A SIMULATION MODEL OF FIRE DEPARTMENT OPERATIONS: DESIGN AND PRELIMINARY RESULTS

Grace Carter and Edward Ignall

THE NEW YORK CITY RAND INSTITUTE
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.
A SIMULATION MODEL OF FIRE DEPARTMENT OPERATIONS: DESIGN AND PRELIMINARY RESULTS

Grace Carter and Edward Ignall

This study was sponsored by the City of New York. Its contents, however, do not purport to represent the official views or policy of the City of New York.

THE NEW YORK CITY RAND INSTITUTE
545 MADISON AVENUE NEW YORK NEW YORK 10022 (212) 758-2244
PREFACE

The simulation presented in this paper was developed as part of a study of the deployment of fire-fighting resources, which is itself a portion of the New York City-Rand Institute research efforts with the Fire Department of the City of New York (FDNY). Other tools in the deployment study which complement the simulation are in various stages of development.

This paper can be viewed as both a presentation of results and a progress report. On the one hand, the simulation is already accepted by the FDNY as a valuable tool, having been used extensively in the Department’s effort to cope with an alarm rate that is almost double that of five years ago. The results of simulation experiments have been used by the Department in reaching a recent decision to add units that work only in peak demand hours, in locating these units, and in changing the number of units sent to an incident during peak hours. The results of some of the simulation experiments which played a part in this decision are given and analyzed in the paper. However, we are still integrating the prescriptions from the analytical models and refining the representation of internal features, particularly travel times and fire growth.
SUMMARY

This Memorandum describes a simulation model designed to compare different policies for locating, relocating, and dispatching fire-fighting units. Simulation experiments were made to compare proposed solutions to workload and response problems to the then current policy in New York City. The proposed policies involved creating new units, some of them to work only in peak alarm rate hours (Tactical Control Units) and reducing the nominal number of fire-fighting units dispatched to street box alarms during those hours (Adaptive Response). We report in this paper a simulation experiment in which one new policy reduced both the total number of responses made by all engines and the average time from first report of the fire to the arrival on the scene of an engine and, if needed, a second and a third engine. This and other simulation experiments are credited with assisting in the adoption and implementation of TCU's and AR in New York City in late 1969.

Several issues concerning the design and use of the simulation model are described. The first of these is the use of internal measures of performance, such as average time from first report to arrival of an engine, in place of global measures, such as loss of life. The second is the integrating of insights gained from analytical models of the various sub-problems (initial dispatch, relocation, etc.) into policies to be tested by simulation. The third is the use of "virtual" measures, a Monte Carlo-like technique for efficient handling of loss of life and other important but rare events.

Aspects of the operations of the Fire Department of the City of New York and their translation into SIMSCRIPT I.5 program are described. In addition, programs that provide input to that simulator and analysis of simulation output are described. Statistical analysis is made of the reliability of the differences in time required for arrival of the third engine under a proposed policy and the current one.
ACKNOWLEDGMENTS

Our knowledge of fire, and thereby most of the essential parts of the simulation, come from the officers and fire fighters of the New York Fire Department. They suffered our presence, worked with us closely for months at a time, and willingly provided data and shared experience. We are especially grateful to Assistant Commissioner P. M. Canick, Chief of Department J. T. O'Hagan, and Chiefs L. J. Harris, L. M. Snyder, F. J. Ronan, and J. J. Cunningham.

The development of the simulation model was very much a joint effort. We were guided and sustained by advice from and discussions with our project leader E. H. Blum, and R. W. Archibald, J. M. Chaiken, L. W. Miller, A. J. Swersey, and W. E. Walker of Rand, and Chiefs O'Hagan, Harris, and Ronan. G. S. Fishman helped us with spectral analysis.
CONTENTS

PREFACE ................................................................. iii
SUMMARY .............................................................. v
ACKNOWLEDGMENTS ................................................ vii

I. INTRODUCTION ...................................................... 1
   Some Background on New York City Fire
   Department Operations ......................................... 2
   Generating Policy Options .................................... 6

II. DESIGN ISSUES ..................................................... 9
   Global vs. Internal Effectiveness Criteria .................... 9
   Edge Effects ...................................................... 11
   Rare Events—Virtual Measures ................................. 12

III. DATA NEEDS AND SOURCES ................................... 14

IV. THE SIMULATION MODEL ........................................ 18
   The Input Program ............................................... 19
   The Simulator .................................................. 19

V. RESULTS AND POST-SIMULATION ANALYSIS ................... 22

REFERENCES .......................................................... 33
I. INTRODUCTION

This paper describes the development of a simulation model for the Fire Department of the City of New York (FDNY). The model was designed as a general research tool for studying aspects of urban fire department operations. It is primarily intended for evaluating and comparing policies that differ with respect to --

1. how many units of each type (pumping engine, ladder truck, etc., each with its complement of men) to send to each incident;

2. which units to send to each incident;

3. given many units busy (at a large fire or several small ones), whether or not to relocate other units into empty fire houses in the area -- and if so which units to relocate where;

4. where to put a new fire house, and where to put the existing ones if we could move them.

To complement the simulation model, analytical models are being developed of these four policy components, each treated as an isolated, and idealized, problem.

In designing the simulation, we were guided primarily by two desires: to understand the complexity of fire department operations, and to be able to compare actual policies, which affect matters of life and death, without having to await full development and use of the analytic work. In the longer term, we expect the basic insights to come from the analytic models. Yet, it was clear to us from the outset that simulation would be the only means that would allow prescriptions from these models to be integrated and tested in realistic situations, while yielding feedback for refining the assumptions of the models.

