PREDICTING THE DEMAND FOR FIRE SERVICE

Jan M. Chaiken
John E. Rolph

May 1971

P-4625

PREDICTING THE DEMAND FOR FIRE SERVICE

Jan M. Chaiken and John E. Rolph

New York City-Rand Institute, New York, New York
The Rand Corporation, Santa Monica, California

I. INTRODUCTION

In early 1968, The Rand Corporation began working for the City of New York on police, fire, health, and housing studies. The work we are reporting is part of an ongoing effort to help the New York City Fire Department use its resources more effectively. Fire alarm rates have tripled in the last decade, creating pressures on the department either to increase the numbers of men and equipment or to find better methods of using the available resources. To make specific recommendations about locating or dispatching fire units, estimates are needed of the distribution of incoming alarms by time, type, and geography. For example, most analytical models which could assist the dispatcher in determining which units to dispatch to the latest alarm require information about (i) the probability that the alarm represents a serious fire and (ii) the probability distribution of incoming alarms in the vicinity of the current alarm 10 or 20 minutes.

---

*Any views expressed in this paper are those of the authors. They should not be interpreted as reflecting the views of The Rand Corporation or the official opinion or policy of any of its governmental or private research sponsors. Papers are reproduced by The Rand Corporation as a courtesy to members of its staff.

This paper was presented as an invited talk at the American Statistical Association Meeting in Detroit, December 1970.

We have been analyzing the records of alarms received by the Fire Department since 1962 to provide estimates of such probability distributions. In this paper we will try to give a picture of the fire alarm distribution, a description of some of the methods we are using for analysis, and finally an example of an application of the estimated alarm incidence rates.

II. QUANTITATIVE PICTURE OF INCIDENCE CHARACTERISTICS

We begin our description of the characteristics of New York City fire alarm incidence with Fig. 1, which shows that the total number of alarms has been increasing rapidly. Even a linear fit to the logs of data through 1967 produces a substantial underestimate of the annual totals for 1968 and 1969. It is therefore useful to split up alarms by type. In Fig. 2 we show the data for Structural fires. The logs could conceivably be fitted either linearly or quadratically for extrapolation, but the quadratic fit proved best for estimating the 1968 and 1969 structural fire totals of 46,600 and 47,200. On the other hand, annual nonstructural fires behave erratically. If brush fires are separated out, you again get a smooth exponential increase (Fig. 3). The 1968 and 1969 totals for nonstructural fires are 79,400 and 81,200. Nonstructural fires with brush removed is 71,200 in 1968. For false alarms we fit a quadratic exponential increase. In fact the extrapolated value for 1968 is almost exactly the actual one of 60,900. So much for the power of crude prediction.

Turning from annual totals to weekly totals we can see some patterns in the false alarms. We fit the logs of the weekly false alarms with a quadratic trend and a 4 harmonic seasonal pattern up to the end of 1967 (Fig. 4).

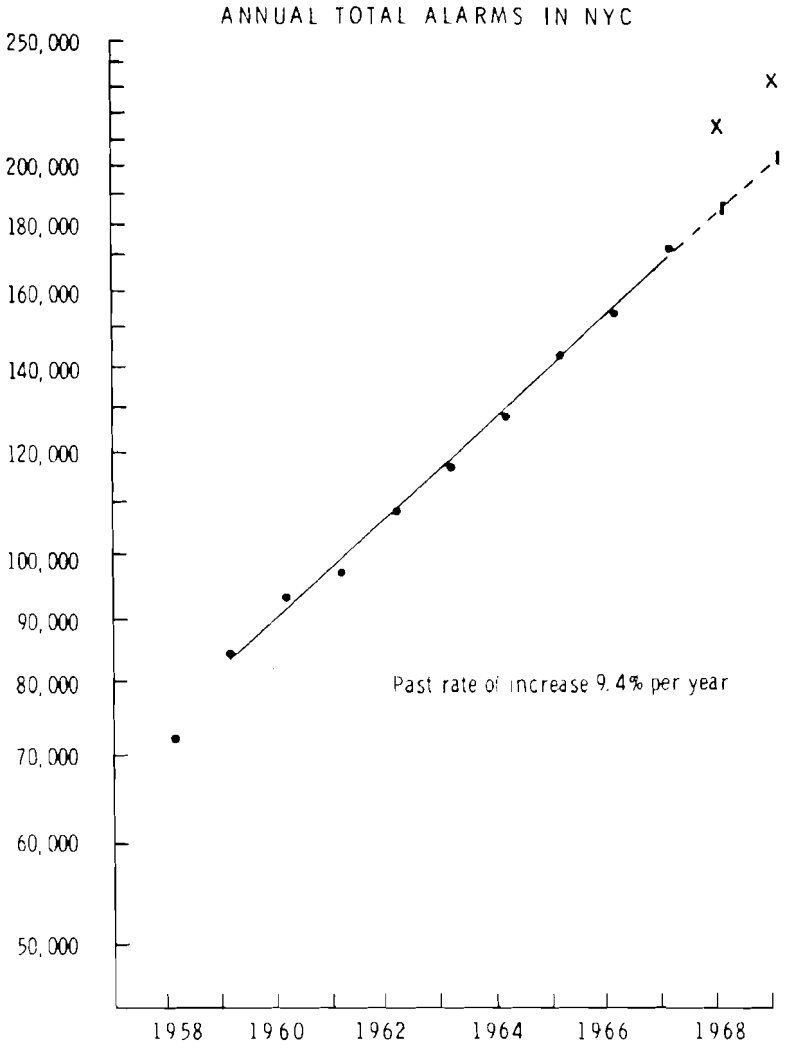
ANNUAL TOTAL ALARMS IN NYC
250,000
200,000
180,000
160,000
140,000
120,000
100,000
90,000
80,000
70,000
60,000
50,000
1958
1960
1962
1964
1966
1968
Past rate of increase 9.4% per year

