Pergamon

Socio-Econ. Plann. Sci. Vol. 31, No. 3, pp. 173–189, 1997

Printed in Great Britain
PII: S0038-0121(97)00013-X 0038-0121/97 $17.00 + 0.00

A Simulation Model for the Analysis and Management of An Emergency Service System

AHMED S. ZAKI† and HSING KENNETH CHENG
School of Business Administration, The College of William and Mary, Williamsburg, VA 23187-8795, U.S.A.

BARNETT R. PARKER
Parker Associates, Chapel Hill, NC 27514-1412, U.S.A.

Abstract—Emergency service systems provide mobile units (crews and/or equipment) to respond to requests for assistance and/or service in emergency situations that occur at any time and any place throughout a specified region such as a metropolitan district or a rural area. This paper presents a simulation model that serves as a tool for resource allocation and management of an emergency service system. The model was tested to study, evaluate, and optimize the allocation of police patrol vehicles to non-homogeneous zones with different demand patterns in the City of Richmond, Virginia. © 1997 Elsevier Science Ltd

INTRODUCTION

Emergency service systems (ERSs) provide mobile units (crews and/or equipment) to respond to requests for assistance and/or service that originate randomly at any time and any place throughout a specified region such as a metropolitan area or rural district. Typical examples of ERSs are ambulance services, fire stations, police patrol cars, bomb disposal units, SWAT teams, road service tow trucks, and emergency repair trucks for gas, electric and water services. Emergency service systems perform a vital role in ensuring and maintaining the safety and well-being of the public. Furthermore, ERSs are highly visible and their effectiveness is of great public concern as evidenced by the recent coverage in widely circulating periodicals such as *The Washington Post* [1] and *U.S. News and World Report* [2]. Mismanagement of ERSs can result in the loss of human lives and valuable resources. Consequently, an efficient model that offers management an effective decision tool for determining the most appropriate equipment allocation, personnel staffing, operating policies and procedures, and short- and long-term planning is of critical importance.

A Washington Post front page article [1] furnishes a good example of the critical importance of effective ERS management. In describing the police emergency system in the City of Washington D.C., the author, Steve Vogel, states that "The disparity in response times across the city points to a failure on the part of the District to move police resources to areas facing the greatest needs and a lack of resolve from elected officials to respond to political pressures that would come with efforts to reallocate crime-fighting resources." In describing one of the reasons for the failure of the response system, the author says that "The primary reason is that the police districts with the most violent crime—which place the heaviest demands on the emergency response system—face the biggest gap between the resources they have been given and their needs on the street."

The optimization of emergency service systems has been receiving considerable attention from several researchers for the last three decades. Two different approaches are used to address the problem. The first finds the optimal locations and/or number of emergency service facilities that

*Author for correspondence. *e-mail*: axzaki@dogwood.tyler.wm.edu

optimize some objective function subject to a given set of constraints. Different choices of the objective function and constraints can result in substantially different emergency service siting models. For example, the earliest work by Toregas *et al.* [3] titled *a location covering model* seeks the least number and location(s) of ambulances such that any request for service has at least one ambulance stationed to respond within a regulated time or distance.

Failure in the earlier work to consider the frequency of demand and the cost of covering remote demand areas led to the 'maximal covering location' problem by Church and ReVelle [4] and White and Case [5]. Given a limited number of ambulances, the maximal covering location problem attempts to find the locations of ambulances that maximize the number of people, or calls for service (CFS), covered by each ambulance within a designated time or distance. Numerous extensions of the maximal location covering problem have been proposed to solve more realistic and practical problems. For instance, backup coverage models were developed by Daskin and Stern [6], Hogan and ReVelle [7], Pirkul and Schilling [8], and Batta and Mannur [9] to account for the possible unavailability of emergency service units. Queuing models were also applied to address the issue of server availability, e.g. the work by Larson [10, 11], Halpern [12], and Benveniste [13]. Capacitated versions of the maximal location covering problem were developed by Chung [14], Current and Storbeck [15], and Pirkul and Schilling [8, 16] to include capacity constraints on the emergency service facility's workload. Heller, Cohon, and ReVelle [17] developed a mixed-integer programming model for emergency medical service systems known as the 'P-median transportation problem (PMTP)'. Their model is capable of handling multiple objectives; it locates P facilities among the nodes of a graph so that the mean response time is minimized while imposing limits on workload to ensure balanced ambulance utilization. ReVelle [18] provides a recent and excellent review of emergency service siting models.

The second approach assumes furnished locations for the emergency facilities as it attempts to find the optimal allocation of personnel and equipment, i.e. that allocation capable of reducing the response time of the emergency service units to, or below, a specified value. Examples of this approach are the optimization models by Cobham [19], Larson [20], and Stevenson [21], and the simulation models of an automated dispatch system for the San Jose Police Department by Adams and Bernard [22], of fire department operations by Carter and Ignall [23], and the New York emergency ambulance service by Savas [24].

We chose to pursue the second approach here since the majority of emergency service systems are already positioned at certain sites and are costly and difficult, if not impractical, to relocate. For example, emergency medical vehicle bases are sometimes constrained to be fire stations where paramedics can rest, perform routine maintenance and re-supply while not on call, as demonstrated by Goldberg *et al.* [25] using the Tucson, Arizona medical vehicles model. Moreover, emergency services in metropolitan areas have branches or secondary sites for their mobile units located at designated locations to ensure coverage of the whole region, e.g. fire stations, police precincts, hospitals, etc. This distribution of facilities within the total area does address, if not optimally, the first approach's objective. Therefore, the issue that actually faces emergency service administrators is the allocation of crews and/or equipment to existing facilities in order to comply with, or achieve, predetermined criteria such as federal, state, or trade regulations, or a desired or competitive service level.

Police emergency service systems characteristics

Prior to our describing key ERSs' characteristics, the following terms must be precisely defined since they are frequently used in this paper. An *incident* causes the emergence of a set of conditions at a location that creates a demand for the services of one or more emergency units at that location. The occurrence of an incident initiates a *call for service* (*CFS*) from a service provider, be it the police, a hospital, a fire station, etc. *Response time* is the time interval from the receipt of a CFS until the dispatch of the first available unit(s) to the scene of the incident.

Response time serves as a surrogate measure for the effectiveness of ERSs because (a) it is easily understood by both analysts and managers, (b) it is accepted by decision-makers, and (c) due to the humanistic nature of emergency systems, and the fact that the caller is in a crisis status, the length of time until the caller receives assistance is crucial and significant. Moreover, response time is considered, in many cities, as the key test of a police department's effectiveness since people seek

quick responses, and the press often focuses on the issue [2]. Therefore, it is one of the primary measures of effectiveness that will be used in this research.

A close examination of the factual environment under which emergency units function reveals that:

1. The travel time to and from the scene of an incident is essentially a function of:
   - The distance between the place where service will be rendered and the location of the emergency unit at the time of it's response to the call. That location may be the garage or parking lot of a police precinct, a hospital, an ambulance unit, or anywhere within the boundary of the precinct/zone area of responsibility. In the case of police emergency vehicles, most, if not all, of the emergency units may, as a crime prevention measure, be patrolling their beats when they are notified to respond to a call.
   - The time of the day; daytime vs night time driving, rush hours, etc.
   - The traffic congestion; peak vs non-peak hours or accidents on the main road.
   - The weather conditions such as rain, snow, icy roads, etc.
2. Incidents differ in their scope, degree of significance, and severity. For instance, a fire in a hospital or a theater, an accident with multiple injuries, or an in-progress street gang fight will require more than one emergency unit and should take precedence over less urgent incidents such as a single car accident with no injuries or noisy neighbors. Consequently, there is a need for the assignment of various response priorities as well as the allocation of different numbers of emergency units to different categories of incidents.
3. If a crisis arises and the needed emergency units are unavailable, it may be necessary to borrow emergency crews and/or equipment from a neighboring zone, county, or even city. This practice is known as inter-zone/city dispatching.
4. Inter-arrival times between demand for services are not always exponentially distributed. It may follow different theoretical or user defined distributions under different conditions. Moreover, the demand for service pattern in an area usually varies with the time of day, season of the year, the type of incident, etc.
5. Real world systems operate in a complex environment that includes ill-defined objectives, and implicit and explicit administrative, legal, and political constraints [26]. A recommended policy or procedure that is acceptable in one environment (state/city) may thus be unacceptable or infeasible in another.