How effective the simulation is in these roles depends on the choices made in modeling the system: how the system's features are represented, the level of detail, etc. An important early decision was a choice not to represent loss of life and fire damage in all its dimensions explicitly in the simulation. We saw the futility of trying to get these ultimate measures -- global measures -- in the model quickly. The relationship between
deployment and loss of life and damage was not known, and obtaining it was clearly a long-term project. Our immediate objectives could be met by substituting several internal measures of performance -- for example, workloads of different units and response times to incidents. And, we soon realized that different policies could be ranked in terms of performance in preventing life loss and damage by examining their response-time characteristics.

SOME BACKGROUND ON NEW YORK CITY FIRE DEPARTMENT OPERATIONS

The following brief sketch of Fire Department operations in New York City should help the uninitiated reader understand some of the problems facing the City, while illustrating the need for some of the features of our simulation model. We begin tracing some aspects of the system by considering a fire which is first reported to the Fire Department by telephone and which turns out to be very serious. We do some injustice, in the details, to how the FDNY actually operates.

Suppose the fire starts in an apartment whose tenants are out. Sometime later a tenant upstairs smells smoke and telephones the Fire Department. The call is received at the appropriate borough dispatching office and the caller is questioned about where the fire is, what is on fire, etc. To permit pre-planned responses, the number of the closest fire alarm box must be determined; ordinarily, this will be within a few hundred feet of the building. A card -- the alarm assignment card -- has been prepared for each box, listing the closest engine and ladder companies, chiefs, and special units in order of their distance from the box.

Using a street index file, the dispatcher finds the alarm box nearest the street address given and obtains the alarm assignment card for that box; this process takes about twenty-five seconds. The first line of the card lists the units which are to respond to the initial alarm. Usually there will be three engines, two ladders, and one chief. The dispatcher checks the availability of the engines and ladders on the first line of the card. If at least one engine and one ladder are available, he "sends the box" over the telegraph system to the fire houses.
The signal received in the houses is the number of the alarm box. By referring to its list of assigned boxes, each company knows whether or not to go, and also where to go. The dispatcher also telephones the closest engine and ladder to give details on the fire -- actual address, apartment number, etc.

At some point after fire units get to the scene, the officer in charge may decide that more help is required and ask that a second alarm be transmitted. The dispatcher checks the availability of the companies on the second line of the alarm assignment card and then sends the second alarm signal to the houses. In reaction to this signal, usually four engines, one ladder, and a chief will go to the fire, and three engines and two ladders will relocate into the houses of some of the seven engines and three ladders now at the fire. As the fire progresses, a third and then a fourth and a fifth alarm may be transmitted before the fire is brought under control. All in all, about 25 engine and ladder units would be at a five-alarm fire at its peak, with about 15 more relocated into the area. The companies which arrived first may work five to ten hours; the average for all 25 might be two or three hours.

A fire of such severity would be unusual; fewer than 1 percent of all incidents require more than three engines and two ladders. In 1968, some 27 percent of all incidents were false alarms; another 35 percent were rubbish fires, car fires, and the like; 18 percent were emergencies (smoky incinerators, leaking gas, etc.); and only 20 percent were fires in structures. On a false alarm, the companies would be back in their quarters about ten minutes after they left; on most rubbish fires, all but one engine would return to quarters right away, and that engine would stay at the fire for about five to twenty minutes. On some minor incidents which are reported by telephone, only one unit would respond.

The proportions of incidents of different kinds vary from one part of the city to another, and within any one geographical region they change with time of day, day of week, and season of the year. For illustration of the variations, we consider two regions that are within a few miles of each other in the Bronx. For the two regions, an approximate breakdown of incidents by number of engine companies required and number nominally dispatched is given
in Table 1 for incidents initially reported by telephone and by alarm box. The proportions for incidents requiring two or more engines have been smoothed, and no really large fires are included. The figures are averages over the entire year.

To illustrate the varying time-of-day effects, consider the periods midnight to 8:00 a.m. and 8:00 a.m. to 4:00 p.m. in the two regions. In the first region, more of the box alarms turned out to be structural fires in the later period -- 16 percent and 21 percent; in the second region, more were structural fires in the earlier period -- 26 percent and 13 percent.

More striking is the time-of-day effect in total alarm rate. In 1968, averaged over the entire year, there were six times as many incidents between 8:00 p.m. and 9:00 p.m. as there were between 5:00 a.m. and 6:00 a.m. On the more active days, this ratio rose to roughly ten.

All of these characteristics of the demand faced by the Department suggest that tailoring the allocation of fire-fighting resources to the time and place more than is done under current practice may be quite helpful.

The need for such tailoring has been emphasized by a steady and substantial growth in the number of calls for service over the past years, coupled with modest change in the number of units. There were 227,000 incidents to which the Department was called in 1968, almost double the 116,000 in 1963. In the same period of time, the number of engine companies increased by 8, from 215 to 223. Further, this increased total workload is not evenly distributed. In 1968, for example, three neighboring engine companies, located within a circle one mile in diameter, made 9100, 6100, and 3900 responses.

All of the above figures are for New York City. The analogous data for other cities can be quite different. In particular, we know that the ratio of box to telephone alarms is higher in New York than in most cities and that sending three engines and two ladders is anything but standard; some cities send more, others send fewer.
<table>
<thead>
<tr>
<th>Incident Type</th>
<th>Number of Engines Nominally Dispatched</th>
<th>Percentage of All Alarms in Region</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Box</td>
<td>Phone</td>
</tr>
<tr>
<td>1 False alarm</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>2 Rubbish fire, etc.</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>3 Minor structural fire, emergency</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4 Structural fire</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>5 Structural fire</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>6 Structural fire</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>7 Structural fire</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td><strong>TOTAL:</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
GENERATING POLICY OPTIONS

The primary use of the simulation model is to examine the effect of different allocation policies on the effectiveness of the FDNY. By an allocation policy, we mean a specification of --

1. how many units of each kind (engine, ladder, chief, etc.) there are and their home locations;

2. a rule for deciding how many, and which, companies to dispatch to each alarm;

3. a rule for deciding when and how to relocate the available companies;

4. the amount of time spent at incidents;

5. the rate at which incidents occur.

