Socio-Econ. Plan. Sci. Vol 5, pp. 491–499 (1971). Pergamon Press. Printed in Great Britain

MANAGEMENT SCIENCE APPROACHES TO THE DETERMINATION OF URBAN AMBULANCE REQUIREMENTS*

WILLIAM K. HALL

The University of Michigan, Ann Arbor, Michigan 48104

(*Received* 9 *February* 1971)

This research paper summarizes a systematic approach to the determination of urban ambulance requirements. Quantitative measures of ambulance system performance and alternative system configurations were developed, and these were examined for the city of Detroit, Michigan. Sampling techniques and data analysis procedures were used to develop an analytical model of the Detroit emergency recovery system, and this model was then utilized to predict the performance of alternative system configurations under varying operating policies. Results and conclusions are presented for the specific system under consideration; extensions of these techniques to other urban emergency systems are also suggested.

THE DEVELOPMENT of effective methods for transporting the victims of emergencies to a hospital is a major problem in contemporary society. Various studies have estimated that 18–20 per cent of the 50,000 annual traffic fatalities in the United States could be avoided with more adequate emergency medical services. Although comparable statistics are not available for medical emergencies (heart attacks, strokes, etc.) and other accidental injuries, it seems clear that the failure to provide rapid, effective care for these emergencies leads to increased morbidity and mortality for many victims.

This paper reports on a study conducted in Detroit, Michigan to consider one aspect of the emergency medical care problem—the determination of the number and placement of ambulances necessary to provide adequate service in an urban area. Increasing the number of ambulances assigned to a service area and placing these vehicles closer to the sites of potential emergencies should reduce the system response time. This reduction will place an emergency victim into contact with appropriate medical care more rapidly, and this rapid care should reduce the extent of victim morbidity and mortality.

Consequently it would appear that the effects of changing the number and placement of recovery vehicles (the recovery system configuration) can be evaluated in terms of changed victim mortality/morbidity indices. Unfortunately, two factors make such evaluation criteria inappropriate. First, victim morbidity/mortality is influenced by many factors in addition to the ambulance system configuration. The quality of ambulance attendant training and medical capability, the victim's pre-emergency physical condition, the severity of the emergency and the quality of medical treatment in the hospital emergency room all influence the victim's condition. Since the effects and interactions between these

* This paper is based upon remarks prepared for the 1971 Systems Engineering Conference of the American Institute of Industrial Engineers, Phoenix, Arizona, February 11–13, 1971. A portion of this research was supported by The University of Michigan Highway Safety Research Institute under U.S. Department of Transportation Contract FH-11-6901.

confounding factors are not well-understood at the present time, it is essentially impossible to separate and measure the effect of the recovery system configuration on indices of morbidity and mortality. Second, all current indices of morbidity/mortality suffer from severe definitional and measurement problems. For instance, victim morbidity/mortality might be measured by the lost productivity due to the emergency, the cost of the emergency, the time spent in the hospital or some other measure of severity. However, productivity is very difficult to quantify, no one is quite sure how to correctly allocate intangible costs to emergencies and some investigators have conjectured that time spent in the hospital is more closely correlated with the extent of the victim's insurance than with the severity of the emergency.

For these reasons, it is necessary to use intermediate measures of effectiveness* to evaluate alternative ambulance system configurations. The measures chosen for this study are directly related to the recovery system response time, and the effects of alternative system configurations on these measures can be summarized as follows: increasing the number of vehicles in the system reduces the probability that no vehicle is available to respond to a medical emergency. Consequently, the average time between emergency reporting and vehicle dispatch is reduced or, alternatively, the likelihood that the emergency must be handled by an auxiliary vehicle from outside the service region is decreased. Improvements in the location of vehicles within the service region potentially have a dual effect: first, with vehicles located closer to potential emergency sites, the average transit time to the scene of emergencies is reduced. Second, this reduction in transit time will result in a reduction in the total time the ambulance is out of service. Ambulances will consequently be available a larger portion of the time, again reducing the time between emergency reporting and vehicle dispatch.