Research motivation

None of the analytical models known to the authors captures all of the above cited features of ERSs. Restricting assumptions such as no priorities for different classes of incidents, exponential service times [21], only one unit viewed as adequate to respond to any call [19], etc. are usually necessary to make the analytical model's formulation practicable. Heller, Cohon and ReVelle [17] observed that many location models rely on deterministic approximation and assume facilities are always available so that demand is always assigned to the closest facility. In optimization models that seek to minimize mean response time, this availability assumption may lead to assigning facilities that, in actuality, are not available thereby producing unrealistic or sub-optimal solutions. For this reason these researchers tested the solution generated by their PMTP model to the EMS system in Baltimore, Maryland using a stochastic simulation of the system. Their rationale was that "the two models work together in a complimentary fashion, addressing two complicating feature of EMS location problems: the multi-objective nature of the problem and the complications arising from dynamic and stochastic nature of the system."

The shortcomings of existing location models for emergency services are succinctly described by ReVelle [18]. He stated that "the environment of these models is not merely random, it is also dynamic, evolving through the day, week, and season. Demand may also exhibit a trend, and new demand areas may evolve over time. None of the probabilistic siting models have yet been extended to changing and evolving environments." Moreover, the models of both approaches fail to consider adequately the stochastic nature of the operating environment together with all or most of the potential contingencies an ERS has to cope with. An expedient approach to model such complex

and stochastic systems is to develop a practical simulation model that can encompass in a more complete fashion the total system and the interrelationships between the system components [10].

With this objective in mind, we present a practical, general, robust, and easy to implement simulation model for the study, analysis, evaluation and optimal allocation of the available resources of an ERS. Moreover, the proposed model runs on a personal computer and does not require a mainframe, as was the case in previous research. It uses one of the PC simulation languages—the ProModel Service and Manufacturing Simulation software. This package was selected because of its ease of use for both end users and analysts, its capabilities for efficient representation of large and complex models, its flexibility in applying sensitivity analysis and dynamically changing model parameters during the simulation runs, its animation capabilities for observing system behavior during the simulated periods, and the adequacy of its generated statistical and graphical outputs. In other words, it provides an efficient tool for the management of an ERS under a dynamic and continuously changing environment.

The following section describes the real world system studied and presents results of our analysis. Next, the simulation model is presented. Model implementation, manipulation and various results are then introduced and discussed. The last section presents the authors' comments and conclusions.

EMERGENCY SYSTEM DESCRIPTION AND ANALYSIS

A police ERS was chosen because its delineation corresponds in many respects to the delineation of most ERSs such as fire stations, emergency medical services, etc. The police emergency dispatching system for the city of Richmond, Virginia, was selected since it represents a significant dispatching system that serves a population of approximately 700,000. Moreover, the Department of Public Safety for the city of Richmond was helpful in providing relevant statistics for the 1990–1991 fiscal year. In this paper, the term *emergency units*, or just *units*, is used to specify the Department's vehicles and/or crews, including any additional necessary equipment when needed.

The setting

For emergency service purposes, The City of Richmond is divided into seven planning zones. The population of each zone has a quasi-homogeneous set of characteristics that is a function of (1) the income level, (2) the type of zone, e.g. residential, industrial, transient, and (3) the ethnic origin of the majority of the residents. Five of the most dynamic zones, portrayed in Fig. 1, are studied and modeled. For simplicity, the true zone numbers [100, 200, 300, 400, and 600] are renamed 1, 2, 3, 4 and 6; respectively.

The characteristics of a zone—its population density, the time of day, e.g. morning or evening and regular and peak hours—affect the rate of occurrence of different types of incidents as well as the travel time to and from the location of the incident. Consequently, the rate of demand for the services of personnel and equipment differs by the season, time of day, and by zone. Incidents vary in severity and scope and thus require the services of different numbers of emergency units. Each planning zone is assigned a certain number of police officers and vehicles (emergency units) commensurate with the anticipated number of calls for service and the type of incidents. The criteria for allocating resources among the zones/districts include each zone's total number of CFSs, its crime rate, its miles of streets, its geographical size, and its percentage of commercial property [1]. As in most other ERSs, police personnel and vehicles are a scarce resource subject to tight budgetary constraints [27].

Response priorities are assigned to incidents based on their degree of severity, e.g. a bomb threat, a fire in a hospital, a homicide, etc. Rules for assigning priorities to calls are usually determined by federal, state, city, and/or trade organizations and are readily available to managers, analysts, and the dispatcher. Table 1 lists the different types of incidents and their associated priority assignments, where priority one indicates the lowest priority and four indicates the highest.

Emergency units are composed of both humans and equipment. The reason for highlighting this characteristic is that personnel morale and equipment durability are critical concerns. As one might expect, different zones have different demand rates and priority profiles which, if not adequately addressed, can adversely affect the equity of the work load distribution and, consequently, morale,

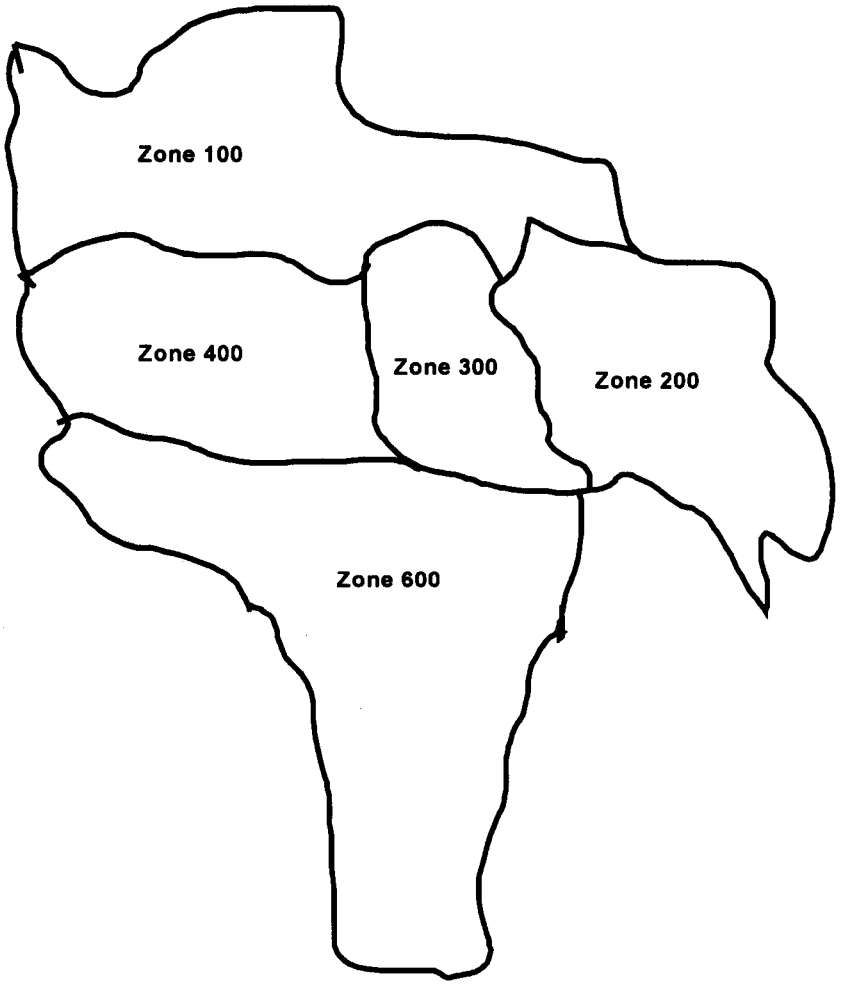

Fig. 1. Partial public safety map for the city of Richmond. *Note*: For simplicity, the zone numbers 100, 200, 300, 400, 600 were renamed as 1, 2, 3, 4, 6, respectively.

the quality of service, and equipment life span. Thus, according to Guttman [2], "sometimes officers handle as many as 28 requests for service in one shift. Often, especially in the summer, so many calls come in that the buffer fills up and officers have to write the extra calls on paper. Requests for 'code 7'—time off to eat—are sometimes turned down because too many calls are outstanding. An indicative measure of a zone's work load is its utilization factor (UF) defined as the expected fraction of time its servers/units are busy.