To the extent that the last two specifications can be controlled by the Department, they are affected mainly by its supervision, prevention programs (building inspections, education, etc.), and technical innovations. With the simulation, we seek to improve the first three components of the allocation policy, given the last two.

Simply stating this definition of an allocation policy raises a particularly important question: Where do we get the allocation policies to try out in the simulation? If we compare a group of policies that are all inferior to some other policy, we will just find the best of a bad group. How can we get a good group to compare?

The initial basis for comparison is the current policy of the FDNY. The first alternative class of policies considered are the modifications of it that have been suggested within the Department. They reflect a distillation of accumulated experience and intuition. As such they also reveal, through the aspects of the policy that are emphasized, what the Department considers to be its problems. For example, many of the policies under consideration in the Department reinforced the conclusion that there was much concern about the increased workload that has accompanied the rising alarm rate and about the concomitant reduction in "coverage" -- having units available to handle the next incident.
The second class of alternative allocation policies have been derived from the analytical models of different subproblems. They treat the following:

1. Which units to send, given the number to be sent: This has been analyzed by means of a many-server queueing model. The analysis indicates that not sending the closest units to selected locations can serve to both equalize workload among units and reduce average response time. In simple situations it prescribes exactly which units go to which boxes if average response time is to be minimized.

2. How many units are sent under the current policy: This has been analyzed by means of a queueing model in which the probability of sending the full dispatch is reduced as the number of units available decreases.

3. How many units to send to a particular incident, without regard to the number of available units or likely future demands: This required a decision analysis involving the assessment of fire chiefs' utility functions.10

4. The number of units working in a region (and hence the need for relocation): This has been analyzed by means of a queueing model which predicts the frequency and duration of periods when more than n units will be working.11

5. Relocations: Integer programming and Markovian decision process formulations are being employed.12,13 Small problems are being solved exactly and, to reduce computations, large ones are being solved approximately.

These models are both prescriptive and descriptive. With the former, their prescriptions can be implemented in several ways, in both the simulation and the real system. First, direct use may be possible. For example, the Department may change the alarm assignment cards in the way that the first model mentioned above advises. Second, it may be possible to imbed an algorithm in the system. For example, if the dispatcher has a computer terminal, he can have a linear programming model of the relocation problem solved at those times when relocation should be considered. Third, if there is no direct prescription and the algorithm requires too much computation, it may be possible to find approximations or heuristics which will imitate the prescriptions well and are easy to implement.

With the descriptive models, one can begin evaluating different policies. For example, consider one traditional method for relieving workload: to establish new units near, or even in the house of, the hard-working units.
There were conflicting theories on the likely effect of this policy, and the disappointing actual experience was confounded with the rising alarm rate. By using first the model that predicts how many units are sent and then the simulation, this situation was clarified. The actual result is that during peak hours the new unit immediately becomes busy, and it is used more for filling out responses than relieving other units. In our view, the importance of this analytical work to the potential successful use of the simulation model cannot be overrated.

OTHER RECENT STUDIES OF EMERGENCY SERVICES

There has been a recent upturn in quantitative approaches to fire and other emergency services. There have been simulations of ambulance and police patrol car operations that are similar in scope to ours. In a study of ambulance service in New York, Savas, and Gordon and Zelin use simulation in treating questions of where to base the ambulances and to which hospital the patient should be transported. In Larson's study, simulation is used to compare different policies for the initial deployment of patrol cars and for their assignment to incidents.

The most noteworthy work in fire is the methodology for locating fire houses that has been developed and applied in Great Britain by Hogg. The use of simulation has been suggested by others, including Rockett and Weitz. Rockett describes the construction of a simulation model of the evolution of "a building fire from ignition to extinguishment." Weitz outlines a simulation which, if realized, would encompass fire prevention, social and economic impact of fire, etc. in addition to fire fighting.
II. DESIGN ISSUES

In this section we discuss several considerations that influenced the design of the simulation model. Continuing the treatment of global vs. internal effectiveness criteria, we describe their interrelationship and construct one useful internal criterion. We discuss a way of simulating a particular region while retaining its interaction with the remainder of the City. Finally, we discuss techniques for obtaining statistically reliable information about loss of life and other important but rare events.

GLOBAL VS. INTERNAL EFFECTIVENESS CRITERIA

Some aspects of Fire Department performance can be observed by watching only the Department's activity, and not its consequences. Examples of this are the number of units responding to incidents and the time it takes each one to arrive; the number of responses made by given units and their work times; etc. We will call these internal measures of performance.

In contrast, global measures have to do with the impact on others of what the Fire Department does. Some global measures are perceived as directly related to the quality of fire service -- the number of people who lose their lives in fires, the damage to buildings and their contents. Other global measures are not so apparent; for example, the effect of a burned-out building on both the lives of the former occupants and the overall housing situation in the City.

Regardless of whether or not the global problem is addressed, the Department would find it very useful to compare allocation policies in terms of internal measures. For one thing, the FDNY is very much interested in reducing the strain on its men and equipment that is implied by retaining traditional policies in the face of the rising demand. Another reason, more important here, is the interrelationship between global and internal measures. An example is the Department's concern with keeping a sufficient number of units available in order to have a small number of instances of either reduced response (fewer than the desired number of units responding) or response by distant units, since either can lead to loss of life or escalation of small fires into large ones.
Let us focus on one global measure: loss of life. How would it be affected if the current policy were changed? In order to find out from the simulation, we 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. It is the second part of this investigation that is so formidable. On the other hand, the first part is easy to obtain in a simulation, and this internal measure can be used as a surrogate or proxy for loss of life and fire damage in the following manner.