Figure 1

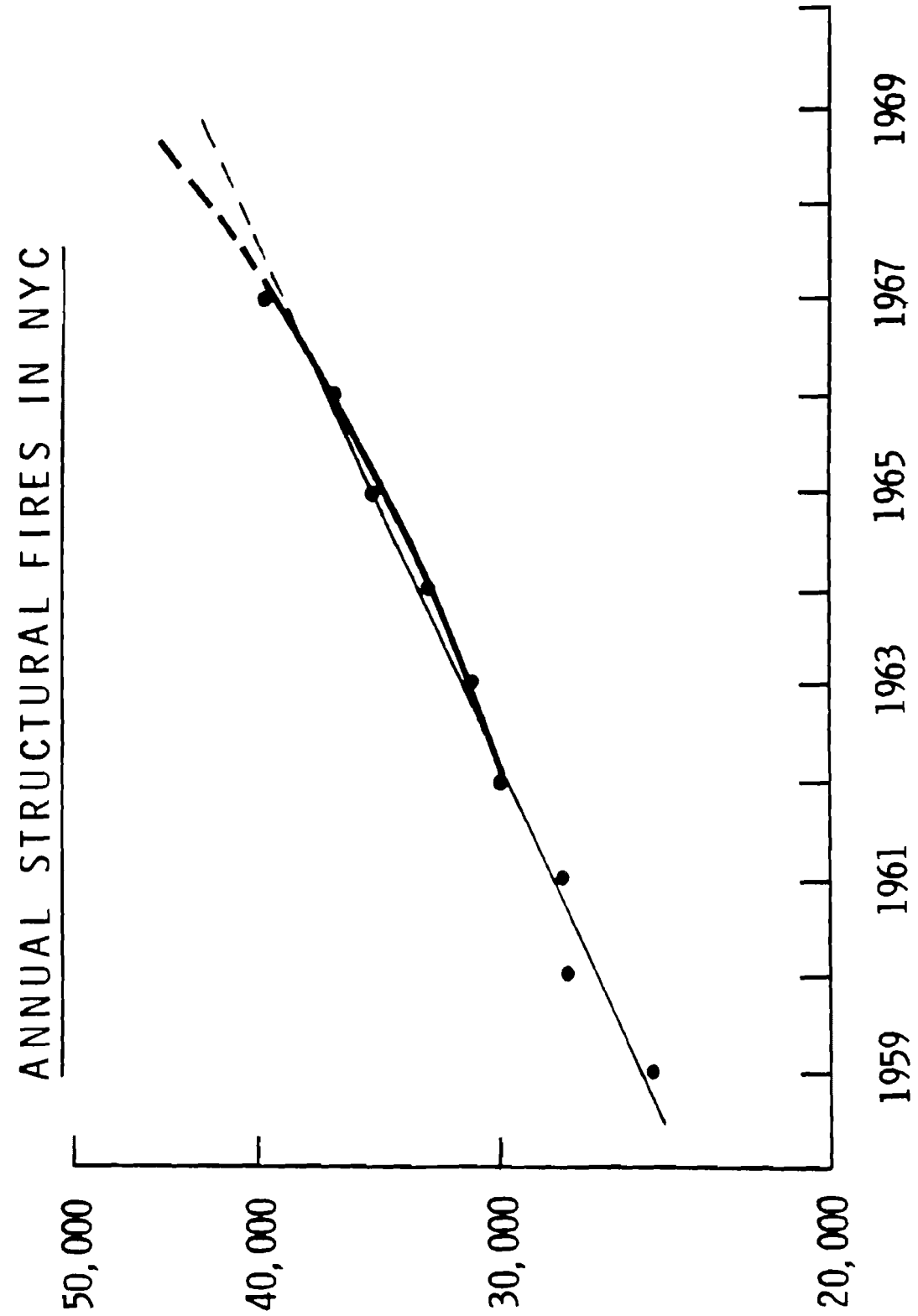
ANNUAL STRUCTURAL FIRES IN NYC
50,000
40,000
30,000
20,000
1959
1961
1963
1965
1967
1969

Figure 2

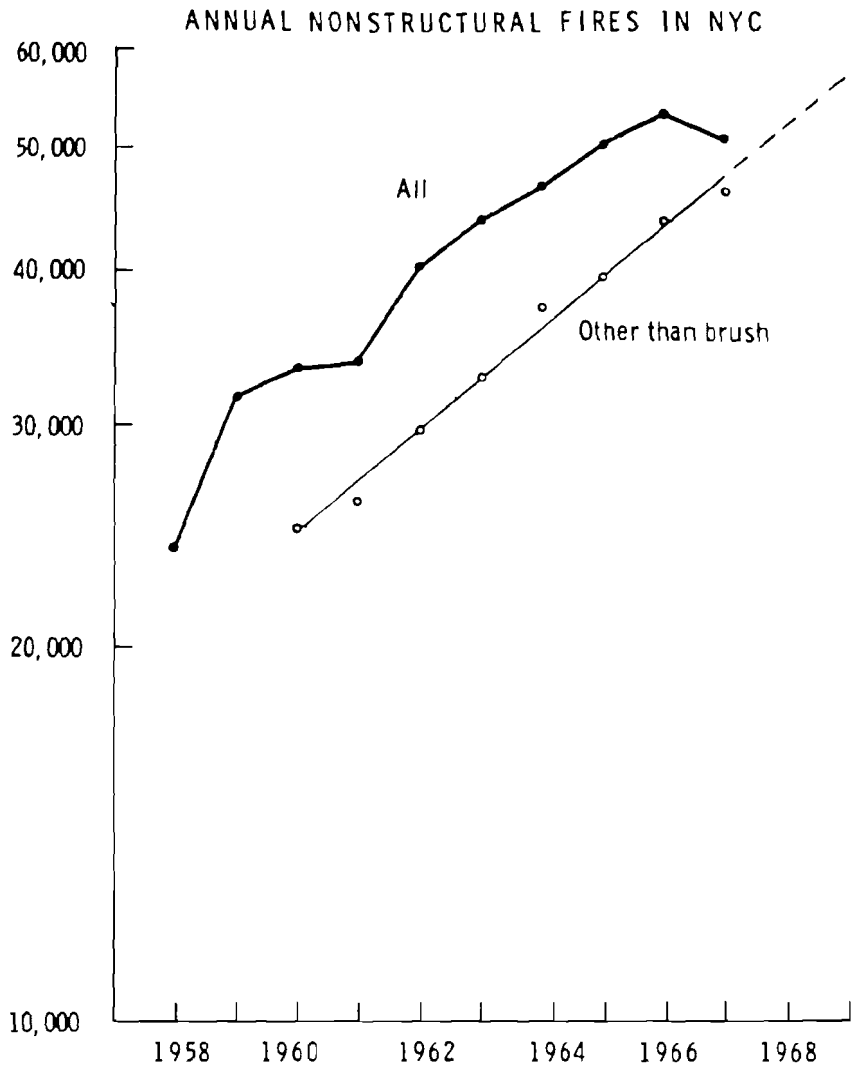
ANNUAL NONSTRUCTURAL FIRES IN NYC
60,000
50,000
40,000
30,000
20,000
10,000
All
Other than brush
1958
1960
1962
1964
1966
1968