The police ERS dispatching process

Personal interviews with the managers and analysts in the Richmond Bureau of Emergency Communications [28] indicated that the dispatching process, at the time of the study, proceeds in the following fashion:

1. Requests for emergency services arrive to the a central location via telephone (a 911 number) or by wireless such as when an emergency unit radios for assistance or backup. The automated answering system routes each call to the zone responsible for providing the emergency service.

Table 1. Priority assignments for different types of incidents

Priority	Types of incidents
1	Abandoned vehicles, parking violation, general advice.
2	Animal bites, loud music, take a report.
3	Larceny just occurred, violence (street or domestic) without injuries, auto accident without injuries, alarms, in-progress call.
4	Officer needs assistance, robberies, auto accident with injury, violence (street or domestic) with injuries.

Table 2. Percentage of arrival rates for all zones

Time Period	Percentage of CFS Arrivals
0100–0300	9.10
0300–0900	8.32
0900–1300	19.76
1300–1900	27.22
1900–0100	35.60

2. Based on the information available to the zone dispatcher when a CFS is received, s/he assigns a priority and schedules the number of units necessary to respond to the call. Calls that originate within a zone are, except under special circumstances, exclusively served by the zone's emergency units.
3. If *all*, or *at least one*, of the required units are available, they are dispatched instantaneously; the rest of the scheduled units, if any, are queued according to their priority and dispatched when they become available *if*, at that time, they are still needed. If no units are available at the time of a request, the call is queued according to its priority until *at least one* of the required units becomes available.
4. The elapsed time between receipt of a call by the dispatcher and his/her issuing a request to dispatch (the dispatching time) is practically negligible. This observation coincides with findings from the simulations of the New York City and Denver fire departments, where 30 s to 1 min were found to be a reasonable estimate [29]. Therefore, in this study, the dispatching time was estimated to be 30 s. If, on the other hand, dispatching requires a non-negligible time, the model is flexible enough to readily accommodate the addition of such an overhead.
5. As a preventive measure, most, or all, of each zone's police units patrol designated beats in their zone to act as deterrents to certain types of crimes. However, any of these units is considered available to the dispatcher to assign to any incident site within its respective zone.
6. After completing the actions necessary to resolve the emergency, the police unit(s), if necessary, return to the zone's main office to file a report and deliver any individuals they may have in their custody. Otherwise, the units(s) proceed to patrol their assigned beat(s) or perform administrative functions such as refueling, eating, etc.

Data analysis

The available data set used here covers the period from January 1st to June 31st, 1991, and consist of:

- Monthly frequency distributions for the number of CFSs by hour of the day and day of the week for each of the six months.
- Overall percentages of the different classes of CFSs and their priorities together with the average time spent on each type.
- Percentages of the different number of units dispatched per CFS for each of the six months.

Using the entire data set, descriptive statistics were computed. Polygons and histograms were then plotted for each of the six months in order to acquire a rudimentary picture of each zone's probability distribution for:

A. The inter-arrival time of CFS.
B. The occurrence rate for each class of CFS in terms of priority level and number of units required to respond.
C. Travel times to and from the site of the incident during different hours of the day as well as different seasons.
D. The processing time for each type of incident at different times of the day.

The analysis showed that, as expected, the hourly arrival rate for CFSs varies for different hours of the day as well as for different days of the week. Furthermore, it showed that even for the same day of the week, the arrival rate differs by month of the year. However, it was possible to identify the ranges for arrival rates during peak and non-peak periods as they apply to all seven zones, as illustrated in Table 2.

Table 3. Demand, travel time and emergency units data

Zone	Intra-zone travel time distribution	% of aggregate calls for service*	Current number of assigned units
1	Normal (4.3, 1.2)		
	For 62% of the zone calls	18.10	9
	Triangular (2,2.75,5)		
	For 38% of the zone calls		
2	Normal (3.5, 0.5)	18.92	12
3	Normal (1, 0.2)	14.65	6
4	Normal (2.1, 0.59)	25.37	16
6	Triangular (2.5, 3.2, 6.0)		
	For 34% of the zone calls	22.96	9
	Uniform (2.2, 0.45)		
	For 66% of the zone calls		

* The daily demand for service distribution was found to be normal (480, 30).

Next, the UniFit II software package [30] was used to expeditiously estimate the theoretical probability distributions for:

- The arrival rate of all CFSs, where the total number of arrivals per month during the period under study was between 14,648 and 18,738.
- The service time for different priorities of calls.
- The intra-zone travel time. For those zones that cover an extensive area, there are two different travel times; one for the close sites and one for the outskirts of the area.

The package considers all possible distributions that could fit the data and conducts several effective heuristics to pick and rank on a scale of 0 (worst) to 100 (best) the 'most fitting' five distributions. It then offers the analyst the option of conducting goodness-of-fit tests (Chi-square and Anderson–Darling) on each, and recommends the most appropriate probability distribution to use as well as the distribution's parameters.

It is relevant to note here that the high rates of demand for all zones studied occurred daily between 9:00 a.m. and 1:00 a.m. (two 8-h work shifts), which will be referred to as the demand *peak hours* (PH) as opposed to the *regular hours* (RH) demand rates that occurred from 1:00 a.m. to 9:00 a.m. Tables 3 and 4 list the values of the model's random variables as derived from the statistical analysis.

THE EMERGENCY SYSTEM SIMULATION MODEL

As previously noted, the objective of this paper is to provide analysts and managers with an effective, flexible, and easy-to-comprehend and implement instrument [31, 32] that will:

1. Serve as an investigative and educational tool for understanding and analyzing the flow of activities and operations throughout the dispatching system in order to identify areas with unacceptably long response times. In this regard, acceptable response times are based upon existing or newly established criteria/regulations.
2. Serve as a diagnostic tool that would enable management/analysts to identify any current or potential problem areas or bottlenecks.
3. Support management in helping determine the most appropriate course(s) of action by offering a means for testing and evaluating the effects of proposed policy changes on key parameters such as response times to all or particular calls, equipment and personnel utilization rates, unit availability, etc.

Table 4. CFS service time and units required to respond for all zones by priority level

Priority level	% occurrence	Service time distribution in 67C:SEPS110minutes	Units required to respond
1	15.32	Uniform (42,2.5)	1 Unit 90% 2 Units 10%
2	52.13	Normal (56.5,8)	1 Units 70% 2 Units 30%
3	31.04	Triangular (50,63,82)	2 Units 100%
4	1.51	Uniform (112,3.5)	3 Units 100%

4. Provide management/users with a tool for forecasting and conducting sensitivity analyses to investigate the impact of proposed or anticipated changes in:
   A. Priority assignments.
   B. The number of vehicles assigned to different zones and shifts.
   C. The number of vehicles assigned to different classes of incidents.
   D. Zone boundaries, assigning territorial responsibility to individual or small groups of vehicles, and replacing mobile units with foot patrols.
   E. Inter-zone dispatching (IZD).
   F. Demand rates.