First, we observe that different types of units perform different functions at a fire, depending mostly on the equipment they carry. Thus, at a particular fire, a ladder company may be able to rescue a person who could not be saved by an engine company. So we certainly have to distinguish between types of units. Moreover, two units working together may be able to take some action which neither could perform alone; therefore, it is important to know the number of each type of unit on the scene at each moment.

On the other hand, there are not such large differences between individual units of the same type that we need to distinguish between them. That is, we may assume that the loss of life at the fire would be identical whether engine company number 1 arrived two minutes after the alarm, with engine company number 2 a minute later, or number 2 arrived at the two-minute mark, with number 1 a minute later.

Thus, for estimating loss of life, we need only observe the vector of response times for engines, ladders, etc. The vector for engines tells the time of arrival (relative to the time of alarm) of the first engine, second engine, etc.

If, at a particular incident, the new policy produces an engine response time vector, every component of which is smaller than the corresponding component for the old policy, and there are no other differences between the two policies, then it is clear that the new policy is as good or better for preventing loss of life, even though we don't know the precise relationship between response time and loss of life. In any event, if we save the response time vectors for each incident simulated, we have a record that is sufficient for analyzing loss of life and fire damage.
We would not expect one policy to be better all the time, so we need
some way of aggregating response time vectors. Suppose that like incidents
were grouped together. For example, let the grouping be those incidents where
the time at which the first ladder arrives is the primary factor determining
whether or not a life will be lost. If \( V_1, \ldots, V_m \) are the response time
vectors of these incidents, we can use them to estimate the distribution of \( V \),
a random variable that is the response time vector to this type of incident
under the given policy. Now if
\[
F_1(x) = P(V \leq x \text{ under policy 1})
\]
\[
\geq P(V \leq x \text{ under policy 2}) = F_2(x) \quad (1)
\]
for every \( x \), then policy 1 is better than policy 2: it will result in as few
or fewer lives being lost, assuming our classification of like incidents is
valid. We can use the observations \( V_1, \ldots, V_m \) under each policy to estimate
\( F_1 \) and \( F_2 \) and to see if (1) is a statistically reasonable conclusion.

What we have said up to this point may suggest that the simulation can
completely ignore the effects of the arrival times of the units -- it cannot.
In those cases where longer response time leads to a larger fire, more units
have to be committed to the incident, and for a longer time. To ignore this
is to impair the validity of simulated response times for the incidents which
occur soon after the one in the question. While estimating this "commitment"
penalty is difficult, it is far easier than estimating damage or loss of life.
Our efforts in this direction are described in Section III.

**EDGE EFFECTS**

In designing the simulation, a key question is the size of the area to
be represented. The FDNY has over 350 engine and ladder units, and simulta-
neously representing the activities of all of them implies a burdensome,
slow-running simulation. While the regions in the city that pose major
problems are relatively small, any group of fire companies that is singled
cut for study is not independent of the others. Since interaction is greater
near the border of the area we wish to study, we term our problem one of
edge effects.
The approach we have taken to minimize these edge effects is the following: First, we distinguish an area that we wish to study and the companies inside it. Then, when an incident occurs inside this area, we follow it in detail; we determine which units to send from the allocation policy and measure all the response times. Incidents outside the area are also treated, but followed in much less detail. The identities of the companies that respond are fixed, regardless of the allocation policy inside the distinguished area, and response time is not measured. The DOWN event is used to treat these outside incidents, as shown in Section IV. We keep track of various workload measures for companies both inside and outside the area. However, we are chiefly concerned with the workloads of the inside companies.

**RARE EVENTS - VIRTUAL MEASURES**

As we have mentioned, two of the most important goals in fire department operation are preventing loss of life at fires and maintaining adequate protective "coverage" at times when serious fires tie up many companies. Certainly the department's performance in these situations is the focus of attention from within and without. Yet these kinds of events are relatively rare. There was loss of civilian life at barely one-tenth of one percent of the 227,000 incidents in New York City in 1968; several serious fires are in progress simultaneously only a few times a year, and these intervals when very few units are available account for less than one percent of the hours in a year. We were thus faced with the prospect of making long simulation runs and getting only a few observations on these rare but important events.

For example, in comparing the new policy to the existing one, it would be reasonable to wish for a 90 percent chance of selecting the new policy if it reduced fatalities by 5 percent, and a 90 percent chance of retaining the existing one if the new one were no better. It is straightforward to show that this would require a sample size of about 500,000 incidents.*

*See section 7.5.3 of Ref. 14.
One possible way of avoiding these large sample sizes is to use what we call virtual measures of performance. Such measures are akin to the variance reduction techniques employed in Monte Carlo and should seem at least vaguely familiar to one conversant with those techniques. Although our investigation of virtual measures is incomplete, we feel they are of sufficient promise to mention them here. Virtual measures depend on the state of the system at any given time, but not on what happens at the actual incidents. For example, we could interrupt the simulation at some point in time and ask what is the expected number of fatalities from alarms reported within the next five minutes? This number clearly depends on --

1. the probabilities of incidents of various types at each of the boxes in our area;

2. which companies are now available; and

3. which of them the allocation policy would send to each incident if it occurred.

Note that we do not cause any of these incidents to occur, we only ask what if each of them did; hence, the term "virtual."