Figure 3

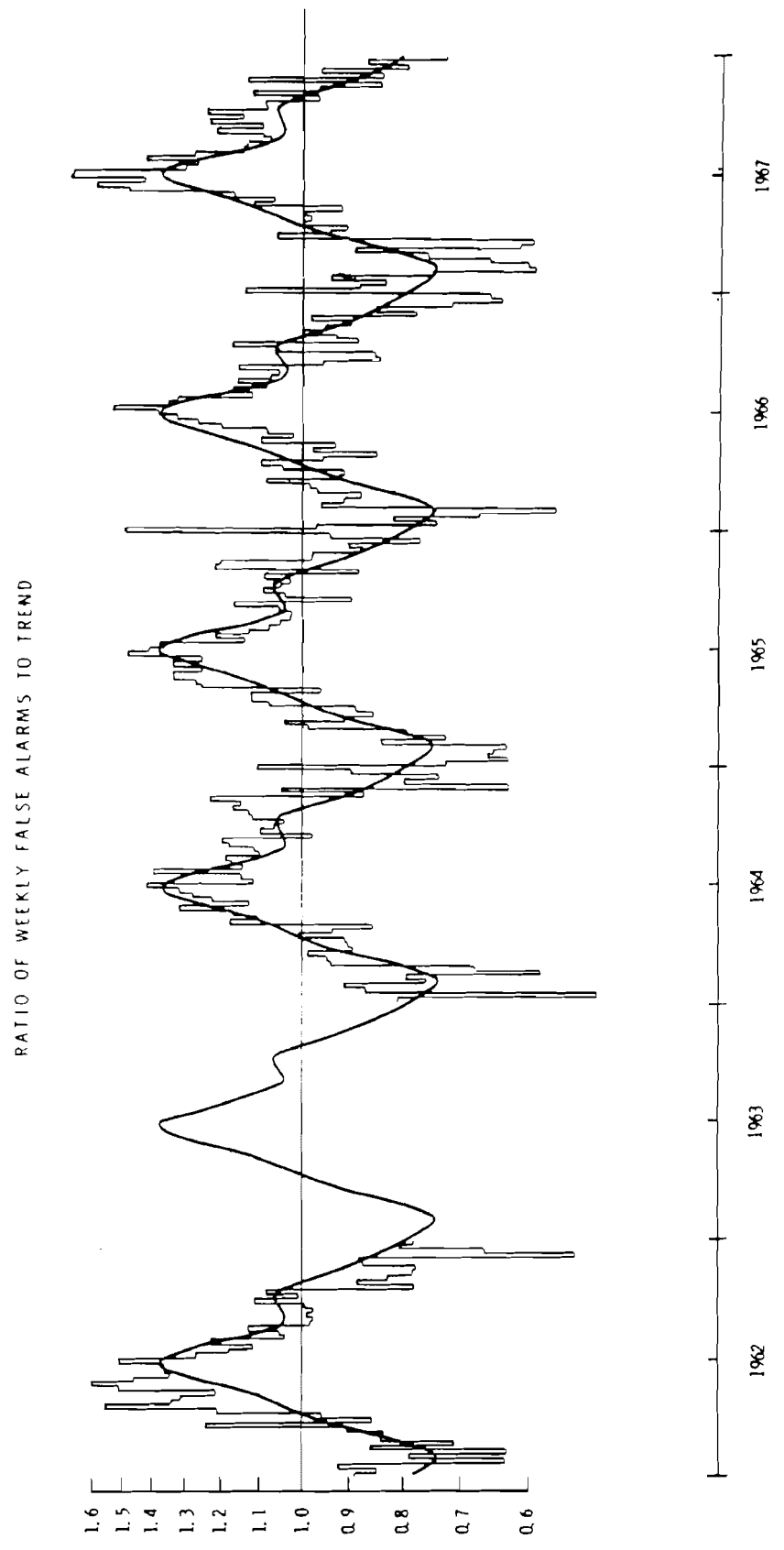
RATIO OF WEEKLY FALSE ALARMS TO TREND
1.6
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
1962
1963
1964
1965
1966
1967

Figure 4

Turning to the pattern of alarms during a day, you can see there may be an order of magnitude difference between the arrival rates at different hours (Fig. 5). It turns out that different types of alarms have different hourly patterns.

Cutting across alarms from a geographical angle produces interesting results. Boxes of extremely high demand are found to be bunched together into small regions. Regions of lower-than-average demand are also geographically well defined, but different neighborhoods can be distinguished by their false alarm ratio and other characteristics.

Main streets are found to have different demand patterns from side streets. In high-incidence areas, the main streets have fewer alarms than the side streets, reflecting a smaller rate of false alarms. In low-incidence areas, the main streets are higher than the side streets (but of course lower than main streets in "bad" areas), perhaps because there are more people around to cause and spot fires. Boxes near vacant land such as parks, playgrounds, cemetaries, have a different behavior from other boxes. Whether or not they have higher false alarm rates than the general neighborhood depends on location.

Weather effects can be observed in a number of ways. For example, the summer of 1969 was extremely wet (for New York) and false alarms and rubbish fires were down.

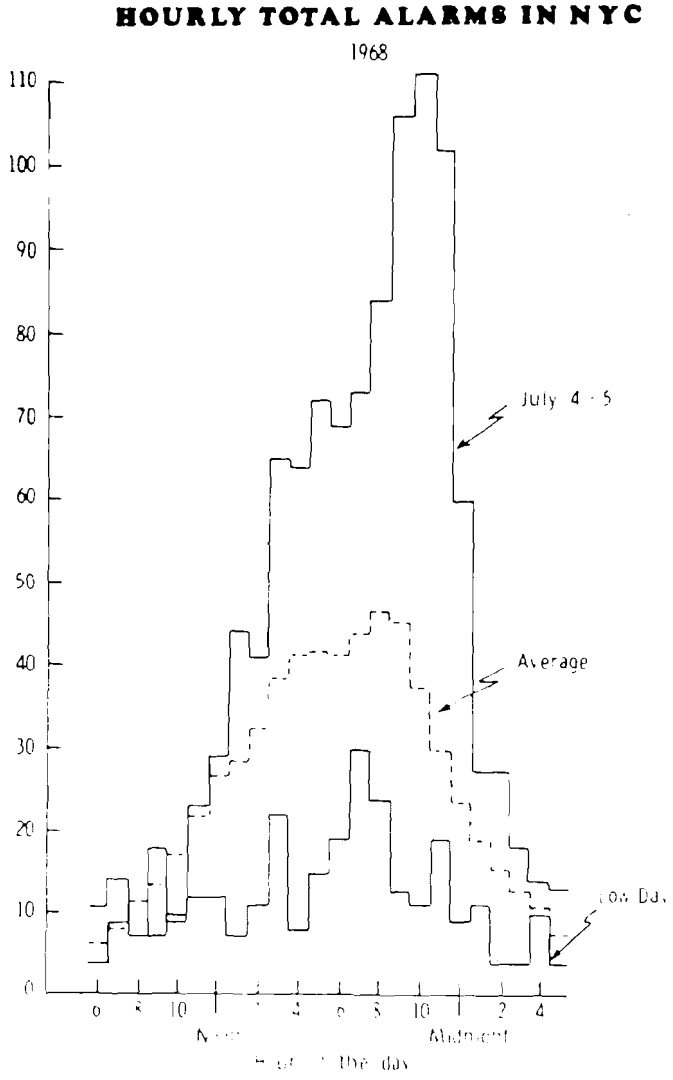
HOURLY TOTAL ALARMS IN NYC
1968
110
100
90
80
70
60
50
40
30
20
10
0
July 4 - 5
Average
Midnight

