A SIMULATION MODEL OF THE NEW YORK CITY FIRE DEPARTMENT: ITS USE IN DEPLOYMENT ANALYSIS

Grace Carter, Edward Ignall,* Warren Walker*

The New York City-Rand Institute

Abstract

This paper describes a simulation model developed as a tool to aid deployment decision-making for the New York City Fire Department. The model, written in Simscript I.5, has been used to evaluate alternative solutions to workload and response problems plaguing the City. These solutions involve new policies for locating, relocating and dispatching fire-fighting units to achieve a more effective utilization of resources. Several specific applications of the simulation are described. In addition, methodological issues concerning the design and use of the model are addressed.

I. INTRODUCTION

Since 1968 a large-scale research program for the Fire Department of the City of New York has been in progress at The New York City-Rand Institute. An overview of that research is contained in [1]. A major part of the program has been an investigation and evaluation of alternative policies for the deployment of fire-fighting resources. This paper describes a simulation model of the responses of fire-fighting units to alarms, which has been one of the tools used for this evaluation. A comprehensive description of the design of the simulation model is given in [2]. Our emphasis in this paper will be on the use of the model. We will describe the various policies which were compared and present the differences in their performance. We will also note new policies which have been implemented based in part on simulation results.

*Also Columbia University.

We had two primary motives for using simulation:

(1) To be able to compare alternative policies without risking lives and property and spending considerable sums of money by trying them in the real world; and

(2) To gain a better understanding of fire department operations by making clear the effects of interactions within the system and the second-order consequences of suggested actions.

The need for such understanding and for the development of new policies had become apparent to the Department as the alarm rate began to grow exponentially in the 1960's. In 1968 the Department responded to 227,000 alarms, more than three times the number responded to in 1956, placing a severe strain on existing fire-fighting resources and increasing the pressures on the City for adding to them. However, a full-time fire company--a pumping engine or ladder truck and a complement of men sufficient to man it around the clock--costs over $600,000 per year. The Fire Department wanted to make more efficient use of existing resources and to determine the most effective ways to add new resources. The simulation became the major tool for evaluating candidate policies.

We present here the results of simulation experiments in which three types of policies were tested:

(1) Dispatching policies: how many and which units of each type (engines and ladders) should be sent to each incident? With some exceptions, in the past the same response was sent to all alarms received by street box in all areas during all times of day. However, box histories showed that the probability that a given alarm signalled a serious fire varied greatly by time and area. With the simulation it was possible to observe the effect of varying the standard response.

(2) Relocation policies: given several units busy in an area (at one large fire or several small ones), which available units should be moved into empty fire houses in the area, when should they be moved, and which empty houses should they fill? The method being used by the Department was not designed to handle the increasingly common instances of several fires in progress in an area simultaneously.

(3) Allocation policies: how many units of each type (engines and ladders) should be located in each area at each time of day, and where should their houses be? Traditional practice in New York City and throughout the country is to have the same number of units on duty 24 hours a day, although, for example, the alarm rate in New York City between 8 p.m. and 9 p.m. is six times as high as between 5 a.m. and 6 a.m., and over 60 percent of the day's alarms are received between 3 p.m. and midnight.

The new dispatching, relocation, and allocation policies tested in the simulation were designed on the basis of the experience of fire officers and analysis by officers and members of the Institute's staff. The purpose of the simulation experiments was to compare the new policies to the traditional ones under controlled conditions to see how much better (or worse) the new ones would be expected to be.

II. MEASURES OF EFFECTIVENESS

A direct way to measure fire department performance is in terms of loss of life and property damage, but we made an early decision in the design of the simulation model not to attempt to do so. The relationships between deployment actions and these direct measures

were not known, and obtaining such measures seemed to be a long and possibly futile project. We felt that our immediate objectives could be met by using several easily calculated surrogate measures of performance such as the response time of units to fires. It developed that different policies could often be ranked in terms of performance in preventing life loss and property damage by examining their performance on the surrogate measures if one assumed only a monotone relationship between them and the direct measures.

By a surrogate or "internal" measure we mean an aspect of the Fire Department's performance which can be observed by watching only the Department's activity and not its consequences. Examples of internal measures are the number of units responding to incidents, the time it takes each one to arrive, the number of responses made by each unit, and the proportion of time each unit spends working.

In order to show the close interrelationship between direct and internal measures let us focus on one direct measure: loss of life. How would it be affected if the current policy were changed? In order to find this out from the simulation we would have to know, first, how the pattern of responses would differ under the new policy; and second, how this changed pattern would affect loss of life. The latter question is quite difficult to answer. On the other hand, the answer to the former is easy to obtain in a simulation, and this internal measure can be used as a surrogate for loss of life and loss of property. The following analysis shows how this may be done.

A useful measure of a policy at a particular incident is the vector of the response times of all units responding to it. The vector of response times for engines gives the time of arrival (relative to the time of alarm) of the first engine, second engine, etc. We divide the vector into two parts depending on whether the unit is an engine company or a ladder company. The vector for each type of unit is then put in order of unit arrival.