Calculation of the expected number of fatalities each time the simulation is interrupted would be very time-consuming, since it requires summation over all the alarm boxes. An alternative that is more promising is simply to record which companies are available each time we interrupt the simulation. (Relocated companies count as the company they are filling in for.) Numbering the companies 1, 2, ..., N, we can treat this as a vector; call it \( \mathbf{a} \). We define \( a_j = 1 \) to mean \( j \) is available and \( a_j = 0 \) to mean it is not. From the point of view of any particular box, the whole coverage or availability vector is not of concern: the state of the five engine houses and three ladder houses closest to the box should be sufficient to get expected loss of life at the box. Several boxes would have the same eight companies so that there might only be on the order of a few hundred "company groups" of interest in the Bronx (which has 57 engines and ladders). Each group would have \( 2^8 \) possible states, so that to get statistically reliable results on the distribution of its availability vector we should have perhaps \( 50 \times 2^8 \) or 13,000 observations on it. In order that these observations be approximately independent, there should be several incidents between each sampling of the coverage vector. Estimating "several" as 4 would mean 50,000 incidents. Since this is a tenth of 500,000, the potential saving appears to be substantial.
III. DATA NEEDS AND SOURCES

In order to run the model, accurate data are needed on --

1. the rate of occurrence of various kinds of incidents at various locations;

2. how many units work for how long at the various kinds of incidents;

3. travel times of the units;

4. the effect of response time on commitment of units at the incident, and on fire damage and loss of life.

We list in this section some actual and potential sources for the data and our efforts so far to tap them. The sources include --

1. the report on every incident made by the battalion chief who is first to arrive, and a punched card summary of the report;

2. the individual units' records of activity;

3. the Department's files on fatal fires;

4. observations and surveys made with the FDNY;

5. a tape that lists the geographic coordinates of every alarm box.

The punched card summaries of the chiefs' reports include date, time, box number, how the incident was reported, type of incident, cause, number of engines (ladders) responding and number working, the chief's time at the incident, etc. With respect to the type of incident, the breakdown is quite fine. For a structural fire, the type of building, its occupancy, and where the fire started are all included on the card. These summaries are available on magnetic tape for all incidents since 1962 (with minor exceptions) and have formed an enviable basis for analyzing and modeling incident occurrences. The full reports, copies of which are filed centrally, include a description of fire-fighting operations and identify all the units involved.

The individual unit records in its journal the box number, a general classification of incident type, time it left its quarters and returned, and the time, if any, spent working at each incident. These journals are retained by the units. They form the basis for our modeling of "how many units work for how long" at each type of incident. It should be clear that the various
times associated with a single incident are not independent. For example, on a structural fire requiring three engines and two ladders, the first-arriving engine and first-arriving ladder usually work about the same length of time; the second engine to arrive usually works less time than the first engine, but longer than the second ladder. Consequently, reconstructing incidents is necessary. Since this requires data that must be collected from records available at the individual fire stations, only about one hundred have been reconstructed. These were selected from chiefs' reports to cover a range of incident types, with emphasis on large fires. In addition, for incidents where only one or two units work and the internal coherence of the set of work times is not a real issue, several hundred individual observations of work times have been collected from unit journals.

With respect to travel time, we have not been so fortunate. At present, time spent traveling to and from incidents is not recorded by the units. To represent travel times accurately requires knowledge of: the distance to be traveled; the area of the city in terms of general traffic conditions, hills, etc., and within each area the time of day; and the weather. The work to date has concentrated on comparing policies under typical conditions. As a result, only exploratory measurements have been made. Travel times have been recorded in some 33 instances of units responding to incidents in lower-income residential areas in the evenings. From knowledge of the unit involved and the incident, actual road distance, Euclidean distance, and right-angle distance -- distance between two points \((x_1, y_1)\) and \((x_2, y_2)\) computed as \(d = |x_2 - x_1| + |y_2 - y_1|\) where the axes are in the direction of the streets in the area under consideration -- have been computed. All three are highly correlated; the travel times vs. road distance are given in Fig. 1, along with the 25 mph line for comparison.

There are over 14,000 street alarm boxes in New York City -- about 80 per square mile in heavily populated areas, 43 per square mile overall. The tape that lists the coordinates of the alarm boxes has been prepared by using a digitizer on maps that have alarm box locations printed on them. This tape makes it easy to represent any portion of the City accurately in terms of distance. In addition, actual incidents from history or a set generated probabilistically can be simulated easily.
FIG. 1 - TRAVEL TIME
In discerning the effect of response time on commitment of units at the fire, our initial goal is an escalation model that will be reasonable enough to assure the validity of response times to future fires. To this end, we have begun to study some 200 fires at which at least two engines and two ladders worked. The chief who had been first to arrive was interviewed soon after the fire and asked for a detailed opinion on what would have happened if the actual response times had been different. An initial hypothesis is that a certain fraction of fires are just not liable to escalation, and the others will escalate if response time is sufficiently unsatisfactory. It appears to be a reasonable approximation to use a threshold criterion on the response time vector to define "sufficiently unsatisfactory" and to have the fire escalate to the next larger incident type, if the threshold is exceeded. And finally, prospects for relating loss of life to response time are much enhanced by the existence of a centrally-filed, detailed report on each fatal fire.
IV. THE SIMULATION MODEL

Our simulation package consists of the following:

1. An input program. It generates the simulated incidents (including fires) that the Department faces.

2. The simulation. Responses to this set of incidents are then simulated for a particular allocation policy. This program produces statistics on company workloads, response times, and coverage. It also produces output files that are available for later analysis. Usually this program is run a number of times, using different policies but the same input file.

3. Post-simulation analysis programs. These programs measure the statistical reliability of the output generated by the simulations and make comparisons between simulation runs.

The input program and the simulator were written in version 1.5 of SIMSCRIPT. At the beginning of the project, GPSS and SIMSCRIPT were the only simulation languages available at our installation. We chose SIMSCRIPT because of the ease with which the program could be divided into changeable subroutines (for the decision rules) and the ability to write FORTRAN compatible output files for use in postsimulation analysis. The faster running-time under SIMSCRIPT reinforced our choice.

It is difficult to say precisely how much programming effort was required, because experiments were conducted with limited early versions of the program, and thus the programming effort was not continuous. The time spent in actual coding and design was between one-half and one man year. The amount of time spent in analyzing the output and designing experiments has already exceeded that figure and is still increasing.

