Using Simulation to Develop and Validate Analytic Models: Some Case Studies

EDWARD J. IGNALL and PETER KOLESAR
Columbia University, New York, New York

WARREN E. WALKER
The Rand Corporation, Santa Monica, California

(Received March 1976; accepted August 1977)

Simulation models are generally costly tools to use in systems analyses. Whenever applicable, a simple analytic model is preferable. However, in many cases the conditions assumed by solvable analytic models do not hold in the real world; hence an analyst would hesitate to use them. A simulation can be used to suggest an appropriate approximate model and to determine how good an approximation a given analytic model is. We show how simulations of New York City's fire and police operations have been used to develop and validate simple analytic models that are now being used to analyze the deployment of resources in these two services.

A SIMULATION model of a large and complex system can be a very useful, but time-consuming and costly tool. Whenever applicable, one prefers to use a simple analytic model yielding closed-form algebraic expressions relating system inputs and outputs. However, in many cases the simplified conditions assumed by solvable analytic models do not hold in the real world, and more realistic models are too complex to solve—hence simulation. The standard use of simulation is direct: to answer a specific question or to obtain a description of the behavior of a system as some of its parameters are changed. In contrast, we illustrate here the use of simulation to confirm that a simpler model may safely be used to describe system behavior and even to suggest the form of such a simpler model. If the analytic model provides an adequate approximation, it can be used more economically than the simulation for future analyses.¹

A preliminary version of this paper was delivered at the 1974 Winter Simulation Conference held in Washington, D. C., January 14th-16th, 1974, and appeared in the proceedings of the Conference.

¹ Note that we will be skipping over the question of whether the simulation is an adequate (valid) representation of reality. We assume that this has already been established. The particular simulation models discussed below have already been tested for “face” validity [19].
Paraphrasing Hamming [9] and Geoffrion [8], we might say, “The purpose of simulation is insight, not numbers.”

Use of simulations to develop and test other mathematical models is conceptually analogous to use of experiments by physical scientists to develop new theory. It is this sometimes overlooked use of simulation models that we focus on through the use of four examples.

The examples are drawn from studies carried out by The New York City–Rand Institute for the New York City Fire and Police Departments. The analytic models are:

1. A model for analyzing police patrol car allocation problems, where the simulation and analytic models were constructed in parallel. One of the chief reasons for building the simulation was to determine how well and under what conditions the analytic queuing model agreed with it (and thus with the real world).

2. A model for estimating fire company response distances, where the analytic model was developed well after the simulation was written. Special simulation runs were made to confirm the validity of the analytic model.

3. A model for predicting the number of fire companies dispatched to an alarm, where the analytic model was suggested (and verified) by an analysis of simulation runs that had been made years earlier for other purposes.

4. A model for estimating the number of fire companies that will be busy at alarms, where the model was developed before the simulation was written but was verified several years later using the results of simulations that were run before the validation effort was begun.

These examples are discussed briefly in Sections 1–4; detailed descriptions are given in [10]. A concluding section briefly discusses the cost and other advantages of using analytic models in these cases.

1. ALLOCATING POLICE PATROL RESOURCES

In [17] a queuing model is proposed to represent the patrol activities of a police command. A patrol car is dispatched immediately to answer a call for service if one is available; otherwise, the call is queued at the dispatching center. Queued calls are served in order of their assigned priority. What is desired is a way of relating queuing delays to $N$, the number of cars assigned to the command, so that $N$ can be chosen rationally.

A queuing model that might be used in this situation is the simple $M/M/N$ priority model of Cobham [6]. It assumes that calls arrive according to a stationary Poisson process, that service times are independent and exponentially distributed, and that each call is served by a single
patrol car. These conditions are not all satisfied in the operating environment of the New York City Police Department (NYPD). While call arrivals are approximately Poisson, call rates, even during a single 8-hour tour of duty, are not constant. Service times are not exponentially distributed and include the time required for a car to travel to an incident, which depends on the number of cars available to dispatch. Moreover, some calls are served by more than one patrol car.