If, at a particular incident, one policy produces a response time vector every component of which is smaller than the corresponding component for another policy, and there are no other differences between the two policies, then it is clear that the former policy is as good or better for preventing loss of life, even though we do not know the precise relationship between response time and loss of life. We would not expect one policy to be better all the time, but we can aggregate the response time vectors by incident type and estimate the distribution of the response time vector for each type. (See Table 1 for one possible breakdown of incidents by type.) The empirical cumulative distribution function (cdf) for a given incident type can then be used to test whether one policy is better than another policy for response to that type of incident. If the grouping we are using for this analysis is those incidents at which arrival time is the

primary factor in determining if a life will be lost, then, if the empirical cdf for policy 1 is better than that of policy 2, policy 1 is the better policy; that is, it will result in as few or fewer lives being lost.

Determining whether the empirical cdf under policy 1 is better than that under policy 2 is generally not an easy task. For example, policy 1 may get the first ladder to incidents faster while making the second ladder response slower than policy 2. In practice we have generally focused on individual components of the vector of response times. For example, we have compared different policies with respect to

- o the distribution of first ladder (engine, arriving unit) response times to different types of incidents
- o the average first ladder (engine, arriving unit) response times to different types of incidents.

We use standard statistical tests to determine if differences in these response time measures between policies are significant.

Other internal measures of interest are coverage (the current proportion of alarm boxes for which nearby units are available) and workload (the number of responses made by fire-fighting units). They are discussed in detail in Section IV.

III. THE SIMULATION MODEL

The simulation package consists of three parts:

1. An input program. Given a probabilistic description of the fire demands in the area to be simulated, this program, written in SIMSCRIPT I.5, generates the set of incidents (exogenous events) which will occur in the simulation. The description consists of:

- o potential incident locations (we used alarm box locations);
- o for each alarm box location, the arrival rate of each type of incident and the proportion of each type which are reported by telephone;
- o a description of each type of incident, which includes the number and kind of companies required to handle the incident and the length of time each company is required.

2. The simulation program. This program, also written in SIMSCRIPT I.5, simulates the Department's response to a given set of incidents under a particular set of deployment policies. The program produces statistics on company utilization, workloads, response times, and coverage. It also produces output files that are available for later analysis. Usually, this program is run several times with the same input file but using different deployment policies.

3. Post-simulation analysis program. These are programs, usually written in FORTRAN, which measure the statistical reliability of the simulation output and make comparisons between simulation runs based on various measures.

The run time of the simulation program increases linearly with both the number of alarms processed and the number of times the simulation is interrupted to obtain samples of the state vector. Approximately 78 samples or 10.6 alarms

can be processed in one CPU second on the IBM 360/65. A typical run might contain 3,000 alarms and 500 samples and thus require 288 CPU seconds.

The simulation was designed to facilitate the process of making policy changes. To write a program which would be able to simulate many different policies without being sure in advance what these policies would be required the use of interchangeable sub-routines and a flexible data base. We present below a brief summary of the organization of the simulation program and a description of the data base which was used in our experiments.

Simulation Flow

The progress of an incident can be traced by following the series of event routines through which it passes. (In what follows the names of event routines will be written in capital letters.) The FIRE first breaks out and, some time later, at an instant predetermined on the input tape, the ALARM is turned in. The program--using a given dispatch policy--decides which companies to send to the alarm and schedules a DISP (dispatch) event to occur after a one-minute delay to allow for the alarm to be processed at the dispatching office and for the men to climb onto the apparatus. In the DISP event an arrival event (FARV) is scheduled for each of the dispatched units. The arrival time depends on the distance between the fire and the unit. The responding (and returning) units are assumed to travel at 20 mph, and a combination of right-angle and Euclidean distances is used to determine response distance.

The first of the FARV events to occur for a particular incident produces a CALIN (Call In) at which the first arriving unit "reports" the condition of the incident to the dispatching office. If too many units have been sent some are directed to return to their houses; that is, the FARV events of the excess companies are cancelled and HARV (House-arrival) events are scheduled for them. During their return home these companies are available for dispatch to other alarms. If the fire is a greater alarm fire, HARLM (higher alarm) events are scheduled, resulting in more DISP and FARV events until enough equipment is at the scene of the fire. The RELS (release from service) events are scheduled at times which depend on the arrival time of the company and the work time parameters found on the input tape. After release, companies proceed back to their fire stations, causing HARV events.

The details of the event routines and the output statistics have been tailored to New York City. However, metropolitan fire operations are sufficiently similar across the country that the basic structure of the simulation should be applicable to other cities.

Data Base

The Bronx, one of New York City's five boroughs, was the subject of the simulation experiments reported on here. Seven incident types were defined based on the number of units

required and the amount of time each unit would work. The incidents ranged from false alarms, which require only a short search by an engine and a ladder to assure that there is no fire, to third alarm fires, at which fifteen companies work for several hours each. For each type of incident work times were treated as constants, made equal to our estimates of mean work-time. The simulation also distinguishes between alarms received by telephone and those turned in from street boxes, since most dispatching policies depend on how the alarm is received. Table 1 lists the seven alarm types together with the number of units each requires and the other incident characteristics.