The run time of the simulation program increases linearly with both the number of alarms processed and the number of the samples of the coverage vector obtained. At each sampling of the coverage vector, some additional computations -- described later in this section -- are made. Approximately 78 coverage samples or 10.6 alarms can be processed in a CPU second on the IBM 360/65. A typical run might contain 3000 alarms and 500 samples of the coverage vector and thus require 288 CPU seconds.
THE INPUT PROGRAM

This SIMSCRIPT 1.5 program converts a probabilistic description of the fire demands in the area to be simulated to a file which specified the characteristics of each alarm to be processed by the simulator. The probabilistic description consists of:

1. Potential locations for incidents — alarm boxes;
2. For each incident location, the arrival rate of each type of incident and the proportion of each type of incident reported by alarm box and telephone;
3. For each type of incident, a description which consists of distributions of —
   a. number and kind of companies required to handle the incident,
   b. length of time each company is required,
   c. delays that do not depend on the availability of companies, such as that from fire outbreak until the alarm is turned in.

This framework allows us to adjust the level of detail to suit the part of the policy that is of concern. For example, to test decision rules for "who to send," simulating a relatively small area and identifying each alarm box individually is appropriate, while to test rules for relocation, simulating a large area (with about 30 or more units) and aggregating individual neighboring alarm boxes into groups of 5 or 10 is appropriate. In the same vein, to compare policies on the basis of the distribution of company workload, we need distinguish incidents only by the number of units required and how long they work and not by whether the incidents are fires or emergencies, etc.

THE SIMULATOR

The events that appear in the simulator are as follows:

1. FIRE — represents the start of the fire.
2. ALARM — the dispatching office is told of the fire.
3. DISP — companies are told to go to the fire (dispatch).
4. FARV — a company arrives at the scene (fire-arrival).
5. CALIN - the dispatching office receives an initial report on the state of the fire from the men at the fire (call in).

6. HALRM - a higher alarm is turned in.

7. ACC - an accident occurs to a traveling vehicle.

8. RELS - a company is released from service at a fire, and starts home.

9. HARV - an arrival at a fire station (house-arrival).

10. AVAIL - a company completes its recovery after a fire, and is again available.

11. DOWN - a company is taken out of service (see Section II).

12. INSRV - a company that was involved in a DOWN event, or an ACC, is restored to service.

13. SMPCV - sample the coverage vector for the present state of the system.

Relocation of companies can be triggered by the DISP event, the CALIN event, or any event which makes companies available. It is one of the causes of the event HARV. The FIRE and DOWN events are scheduled on the input file prepared by the input program. All other events are scheduled endogenously by the simulator.

The progress of an incident through the system can be traced as follows: the fire first breaks out, and at a time predetermined on the input tape, the alarm is turned in. The program uses the allocation policy to decide which companies to send to the alarm and schedules a DISP event. The time between an ALARM and DISP event depends on how many alarms are currently being processed by the dispatcher. At the DISP event, arrival at the incident (FARV) is scheduled for each of the companies selected to respond. For each of these companies, the travel times are computed on the basis of the distance to be traveled (see Section III). This distance need not be from a fire station to the incident, since the company may already be traveling (returning from another incident, relocating, etc.).

The first of the FARV events to occur produces a CALIN, which reports the condition of the incident. If too many units have been sent, the FARV events of the excess companies are canceled and HARV events are scheduled for them. During their return home, these companies are available for dispatch to other alarms. If the fire is a serious one, and not enough equipment has been dispatched, another DISP event will be scheduled. If the fire is a greater
alarm fire, HARLM events are scheduled, resulting in more DISP and FARV events until enough equipment is at the scene of the fire. The RELS 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. For most fires, the companies are available for further dispatch at their RELS events. However, if the fire was serious, they may require a period of recovery back at the fire stations, and AVAIL events are scheduled.

The simulator produces statistics on company workload, response times, and availability. The workload measures, currently being computed for each company, are number of responses made, number of incidents worked, and time spent in traveling, at work, in recovery, etc. Response times are measured as a vector time from alarm to arrival of first engine, second engine, first ladder, etc. These vectors are aggregated by location and type of fire to produce means, standard deviations and histograms. At times when the coverage vector is sampled, some coverage measures are computed. The first of them is the probability that, if an incident were to occur just then, the first engine to arrive at it would be the one listed first on the alarm assignment card. The second measure is the probability that the first engine to arrive would be either the first or the second on the card, etc. These measures then reflect both which units are available at the time and the relative distribution of incidents in the area.

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 in other cities.
V. RESULTS AND POST-SIMULATION ANALYSIS

In an effort to reduce workload and to maintain coverage, the FDNY conceived several alternatives to its then current allocation policy. Of those which held the most interest, one proposed to --

1. establish new units to work mainly during the hours of peak demand;

2. send two engines and one ladder to incidents reported from street alarm boxes during those hours;

3. treat each new unit and the busy one with which it would be paired as a two-section company, meaning that both would never go to the same incident (if both were engines closest to an incident, and two or more engines were desired, only one of them would go).

Let us go through some of the rationale underlying this policy and the questions about it that were answered by simulating it. It has been customary in paid fire departments in the United States to have around-the-clock companies and to staff them uniformly throughout the day. However, the idea of matching resources more closely to demand by having evening-only companies was clearly not new; what was new was that with the simulation we had a way of evaluating. It would seem clear that three units working only the peak eight hours of the day would relieve the workload of existing units more than one around-the-clock unit would. However, as the queueing model described in Section III revealed, it is precisely during the peak hours that new units are drawn more to "filling out" responses (that otherwise would have been less than 3 and 2), than to relieving the original units. Consequently, the issue was not so clear.

Similarly, it seemed clear that, as a result of adding new units and modifying the nominal response to street box alarms, the total number of responses per unit would go down. And, the concept of the two-section company seemed to ensure that more relief would go to the busy unit. What the total amount of relief would be and how in detail it would be distributed among the units, however, were not clear.