The objective in this study was to develop a mathematical model to quantify the qualitative effects of alternative recovery system configurations introduced above. Such a model can then be utilized to predict changes in ambulance system performance which can be attributed either to increasing the number of recovery vehicles or to changing the position of these recovery vehicles within the service region.

It was felt that a mathematical model designed to examine these effects would be highly dependent upon the characteristics of the emergency occurrence and service processes in the area under consideration. Therefore, data on these processes in the city of Detroit were collected and analyzed. The data analysis revealed several subtle and important features of this process which were utilized in the model development. The resulting analytical model was then used to study a wide range of alternative system configurations.

Two alternative approaches to this problem have recently appeared in the literature. Bell and Allen [1] have developed a general queueing model for exploring the effects of changing the number of vehicles in a service region. The data analyses conducted in Detroit indicated that the special assumption of this model were inappropriate for the service region under consideration. Savas [8], in a study of the New York City ambulance system, used simulation as a basis for recommending several changes in the system configuration. The analytical model constructed in this study offers an interesting and useful alternative to these simulation efforts.

The data analyses and the subsequent model development are summarized in the next

* The reader is referred to Hall and O'Day [3] for a systematic approach to the development of intermediate measures in highway safety counter-measure programs.

two sections. Then various prediction of system performance drawn from the model are discussed, and recommendations based upon the model analysis are summarized.

While these numerical results and recommendations from this study apply specifically to the city of Detroit, the underlying data analysis and model-building should be useful to other urban areas contemplating changes in their public ambulance systems. In addition, the concepts underlying the analysis of this study can be applied to study requirements for police, fire and other emergency systems in an urban area.

DATA ANALYSIS

The city of Detroit and most major cities in the United States are divided into police precincts for emergency command and control. Since the police operate the ambulance services in Detroit (and in one-half of the fifteen largest cities in the U.S.), the precincts also serve frequently as basic geographic units for ambulance allocation.

Two precincts in Detroit were chosen as sampling units for this study. These were selected as a representative sample of the thirteen Detroit precincts in terms of demographic characteristics and accident experience. One of the precincts was a densely populated "inner city" precinct and the other was a more sparsely populated "outer city" precinct.

In these two precincts exhaustive samples of all dispatches for emergency medical runs were obtained and analyzed for the one month period November 18–December 15, 1968. Police emergency runs were similarly analyzed for the same time period in order to make inferences on the operation of a "dual function" police–ambulance system—where police vehicles respond to medical emergencies and police calls on a first come-first served basis. (A situation which prevails in most urban areas with a police-administered public ambulance system.)

While data from this one month time frame precluded direct inferences on long term characteristics of the emergency occurrence and service processes, the series were long enough to facilitate an examination of potential hourly, daily and weekly trends or cycles. Results from these examinations provided some indications as to the structure of the long term series, and model parameterization was then utilized to study system response to changes in such structure.

In each precinct the sequence of times between emergency occurrences was partitioned into two sequences—the times between medical emergencies and the times between police emergencies. These sequences were subjected to statistical analyses in order to make inferences on the probabilistic nature of the processes underlying the realizations. Analyses to determine the extent and nature of the interactions between the two sequences were also conducted.

These analyses revealed that the medical emergency occurrence process in the study precincts had no observable trends or cycles within the one month time frame. On the other hand, the police emergency occurrence process had a significant daily cycle. Further analysis revealed that this daily cycle arose because of the higher intensity of police calls between 0800–2400, and consequently the mathematical model was developed to predict performance under these more severe operating conditions. Statistical analyses* indicated that the times between medical emergencies followed a Gamma distribution with a mean time between emergencies of 101 min and with a coefficient of variation of 1.15. Similar

* The majority of the statistical analyses were performed using an IBM 360/67 computer program for the statistical analysis of series of events (SASE II). This program is documented by Cox and Lewis [2] and Lewis [7].

analyses indicated that the times between police calls in the 0800–2400 time interval followed a Poisson process with a mean time of 17.2 min. No significant inter-precinct differences were found in the occurrence processes for either class of emergencies.