As a result, although we wanted to use the simple $M/M/N$ model to analyze deployment options for the NYPD, we first had to verify that, despite the above-mentioned variations from the model, it still produced predictions of sufficient accuracy. To make appropriate tests, we wrote a detailed police patrol simulation of a single police precinct [16]. The simulation included the complexities mentioned above as well as others, and used actual call histories in the precinct for arrivals and service times. We compared simulation results to those obtained from the queuing model with the same average call rate and the same average service time. The results, described below, were sufficiently close to give us and the Police Department confidence that the queuing model could be used instead of the simulation model.

Based on the call rate, average service time, and number of servers, the queuing model gives the probability distribution of the number of calls being serviced and the number waiting to be dispatched. From this distribution one can obtain $q$, the probability that all $N$ patrol cars are busy, and $\bar{D}$, the average time a call will spend in queue before being dispatched (and other performance measures).

To test this model, we used it to calculate $q$ and $\bar{D}$ as functions of $N$ and then compared the results to those obtained from the simulation model. One NYPD precinct, the 71st Precinct in Brooklyn, was chosen for study because a rich set of data on its operations was available. We considered each of the three shifts or “tours” worked by the policemen: tour 1—midnight to 8:00 a.m.; tour 2—8:00 a.m. to 4:00 p.m.; tour 3—4:00 p.m. to midnight.

The average service time was approximately the same for all tours. The queuing model was used to analyze conditions for tours 1 and 3 with different numbers of cars on duty. The simulation was then run for these values of $N$, using as input a historical stream of calls for a given tour. We used the actual time each call was received and its actual service time, location, and priority. (The input stream for the simulation of a given tour was prepared from computerized records maintained by the NYPD by concatenating all of the calls received during that tour during July and August 1972. For example, when simulating tour 3, the last call before midnight on one day would be followed by the first call after 4:00 p.m. on the following day.)
A comparison of the results from the simulation and the queuing model is given in Figures 1 and 2. Figure 1 shows the percent of time that all patrol cars are busy. The results are remarkably similar, with the queuing predictions being consistently slightly lower than the simulation re-
sults. (We predicted this difference because the simulation makes multiple car dispatches while the queuing model assumes that one car is sent to each call.) Figure 2 plots the average queuing delay as a function of the number of patrol cars on duty. The results, again, are quite close. Because of the type of applications we had in mind, we were particularly interested in finding out whether the queuing model would predict the “elbow” in these functions—the point at which the curves begin to rise steeply and performance begins to degrade badly. It appears to do so.

The Police Department accepted the fact that the queuing model represents a reasonable approximation to the dispatching and service activities of the patrol force. The model has been imbedded in a computer program called the Patrol Car Allocation Model [5], which the Police Department has begun to use as an aid in determining the number of patrol cars to assign to duty during each tour in each police precinct.

2. ESTIMATING FIRE COMPANY RESPONSE DISTANCES

Kolesar and Blum [14] derived an inverse square-root relationship between the average distance traveled by fire companies responding to calls in a region and the number of locations from which they respond. The relationship was derived under idealized conditions: an infinitely large region in which the units are located either uniformly on a grid or purely at random, and in which calls for service are distributed homogeneously in space, while emergency vehicles travel along simple response paths. However, to have practical usefulness, it was important to show that the relationship provided a reasonable approximation to actual average response distances under more realistic conditions.

An existing simulation of New York City fire fighting operations was used to test the validity of the model for fire company responses. The simulation program is documented in [1], its design and development are described in [2], and its use in policy analysis is described in [3]. Two versions of the square-root relationship were to be tested by simulation:

1. The expected response distance of the closest available unit \(ED(N)\) to a fire alarm occurring when there are \(N\) available companies in a region of area \(A\) is given by

\[
ED(N) = k \sqrt{A/N}. \tag{1}
\]

2. If there are \(n\) companies located in a region of area \(A\) and if, on the average, \(b\) are busy, then the average response distance for the first-arriving company to alarms in the region is given by

\[
\bar{D}(n) = k \sqrt{A/(n-b)}, \tag{2}
\]

where \(n-b\) is the long-run average of \(N\) and \(\bar{D}(n)\) is the long-run average of \(ED(N)\). Since \(n\) is a major policy variable under management’s control, this relationship is of more general interest than (1).
Verifying these relationships by gathering empirical data from Fire Department operations would have been an extremely difficult task. In fact, verifying (2) would have required the Department to vary the number of units operating in an experimental region at different times—an unthinkable procedure, especially if the changes meant using so few companies that lives and property were endangered. Instead, by means of simulation these tests were able to be made safely and economically without modifying Fire Department operations.