To enhance the practicability and portability of the model so that it may be directly applicable or readily adaptable to other types of ERSs, we designed it to accommodate the following eventualities:

1. The number of units regularly assigned to each zone (see Table 3). However, that number can be increased or decreased in order to accommodate the fluctuations in demand during peak and non-peak hours.

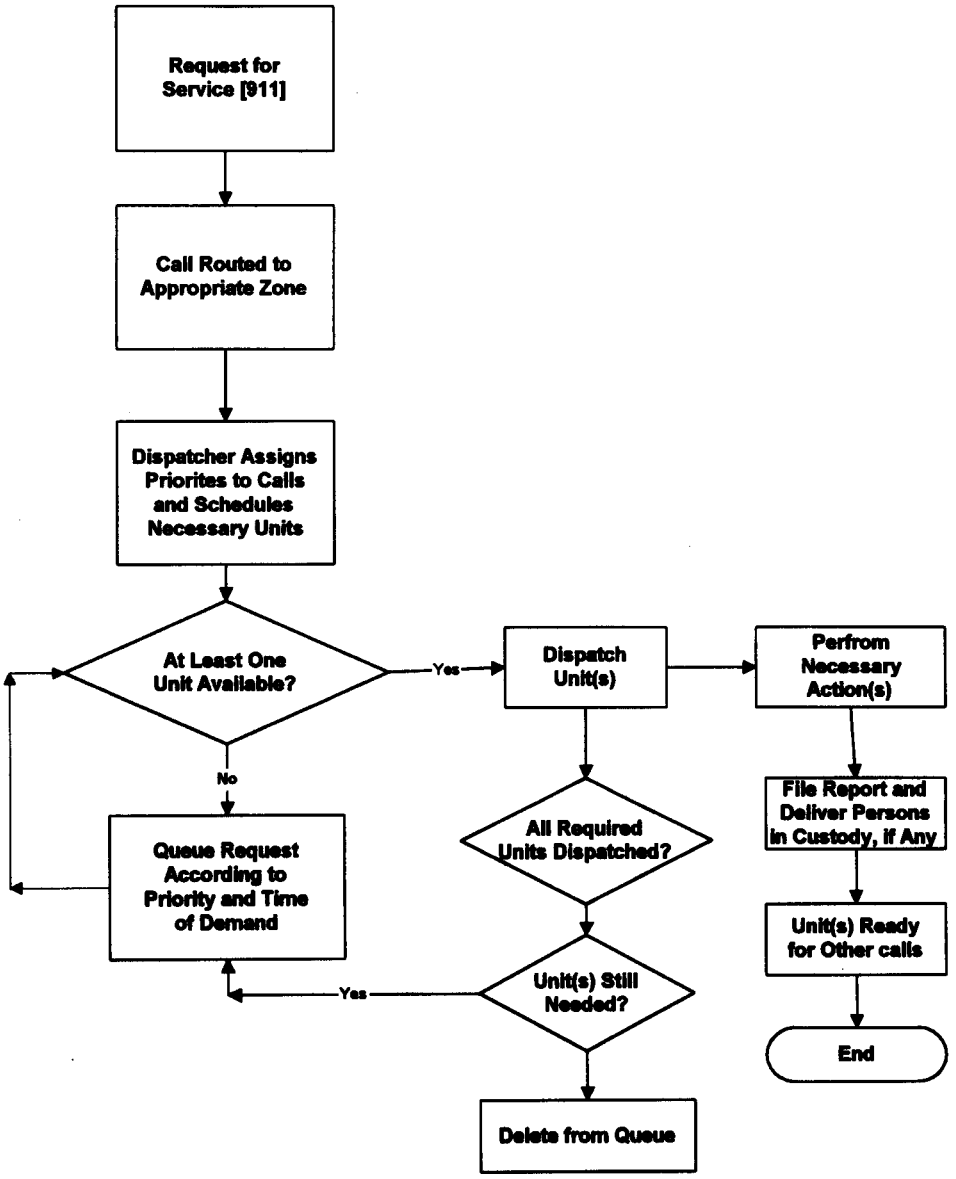

Fig. 2. Flow diagram of vehicle dispatching without inter-zone dispatching.

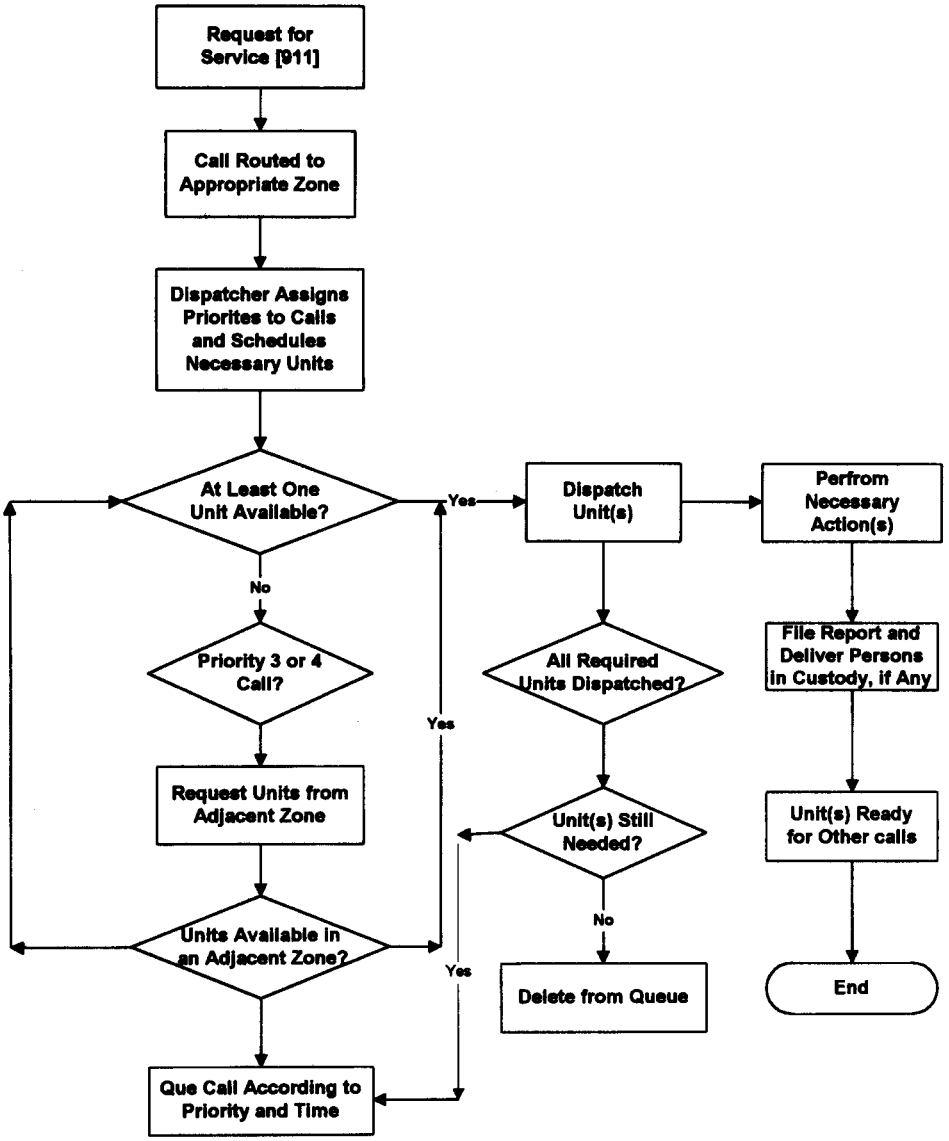

Fig. 3. Flow diagram of vehicle dispatching with inter-zone dispatching.