In analyzing the dual-function police ambulance system, it was necessary to develop a model for the combined occurrence process of police and ambulance calls. It was assumed that these processes might be interdependent (as in the case where a certain portion of police calls give rise to ambulance calls), and hence a semi-Markov model with a bivariate state space (indicating whether the call was a police or ambulance call) was postulated for this combined occurrence process.

Various statistical inference techniques were used to estimate the transition distributions in this process. These analyses indicated that state transitions in the hypothesized semi-Markov emergency occurrence process were Bernoulli with independent, negative–exponential times between transitions.

Police call and ambulance call service times were analyzed separately to determine the stochastic characteristics of these processes. The relationship between transit distances and transit times for ambulance calls was analyzed in detail because of the importance of this relationship on vehicle distribution policies. Regression analyses on transformed and untransformed data indicated that only 25 per cent of the variation in transit times was explained by transit distance. Consideration of other factors—weather conditions, traffic conditions and type of emergency, did little to explain the residual variation.

Further analysis of the distance–time relationships revealed that times were significantly longer for ambulance runs greater than 1 mile than for ambulance runs less than 1 mile. No other significant differences were found. One hypothesized explanation for this difference is that ambulance response characteristics differ for "nearby" and more distant emergencies. A significant portion of the transit time to nearby emergencies is apt to involve the performance of certain "fixed" tasks—acknowledgement of the call, determination of the route and initiation of the run. Travel to such emergencies will generally proceed over city streets. On the other hand, when longer transit distances prevail, expressway travel at higher speeds is more likely. A second hypothesized relation for the lack of a more precise distance–time relation is behavioral. The times to the scene of emergencies clustered about 5 min, and it is possible that the police ambulance attendants adjust their rate of response to this target time—a target which seems to be acceptable to the community being served. Regardless of the explanation, extensive analysis of the distance–time relationship did indicate that a "two-level" mean service time model was sufficient for medical emergencies. Statistical analyses of the resulting data indicated that the service time processes all follow a negative exponential distribution, with mean times of 18.4 min for ambulance calls having transit distances of less than 1 mile, 20.5 min for ambulance calls having transit distances greater than 1 mile, and 38.3 min for police calls.

MODEL DEVELOPMENT

Based upon the data analysis of emergency occurrence and service processes, a mathematical model of the urban ambulance system was developed utilizing certain concepts from queueing theory. This model incorporated the following important characteristics the system:

1. A semi-Markov arrival process corresponding to the joint occurrence of police and ambulance calls over time.

2. A multi-function stochastic service process depending both upon the type of emergency being served and upon the particular vehicle assigned to the emergency (since vehicles within 1 mile of a medical emergency had a significantly different service rate than those further than one mile from the emergency).
3. A multiple channel service system (with channels corresponding to recovery vehicles assigned to the system) incorporating a control mechanism which selectively routes emergencies to vehicles for service. This control mechanism is important since all police calls will not be routed to dual-function vehicles. Instead, a significant proportion will be routed to scout cars within the precinct. Moreover, whenever possible it is desirable to route medical emergencies to available recovery vehicles which are within 1 mile of the emergency. Furthermore, it may be desirable to develop a protective dispatch policy, which routes all police calls to scout cars in certain situations, thereby maintaining the availability of dual function vehicles for ambulance service. The control mechanism was developed to allow all of these aspects of system operation to be examined.

These three characteristics of the dual function ambulance system suggest modeling the problem as a multi-function stochastic service system with state-dependent server selection. The mathematical analysis of the resulting model for this system is complex, and this development is not presented here.* However, the general structure of this model is presented in some detail. Specifically, consider the following conceptualization of the dual function urban ambulance system (Fig. 1).