Because there would be more units and each original unit would be available more of the time, the first two engines and the first ladder should arrive sooner on the average. What would happen to the response time of the third
engine and the second ladder was quite unclear: street box alarms certainly could not even get them dispatched until a unit was on the scene; on the other hand, both street box and telephone alarms often did not get them on initial dispatch under the original policy anyway, and now telephone alarms would have more chance of getting them initially.

In answering these questions, separate runs were made for engines and for ladders, and to be specific we concentrate here on the engine simulations. The incident types were distinguished on the basis of the number of companies each allocation policy would nominally dispatch and the number that would work, leading to fourteen types of incidents. For each type, work times were treated as constants, made equal to our current estimates of mean work-time vectors (Table 2). We should note here that variability in work-time seems to be at most a second-order effect. It has been proved for the queuing model which treats the question of which units to send, that workload and response time depend on the work-time distribution only through its mean.

We present here some results used in actual policy deliberations. When these policies were being considered, key substantive parts of the model had been completed, but the geographical representation was still that used originally: a square 6 miles on a side, divided into 36 squares by a 6 by 6 grid. (The tape that lists alarm box coordinates was not yet available when these runs were made.) At the center of each of these squares is an engine company house. In addition, 28 engine companies are located around the outer perimeter of the area. Each of the 36 squares is also subdivided into 9 smaller squares by a 3 by 3 grid; the centers of each of these 324 smaller squares were the potential incident locations — the hypothetical alarm boxes (Fig. 2).

The engines are assumed to travel at 20 mph, and right-angle distance is used. There is assumed to be a one-minute delay to get a message through the dispatching office and start a unit on its way.

The area is divided into regions of high, medium, and low demand, as indicated by the shading in Fig. 3. The ratio of total alarm rate per box in the three regions was assumed to be 20:5:1. The proportions of the various incident types are given in Table 2; they are modeled on two of the battalion districts in the Bronx. The total alarm rate was 16.7 incidents/hour, roughly comparable then to a summer evening in the Bronx.
Table 2

CURRENT AND PROPOSED POLICIES: Differences in Number of Units Dispatched and Characteristics of Incidents Faced

<table>
<thead>
<tr>
<th>Incident Type</th>
<th>Number of Engines Nominally Dispatched</th>
<th>Typical Work Times (Min)</th>
<th>Percentage of All Alarms in Region</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Current</td>
<td>Proposed</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Box</td>
<td>Phone</td>
<td>Box</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

TOTAL: 70% 30% 37% 63%
FIG. 2 - COMPANY AND ALARM BOX LOCATIONS

FIG. 3 - AREAS OF HIGH, MEDIUM AND LOW DEMAND
The allocation policy compared to the then current one adds four new engines, one at each of the four houses in the high-demand region. Within each of the resulting two-section companies, each of the two units makes an equal number of responses, although this is not necessary. How the two policies differ in the number of engines nominally sent is shown in Table 2. Under both policies, relocation of units occurs on second and higher alarms.

The input program prepared a sequence of incidents covering a 240-hour period, equivalent to an "end-on-end" month consisting of evenings. The incidents were generated assuming -- as sketchy data appear to verify -- that the various incident types at the various boxes are realized as independent Poisson processes. In these particular runs, there were 4,141 incidents in the resulting sequence. The same sequence of incidents was faced by both policies.

Using the same incident sequence, rather than using different and independent sets of incidents for each policy, has significant advantages in this kind of simulation. In general, when the same set is used, any difference in the way two policies perform is plainly a result of intrinsic differences between them. It is not confounded with their facing different situations, as in the alternative procedure, and in most "real" experiments. This ability to confront each of the policies with the same incidents in the same places at the same times is in fact one of the biggest virtues of digital simulation. In particular, to compare two policies as effectively using different incidents would require almost four times as many observations; according to our calculations that follow, the variance of the estimator of the mean difference between two policies in time-to-third-engine when n incidents are simulated is approximately 260/n when the same set of incidents is used, and is approximately 1,000/n if different sets are used. Consequently, n must be almost four times as large in the latter case to achieve the same variance as in the former.

In these runs, responses by the engines inside the area totaled 7,940 under the current policy and 7,213 under the proposed one, with the four new engines making 1,454 responses and the original 36 making 5,754. To examine
the way the reductions were distributed under the proposed policy, we
classified the original 36 engines into four groups:

1. Four original companies in the high-incidence area.

2. The eight companies in the medium-incidence area which
were closest to the high-incidence area.

3. The other four companies in the medium-incidence area.

4. The twenty companies in the low-incidence area.

The average number of responses under the two policies are displayed by group
in Fig. 4, where one box in each group is highlighted.

Average response times are given in Table 3. The times to second engine
and third engine shown are only for those incidents that needed them, since
the times are irrelevant in other cases. (Virtual response times to second
and third engine could have been calculated but, in this set of runs, were not.)
The improvements observed in time to first and second engine under the proposed
policy were expected. The improvement observed in time to third engine is
more interesting, since there was no obvious prediction. The next question is
whether or not this improvement is statistically significant.