2. All or some of the demand rates, service and travel times can be, and usually are, stochastic. Their probability distributions can take any shape and may be defined in a closed form, by a random number generator, or in the form of a table look-up.
3. Response to calls for service is based upon the priority assigned to each call by the dispatcher. If there are no units available to respond to a call, it is queued according to its priority. That is, a high priority call will be queued ahead of a lower priority call even if the latter was received earlier.
4. If a request for service requires more than one emergency unit and only one is available, then the available unit is dispatched and a request for the rest of the originally requested units is queued and dispatched when those units become available, *if* there is still a need for their services. Since the average service time for 98.49% of the emergency calls ranges from 42 to 56 min, as shown in Table 4, it is reasonable to assume that backup units will no longer be needed if they were not dispatched within a maximum of 30 min from the time of their request. Of course, in real world situations, some backup units might be deemed unnecessary after

less or more than 30 min depending on each specific incident. Increasing or decreasing this time constraint requires changing its value in only five statements.

5. An additional travel time was added to the service time of each call to give the assigned unit(s) the opportunity to perform administrative chores such as filing a report or delivering individuals who might be in the unit's custody. Immediately after that, the unit is considered available and ready to be dispatched to respond to another CFS.
6. All crews and equipment (units) assigned to a shift, including those on routine duty or patrolling their beats, are devoted exclusively to servicing emergency calls.
7. Initially, the model is designed on the assumption that personnel absenteeism or equipment breakdowns can be replaced instantaneously. In later runs, this assumption will be relaxed and the model will simulate scenarios where, in each zone, randomly generated numbers of emergency units will be down for random periods of time during both regular and peak hours.

The current system does not allow the dispatcher to preempt any of the emergency units that have been allocated to an incident in order to allocate them to a higher priority or a more severe incident. This conclusion was made after discussions with the police dispatching system managers revealed that preemption is usually impractical and difficult to administer. They would thus not recommend it, save exceptional situations, as in the case of a disaster.

Some of the difficult and delicate questions that preemption raises include the following:

A. What criteria should be used to justify preemption?
B. Who is to decide which call to preempt and whether to appropriate all or some of its assigned units?
C. What are the benefits and costs to both the preemptor and the entity that has lost one or more resources assigned to it due to preemption?

For these reasons, preemption was considered as a non-viable option and not included in the model.

Flow diagrams of the police ERS excluding and including inter-zone dispatching (IZD) are shown in Figs 2 and 3; respectively.

Model validation

Before testing the model's validity, i.e. evaluating how well it represents the germane aspects of the real world ERS, the following issues needed to be resolved:

1. The run length: To produce a representative imitation of the real world system, the experiment was designed to ensure that all possible events and contingencies in the real world would occur with the same frequency during each simulation run. Examination of the data revealed that all types of incidents occurred during a 24 h run since none of them has a rare probability of occurrence.

Table 5. Performance measures using different inter-zone borrowing policies

Zone	Regular policy using current unit assignment (1)	All zones can borrow from adjacent zones (2)	Only Zones 6 and 3 can borrow from their adjacent zones (3)	Regular policy using different unit assignment (4)
Unit/zone assignment				
Z1	11	11	11	10
Z2	12	12	12	11
Z3	6	6	6	8
Z4	16	16	16	13
Z6	9	9	9	12
% of CFSs With response time = 0	79.55	60.10	75.85	84.81
Average response time for all other CFSs	4.69	9.49	3.63	1.93
% non-response within 30 m	0.34	0.31	0.13	0.15
Utilization factors				
Zone 1	0.55	0.70	0.62	0.56
Zone 2	0.52	0.74	0.66	0.56
Zone 3	0.69	0.76	0.62	0.55
Zone 4	0.50	0.70	0.65	0.61
Zone 6	0.82	0.83	0.69	0.59

Table 6. Performance measures for different policies using 54 units*

Poilicy Significant parameters	Regular (1)	IZD (2)	Shifts (3)	Shifts except for Zones 4 and 6 (4)
Number of incoming CFSs	6692.70–6771.17	6702.05–6775.15	6678.18–6746.35	6671.16–6742.30
% of CFSs with response time = 0	82.73–82.79	81.20–81.36	81.88–82.02	81.27–81.28
For CFSs with response time > 0, average response time	2.26–2.48	2.77–3.19	2.53–2.74	2.59–2.82
% of CFSs aborted (> 30 min)	0.16–0.37	0.09–0.27	0.17–0.39	0.26–0.48
Utilization factors by zone				
Z1	0.58–0.60	0.62–0.63	0.62–0.64	0.62–0.63
Z2	0.55–0.56	0.62–0.63	0.56–0.58	0.56–0.58
Z3	0.54–0.56	0.52–0.53	0.57–0.59	0.58–0.60
Z4	0.62–0.63	0.66–0.67	0.63–0.65	0.61–0.63
Z6	0.62–0.63	0.57–0.58	0.63–0.65	0.62–0.63

*The values listed represent the upper and lower bounds for the 99% confidence intervals.

In addition, a 24 h day includes both regular and peak time conditions. Since each 24 h run required about 20 s of CPU time, using a 486/33 MHz, the author opted to run the model for 336 h, i.e. for 14 days, so as to observe the system performance over a relatively long period and thus detect the impact of continuous daily operations under unsteady conditions. In addition, a long run would minimize the effect of the initial conditions.

2. The initial conditions: If the simulation is started with no entities or activities at simulation time zero (initial conditions), it may take a certain period of time to reach a state when it truly mimics the real world system, or the steady state. Since the police ERS, like most other ERSs, has a demand rate that varies with time of day, as do the type of incidents and the number of available units, there is no variable(s) to observe in order to predict the time period necessary for the system to truly mimic real world conditions. Yet, if one were to use a data sample of the system states at midnight, it would provide a set of actual observations of the system's operation. To this end, we sampled seven midnight states from every one of the six months; i.e. 42 observations, and acquired the necessary data that were subsequently used as initial conditions.

Three 14-day runs, each using a different set of random number seeds, were performed. Each run's output was compared with the available historical data and also discussed with the analysts at the Department of Public Safety. Comparison of the historical and model output data, together with the analyst's comments, were extremely useful. It turned out that the probability distribution of the processing time for priority 4 incidents underestimated the real world time. The reason appeared to be that the historical data did not include the time necessary to prepare and file the detailed reports required in such cases. The three additional runs performed after the modification

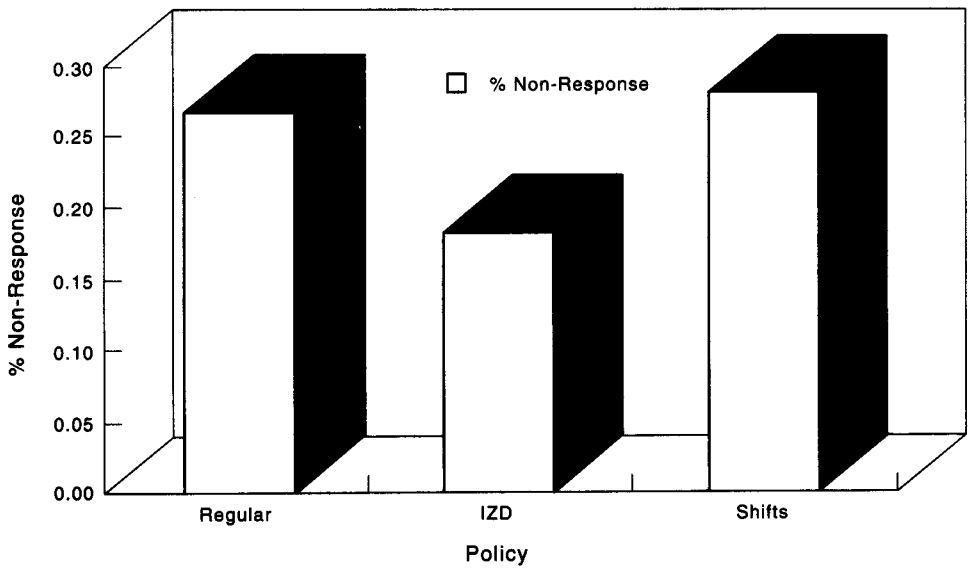

Fig. 4. Non-response percentages for different dispatching policies.