Figure 5

Weather variables other than percipitation were also considered. Figures 6 and 7 show plots of hourly alarms in 1968 against temperature and relative humidity. Some variables do seem to have an effect. These plots, together with other methods, were used for selecting weather variables for the analysis.

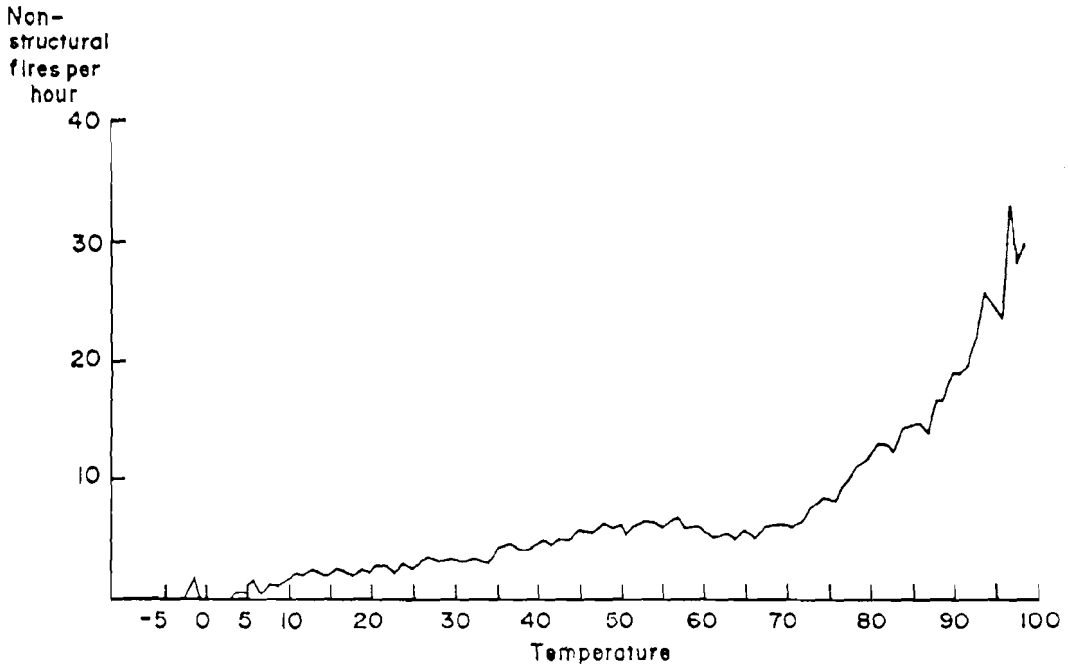

Figure 6

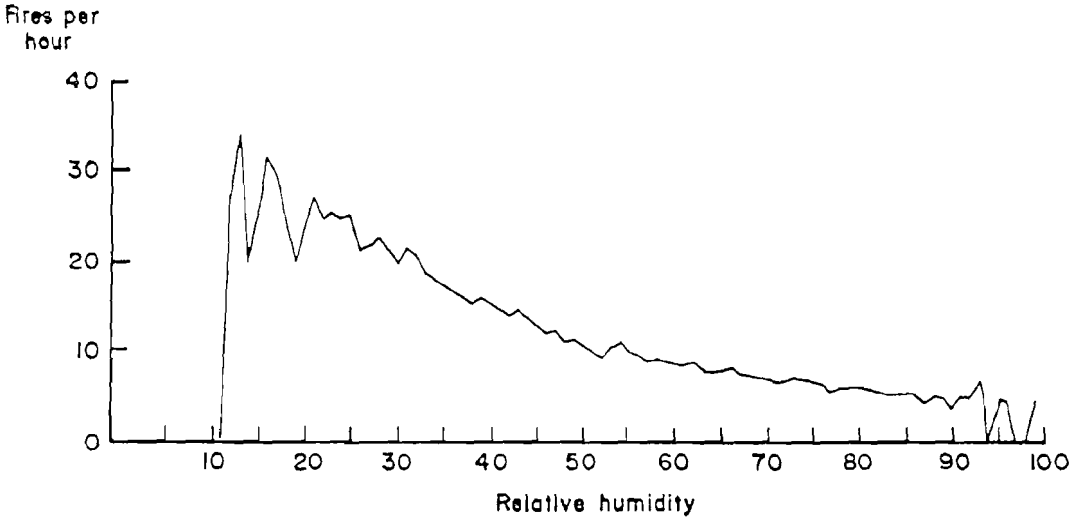

Figure 7

III. METHODOLOGY

We seek a method for short term prediction of the incidence rates for various types of fire alarms as a function of location, time (year, season, day of week, hour), method of reporting (box, phone, etc.), and weather conditions (termperature, relative humidity, precipitation). Since we are merely extrapolating past data and not making causal statements, we can't hope to look more than a year ahead in detail with any confidence. Naturally we can't predict occurrences such as the tremendous increase in rubbish fires which took place during a strike by sanitation men or the large number of false alarms and fires the night Martin Luther King was murdered.

Although we have records for 1 1/2 million alarms, there are still problems related to sample size. The various marginal tabulations of the data such as those shown in the figures give one an idea of where to start. Next comes the problem of putting all these factors together. It is basically a problem of what level of disaggregation and smoothing works for which variables. We will give a rough outline of the approach and then discuss in detail several parts of the work.

On both conceptual and empirical grounds we begin by assuming a Poisson-type probability model of fire alarms. That is, given the type of alarm and method of reporting, the occurrence of that type alarm is viewed as an independent

Poisson point process whose parameters vary with time, geography, and weather. The conceptual grounds are merely the Poisson assumptions--independence over time and geography and "no multiple hits." For a methodical arsonist or a carload of kids riding down the street pulling alarm boxes, this does not hold, but nearly all the data are found to confirm these assumptions. One benefit we reap from the Poissonicity is that given a region, a set of types, and a time period, the number of alarms has a Poisson distribution whose parameter is the integral of the intensity function over the region, time period, and summed over the types.