To reduce computer storage requirements, the 2,500 alarm boxes in the Bronx (roughly one on every second street corner) were gathered into 358 relatively homogeneous box groups. Each of these box groups is then simulated as if it were a single alarm box. The location of each group was defined as the centroid of the boxes composing it. The box groups were then assigned to one of two sets based on their location, dividing the Bronx into two disjoint regions. The buildings in Region 1 (the South Bronx) are older, the region is more densely populated, and it has a very high rate of fire alarms. Region 2 has a lower population density and fewer fire alarms. In both, about 40 percent of the street box alarms are false while less than 5 percent of the telephone alarms are false. Some other regional characteristics are:

Table 1. INCIDENT CHARACTERISTICS

Incident type	No. of units required		Average work times (mins.)		Percentage of all alarms in region (1968)			
					Region 1		Region 2	
	Engines	Ladders	Engines	Ladders	Box	Phone	Box	Phone
False alarms	1	1	5	5	24.2	2.3	12.6	3.0
Easy emergencies, non-structurals, and transportation fires	1	1	18	18	23.9	15.7	12.0	29.3
Hard emergencies and easy structural fires	1	1	18	18	13.1	17.1	7.4	32.1
Structural fires	2	1	75,45	60	1.02	1.08	.51	1.53
Structural fires	3	2	150,105,60	150,90	.60	.63	.30	.90
Structural fires	7	3	240,180,120,90, 90,60,60	180,135,105	.12	.13	.06	.18
Structural fires	11	4	360,300,270,240, 240,180,150,120, 120,90,90	330,270,180, 135	.06	.06	.03	.09
Total					63%	37%	33%	67%

Region 1

- o over 2/3 of the alarms in the Bronx
- o covers 1/4 of the borough's area
- o almost 2/3 of the alarms in the region are reported by street box

Region 2

- o less than 1/3 of the alarms in the Bronx
- o covers 3/4 of the borough's area
- o 1/3 of the alarms in the region are reported by street box

The exact proportions used in the experiments were based on analysis of 1968 incidents. A percentage breakdown of incidents by type for each region is included in Table 1.

Since modelling the incidence of box and telephone alarms for each type of incident at the location of each box group as independent Poisson processes yields good fits to the observed data (see [3]), we have used this Poisson assumption in generating alarms. For computational ease, we generate independent exponential random variables for the times between successive incidents in the entire borough. The type and location of each incident is then determined by matching random numbers to conditional probabilities. Specifically, for each incident, we let it happen at a particular alarm box with probability equal to the proportion of all 1968 Bronx incidents which occurred there. This location also determines the region in which the incident occurs. Given the region assignment, we let the incident be a box or telephone alarm with the probability appropriate to the region. The incident is then assigned a type using a probability appropriate to the region and how it was reported.

IV. POLICY ANALYSIS

In the last three years we have made many simulation runs to evaluate the effects of various deployment policies. Many of them have led directly to the implementation of new policies by the New York City Fire Department; others have indicated deployment changes which are now being considered. We describe some of these simulations in this section.

A. INITIAL DISPATCH AND ALLOCATION POLICIES

This series of simulation experiments considered a way of reducing the heavy workload being experienced by some companies, without either sacrificing fire-fighting effectiveness or making a large investment in new fire companies. The solution examined consisted of adding a small number of new full-time or part-time companies and modifying the dispatching policy to send fewer companies to some alarms. The potential locations for the new units were existing fire houses. The locations used were determined by practical conditions, such as space in the fire house for men and equipment (part-time units needing less), and the need for help, as measured by the workload of the current units.

New York City's dispatching policies employ alarm assignment cards, one of which is associated with every alarm box in the City (a sample card is shown in Fig. 1). The first half of the card lists the engine companies and ladder companies in increasing order of distance from the

3311	CRESTON AVENUE and 192nd STREET							BRONX
ENGINE CO'S	Marine Co.	Res. Co.	LADDER CO'S	D.C.	B.C.	Special Apparatus	Covering Assignment	COMPANIES TO CHANGE LOCATION 11/67
								ENGINE / LADDER
48 75 79			33 37	7	19		B.C. 15	
81 88 42			46				D.C. 6	50-75 38-79 / 49-33 32-37
43 46 62 95		3	38		18			41-46 90-62 67-95 /
92 45 68 93			27					83-92 94-45 80-63 59-93 / 19-27
82 71 60 69			36					96-82 35-71 53-60 40-69 / 34-36

Fig. 1. A Typical Alarm Assignment Card

alarm box. The traditional dispatching policy for box alarms is to send whoever is available of the first three engines and two ladders listed on the card for the box, "special calling" companies if necessary to assure a response of at least one engine and one ladder. The new dispatching policy to be tested, called adaptive response (AR), would send exactly two engines and one ladder to alarms reported from selected street boxes. (In the simulation, the change to AR required only the rewriting of the ALARM subroutine.)