To validate and test the two relationships, seven runs were made with the model of fire fighting operations in the Bronx (Table I). In each case the alarms were distributed probabilistically among 358 locations accord-

<table>
<thead>
<tr>
<th>Simulation No.</th>
<th>Alarm rate (alarm/hour)</th>
<th>No. of active ladders</th>
<th>No. of active engines</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>31</td>
<td>37</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>31</td>
<td>37</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>24</td>
<td>37</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>20</td>
<td>37</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>20</td>
<td>37</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>12</td>
<td>37</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>12</td>
<td>37</td>
</tr>
</tbody>
</table>

ing to actual 1968 alarm patterns. The engine locations were those actually being used in the Bronx in 1971. The ladder locations in the 12-, 20-, and 24-ladder cases were subsets of the 27 actual ladder locations then. In the 31-ladder case, 4 additional ladder locations were added in the south Bronx in carefully chosen places. Each simulation consisted of an extended time period during which the alarm rate and number of active units were unchanged. The simulation durations were chosen so that, in each case, about 3,500 alarms were handled. This sample size was selected after statistical analysis of the random variation in simulation output statistics. The results produced should be interpreted as estimates of performance of steady-state behavior under the conditions simulated.

Relationship (1)

First, we consider the validation of relationship (1) between $ED(N)$ and $N$. The simulation program recorded the response distance and the number of companies of each type (engines and ladders) available at the instant of dispatch for each alarm. These data were accumulated separately for two regions of the Bronx; the south Bronx (a small region with a high incidence of alarms) and the rest of the Bronx (called the north Bronx). The data were collected for the closest engines and ladders to each alarm,
as well as for the second and third closest units for those alarms to which such units were dispatched.

Each simulation run provided data for an independent test of relationship (1). Figure 3 is a plot of data from a typical simulation. The curve was fitted to the data by nonlinear regression. The closeness of the curve to the observed data points and the closeness of the estimated exponent to \(-0.500\) suggest the validity of the square-root hypothesis. Reference [14] provides results obtained from other simulation runs that confirm

\[ D = \text{Average Response Distance (miles)} \]

\[ D = 1.83N^{-0.501} \]

* The Average of n Observations

**Figure 3.** Simulated average response distances vs. the number of available companies (closest engine and ladder companies in the south Bronx).

the validity of the square-root model and describes the regression in detail.

**Relationship (2)**

We now turn attention to validation of relationship (2) between long-run average response distance, \(\bar{D}\), and \(n\), the number of companies assigned to the region. The results discussed above indicate that the square-root model describes the relationship between average response distance and number of units available when an alarm occurs. However, this does not assure that a square-root law describes the relation between long-run average response distance and the average number of companies available
to respond to an alarm. On the contrary, if the square-root relationship holds for the former, it cannot hold exactly for the latter since the inverse square-root function is convex, and for a convex function \( f(\cdot) \) of a random variable \( X \), \( Ef(X) > f(EX) \) (Jensen's inequality). We wanted to determine whether the square-root relationship provides an adequate approximation.

Figure 4 displays the simulated long-run average response distance for closest ladders versus average numbers of ladder units available for south Bronx ladders. The data were generated by the seven simulations listed in Table I. In this case each of the plotted points represents the results of an entire simulation run. Figure 4 contains regression fits of two functions:

\[
\bar{D} = k \left( \text{average number of available ladders} \right)^{1},
\]

\[
\bar{D} = \alpha \left( \text{average number of available ladders} \right)^{\beta}.
\]

The closeness of the fitted curves to the data points, the near coincidence of the two curves and the closeness of the estimated exponent \( \beta \) in the more general model to \(-0.500\) all confirm the validity of the model. Reference [14] provides additional details of these tests.
Uses of the Square-Root Model

The distance that responding fire units travel to reach fire alarms is an important measure of the service being provided by the Fire Department. By being able to predict the average response distance in a region as a function of alarm rate, the number of units assigned to the region, and other measurable parameters of the region, allocation policies can be evaluated quickly without the use of simulation. Among the many questions that the Fire Department of New York has already used the square-root model to answer are:

1. What will be the effect on response distance of removing a company from a region?
2. How should the number of units on duty be varied over the day (as the alarm rate varies) to maintain a given average response distance in a region?
3. How many fire units will be required in the future under projected alarm rates to maintain desired average response distances?