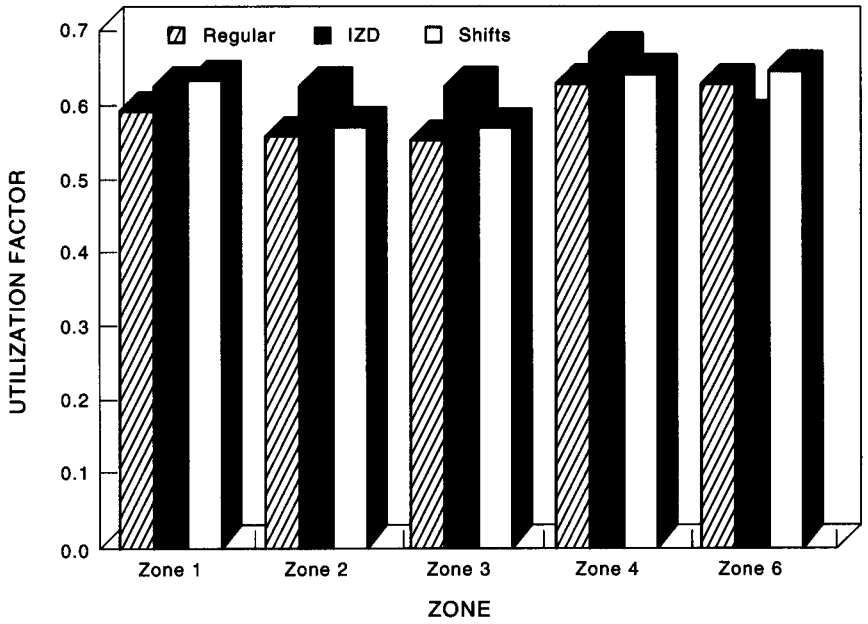

Fig. 5. Comparison of utilization factors under different dispatching policies.

of the priority 4 density function showed that the simulation model's results did, indeed, agree with those of the real word system.

SIMULATION MODEL IMPLEMENTATION

Since one of the emergency response system variants that this research investigated is inter-zone dispatching, several pilot runs were conducted to explore the feasibility of IZD between *all* or *certain* zones as a measure for enhancing performance of the current ERS. For simplicity, we refer to the model where IZD is not exercised as the 'regular' model, and that where IZD is an option for *all* or for *some* of the zones as the 'IZD' model. It should be pointed out here that:

1. A zone can borrow one or more units from a *bordering* zone if it has no units available to respond to a CFS that has a priority greater than or equal to 3.
2. Since there are no statistics concerning the travel time between different zones, the inter-zone travel time is computed as the sum of the travel times in the two affected zones.

Table 5 compares results of several 14-day pilot runs using regular and IZD policies. Comparing the statistics in columns (1) and (2) demonstrates that allowing IZD for all adjacent zones resulted in substantially decreasing (by 24.45%) the percentage of CFSs with a zero response time and more than doubled the average response time for the rest of the CFSs. Since the utilization factors (UFs) of Zones 3 and 6 in the regular model were high relative to the other three zones, the authors experimented with allowing only Zones 3 and 6 to borrow units from their adjacent zones. Contrasting the results in columns (2) and (3) of Table 5 shows that all performance statistics were considerably enhanced. In particular, the UFs became more equitable.

The logical inference from the Table 5 data is that the real issue is the unit/zone assignment. Thus, Zone 6, which has the second highest CFS arrival rate (22.96%), is assigned a relatively smaller number of units than are those zones that have lower demands for service. Also, the six units assigned to Zone 3 are not commensurate with its rate of demand (14.65%) relative to the other zones. Consequently, Zones 3 and 6 had to request more units from Zones 1, 2, 3, and 4 than they lent. According to the model-generated statistics, during the 14 simulated days Zone 3 borrowed a total of 630 units (212 from Zone 1, 188 from Zone 2, and 230 from Zone 4), while Zone 6 borrowed a total of 702 units (310 from Zone 2 and 392 from Zone 4). On the other hand, Zone 3 lent only 243 units (38.6% of the number of units it borrowed) to both Zones 2 and 4, whereas Zone 6 lent 223 units (30.3% of the number of units it borrowed) to Zones 2 and 4. It

is thus reasonable to deduce that such an imbalance in the number of units borrowed by Zones 3 and 6 is equivalent to decreasing the number of units assigned to Zones 1, 2, and 4. This conclusion also explains the increase in average response times and the UFs for Zones 1, 2, and 4, as depicted in Table 5.

To ascertain the validity of this conjecture, several more two-week 'what-if' runs were conducted where units from zones with relatively low UFs were reallocated to zones with high UFs and long response times in order to find the most appropriate unit/zone assignment. The configuration, listed in column (4) of Table 5, yielded the best overall performance, indicating that it is the most appropriate unit/zone assignment. It compares favorably with the two IZD policies as well as the regular policy. In addition, the UFs for all zones decreased and became fairer. Challenging the model, the authors experimented with IZD between all adjacent zones using the new unit/zone assignment; however, results generally compared unfavorably with those listed in Table 5.

The key benefits gained from the pilot runs are considered to be as follows:

1. The current unit/zone assignment need not be further investigated. Subsequent experiments will use the unit/zone configuration shown in column (4) of Table 5.
2. The pilots showed that if IZD is to be investigated, then it should only be available to Zones 3 and 6.
3. They also demonstrated the usefulness of a simulation model to serve as an investigative, educational, sensitivity analysis, as well as a diagnostic tool.

The model was then used to explore in detail two scenarios. The first employs the regular policy while the second pursues the IZD policy. In order to augment validity of the simulation results, every experiment or run in this research consisted of 30 14-day replications. Scenario results were analyzed, while modified versions of the policies were inferred to search for the optimal policy and unit/zone assignment. The only imposed constraint, as is the case in most other systems, was a budgetary constraint. Hiring additional police officers or buying extra vehicles is considered highly unlikely. Thus, only reassigning emergency units between zones was initially investigated.

Columns (1) and (2) in Table 6 summarize results of the experiments under the two scenarios. However, one should be careful when interpreting the values of the utilization factors (UFs) in the specific case of police ERS since most, if not all, of the police response units are usually performing routine patrols in their zone as a crime prevention measure. When, for example, the model calculates average UFs for a zone's emergency units to be 0.62, it should be interpreted as the crews and the vehicles of that zone actually being en-route to, or processing, emergency calls 62% of the time. During the rest of the time, they are not literally idle; all, or most of, that 38% of the time is being spent performing preventive patrolling. Although routine patrolling is not as demanding as responding to and processing emergency calls, it still involves some tedious work.

The data in column (2) of Table 6 clearly indicate that inter-zone dispatching did not improve system performance. As a matter of fact, when compared to the figures in column (1), where IZD was not an option, almost all performance measures, including the UFs, declined. However, the percentage of aborted CFSs, i.e. non-responses to a CFS, decreased by 47%. This is a critical performance measure that should be explicitly considered when evaluating scenarios and policies. Failure to respond to some emergency calls, especially when many of them may be life-threatening,

Table 7. Performance measures for different policies using 56 units*

Policy Significant parameters	Regular (1)	IZD (2)	Shifts (3)
Number of incoming CFSs	6685.12–6757.15	6705.42–6724.91	6680.09–6734.17
% of CFSs with response time = 0	85.45–85.52	82.52–83.48	84.76–84.87
CFSs with response time > 0: average response time	1.82–1.98	2.31–2.56	1.96–2.18
% CFSs units aborted (> 30 min)	0.14–0.31	0.11–0.25	0.15–0.37
Utilization factors by zone			
Z1	0.58–0.60	0.62–0.64	0.60–0.61
Z2	0.54–0.56	0.61–0.63	0.56–0.58
Z3	0.55–0.56	0.52–0.53	0.57–0.58
Z4	0.57–0.58	0.63–0.64	0.58–0.60
Z6	0.56–0.58	0.53–0.54	0.59–0.60

*The values listed represent the upper and lower bounds for the 99% confidence intervals.