We simulated because naive calculations of the effects of these proposals were inadequate and simulation would permit more precise calculations. For example:

(1) Under the traditional dispatching policy as few as one engine and one ladder might be sent to a box alarm because other units were unavailable. Therefore, the effect on workload of adding units without changing the response policy was hard to calculate. The same work would not be split among a larger number of units. Instead, more units would be available on the average, so the total number of responses would go up, not necessarily reducing any company's workload.

(2) Sending exactly two engines and one ladder under adaptive response might not reduce any company's workload either. In the case of engines, for example, even though some box alarms received three engines under the traditional policy some also received only one. Thus, it was even possible that adaptive response, at least during busy periods, might actually increase the total number of engine responses.

(3) The effect of adaptive response on response time was also unclear. If availability were increased, then reductions in first and second engine and first ladder response times would be expected. If a third engine or second ladder were needed at an incident at an AR box, one would guess that its response time would go up, since these alarms would always have to wait for the first arriving unit to request additional help. However, the overall average third engine or second ladder response time could end up being reduced since, even under adaptive response, telephone alarms which sound serious are dispatched

the full complement of three engines and two ladders. Reducing response to potentially less serious box alarms means a greater chance of having a nearby third engine available for a serious telephone alarm. Also, even though the initial dispatch of the third engine (when needed) at AR boxes is delayed by the amount of time it takes the first unit to arrive, the third engine will, on the average, be closer than under the traditional policy and might sometimes get to the fire faster.

We simulated the adaptive response policy for several different specified numbers of fire companies and four different alarm rates, 5, 13-1/3, 21 and 30 alarms per hour in the borough. These alarm rates roughly correspond to the average alarm rates for early morning, midday, evening and a peak evening. All runs at the same alarm rate used the same sequence of incidents (numbering about 2,000), so that true differences between policies were not obscured by their facing different alarm realizations. The results are given in Tables 2 and 3.

The most important interpretation of these results is that, at high alarm rates, the adaptive response policy apparently dominates the traditional one for ladders. That is, we see that at 30 alarms per hour and 12 ladders, the average time to first and second ladder both decrease, and the responses per hour per ladder decrease. (We say apparently because neither response time reduction is, by itself, statistically significant.)

For engines, we see that under adaptive response at either of the two high alarm rates, all three response times decrease (with a statistically significant reduction in second engine time). However, the average number of engine responses per hour increases, which implies that engine availability would be so low under the traditional policy that, on the average, fewer engines were dispatched than under adaptive response.

From Tables 2 and 3 we also note that under either the traditional policy or AR, adding new units has a greater effect on response times at high alarm rates than at low ones. Under the traditional policy, for example, reduction in the average first ladder response is about one second (.025 minutes) per ladder added at 5 alarms per hour; 5 seconds per ladder added at 13.5 alarms per hour; and 16 seconds per ladder added at 30 alarms per hour.

As we supposed, the workload reductions for busy units when companies are added under the traditional response policy turn out to be less than what might naively be expected. For example, for ladders at 30 alarms per hour, we have 2.072 responses per ladder per hour with 12 ladders. When three ladders are added, if the same work were to be redistributed, we would expect (12/15) x 2.072 or 1.658 responses per ladder per hour. However, the simulation results in 1.942 responses per ladder, indicating that the main effect of the new ladders on the original ones is to make them available to answer alarms that previously received one ladder or a ladder from outside the region.

Table 2. ADAPTIVE RESPONSE SIMULATION TEST
REGION 1 RESULTS: ENGINES

Bronx alarm rate (alarms/hr.)	No. of engines	Response times (mins.) to (without AR/AR) First engine	Second engine (when needed)	Third engine (when needed)	No. of responses/hr. per engine
5	18	2.30/	3.26/	4.32/	.533/
	19	2.30/	3.26/	4.28/	.474/
13-1/2	18	2.56/2.55	3.55/3.43	4.81/5.35	1.174/1.079
	19	2.53/2.52	3.53/3.39	4.79/5.25	1.136/1.028
	20	2.44/2.42	3.43/3.37	4.75/5.27	1.102/.986
	21	2.41/2.39	3.42/3.33	4.72/5.16	1.068/.943
21	18	2.92/2.89	4.47/4.07	6.13/6.05	1.649/1.657
	21	/2.62	/3.78	/5.80	/1.468
30	18	3.57/3.57	6.12/5.33	8.07/8.05	1.829/2.224
	21	3.13/3.10	5.05/4.62	6.76/6.75	1.940/2.041
Range of no. of incidents		1820-2208	50-78	25-31	
Range of raw std. dev. of indicated response times		.81-1.93	1.72-2.39	2.56-2.74	