Statistical Methods

Fixing on a type of alarm and method of reporting (call it "a") we wish to separate out the effects of time variation and weather variation on the alarm rate. The two factors are intertwined since weather varies with season and time of day as well as year by year.

We first control for time effects and then estimate weather effects conditional on the time effect. Thus our predicted number of "a" alarms is

$$P_a(t, w) = P_a(t) + f(0_w(t-s) - P_w(t-s);\ s = 0, 1, \ldots, k)$$

where

$P_a(t)$ = predicted fire alarms of type a at time t, given normal weather

$P_w(t)$ = predicted weather at t

$O_w(t)$ = observed weather at t,

f = a function of residual weather which may also depend on values prior to t.

Specifically, using least squares as the method of fitting the model, we look at

$$O_a(t) - P_a(t) = \sum_{s=0}^{k} \gamma_s [O_w(t-s) - P_w(t-s)] + \text{error}$$

for fitting γ_s, where $O_a(t)$ is the actual number of alarms at t. Naturally we want a stable variance, so a first step is to transform the observations so that var $(O_a - P_a)$ does not depend on P_a. Graphically this is done by a plot of the predicted values $P_a(t)$ against the residuals $R_a(t) = O_a(t) - P_a(t)$.

Plotting log $(P_a(t))$ on the x-axis and log $(|R_a(t)| + \epsilon)$ on the y-axis (ϵ small), the approximate transformation for stabilizing the variance is

$$g(O_a(t)) = \begin{cases} O_a(t)^{1-\theta} & \text{if } \theta \neq 1 \\ \log(O_a(t)) & \text{if } \theta = 1 \end{cases}$$

where θ is the approximate slope of the graph.* Using this method on false alarm data yielded a log transformation. For temperature and relative humidity, no transformation was indicated (i.e., $\theta = 0$).

Turning back to the disaggregation problem we mentioned earlier, geography is interesting. Let $X(t, r, j, i)$ be the number of type j alarms received in time unit t, by reporting method r, at location i, and let

$$X(t, r, j) = \sum_{i=1}^{I} X(t, r, j, i).$$

Then both $X(t, r, j)$ and $X(t, r, j, i)$ are assumed to be Poisson random variables with parameters $\lambda(t, r, j)$ and $\lambda(t, r, j, i)$ respectively. Let

$$p(t, r, j, i) = \frac{\lambda(t, r, j, i)}{\lambda(t, r, j)} .$$

It then follows that, conditional on $X(t, r, j) = n$, the random vector $[X(t, r, j, 1), \ldots, X(t, r, j, I)]$ has a multinomial distribution with parameters n and

$$\underline{p}(t, r, j) = [p(t, r, j, 1), \ldots, p(t, r, j, I)]$$

---

*See J. W. Tukey "On the Comparative Anatomy of Transformations," Ann. Math. Statist., 28, 1957, pp. 602-632.

Testing for geographical effect on time pattern amounts to testing whether $\underline{p}(t, r, j)$ depends on t and can be done via a chi-square test on each contingency table, given r and j.

We constructed neighborhoods in Brooklyn for this, and our tests showed mixed results. Seasonal patterns vary geographically for virtually all types of fire and all methods of reporting. Hour of the day patterns did not vary significantly by geography except for

1. nonstructural fires
2. false alarms reported by box
3. transportation fires on weekdays reported by phone.

We will skip over the actual time fitting other than to say that using logs seems to work well, suggesting a multiplicative model with exponential growth. We are still working to improve fits.

Future Work

The Fire Department is designing a computerized management information and control system which will use all this information, but it is not slated to be operational for at least two years. Thus producing detailed predictions now is unnecessary. Our main aim is to develop sufficiently accurate prediction methods so they can be used in the system. Other aspects of the design problem which we have not covered here are updating and outlier identification, which are also being studied.

The uses for predicted incidence range from the long-term problems of locating new station houses and determining future personnel needs, to selection of relocation and prepositioning strategies (which have a planning horizon of several hours), to actual dispatching of units to incoming alarms. Next we look at a specific dispatching problem.

IV. APPLICATION OF INCIDENCE PREDICTION*

We look at the very simple question: given the location of units and number of units to be dispatched, which ones should be assigned to a particular alarm? To clearly illustrate the effect of geographic differences in incidence rates, we treat a simple example in which there are only two units.

The approach is to select a geographical district or response area for each unit to serve. The unit responds to all alarms inside its response area unless it is busy with another call. When one unit is busy the other unit responds unless it is also busy. The objective is to draw boundaries to response districts so as to minimize average response time and more evenly distribute the work between units. Clearly these objectives can conflict in some situations. The importance of average response time is self-evident. Equalizing workload is a secondary objective for the simple reason that overfatigued units are less efficient and may even need to be replaced or supplemented. The results derived from the model include:

1. Formulas for the workload of each unit and the average response time to all incidents as functions of the response districts.

---

*This section is based on G. M. Carter, J. M. Chaiken, E. Ignall, Response Areas for Two Emergency Units, New York City-Rand Institute, R-532, March 1971.

2. A determination of the district boundary which minimizes average response time--frequently different than the commonly used boundary which is equidistant from the home locations of the two units. The equidistant boundary is the one usually used when one is ignorant of incidence rates.

3. 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 more equal distribution of work between units.

We give some illustrations of boundaries and then characterize some of the results for you.

Example 1 (Fig. 8)

Consider a rectangular region B with units symmetrically located on the x-axis. We assume that the streets are parallel to the x and y axes, and that a unit begins each response at its home location. The units travel at constant speed v_1 on streets parallel to x-axis and at constant speed v_2 on streets parallel to y-axis. The total time from location 1 to (x, y) is

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

and from location 2 to (x, y) it is

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

The y-axis divides the region into closest unit response areas but may not give minimum average response time. To understand this, consider Fig. 8 where the arrival rate of calls in B_2 is much higher than that in B_1. If an incident occurs at a and unit 2 is dispatched, there may be a good chance that another alarm will arrive while unit 2 is busy. If such an alarm arrives, the chances are good that it will be in B_2 rather than B_1. Suppose it occurs at b. Then unit 1 will travel a large distance to b to service the alarm. Had we originally sent unit 1 to a, then unit 2 could have gone to b and it is clear that the average travel time for these 2 dispatches is lower than for the original 2 dispatches. Thus it may have been better to have a in the response area for unit 1.