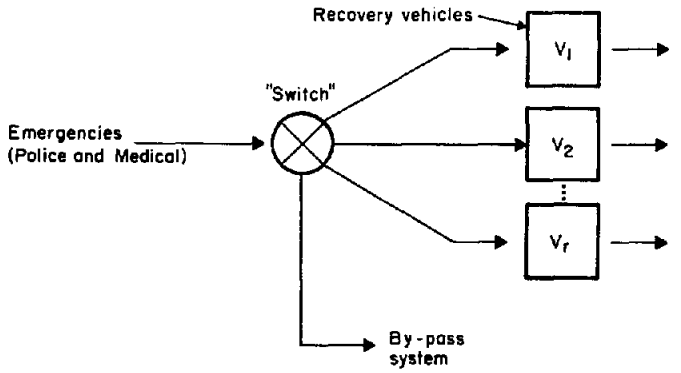


FIG. 1. Conceptual model for an urban ambulance system.

The sequence of emergencies serving as inputs to the recovery system can be described by the bi-variate random sequence (X_n, Z_n), where X_n is the time between the nth and $(n-1)$st emergency and Z_n is the type of the nth emergency ($Z_n = 1$ if the emergency is a police call and $Z_n = 0$ if the emergency is an ambulance call). This bi-variate process is the semi-Markov emergency occurrence process identified in the data analysis.

Assume that r recovery vehicles have been allocated to the system. Each of these vehicles will be in one of four "states" at the instant of the nth emergency occurrence.

* The reader is referred to Hall [5, 6] for a complete discussion of this model development and analysis.

Let S_n^j be a random variable denoting the state of vehicle j at this instant of time, and let the states of this vehicle be described as follows:

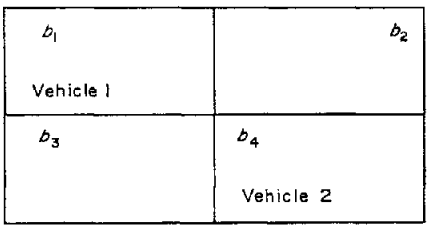
$$S_j^n = \begin{cases} 0 & \text{Vehicle } j \text{ is idle (available).} \\ 1 & \text{Vehicle } j \text{ is serving a police call.} \\ 2 & \text{Vehicle } j \text{ is serving an ambulance call which requires less than 1 mile transit distance to the scene.} \\ 3 & \text{Vehicle } j \text{ is serving an ambulance call which requires more than 1 mile transit distance to the scene.} \end{cases}$$

The states of all vehicles at the instant of the nth emergency occurrence can then be denoted by the vector-valued random process $(S_n^1, S_n^2, \ldots, S_n^r)$.

The switch which assigns emergencies to vehicles for service can also be described by a two-state process—the number of the recovery vehicle selected and the type of service rendered by this vehicle. Let Y_n denote the assignment of the nth emergency. The Y_n is a random variable which takes on the value $(k, 1)$, where k is the vehicle selected, $k = 1, 2, \ldots, r$, and 1 is the type of service required, $1 = 1, 2, 3$.* In addition to these values, Y_n is allowed to take on the value zero when emergencies are not routed to one of the recovery vehicles—when they "bypass" the recovery system and are serviced by auxiliary vehicles.

Observe that the selection of a particular $(k, 1)$ combination depends upon the type of emergency Z_n and upon the state of all r recovery vehicles $(S_n^1, \ldots, S_n^r)$. That is, the probability distribution associated with the switch Y_n is a conditional distribution of the form $\Pr[Y_n | Z_n, S_n^1, \ldots, S_n^r]$. Probabilities can be assigned in this conditional distribution in a manner illustrated in the following example.

Consider an over-simplified service area divided into four scout car territories with two single function ambulances assigned to the service region. The probability that a medical emergency occurs in the jth scout car territory is denoted by $b_j, j = 1, 2, 3, 4$. The two ambulances are placed in territories 1 and 4. A schematic diagram of the situation is presented below:



In this example it is assumed that all travel distances are less than 1 mile unless the vehicle travels between scout car territories one and four. In this latter situation, the transit distance is assumed to be greater than 1 mile. When both vehicles are idle and equidistant from an emergency, a random vehicle selection mechanism is used.