Table 3. ADAPTIVE RESPONSE SIMULATION TEST
REGION 1 RESULTS: LADDERS

Bronx alarm rate (alarms/hr.)	No. of ladders	Response times (mins.) to (without AR/AR) First ladder	Second ladder (when needed)	No. of responses/hr. per ladder
5	12	2.58/	4.02/	.547/
	14	2.53/	3.81/	.480/
13-1/2	12	2.93/2.90	4.42/5.07	1.196/.969
	13	2.79/2.77	4.19/4.90	1.143/.900
	14	2.79/2.78	4.01/4.80	1.086/.846
	15	2.67/2.66	3.90/4.61	1.046/.794
21	12	3.47/3.44	5.67/6.12	1.678/1.473
	15	/2.99	/5.44	/1.238
30	12	4.45/4.37	8.11/7.94	2.072/1.927
	15	3.61/3.55	6.82/6.72	1.942/1.710
Range of no. of incidents		1820-2211	50-78	
Range of raw std. dev. of indicated response times		.97-2.07	1.94-3.41	

Tables 2 and 3 can also be used to compare the benefits derived from adding part-time companies to those derived from adding full-time companies. For example, assuming that a day consists of three 8-hour periods at each of the last three alarm rates (with these high alarm rates meant to correspond to what can be expected in the near future), we can compare adding one 24-hour unit to adding three that work only eight hours each evening. The average daily workload would be 35 responses per ladder and 40 per engine. Under AR, the reduction in average daily responses per region 1 ladder would be about 1.67 if one ladder is added around the clock (1.67 = 8 hours x $\sum_{i=1}^{3}$ (responses/hr/ladder in period i with 12 ladders - responses/hr/ladder in period i with 15 ladders)/(15 - 12) = 8 x (1/3)[(.969 - .794) + (1.473 - 1.238) + (1.927 - 1.710)]).

If three ladders were added in the evening, the reduction would be about 1.74 response per ladder per day (= 8 x (1.927 - 1.710)). A similar calculation for engines shows that the reduction in average daily responses for region 1 engines is about 1.35 for one full-time engine and about 1.46 for three evening only engines.

Overall, we see that by adding new companies and using adaptive response during the evening hours we can get both a reduction in company workload and an improvement in average response time relative to the traditional policy. Encouraged by the results of these simulation experiments, the New York City Fire Department adopted AR in part of region 1 in the evenings and added several part-time and full-time fire-fighting units in late 1969.

B. RELOCATION POLICIES

One aspect of deployment is the relocation of available fire companies to fill holes in coverage created when one large fire or several small fires are being fought simultaneously in a single area of the city. Currently in New York City the alarm assignment cards are used to specify predetermined relocations based upon houses made empty when companies are working at an alarm at a particular box. (See Fig. 1. The right-hand side of the card lists the relocations.) The relocations specified are based on the assumption that the alarm at that box is the only alarm in progress in the general area and, therefore, that each company specified to relocate is available to do so.

This method of pre-planned relocations breaks down when, as is an increasingly common occurrence, several incidents are in progress simultaneously in one area. An algorithm was developed [4] which replaces the system of predetermined relocations by a system which determines relocations based on current information on incidents in progress and current unit availability. It was designed for use in the Department's new on-line computerized Management Information and Control System.

We used the simulation to aid in designing the new algorithm and to compare the performance

of the new algorithm to the system currently being used. The input program was used to prepare a sequence of 3620 incidents covering a 180-hour period of constant high alarm rate--equivalent to three weeks of evening periods placed end to end. We did not want to look at low alarm periods of the day since few relocations would be required during these periods, and little difference could be seen between policies. Again, for control purposes, the same sequence of incidents was faced by both policies. The adaptive response policy described above was used to determine the initial dispatch to alarms.

We compared the results of these two simulations using three different measures of performance: coverage, workload, and response times.

Coverage

Fire is a random phenomenon, and since the Department cannot be sure where the next alarm will come from, it tries to position companies so that, no matter where the next fire occurs, there will be units available close by. The fire houses located throughout the city provide this protection when they are occupied, but, when fires are in progress, some houses become empty and the "coverage balance" is upset. Relocations are used to correct the imbalance. This is the most important reason for making relocations.

We measured the degree to which the current and the proposed relocation policies succeeded in providing adequate coverage by sampling, at 15-minute intervals, the proportion of alarm boxes which had at least one of their two closest ladder companies available and the proportion of boxes which had at least one of their three closest engine companies available. For coverage purposes the higher the proportions the better the relocation policy.

We found that the proposed algorithm was able to improve coverage considerably. Table 4 presents the empirical cumulative distribution functions for the coverage measure under each of the relocation policies.

Table 4. RELOCATION SIMULATION TEST RESULTS: EMPIRICAL CUMULATIVE DISTRIBUTION FUNCTIONS FOR COVERAGE

$F(x) = P$ (<x percent of boxes have at least 1 of 2 closest ladders available)

x	$F_C(x)$ (Current policy)	$F_P(x)$ (Proposed policy)
35	.000	.000
40	.001	.000
45	.006	.001
50	.007	.006
55	.011	.006
60	.024	.008
65	.043	.018
70	.063	.032
75	.096	.064
80	.152	.113
85	.257	.195
90	.423	.338
95	.622	.555
100	.944	.944

The empirical cdf evaluated at x is the proportion of the samples at which <x percent of the alarm boxes have at least one of their two closest ladders available. The lower the value of this proportion for a given value of x, the

better the policy that produced it is for coverage. Letting $F_C(x)$ be the empirical cdf from the current relocation policy and $F_p(x)$ be the empirical cdf from the proposed relocation policy, we see from Table 3 that

$$F_C(x) \geq F_p(x) \text{ for all } x.$$

These results indicate that, as far as coverage is concerned, the proposed relocation policy is at least as good as the current policy and may be considerably better. The proportion of the time that less than 90 percent of the boxes have at least one of their two closest ladders available drops from .42 to .34, a 19 percent decrease.