The time units spend traveling to reach alarms is another important measure. A travel-time model, based on the square-root model for distances and the results of a field experiment in which fire company responses were timed, has also been developed and tested in the simulation by Kolesar [13].

3. PREDICTING THE NUMBER OF UNITS SENT TO A FIRE ALARM

In New York City, dispatching rules for an incoming alarm are governed by the “alarm assignment card” for the fire alarm box closest to the alarm. The first line of an alarm assignment card contains the names of the three closest engine companies and the two closest ladder companies. The traditional policy for alarms turned in by box had been to send whichever first line companies were available, and “special call” companies further down on the card if necessary to assure a response of at least one engine and one ladder. As a result of this policy (which we call a “New York 3” dispatching policy for engines and a “New York 2” policy for ladders), the number of engines and the number of ladders actually sent to a box alarm were random variables between 1 and 3.

We were concerned with predicting how the actual number of units dispatched depended on the average unit availability in the surrounding region. By analyzing simulation runs that had been made for other purposes, we derived a simple relationship between the number of units sent and the average unit availability when a “New York” dispatching policy is used [11].

In this section we discuss the use of simulation, deriving and verifying
this relationship for the New York 2 policy for ladders. In similar ways we have derived and tested relationships for other dispatching policies.

Let "a" be the average fraction of the time a ladder company is available. Define $P$ as the probability that 2 ladders are dispatched to an incident under the New York 2 policy during a period in which the average availability is $a$. A good fit to the simulation data was found by using:

$$P = a^2$$  \hspace{1cm} (3)

TABLE II
DERIVATION OF RELATIONSHIP BETWEEN NUMBER OF LADDERS DISPATCHED AND AVERAGE AVAILABILITY

<table>
<thead>
<tr>
<th>South Bronx</th>
<th>North Bronx</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed availability</td>
<td>Observed and predicted (from availability) percent of NY2 alarms receiving 2 ladders</td>
</tr>
<tr>
<td>.952</td>
<td>86.8/90.7</td>
</tr>
<tr>
<td>.951</td>
<td>78.2/90.4</td>
</tr>
<tr>
<td>.877</td>
<td>75.7/76.9</td>
</tr>
<tr>
<td>.875</td>
<td>75.0/76.7</td>
</tr>
<tr>
<td>.871</td>
<td>73.5/75.9</td>
</tr>
<tr>
<td>.869</td>
<td>72.5/75.6</td>
</tr>
<tr>
<td>.776</td>
<td>52.7/60.3</td>
</tr>
<tr>
<td>.666</td>
<td>41.3/44.3</td>
</tr>
<tr>
<td>.623</td>
<td>33.8/38.8</td>
</tr>
</tbody>
</table>

This relationship would be true if:

1. The average availability were the same for every ladder company in the region;

and

2. The event that any particular ladder company is available were independent of the status of all other ladder companies.

Neither of these conditions is true, yet the relationship appears to be a good approximation. In fact, the validity of relationship (3) was discovered in the course of attempting to see how poor $a^2$ is as an estimate of $P$.

The relationship was developed from a set of nine simulation runs that had originally been made to test the effects at different alarm rates of both adding new companies to an area and modifying the dispatching policy (see [3]). The simulation model mentioned in Section 2 was used. For each simulation run we calculated the fraction of time each ladder company
was available. We then obtained the average \( a \) of these availabilities over all the ladder companies in the region. For all incidents in the region for which the dispatching policy was New York 2, we determined the proportion that received two ladders as their initial dispatch (these data had been part of the normal output from all simulation runs). The results in Table II show that \( P \) was very close to \( a^2 \).