For instance, consider $\Pr(Y_n = (1,2) | Z_n = 0, S_n^1 = 0, S_n^2 = 0)$. Here both vehicles are idle. Thus vehicle one is dispatched on a "short" run if the emergency is in territory one (with probability b_1). Vehicle one is also dispatched on a "short" run one-half of the

* Taking on the value $1 = 1$ when the call is a police call, $1 = 2$ when the call is an ambulance call requiring less than 1 mile transit distance to the scene and $1 = 3$ when the call is an ambulance call requiring more than 1 mile transit distance to the scene.

time when the emergency is in regions two or three. The resulting conditional probability is $b_1+(b_2+b_3)/2$.

An alternative switching model was developed for police emergencies ($Z_n = 1$). Since there is some probability that a police call will be serviced by a non-recovery vehicle even when one or more recovery vehicles are idle a non-zero conditional probability was assigned to this event. Two alternative dispatch "strategies" were assumed. In the first a constant probability was assumed for all levels of recovery vehicle availability. In the second, it was assumed that as recovery vehicles are assigned to emergencies, the idle recovery vehicles are "protected" by reducing the probability that they will be assigned to service a police call. For both strategies, in the event a police call is accepted by the recovery system, it was assumed that it is equally likely that any available vehicle will service this call.

The exact probability a police call will be serviced by one of the dual function vehicles when one or more of these are available is difficult to estimate in the existing recovery system in Detroit, and it probably varies under alternative operating situations. Hence, this parameter was used as an independent variable, and results were derived for a range of potential values. Specifically, the probability of accepting a police call when two recovery vehicles are available in a two vehicle system was allowed to vary from zero (a single function pure ambulance system) to 0.5.

Once an emergency has been assigned to a vehicle and the type of service has been selected, service on this emergency proceeds according to a negative exponential distribution as identified in the data analysis. Consequently "service transition probabilities" of the form $\Pr(S_n^1, S_n^2, \ldots, S_n^r / Y_{n-1}, S_{n-1}^1, \ldots, S_{n-1}^r, X_n)$ were computed as products of exponential functions.

Furthermore, it can be shown (Hall and Disney [4]) that the stochastic process $(S_n^1, S_n^2, \ldots, S_n^r, Z_n)$ is a Markov chain. Transition probabilities for this chain can be easily found by manipulating the conditional switching probabilities, the service transition probabilities and the probabilities associated with the arrival process. From these transition probabilities, equilibrium probability results were derived describing the system performance as a function of alternative system configurations and operating policies.

MODEL ANALYSIS AND RESULTS

The analytical techniques described above were programmed into a Fortran IV computer program for use on The University of Michigan IBM 360/67 computing system. The resulting program was then utilized to generate numerical results for alternative systems. Four allocation policies were examined—recovery systems with one, two, three and four vehicles assigned to a police precinct.* For each of these allocations, three distribution policies were examined:

1. Assigning the vehicles to the sub-regions of heaviest demand for ambulances.
2. Assigning the vehicles "uniformly" throughout the precinct on a geographic bases.
3. Assigning all the recovery vehicles to a single fixed station within the precinct (the precinct police station).

* In utilizing the computer program, it was found that the four vehicle system leads to very time consuming and expensive computation. Hence this system was approximated by two-vehicle systems operating independently in adjacent half-precincts. This approximation should closely represent actual operating policy in such a system.

Numerical results derived from the equilibrium distributions were then used to characterize system performance for each of these policies.

It was found that the probability that one or more vehicles are available at the instant of a medical emergency varies from 0.79 for a single vehicle function system to approximately 1 for a four vehicle system. The exact results are given below:

TABLE 1

No. of vehicles	Probability one or more vehicles are available
1	0.789
2	0.976
3	0.998
4	1.000

The vehicle availability declines as a function of the percentage of police calls accepted for service as shown in Fig. 2. However, the rate of decline is less for allocations with a larger number of vehicles.