Response Time

In the proposed relocation algorithm coverage is the criterion used to determine which of the empty houses should be filled. To determine which of the available companies should be moved to fill those houses a response time criterion is used. Of course, coverage and response time are related measures; coverage representing latent response times. But they are far from equivalent.

The proposed relocation algorithm led to small but consistent improvements in response time. For example, we found a 1 percent reduction in average time to first arriving ladder and a 6 percent reduction in average time to the second ladder. (The simulation runs were too short for these and similar differences to be statistically significant.)

The behavior in the tails of the response time distributions is also important. It was hoped that the new relocation policy would reduce the probability of having large response times to fires. The simulation indicates that this will happen:

Let $G_C(x)$ = fraction of all alarms which have a first ladder response time $\leq$ x minutes using the current relocation policy

$G_p(x)$ = fraction of all alarms which have a first ladder response time $\leq$ x minutes using the proposed relocation policy.

We found that $G_C(x) \leq G_p(x)$ for all values of x.

The inequality also holds when we look at second ladder response times rather than first ladder. Differences between the two policies are greater than for first ladders and are often meaningful. For example, the proportion of alarms at which the second ladder arrives in more than 8 minutes was reduced from .15 to .08.

Workload

There is a wide disparity in workload among fire-fighting companies in New York City. There are areas of the city in which companies respond to over 6000 alarms a year, while several miles away other companies respond to fewer than 2000 alarms. Relocation, as an auxiliary effect, can help balance workload. Generally, the empty houses to be filled will be the houses of the busy companies. If a company's workload were considered in choosing companies to fill these houses, low running companies could be given increased work. The current relocation policy does not consider workload; the proposed one does.

The simulation showed that the proposed policy would lead to significant shifts in workload. For example, one busy company's workload was reduced 10 percent while a slow low running company's workload was increased by 17 percent.

After these initial tests of the proposed relocation policy were made, to determine if the policy would work at least as well as the present policy, we used the simulation again to test the proposed policy on a real scenario which represented the alarms received on one of the worst evenings ever experienced in the Bronx. No change was required in the simulation program to test this case; it was sufficient to bypass the input program and use the real data as input to the simulation program. We then compared the simulation's output with the actual results from that evening. The results are reported in [4].

C. OTHER ALLOCATION POLICIES

One approach to matching the fire-fighting resources on duty during a given period to the demand for their services is to create "part-time" companies which operate only during the periods of high alarm incidence. This approach was described above. An alternative approach is to use existing full-time units, but revise the firemen's work hours to have the on-duty periods for the two shifts (or platoons) of firemen overlap or run concurrently.

Currently one platoon of firemen works from 9 a.m. to 6 p.m. (9 hours) and the second platoon works from 6 p.m. until 9 a.m. (15 hours). Under the proposed work-chart (called the "concurrent two platoon" schedule) one platoon would work from 3 p.m. to midnight (9 hours) and the second platoon (manning a separate piece of apparatus) would work from 9 a.m. to midnight (15 hours). This schedule would place on duty in the concurrent company's house the following number of units over the day:

midnight-9 a.m.	0 units
9 a.m.-3 p.m.	1 unit
3 p.m.-midnight	2 units

Assuming there are nearby units manned in the usual way, this schedule provides a time distribution of units which closely matches the time distribution of alarms. For example, two units in an area, one a concurrent, would provide 1, 2, and 3 units on duty in the area in the three time periods.

The impact of such a reassignment of manpower was hard to predict without simulation, particularly if it were tried in conjunction with a change in the response policy for street box alarms (namely a change to the adaptive response policy previously described). Between 9 a.m. and 3 p.m. there would be no difference in workload or availability since there would be no change in the number of active companies. However, between midnight and 9 a.m. there would be fewer units on duty than before, and the increase in response time to alarms occurring during these hours would have to be compared to the reduction in workload and response times produced during the hours of 3 p.m. to midnight.

In addition, the magnitude of increase in workload for the concurrent company would have to be examined.

We ran the simulation using three different alarm rates:

- o 5 alarms per hour, representing the period midnight to 9 a.m. (period 1)
- o 10 alarms per hour, representing the period 9 a.m. to 3 p.m. (period 2)
- o 20 alarms per hour, representing the period 3 p.m. to midnight (period 3)

The traditional initial dispatch policy was used for rates of 5 and 10 alarms per hour and the adaptive response dispatch policy was used for 20 alarms per hour (since adaptive response had been shown to be most effective at high alarm rates and when extra companies were on duty). The number of ladders on duty in region 1 (the same high activity region as in the first set of simulations) was varied from 10 to 19 and the number of engines from 12 to 25. The simulation was run for 22 different combinations of alarm rate and number of units on duty. In each case a run length of 3620 incidents was used. The same sequence of incidents was used for the 5 and 10 alarms per hour runs (although the alarms were occurring faster at 10 alarms per hour). But, for 20 alarms per hour, a different probability distribution of incident types was used (since false alarms and rubbish fires represent a significantly higher percentage of alarms in the evening than at other times).