The relationship was then validated by analyzing the results on a different set of simulations that had also been run previously but had not been examined during the derivation of the relationship. The original

| TABLE III |
| Verification of Relationship between Number of Ladders Dispatched and Average Availability |
| South Bronx | North Bronx |
| Observed availability | Observed and predicted (from availability) percent of NY2 alarms receiving 2 ladders | Observed availability | Observed and predicted (from availability) percent of NY2 alarms receiving 2 ladders |
| .945 | 83.8/89.2 | .849 | 72.2/72.1 |
| .882 | 70.2/77.8 | .705 | 50.2/49.7 |
| .948 | 83.8/89.8 | .847 | 71.9/71.7 |
| .888 | 69.3/78.9 | .699 | 49.3/48.9 |
| .738 | 42.5/54.4 | .494 | 24.5/24.4 |
| .745 | 45.0/55.5 | .485 | 24.2/23.6 |
| .689 | 38.1/47.4 | .469 | 13.9/21.9 |

objective of these runs had been to study the effects of matching the number of fire engines on duty more closely to the time-varying alarm rate. Results of these simulations were analyzed and compared to the results predicted by the relationship (see Table III). On the basis of this comparison, we were able to conclude that the relationship did provide a useful approximation to the actual field-dispatching behavior.

Since availability can be predicted from the alarm rate, service time distribution, and number of companies stationed in the region, we have used this relationship instead of the simulation to analyze the effects of various Fire Department deployment options on the number of units dispatched.

4. ESTIMATING THE NUMBER OF FIRE COMPANIES WORKING AT ALARMS

Chaiken [4] proposed a queuing model of the number of fire companies busy working at alarms. The model gives the long-run probability dis-
tribution of the number of busy companies. Its assumptions are much closer to reality than those of the models in Sections 1 and 2—nonexponential service times are allowed and there are no assumptions about the geographical distribution of alarms or companies. Therefore, the model had been used for planning purposes in New York City before we compared its predictions with simulation results. We discuss the model and its assumptions and how we validated it using results from simulation experiments carried out several years earlier for other purposes.

In response to a fire alarm, several units are immediately dispatched to the scene. Should the situation require more units, they are subsequently dispatched. Units are released from service one at a time or in groups until the incident is over. The queuing model separates alarms into types according to the number of units required to serve the alarm and the distribution of service times for each stage of service. A stage of service consists of a period of time when a fixed number of units is committed to serving the alarm. Thus a new stage begins whenever an additional unit is dispatched or a unit completes service. Permitting some stages to have zero duration allows units to be dispatched or released in groups of two or more.

Important assumptions of the model are:

1. Infinitely many units are available (this is a reasonable assumption since fire departments will generally send as many units as are required even if it is necessary to enlist the assistance of nearby cities to obtain a sufficient number);
2. Different types of alarms in the region are generated according to independent Poisson processes;
3. Alarm and service rates do not vary with time;
4. Service times for different stages of an alarm of a given type are statistically independent of each other and of the number of units already assigned to previous alarms.

Chaiken's model is a generalization of Erlang's formulas [7], which Khintchine [12] proved valid for the case of an infinite server queue and general service times. Chaiken finds closed-form expressions for the probability distribution of the number of busy servers. These expressions depend upon easy-to-compute convolutions that involve alarm rates and the worktime distributions of the various stages. The expressions can be used to decide how many emergency units to locate in any specified geographical region. Enough units can be assigned to the region so that the probability that more than that number are busy does not exceed a certain threshold, say 2%.

Validating the Model

To test the model with empirical data would be very difficult, if not impossible. An important problem in this regard is that the model gives
**DATA:** ENGINE COMPANIES  
THE NORTH BRONX  
$\lambda = 5$ ALARMS PER HOUR

**Figure 5.** The probability that $n$ engine companies are busy (north Bronx, $\lambda = 5$).
steady-state results for a system with a stationary alarm rate, yet the actual alarm rate varies throughout the day. We decided to test the queuing model by comparing its results to those derived from the simulation model mentioned in Section 2. The simulation incorporates several characteristics of actual operations not treated in the queuing model. For example, in the simulation, service times for fires are somewhat state dependent because travel times and the number of units actually dis-

![Simulation Results Diagram](image)

**Figure 6.** The probability that \( n \) engine companies are busy (south Bronx, \( \lambda = 30 \)).

patched depend on the number available when the alarm occurs. In addition, there is a finite number of units in the simulation while the queuing model assumes an infinite number. Units can be dispatched across region boundaries, and there is a "relocation" procedure that will temporarily reassign units to firehouses other than their home houses if protection in a region is too low [15].