We first look at the times to third engine under each of the policies
considered individually. Histograms, which summarize the response times to
third engine for the 49 incidents that needed three or more engines, are given
in Figs. 5 and 6. Letting \( T_1, \ldots, T_n \) and \( U_1, \ldots, U_n \) be the times under the
current and proposed policies, respectively (\( n = 49 \) here), we calculate
\[
\bar{T} = (T_1 + \cdots + T_n) / n = 7.26,
\]
\[
\bar{U} = 6.57,
\]
\[
\frac{S_T^2}{T} = \frac{(T_1 - \bar{T})^2 + \cdots + (T_n - \bar{T})^2}{(n - 1)} = 7.3,
\]
\[
\frac{S_U^2}{U} = 4.3.
\]

Now if the \( T_i \) are independent, \( \frac{S_T^2}{n} \) is a good estimate of \( \text{Var}(\bar{T}) \). It seems
reasonable to suppose that they are independent, for the average time between
fires that require three or more engines is 4.8 hours, and the average duration
of such events is 3 hours. Considering the geographical dispersion of these
FIG. 4 - AVERAGE NUMBER OF RESPONSES PER ENGINE COMPANY, BY GROUP
Table 3
RESPONSE TIME CHARACTERISTICS FOR THE TWO POLICIES

<table>
<thead>
<tr>
<th>Policy</th>
<th>First</th>
<th>Second</th>
<th>Third</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Policy</td>
<td>3.21</td>
<td>5.27</td>
<td>7.26</td>
</tr>
<tr>
<td>Proposed Policy</td>
<td>2.71</td>
<td>4.48</td>
<td>6.57</td>
</tr>
</tbody>
</table>

FIG. 5 - ORIGINAL POLICY: TIME TO THIRD ENGINE

Average Time to Third Engine when needed is 7.26 mins.
FIG. 6 - PROPOSED POLICY: TIME TO THIRD ENGINE
alarms, and that the average engine at work at such fires stays much less than 3 hours, there is good reason to suppose that our samples of these response times are independent. To test this hypothesis, the standard test of runs above and below the median was used on one of the simulation runs. If our samples came from an independent process, one would expect 25.4 runs with a standard deviation of 3.4. The observed number of runs was 23, within one standard deviation of that expected. Hence, we accepted the hypothesis that the observations of time to third engine were independent.

If the $T_i$ had been times to first engine, they would not be independent of each other. Times to first engine at succeeding incidents are positively correlated. During periods when incidents turn out to be heavy, response times to a sequence of incidents will all tend to be higher than average, since the closest company is, in each case, more likely to be unavailable; conversely successive response times tend to be below average during quiet times. As a consequence of the positive correlation, $S_T^2/n$ would be an underestimate of $\text{Var}(\bar{T})$ and could lead to, for example, unjustifiably narrow confidence intervals. This problem is well known, and we have used the approach of Fishman. 17 We calculate the spectral density function of the series $T_1, \ldots, T_n$ and use the spectrum at zero to estimate $\text{Var}(\bar{T})$. For time to first engine under the original and proposed policies, these variances are .0025 and .0009 and the standard deviations are .05 and .03.

Returning to the situation at hand, if $\bar{T}$ and $\bar{U}$ were independent (that is, if the two policies faced independent sets of incidents), an estimate of the standard deviation of $(\bar{T} - \bar{U})$ would be $\sqrt{(7.3 + 4.3) / 49} = .49$. Now $(7.26 - 6.57) / .49 = 1.4$, so that, assuming normality of the difference, the difference would appear to be significant at the 10 percent level but not at the 5 percent level. However, the two policies were run against the same sequence of incidents. In Fig. 7 we show the histogram of the differences in response time. The large fraction of zeroes -- usually meaning the same engine was dispatched under both policies -- implies that the individual differences $D_i = T_i - U_i$ are not normally distributed. However, $D$ should be approximately normal and the standard deviation of $D$ is approximately $\sqrt{S_D^2/49} = \sqrt{3.3/49} = .25$. Since $(7.26 - 6.57)/.25 = 2.8$, we see that, in fact, the difference is significant at the 1 percent level.
Thus, on the basis of these two simulation runs, the new policy could be expected to:

1. Reduce the total number of responses made, with the bulk of the reduction going to those units which originally responded most.*

2. Improve the distribution of the response time vector, component by component, and thus reduce both loss of life and fire damage.*

The policy adopted and now being put into practice by the Fire Department differs somewhat from the one described above. For example, not all street alarm boxes are designated to receive two engines and one ladder, and those that are so designated receive exactly two-and-one, as opposed to the possibility of one-and-one under a nominal two-and-one dispatching policy. In addition, the areas affected received around-the-clock units as well as peak-hour units.

The results of the runs described, and of other runs that treated related policies, have been credited with assisting the policy choice. As used perceptively and imaginatively by those involved in setting the policy, these results appear to have permitted a better decision to have been made than would have been made without them. In the hands of an interested and willing department, they appear to have helped resolve uncertainties that had troubled promising policies, and to have provided new bases on which to build future intuition.

Because important decisions obviously could not wait for the model to be finished, the simulation used for these runs was not complete. We saw at the outset, however, that incomplete versions of the simulation would have to be used, and to be relied upon. Therefore, we staged its development to focus early on gaining insight, testing representations, and validating policy-oriented conclusions, even at the expense of both comprehensiveness and detail. We thus felt that the conclusions yielded by the incomplete model could be safely used in formulating policy. Determining the extent to which these conclusions hold in detail, awaits both simulation runs using the "real" geography of the City and the new policy and the Department's actual experience, and will be reported in another paper.

*Since all incidents received the number of units needed, usually more quickly than before, the responses eliminated were all unneeded. Thus, the new policy could be expected both to enhance the actual service provided and to reduce needless running.
REFERENCES


Rand maintains a number of special, subject bibliographies containing abstracts of Rand publications in fields of wide current interest. The following bibliographies are available upon request:

Africa • Arms Control • Civil Defense • Combinatorics
Communication Satellites • Communication Systems • Communist China
Computing Technology • Decisionmaking • East-West Trade
Education • Foreign Aid • Health-related Research • Latin America
Linguistics • Long-range Forecasting • Maintenance
Mathematical Modeling of Physiological Processes • Middle East
Policy Sciences • Pollution • Procurement and R&D Strategy
Program Budgeting • SIMSCRIPT and Its Applications • Southeast Asia
Systems Analysis • Television • Urban Problems • USSR
Water Resources • Weather Forecasting and Control

To obtain copies of these bibliographies, and to receive information on how to obtain copies of individual publications, write to: Communications Department, Rand, 1700 Main Street, Santa Monica, California 90406.