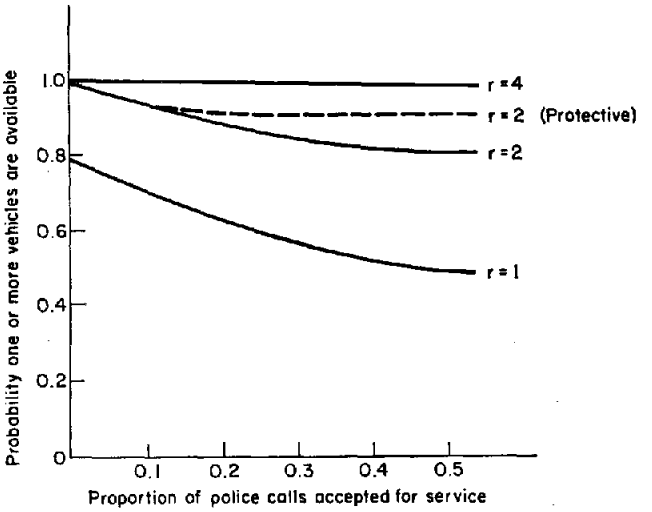


FIG. 2. Recovery vehicle availability in a dual function ambulance system.

The idea of utilizing the protective dispatch policy discussed earlier to reduce this decline in vehicle availability was studied for the two vehicle systems. Results for this policy are shown in the dashed line of Fig. 2, and they indicate that 90 per cent vehicle availability can be obtained under a protective policy, even when 50 per cent of the incoming police calls are accepted for service when both vehicles are idle.

The probability that one or more vehicles are available at the instant of a medical emergency was found to be almost completely insensitive to the placement of vehicles within the precinct. When the ambulances are placed in the subregions of highest demand, the probability that a vehicle is available to respond to an emergency requiring less than 1 mile transit distance is slightly higher than the corresponding probability for the "uniform" distribution policy, and it is much higher than the corresponding probability for the single station system.

However, the improvement decreases as the number of vehicles increases. For example, the percentage increase in the probability of responding to an emergency requiring less than 1 mile transit distance when vehicles are assigned to sub-regions of maximum demand is tabulated below:

TABLE 2

No. of vehicles	Improvement over "uniform" policy (%)	Improvement over "single station" policy
1	28.4	–
2	15.2	80.5
3	11.7	69.4

On the basis of these numerical results it was recommended to the city of Detroit that two recovery vehicles be assigned to a police precinct. A single ambulance system does not appear to provide adequate vehicle availability, and three and four vehicle systems provide only marginal increases in availability over the two vehicle systems. Based on qualitative considerations, 95 per cent vehicle availability in a single function two-vehicle system and more than 90 per cent availability in a dual function two-vehicle system with a protective dispatch policy were felt to be reasonable for the type of medical emergencies and relatively short ambulance service times encountered in the urban area under consideration.

A "multiple" station distribution policy was also recommended. Because of the similarities in performance characteristics for the "highest demand" and "uniform" policies, the choice between these two was left to city decision-makers on the basis of other considerations.

REFERENCES

1. C. E. BELL and D. ALLEN, Optimal Planning of an Emergency Ambulance Service, presented at Operations Research Society of America 35th National Meeting (1969).
2. D. R. COX and P. A. W. LEWIS, Computer program for statistical analysis of series of events, *IBM Res. J.* (1965).
3. W. K. HALL and J. O'DAY, Causal chain approaches in the evaluation of highway safety countermeasures, *J. Saf. Res.* **3**, 9–20 (1971).
4. W. K. HALL and R. L. DISNEY, Systems of queues in parallel under a generalized channel selection rule, *J. appl. Probability* to appear.
5. W. K. HALL, Multi-function stochastic service systems with state-dependent server selection, submitted for publication.
6. W. K. HALL, A queueing theoretic approach to the allocation and distribution of ambulances in an urban area, submitted for publication
7. P. A. W. LEWIS, A computer program for statistical analysis of series of events, *IBM Systems J.* 202–225 (1965).
8. E. S. SAVAS, Simulation and cost–effectiveness analysis of ambulance service in New York, *Mgmt Sci.* **15**, 608–627 (1969).