Table 8. Comparative performance measures for different scenarios

Policy	Regular		IZD		Shifts	
Significant parameters	Using 54 units	Using 56 units	Using 54 units	Using 56 units	Using 54 units	Using 56 units
% CFSs with response time = 0	82.73–82.79	85.45–85.52	81.20–81.36	82.52–83.48	81.88–82.02	84.76–84.87
CFSs with response time > 0: average response time	2.26–2.48	1.82–1.98	2.77–3.19	2.31–2.56	2.53–2.74	1.96–2.18
% CFSs aborted (> 30 min)	0.16–0.37	0.14–0.31	0.09–0.27	0.11–0.25	0.17–0.39	0.15–0.37
Utilization factors by zone						
Z1	0.58–0.60	0.58–0.60	0.62–0.63	0.62–0.64	0.62–0.64	0.60–0.61
Z2	0.55–0.56	0.54–0.56	0.62–0.63	0.61–0.63	0.56–0.58	0.56–0.58
Z3	0.54–0.56	0.55–0.56	0.52–0.53	0.52–0.53	0.57–0.59	0.57–0.58
Z4	0.62–0.63	0.57–0.58	0.66–0.67	0.63–0.64	0.63–0.65	0.58–0.60
Z6	0.62–0.63	0.56–0.58	0.57–0.58	0.53–0.54	0.63–0.65	0.59–0.60

is much more grave than responding a few minutes late. Therefore, since performance figures using the IZD policy did not significantly deteriorate, the authors strongly advocate that it be considered the most appropriate one, given current resources.

Observing that the UFs for all zones are relatively low, the authors decided to further investigate if decreasing the number of emergency units assigned to each zone by one unit during the non-peak hours of 1:00 a.m. to 9:00 a.m. would result in a detrimental impact on ERS performance. The results depicted in column (3) of Table 6 show that releasing those units during non-peak hours did not have a perceptible negative impact on ERS performance. Moreover, this policy, which will be referred to as the 'shifts' policy, benefits the zone's human resources and equipment in that they then have an eight hour relief per day.

Since the UFs of zones 4 and 6 units are a little higher than those for the other zones, the authors ran the model again without letting these zones release one unit during non-peak hours. As expected, the UFs were rendered more equitable. However, one must be careful when utilizing the word 'equitable' in this or similar instances since the personnel in Zones 1, 2, and 3 will acquire relief time while those in Zones 4 and 6 will not. However, both alternatives of the shifts policy produced relatively high non-response percentages.

The dispatching policy that appears to be the most appropriate is IZD since, as portrayed in Fig. 4, it produces the minimum percentage of non-responses to emergency calls, a crucial yardstick when evaluating the effectiveness of ERS.

Since Table 6 and Fig. 5 show that the UFs in Zones 4 and 6, under all the investigated policies, have been consistently higher than those in the other zones, it is logical to investigate the possibility of increasing the total number of emergency units by assigning an additional unit to each of Zones 4 and 6. That notion was motivated by the administration's 1995 decision to considerably increase the number of police officers deployed in the nation's streets. Three more experiments, one for each of the regular, IZD, and shifts policies were performed. The performance measures are displayed in Table 7.

The results in Table 7 corroborate conclusions reached from the analysis of Table 6's data. It also shows that adding those two extra units not only lowered the UFs of Zones 4 and 6 so that the UFs of all zones became more uniform, but also improved all other performance measures for all three policies.

Table 8 contrasts performance measures for the three policies when 54 and 56 units are available. The figures clearly indicate that adding an extra unit to each of Zones 3 and 4 increases the percentage of units that has a zero response time, and the average response time for all other CFSs. Further, it decreases the percent of CFSs aborted, while rendering the UFs more equitable for all three policies. Comparison of columns (2) and (6) reveals that permitting each zone to relieve one unit during the non-peak hours does not have a discernible impact on system performance, while

Table 9. Probability distribution of crew/vehicles downtimes

Down time duration (hours)	Probability of occurrence
24	0.05
16	0.10
8	0.15
4	0.20
2	0.25
0	0.25

Table 10. Comparative performance measures for different scenarios* with random occurrence and random duration downtimes

Policy Significant parameters	Regular (1)	IZD (2)	Shifts (3)
Number of incoming CFSs	6696.73–6760.13	6700.51–6743.76	6672.52–6752.08
% CFSs with response time = 0	84.81–84.59	81.98–82.02	84.66–84.74
CFSs with response time > 0:average response time	1.94–2.12	2.48–2.68	2.00–2.26
% CFSs aborted (> 30 min)	0.15–0.43	0.11–0.28	0.14–0.34
Utilization factors by zone			
Z1	0.59–0.61	0.64–0.65	0.60–0.63
Z2	0.56–0.57	0.63–0.64	0.58–0.60
Z3	0.56–0.58	0.52–0.54	0.58–0.59
Z4	0.58–0.60	0.64–0.65	0.59–0.61
Z6	0.58–0.59	0.55–0.56	0.59–0.61

*The values listed represent the upper and lower bounds for the 99% confidence intervals.

it would be beneficial for the crew, the equipment, and morale. It also allows one unit to be available for contingencies in cases of equipment or personnel down time, or other possible contingencies. It is noteworthy that the IZD policy consistently provides the least percentage of non-responses to CFSs. Using 99% confidence intervals mid-points, the percentage of non-response is 20% less than under the regular policy and 31% less than under the shifts policy. The price of following the IZD policy is a 2.5% decrease in the percentage of CFSs with zero response time, and a 22 s increase over the minimum average response time for the other two policies.

Further to these considerations, the flexibility of dynamically reallocating the emergency units to handle contingencies is important. ERSs can be short-handed everywhere-that is a given in any major department-but one must still be able to move resources to where they are needed most. Being undermanned is insufficient cause for not effectively managing available resources [1].

Finally, in view of tighter budgets during last few years, cities and municipalities have not allocated the funds necessary to hire additional police officers or to renew their vehicles and equipment. As a result of heavy and demanding usage, the potential for personnel absenteeism or need for vehicle maintenance or breakdowns has increased. To investigate the consequences of breakdowns, random downtimes for the emergency units were simulated. The duration of the down times, presented in Table 9, correspond to work shifts and ranges from three shifts (24 h), in the case of a significant breakdown, to one fourth of a shift (2 h) as when a vehicle performs routine maintenance, or refuels.

Table 10 lists results of the three experiments using the regular, IZD, and shifts policies. Although both the regular and shifts policies produced slightly better performance outcomes, the IZD policy remained prevalent since its percentage of non-responses to CFSs was considerably lower than those of the other two policies.

The decision as to which is the most appropriate policy is situation-dependent. All three policies have both strong and weak points. Decision maker(s) might therefore wish to assign weights to each performance measure in order to reach a more rational decision. However, managers of ERSs should realize that they are working under very dynamic environments. Migration of people of different ethnic origins, religions, and cultures changes city/zone characteristics. Further, such movement is usually accompanied by the departure or arrival of certain commercial types of businesses. Such changes subsequently impact the types and frequencies of crime and the demand rates for CFSs, while requiring redistribution of available resources in order to cope with the new population's structure. It is thus necessary that ERS managers regularly collect data concerning the composition of, and changes in, the regional population and its businesses, as well as changes in the types and rates of crime by district, by time of the day and day of the week.