Results from 16 of the simulation runs are presented in Table 5. They may be used in several ways to develop allocation policies which match service to demand. For example, suppose we wish to determine the number of concurrent companies to create so as to minimize the average response time to all serious fires (fires requiring the services of two or more ladder companies). The rate of occurrence of serious fires is not the same throughout the day. The percentage of all alarms in region 1 which are serious, the average number of serious alarms which occur per hour, and the percentage of the day's serious alarms which occur in each of the 3 periods of the day are given below:

Period (i)	Serious Percentage	Serious alarms per hour	Percentage of day's serious alarms
1	5	.25	22
2	4	.40	24
3	3	.60	54

To evaluate alternative concurrent two-platoon policies we look only at allocations which leave 21 engines and 15 ladders in region 1 (the number currently located there) between 9 a.m. and 3 p.m., take away a certain number of units (n) between midnight and 9 a.m., and add this number of units to the region for the 3 p.m. to midnight shift. For example, one such policy (for n = 3) locates, respectively, 18, 21 and 24 engines in region 1 for the three periods of the day (which means transforming three full-time engine companies into concurrent two-platoon companies).

To determine the n which minimizes the average first engine response time to serious fires in region 1 we calculate, for each n:

Table 5. CONCURRENT TWO-PLATOON SIMULATION TEST
REGION 1 RESULTS

Alarm rate (alarms/hour)	Engines					Ladders			
	No. in region	Workload (responses/hour)	Response times (mins.) 1st E	2nd E	3rd E	No. in region	Workload (responses/hour)	Response times (mins.) 1st L	2nd L
5 (midnight-9 a.m.)	12	.720	2.41	3.80	4.90	10	.623	2.57	4.17
	14	.642	2.31	3.51	4.74	11	.577	2.47	3.99
	17	.538	2.28	3.51	4.33	12	.549	2.46	3.87
	18	.515	2.20	3.25	4.17	13	.515	2.40	3.72
	21[a]	.448	2.17	3.17	3.83	15[a]	.443	2.33	3.53
10 (9 a.m.-3 p.m.)	12	1.261	2.76	4.47	5.86	10	1.119	2.94	4.92
	14	1.152	2.58	4.10	5.39	11	1.050	2.78	4.67
	17	.996	2.45	3.85	4.80	12	1.018	2.73	4.44
	18	.956	2.33	3.61	4.71	13	.962	2.62	4.29
	21[a]	.849	2.27	3.43	4.16	15[a]	.838	2.47	3.84
20 (3 p.m.-midnight)	12	2.075	3.45	4.79	7.61	10	1.631	3.66	6.51
	14	1.958	3.15	4.45	6.58	11	1.541	3.43	6.16
	21[a]	1.464	2.50	3.62	5.31	15[a]	1.283	2.77	5.04
	24	1.308	2.35	3.32	4.82	17	1.158	2.66	4.75
	25	1.260	2.32	3.29	4.90	18	1.105	2.56	4.64
						19	1.053	2.50	4.50

[a]The number of units currently located in Region 1.

$$S(n) = .22R_1(21-n) + .24R_2(21) + .54R_3(21+n)$$

where $S(n)$ = the average first engine response time to serious fires with n concurrent engine companies

$R_i(k)$ = the average first engine response time in time period i with k engine companies located in region 1.

The values of $S(n)$ for n = 0, 3 and 4 are tabulated below:

n	S(n)
0	2.372 minutes
3	2.298
4	2.299

We see that creating 3 (or 4) concurrent engine companies would reduce first engine response time to serious fires by about five seconds.

The use of concurrent companies has another effect on response times, serving to reduce the wide spread in average response times over the day. For example, in the above case (n = 3 for engines) the spread in average response time to the third engine is reduced from 1.48 minutes (n = 0) to .66 minutes without seriously degrading response times during the early morning hours (the first and second engine response times during period 1 remain better than during either of the other periods).

The other major effectiveness measure which is affected by the creation of concurrent companies is workload. Continuing the example used above, suppose 3 concurrent engine companies are created. As a result, 18 engines remain on duty in period 1 with each one making more responses. The average number of runs made per engine company during this period is increased from 4.0 to 4.6. But, by having 24 units on duty in period 3

instead of 21 the average number of runs per engine company during this period is reduced from 13.2 to 11.8, lightening their burden during the busiest time of day, when they need it most.