In fact the dividing line which minimizes average response time is a vertical line to the right of the y-axis. For this example it is the line

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

where λ is the arrival rate of alarms in B, μ is the service rate of alarms, and T_i is the expected response time if every arriving call were to be serviced from the

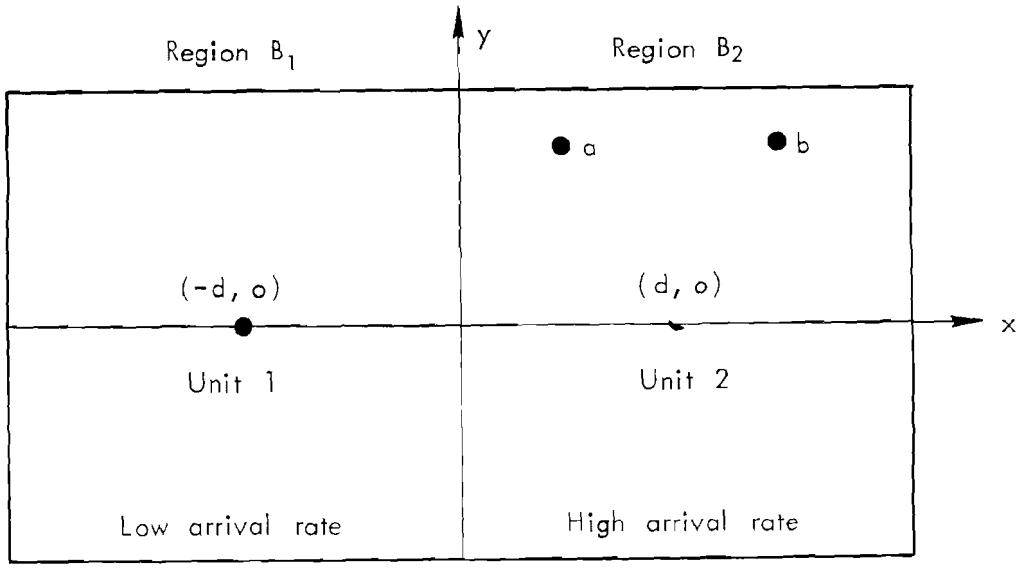
Region B_1
y
Region B_2
a
b
(-d, o)
(d, o)
x
Unit 1
Unit 2
Low arrival rate
High arrival rate

Figure 8

location of unit i. ($T_1 > T_2$ since more alarms are on the right.) Note that the optimal line depends not only on the distribution of alarms geographically (which enters into the calculation of T_1 and T_2) but also on the total expected alarm rate. Thus, in applications it may be necessary to have different response areas at different times of day.

It turns out that as the dividing line moves to the right from the y-axis to the optimal line, the average response time decreases to the minimum. It is clear that for some of these dividing lines the average response time is reduced compared to the closest-unit division and the workload balance is improved. We say these lines dominate the closest unit division.

Example 2.

Even if the alarm rate is uniform across the region but the units are unsymmetrically located similar considerations apply--Fig. 9. The optimal dividing line is

$$x = d \cdot \frac{\lambda}{\lambda + \mu} \frac{b}{2a + b} .$$

Example 3 (Fig. 10) Units Not on Same Street

The line aa is the equidistant one (assuming $v_1 = v_2 = v$). For the case of $T_1 - T_2 =$, ϵ small, then the optimal dividing line is bb where the translated distance is