In light of these considerations, it would thus appear that the key to effectiveness is to continuously manage, control, and allocate available resources wherever they are most needed. We believe that is where generalized and portable simulation models, such as the one presented in this paper, might prove useful to management as a decision support tool. This is due to the fact that it uses feedback to constantly evaluate the ERS, thus allowing for indication of potential trouble areas and points possibly requiring corrective measures.

CONCLUSION

This paper has presented a versatile, easy to implement, and, most significantly, practical simulation model for resource allocation and management of an ERS. Although different types of ERSs have their idiosyncrasies, they do have many common characteristics. To enhance the proposed model's compatibility, several options were included. However, care was taken not to render the model too complicated to comprehend, use, or adapt by system administrators or end users. The availability of such decision models, i.e. those that are easy to use and inexpensive to obtain, will likely be of great value to management both for control of short-term operations as well as long-term planning.

A basic assumption in the paper is that a short elapsed time exists between the dispatcher's receipt of a CFS and his/her issuing a request to dispatch an emergency unit; thus, the dispatching time was set at 30 s. This assumption was confirmed by previous research [29]. However, due to recent substantial changes in the composition and size of populations in several large cities, as well as the existence of tight budgets, this assumption is now likely to be less valid. In Washington D.C., for example, Vogel [1] observes that "The 911 communications system gets hit from both ends: Not only are there not enough cars on the streets, there are also not enough operators and dispatchers to handle incoming calls." Elsewhere, the total calls to Columbus, Ohio's system jumped from 222,000 during July 1987–June 1988 to 310,000 in 1994–1995, while in New York City, the number of 911 calls is expected to grow to 12.5 million by the year 2005. Recent evidence indicates that the volume of 911 calls in Los Angles rose 70% during the past twelve years [2]. At the same time, due to tight budget constraints, the number of operators and dispatchers has not grown to accommodate these increases.

Other factors [2] that will likely adversely affect the 911 communication system are:

1. The 911 number is easy to remember and the public is calling in for anything. In some places, 90 percent of the calls are for non-emergencies.
2. The proliferation of cellular phones. Every year, almost 18 million additional 911 calls are made from cellular phones. In California, cellular 911 calls to the Highway Patrol jumped from 29,000 in 1985 to 2.8 million in 1996.
3. The high cost of renovating the 911 communications systems. In 1992, Los Angles approved a $325 million bond issue for a massive 911 upgrade. New 911 systems in Chicago and New York cost $217 million and $156 million, respectively.

The effectiveness of an ERS is a function of the efficiency of the 911 dispatchers and the efficient allocation of emergency resources. Whiles researchers have addressed the second issue, more research should be directed towards optimization of the responsiveness of the communications system itself. In this regard, vast technological advances have been achieved since the establishment of the 911 system in February 1968. These can almost certainly help rejuvenate the current ailing system so that it can handle more calls, and possibly redirect calls to different resources based upon their degree of severity.

Acknowledgements—The authors wish to thank the anonymous referees for their comprehensive and helpful comments. We also wish to thank the personnel in the Department of Public Safety for the City of Richmond for their assistance in providing us with very orderly and elaborate statistics.

REFERENCES

1. Vogel, S., A Lifeline Unravels. *The Washington Post Magazine*. June 2, 1996, No. 180. Washington D. C., pp. 11–15.
2. Witkin, G. and Guttman, M., Please hold. *U. S. News and World Report*. June 17, 1996, pp. 31–38.
3. Toregas, C., Swain, C., ReVelle, C. and Bergman, L., The location of emergency service facilities. *Operations Research*, 1971, **19**, 1363–1373.
4. Church, R. and ReVelle, C., The maximal covering location problem. *Papers of the Regional Science Association*, 1974, **32**, 101–118.
5. White, J. and Case, K., On covering problems and the central facilities location problem. *Geographical Analysis*, 1974, **6**, 281–293.
6. Daskin, M. and Stern, E., A hierarchical objective set covering model for emergency medical service development. *Transportation Science*, 1981, **15**, 137–152.
7. Hogan, K. and ReVelle, C., Concepts and applications of backup coverage. *Management Science*, 1986, **32**, 1434–1444.

8. Pirkul, H. and Schilling, D., The siting of emergency service facilities with workload capacities and backup service. *Management Science*, 1990, **34,** 896–908.
9. Batta, R. and Mannur, M., Covering-location models for emergency situations that require multiple response units. *Management Science*, 1990, **36,** 16–23.
10. Larson, R., A hypercube queuing model for facility location and redistricting in urban emergency survival. *Computers and Operations Research*, 1974, **1,** 67–95.
11. Larson, R., Approximating the performance of urban emergency service systems. *Operations Research*, 1975, **23,** 845–868.
12. Halpern, J., The accuracy of estimates for the performance criteria in certain emergency service queuing systems. *Transportation Science*, 1977, **11,** 223–242.
13. Benveniste, R., Solving the combined zoning and location problem for several emergency units. *Journal of the Operational Research Society*, 1985, **36,** 433–450.
14. Chung, H., Recent applications of the maximal covering location planning (M.C.L.P.) model. *Journal of Operational Research Society*, 1986, **37,** 735–746.
15. Current, J. and Storbeck, J., Capacitated covering models. *Environment and Planning B*, 1988, **15,** 153–164.
16. Pirkul, H. and Schilling, D., The maximal covering location problem with capacities on total workload. *Management Science*, 1991, **37,** 233–248.
17. Heller, M., Cohon, J. and ReVelle, C., The use of simulation in validating a multiobjective EMS location model. *Annals of Operations Research*, 1989, **18,** 303–322.
18. ReVelle, C., Siting ambulances and fire companies: New tools for planners. *Journal of the American Planning Association*, 1991, **57,** 471–484.
19. Cobham, A., Priority assignment in waiting line problems. *Operations Research*, 1954, **2,** 70–76.
20. Larson, R., Models for the allocation of urban police patrol forces. M.I.T. Operations Research Center Technical Report No. 44, Boston, 1969.
21. Stevenson, K., Operational aspects of emergency ambulance services. M.I.T. Operations Research Center Technical Report No. 61, 1971.
22. Adams, R. and Barnard, S., Simulation of police field forces for decision-making in resource allocation. *Law Enforcement Science and Technology III*, IIT Research Institute, Chicago, Port City Press, 1970.
23. Carter, G. and Ignall, E., A simulation model of fire department operations: Design and preliminary results. *IEEE Transactions on Systems Science and Cybernetics*, 1970, Vol. SSC-6, 282–293.
24. Savas, E., Simulation and cost-effectiveness analysis of New York's emergency ambulance service. *Management Science*, 1969, **15,** 608–627.
25. Goldberg, J., Dietrich, R., Chen, R. J., Mitwasi, M., Valenzuela, T. and Criss, E., Validating and applying a model for locating emergency medical vehicles in Tucson, AZ. *European Journal of Operational Research*, 1990, **49,** 308–324.
26. Chaiken, J. and Larson, R., Methods for allocating urban emergency units: A survey. *Management Science*, 1972, **19 December,** 110–130.
27. Gass, S., Models in law enforcement and criminal justice. In *A Guide to Models in Government Planning and Operations*, Edited by Gass, S and Sisson, R., Sugar Books, 1975, pp. 231–275.
28. Hobgood, L. Jr., Interviews and correspondence, Department of Public Safety, Bureau of Emergency Communications, City of Richmond, (1993–1994).
29. Walker, W., Chaiken, J. and Ignall, E. (ed.) *Fire Department Deployment Analysis: A Public Policy Analysis Case Study*. The Rand Fire Project, North Holland, New York, 1979.
30. Law, A. and Vincent, S., *UniFit II User's Guide*, Averill M. Law and Associates, Tuscon, AZ, 1991.
31. Law, A and Kelton, W. D., *Simulation Modeling and Analysis*. McGraw Hill, Inc., New York, 1991.
32. Pritsker, A., *Introduction To Simulation and SLAM II*, Third Edition, Systems Publishing Corporation, West Lafayette, Indiana, p. 707.