The simulation has a set of incident types based on the number of units required by the incident and their service times. We were therefore able to designate the stages and stage durations of the incidents according to the structure of the queuing model. The simulation that had been con-
structed years before we carried out this validation exercise was clearly not designed with this test in mind. Its printed output reports did not include the data we needed for the test. Fortunately, an output tape had been created from each simulation run that recorded the status of every fire-fighting unit at 15-minute intervals of simulated time. Using tapes from past simulation runs, we were able to create histograms of the number of busy units.

The output tapes from four simulation runs, each for a different alarm rate (5, 10, 20, and 30 alarms per hour), produced 24 sets of test data when engines were considered separately from ladders and the Bronx was considered as a whole as well as in two parts (north and south). For each of these data sets we estimated the parameters of the queuing model from the simulation data and then computed the “theoretical” distribution of the number of busy units.

The results in each case were extraordinarily good. Plots of the theoretical frequencies and those produced by the simulation show a very close correspondence. Figures 5 and 6 are a sample of the results. We remark that although the fits are quite close “by eyeball,” they fail the chi-square goodness-of-fit test. The reason is that the sample sizes are very large: they range from 430 to 2871 observations. However, the issue should be whether the fit is close enough for policy analysis purposes. We feel that these comparisons validate the model for the decision-making applications for which the queuing model was created.

5. CONCLUSIONS

The four examples we have just discussed illustrate the use of simulation to develop and validate analytic models. The reason for doing so is to give the potential user of the analytic model confidence that it is a safe substitute for the more accurate simulation model.

What are the advantages of analytic models over continued use of simulation models in these and other cases? In a narrow sense, the analytic models are cheaper to use. Calculations with a square-root model can be made using a desk calculator; the queuing models would require a few minutes at the terminal of a time-shared computer system. Thus, performance characteristics for a range of policy choices can be produced quickly for at most a few dollars of computer time. In contrast, the simulation models would typically require several runs at perhaps $10 to $100 each, and a turnaround time of several hours or days for each run.

In a broader sense, the analytic models offer even more substantial advantages. Two of the most important are:

1. They can be imbedded in other models. For example, the square-root model is an integral part of procedures for choosing the number of
fire companies [18] and the number of police patrol cars [5] to assign
to different regions of the city.

2. They require far less detailed input than simulation models, which
saves both time and money. For example, the preparation of a simu-
lation input tape for two months of police patrol activity (see Section
1) in another precinct in New York or in another city might take
several weeks. Abstracting the essential information needed for the
\( M/M/N \) queuing model might take a few minutes if summary
statistics on call for service and service times were available, as they
often are.

REFERENCES

Description,” R-1188/2, The Rand Corporation, Santa Monica, Calif.,
1974.

2. G. Carter and E. Ignall, “A Simulation Model of Fire Department Op-

York City Fire Department: Its Use in Deployment Analysis,” P-5110-1,
The Rand Corporation, Santa Monica, Calif., 1975.

4. J. M. Chaiken, “The Number of Emergency Units Busy at Alarms Which
Require Multiple Servers,” R-531, The Rand Corporation, Santa Monica,
Calif., 1971.


70–76 (1954).


8. A. Geoffrion, “The Purpose of Mathematical Programming Is Insight, Not

9. R. Hamming, Numerical Methods for Scientists and Engineers, McGraw-Hill,

and Validate Analytical Emergency Service Deployment Models,” P-5473,
The Rand Corporation, Santa Monica, Calif., 1975.

11. E. Ignall and R. Urbach, “The Relationship between Fire Fighting Unit
Availability and the Number of Units Dispatched,” P-5420, The Rand
Corporation, Santa Monica, Calif., 1975.