UNITS NOT SYMMETRICALLY LOCATED

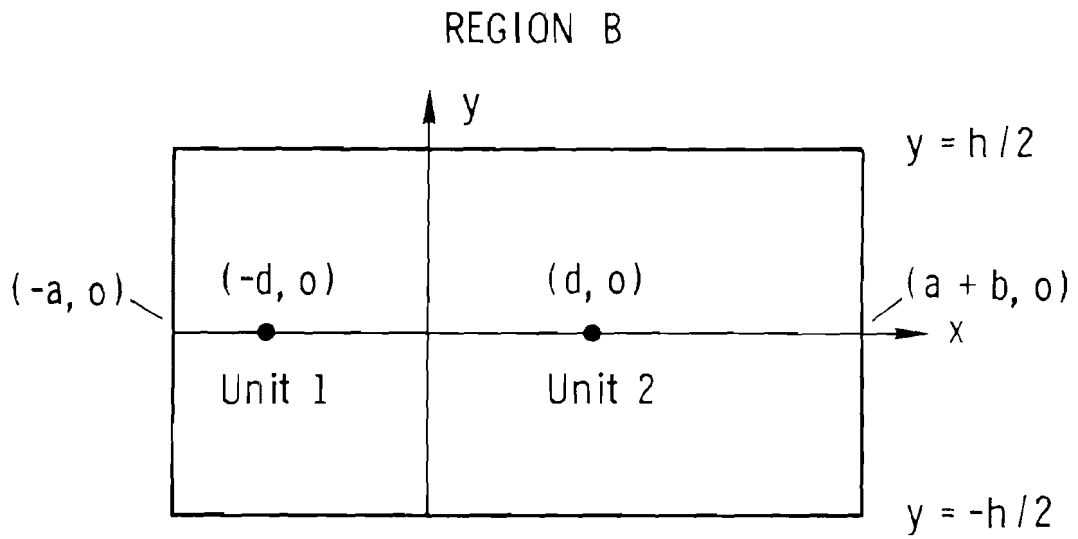

Figure 9

UNITS NOT ON THE SAME STREET

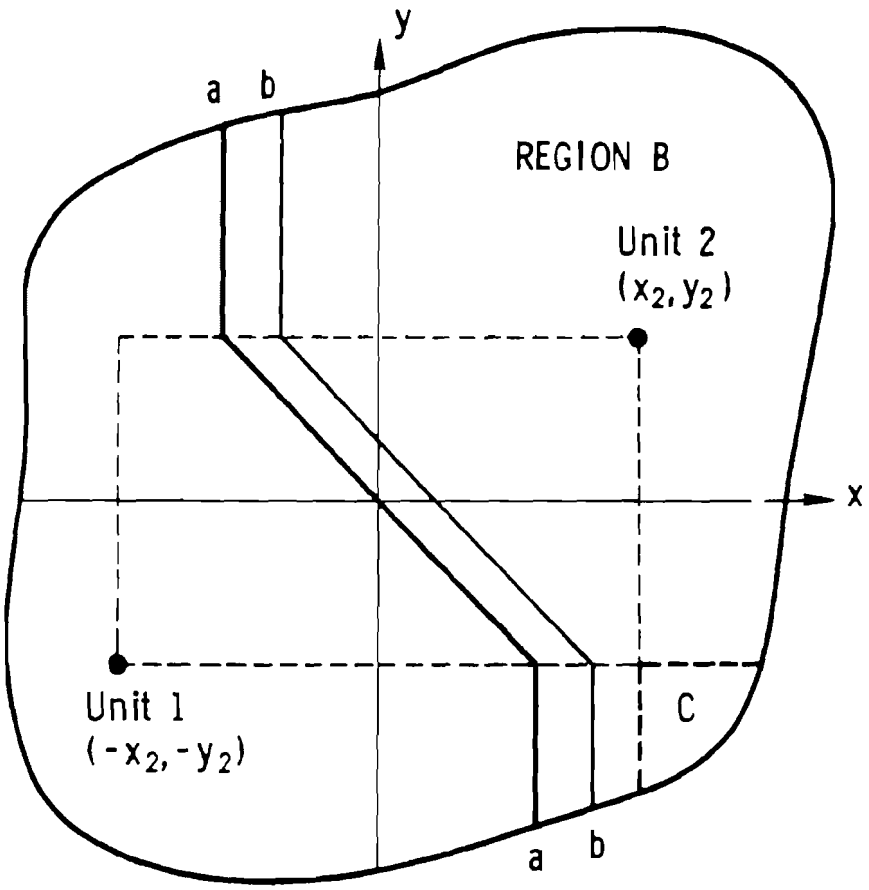

Figure 10

$$\frac{1}{2} v \frac{\lambda (T_1 - T_2)}{\lambda + \mu} .$$

We turn now to the more general situation. Let A be the response district for unit 1. We wish to find the average response time $\bar{T}(A)$ for the whole region B as a function of A. After considerable calculation a useful formula for $\bar{T}(A)$ is

$$\bar{T}(A) = \frac{P_{oo}}{\lambda (B)} \int_A (t_1(\underline{x}) - t_2(\underline{x}) - s_o) d\lambda (\underline{x}) + \alpha$$

where P_{oo} is the steady state probability of both units being available;

λ is the arrival rate measure, thus the total arrival rate in B is $\lambda (B) = \int_B d\lambda (\underline{x})$;

$t_i(x)$ is the expected travel time from location of unit i to location $\underline{x}$;

$$s_o = \frac{\rho}{\rho + 1} (T_1 - T_2), \text{ where } \rho = \frac{\lambda (B)}{\mu} ;$$

$$\alpha = P_{oo} \left\{ \frac{\rho + \rho^2/2}{1 + \rho} T_1 + \frac{1 + \rho + \rho^2/2}{1 + \rho} T_2 + \rho^2 \tau/2 \right\},$$

where τ is the average response time if a unit outside B responds, and thus α is independent of A.

The form of this equation suggests that points $\underline{x}$ such that $t_1(\underline{x}) - t_2(\underline{x}) = s_o$ play an important role. Note since $t_j(\underline{x})$ can be any bounded nonnegative function on B integrable w.r.t. λ, $t_j(\underline{x})$ can be thought of as the utility of responding from location j to $\underline{x}$.

Define the sets

$$X(s) = \{x \in B: \ t_1(\underline{x}) - t_2(x) < s\}$$

and

$$Y(s) = \{\underline{x} \in B: \ t_1(x) - t_2(\underline{x}) \leq s\}.$$

Note that calculating these sets does not depend on knowledge of the arrival patterns of alarms. One can show:

1. For any set $A \subset B$, $\bar{T}(A) \geq \bar{T}(X(s_o))$ and further for $X(s_o) \subseteq A \subseteq Y(s_o)$

$$\bar{T}(X(s_o)) = \bar{T}(A) = \bar{T}(Y(s_o)).$$

A consequence is that one can choose on other grounds (workload or convenience) which A to take if $X \neq Y$.

Recall Example 3 (see Fig. 11); suppose that s_o is such that the set $E = \{\underline{x} \in B: \ t_1(\underline{x}) - t_2(\underline{x}) = s_o\}$. The above assertion says that any optimal response area for unit 1 must contain all points to the left of the shaded and heavy lines and may contain some points in E. In practical cases s_o may not be easy calculable, so one can describe the whole family of sets and by trial and error select the best.

AMBIGUOUS LOCATION OF DIVIDING LINE

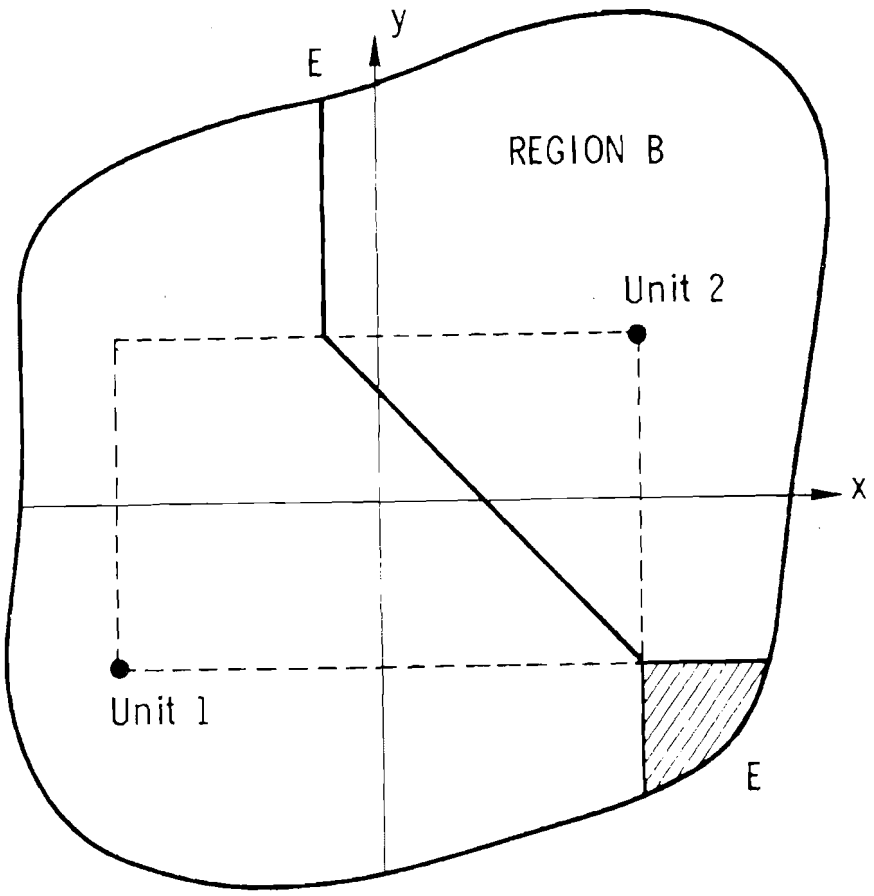

Figure 11

2. $g(s) = \bar{T}(X(s))$ monotonically increases as s moves away from s_o.