Although it is interesting to look at the effects of concurrents on average workload, their principal usefulness is to reduce the workload of specific busy companies. The workload effect from creating concurrents will vary widely depending on which companies are chosen to become concurrents where they are deployed and what the response policy is. If the companies made into concurrents are low running companies, and they are stationed with high running companies, splitting their responses to alarms, then a better balance in company workload would be obtained.

The simulation can be used to assess the workload effect of concurrents on each individual company. In particular, Table 6 shows the distribution of work among the 21 engine companies in Region 1, and the effect of creating three concurrent engine companies. The average number of responses per hour has been tabulated for each company for each time period with and without the three concurrent companies. For each company the average number of daily responses has been calculated from the hourly averages. The three companies chosen to be made into concurrents are each co-located with another engine company at present, so no engine house was left vacant during period 1. Their partners are indicated by a single asterisk in Table 6.

The results indicate that several of the engine companies (principally those which gain a partner during period 3, indicated by a double asterisk in Table 6) would obtain a significant reduction in period 3 workload. For example, the workload of engine 10 in period 3 is reduced 32 percent, from 1.42 responses per hour to .96 responses per hour. Since it experiences only a small increase in workload in period 1, its average daily responses drop 17 percent, from 23 to 19. However, most of the 18 regular companies do not get so much relief.

The three concurrent companies become hard working companies. Since they do not work during period 1 and double up during period 3 their workload, which had averaged 14 percent less than that of the other 18 companies, becomes almost 16 percent higher than the new reduced average workload of the 18 companies. One concurrent company, which had been the lowest running company of all 21 units, becomes thirteenth lowest. The other two concurrent companies have their rankings increased from 10th to 17th lowest and from 12th to 18th lowest (or fourth highest).

The concurrent two-platoon system has not yet been implemented. There is a natural reluctance of the fire-fighters to change their working hours, which are now guaranteed by a provision in the state constitution. But, chances for implementation grow as the city's budget problems and demands for increased productivity grow.

Table 6. EFFECT ON WORKLOAD OF CHANGING 3 PERMANENT ENGINE COMPANIES TO CONCURRENT COMPANIES

Engine identification	Average number of responses/hour (without concurrents/with 3 concurrents)			
	Midnight-9 a.m.	9 a.m.-3 p.m.	3 p.m.-midnight	Daily average
Regular companies				
1**	.54/.55	.96/.96	1.48/1.08	23.93/20.46
2	.61/.61	1.06/1.06	1.73/1.68	27.39/26.95
3*	.43/.77	.87/.87	1.48/1.39	22.41/24.66
4	.35/.36	.70/.70	1.41/1.32	20.04/19.32
5	.35/.36	.70/.70	1.41/1.32	20.04/19.32
6	.53/.54	.99/.99	1.59/1.49	24.96/24.20
7	.34/.35	.69/.69	1.28/1.23	18.72/18.36
8	.34/.35	.69/.69	1.28/1.23	18.72/18.36
9*	.35/.62	.72/.72	1.48/1.33	20.79/21.87
10**	.51/.52	.96/.96	1.42/.96	23.13/19.08
11	.76/.77	1.31/1.31	1.97/1.99	32.42/32.69
12	.83/.82	1.39/1.39	2.03/2.02	34.09/33.99
13	.30/.31	.60/.60	1.12/1.09	16.37/16.15
14	.45/.46	.87/.87	1.56/1.42	23.31/22.14
15**	.66/.66	1.17/1.17	1.77/1.28	28.88/24.47
16	.37/.37	.74/.74	1.27/1.20	19.20/18.57
17	.37/.37	.74/.74	1.27/1.20	19.20/18.57
18*	.27/.50	.54/.54	1.11/1.04	15.66/17.10
Average: Regular companies	.46/.52	.87/.87	1.48/1.35	22.74/22.01
Concurrent companies				
19a	.27/--	.54/.54	1.11/1.04	15.66/24.12
b			--/1.28	
20a	.35/--	.72/.72	1.48/1.33	20.79/26.01
b			--/1.08	
21a	.43/--	.87/.87	1.48/1.39	22.41/26.37
b			--/.96	
Average: Concurrents	.35/--	.71/.71	1.36/2.36	19.62/25.50

*Presently one of the 3 companies to be made into a concurrent company is co-located with this company and shares its responses. When three concurrent companies are created one of them will be co-located with this company for 15 hours (9 a.m.-midnight) and will share its responses.

**When 3 concurrent companies are created one of them will be co-located with this company for 9 hours (3 p.m.-midnight) and will share its responses.

REFERENCES

1. Blum, E. H., "The New York City Fire Project," in Analysis of Public Systems, A. Drake, R. Keeney, P. Morse (eds.), M.I.T. Press, 1972.

2. Carter, G., and E. Ignall, "A Simulation Model of Fire Department Operations," IEEE-System Science and Cybernetics, Vol. 6, No. 4, October 1970.

3. Chaiken, J., and J. Rolph, "Predicting the Demand for Fire Services," The New York City-Rand Institute, P-4625, May 1971.

4. Kolesar, P., and W. Walker, "An Algorithm for the Dynamic Relocation of Fire Companies," The New York City-Rand Institute, R-1023, 1972.