the theorem, one must determine which side of the equal-response time dividing line has more than half of the alarms, and which unit has the lower average response time to all the alarms.

VII. SENSITIVITY OF FINDINGS TO MODEL ASSUMPTIONS

We have noted previously that certain assumptions used in our model are somewhat unrealistic and ought to be eliminated. Among the most important ones are:

(1) The assumption that the units inside the region B may not respond to outside incidents, but outside units do respond into B.

(2) The assumption that exactly one unit serves each incident.

(3) The assumption that total service time (including travel) is independent of the location of the incident and the unit which serves.

The authors have developed several models which eliminate these restrictions, at least partially, but none of them is wholly satisfactory. Handling the first and second problems is best accomplished by treating systems with more than two units. When the total number of units is sufficiently large, the number of incidents which must be answered from outside the region becomes negligibly small. We have been able to make numerical calculations of state probabilities for specific examples of such models, but at present we have neither analytical formulas nor the counterparts of the theorems in this paper. We plan to report these findings at a later date.

We have also approached the first problem by assuming that calls arrive from outside B at a certain rate for unit 1 and at another rate for unit 2. Under the assumptions that these outside alarms arrive according to a Poisson process and have the same service-time distribution as inside calls (neither of which is completely realistic), it is possible to demonstrate that all our findings remain unchanged except for the definition of s_0 and a minor change in ΔW.

The third problem should be handled by assuming that the region B is divided into small subregions. In each subregion, alarms arrive according to a Poisson process, and the service-time distribution depends on the particular subregion in question and the unit which responds. Such a model becomes analytically complex once the number of subregions

exceeds three or four, but we have been able to calculate numerical results for average travel time and workload balance in simple cases.

All of our numerical results have the same qualitative interpretation as the simpler model. Namely, it is possible to reduce average travel time and improve workload balance by moving the dividing line toward the side of B which has the most alarms. However, the following differences are observed:

- Comparing the dividing line ℓ_1 which produces equal workloads for the simple model with the analogous line ℓ_2 for the model in which service time depends on travel time, we find that ℓ_2 is nearer to the closest-unit dividing line than ℓ_1. This is not surprising, since the dividing line which causes both units to respond to the same number of incidents will cause the unit which travels furthest (on the average) to be busy a greater fraction of the time.
- There is a smaller range of dividing lines which cannot be dominated.

The significance of these approximate results is that, in a practical situation, one can estimate the location of "good" dividing lines using the simple model and then apply simulation models [1] to obtain realistic appraisals of the changes to be expected in response time and workloads.

REFERENCES

1. Carter, G., and E. Ignall, "A Simulation Model of Fire Department Operations," IEEE System Science and Cybernetics, 6, 282 (1970).

2. Chaiken, J. M., "Estimating Numbers of Engine Companies Needed in NYC Fire Divisions," New York City-Rand Institute, R-508, to appear.

3. Chaiken, J. M., and E. Ignall, "An Extension of Erlang's Formulas Which Distinguishes Individual Servers," New York City-Rand Institute, R-567, March 1971.

4. Chicago Police Department, Operations Research Task Force, Al Bottoms, Project Director, "Quarterly Progress Reports".

5. Cooper, L., "Heuristic Methods for Location-Allocation Problems," SIAM Review, 6, 37-53 (1964).

6. Gass, S. I., "On the Division of Police Districts into Patrol Beats," Proceedings, 1968 ACM National Conference.

7. Hogg, J. M., "The Siting of Fire Stations," Operational Research Quarterly, 15, 261-70 (1964).

8. Larson, R. C., "Operational Study of the Police Response System," Technical Report No. 26, Operations Research Center, MIT (1967).

9. Larson, R. C., "Models for the Allocation of Urban Police Patrol Forces," Technical Report No. 44, Operations Research Center, MIT (1969).

10. Larson, R. C., and K. A. Stevenson, "On Insensitivities in Urban Redistricting and Facility Location," New York City-Rand Institute, R-533, March 1971.

11. McEwen, T., Project Director, "Allocation of Patrol Manpower Resources in the Saint Louis Police Department," Vols. I and II (1968).

12. Morse, P. M., Queues, Inventories, and Maintenance, John Wiley, New York (1967).

13. Savas, E. S., "Simulation and Cost-Effectiveness Analysis of New York's Emergency Ambulance Service," Management Science, 15, B-608-27 (1969).

14. Scott, A. J., "Location-Allocation Systems: A Review," Geographical Analysis, 2 (1970).

15. Strauch, R. E., "When a Queue Looks the Same to an Arriving Customer as to an Observer," Management Sciences: Theory, 17, 140 (1970).

16. Teitz, M. B., "Toward a Theory of Urban Public Facility Location," Papers of the Regional Science Association, 21, 35-51 (1968).