Moving to the tradeoff between decreasing average response time and equalizing workload, we introduce the notion of dominance. We say a response area A for unit 1 dominates area A' iff

$$\bar{T}(A) \leq \bar{T}(A')$$

and

$$\Delta W(A) \leq \Delta W(A') \ ,$$

with at least one inequality strict, where

$$\Delta W(A) = \frac{P_{oo} | 2\lambda(A) - \lambda(B) |}{(\mu + \lambda(B))}$$

is the difference in workload between units 1 and 2.

Example

Table 1

VALUES OF PARAMETERS FOR EXAMPLE (FIG. 9)

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	$\lambda(A)$	$\int_A (x+2)\, dxdy/4h$

Consider the geographical arrangement shown in Fig. 8 with the parameters shown in Table 1. Here s_o = 11/480 hours and the dividing line which minimizes average response time is a vertical line at x_o = 11/48 miles. One can see where the tradeoffs are as a function of the dividing line in Fig. 12.

For the general case it is possible to characterize the response areas which cannot be dominated and then there is no reason to consider any others. There are two results.

1. A set must have two properties in order not to be dominated by any other:
   (a) it must lie between some X(s) and corresponding Y(s)
   (b) It must lie between the "minimum response time" area and the "equal workload" area.

Note that it can happen that X(s) and Y(s) differ considerably, so that it is a bit tricky to compare two sets between them to see if one dominates the other.

2. 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.

Conclusion

The X(s) play an important role and the sets Y(s) can be determined directly from them. No knowledge of the alarm arrival pattern is needed. Usually when alarm rates

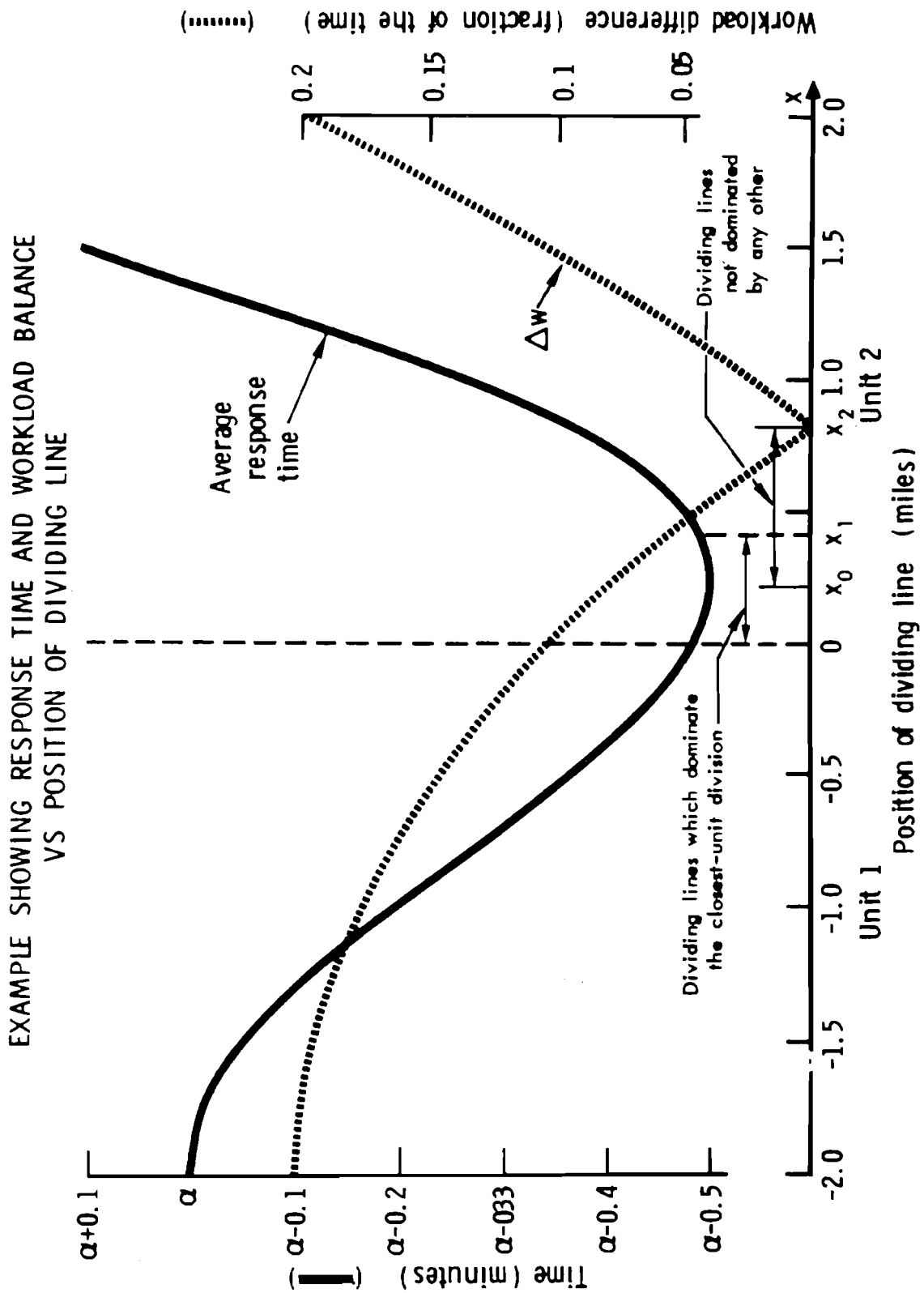
EXAMPLE SHOWING RESPONSE TIME AND WORKLOAD BALANCE
VS POSITION OF DIVIDING LINE
Average response time
Time (minutes) (——)
α+0.1
α
α-0.1
α-0.2
α-033
α-0.4
α-0.5
Workload difference (fraction of the time) (.....)
0.2
0.15
0.1
0.05
Δw
Dividing lines not dominated by any other
Dividing lines which dominate the closest-unit division
-2.0
-1.5
-1.0
-0.5
0
x_0
x_1
x_2
1.0
1.5
2.0
x
Unit 1
Unit 2
Position of dividing line (miles)

Figure 12

vary substantially over small distances the closest unit division is not a good candidate.

Some assumptions of the model are unrealistic:

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

(2) Exactly one unit is assumed to serve each incident.

(3) Total service time (including travel) is assumed independent of the location of the incident and the unit which serves.

Assumptions (1) and (2) are best eliminated by treating systems with more than 2 units. Numerical calculations have been done for this case, but the corresponding theorems have not been proved. Assumption (3) can be removed by subdividing regions. Numerically the results agree with the simpler model, although analytical formulas have not been derived.

Since the NYFD sends a single Batallion Chief to every incident the one unit server is appropriate for application to this case. Redesign of the chief's response boundaries is currently underway, but the new designs have not yet been implemented.