AN EMERGENCY MEDICAL SERVICES SIMULATION MODEL FOR BALTIMORE CITY: AN OVERVIEW

Miriam Heller, Kathleen B. Hogan, Perry A. Appino,
Jared L. Cohon, Charles S. ReVelle

The Johns Hopkins University
Baltimore, Maryland 21218

Emergency medical services, particularly in urban areas, are hardpressed to respond to the physical and demographic changes observed in many cities. Spatial and temporal demand changes must inevitably alter the quality of service in urban areas unless there is an ongoing response to these changes. Municipalities face the complex task of making long-range location decisions coupled with shorter-term deployment and dispatching strategies to provide the best possible service. These decisions must be made under increasingly limited budgets. Simulation is a very useful methodology for assessing the performance of emergency systems under various scenarios of demand for and supply of services.

The purpose of this paper is to describe a simulation model developed under a contract with the Baltimore City Fire Department. The model, which was built to evaluate possible modifications in the ambulance system, addresses, among other things, demand in terms of the appropriate medical response, the multiple legs of each ambulance run, and the high work load on each vehicle.

BACKGROUND: EMERGENCY MEDICAL SERVICES IN BALTIMORE

A brief overview of the Baltimore City system is necessary before examining the simulation model in detail. The emergency medical services are comprised of 16 medic units (ambulances) and four reserve medic units. All medic units are housed in Baltimore City fire stations. Of the 50 city fire stations, 22 can accommodate medic units. Presently, 16 fire stations serve as the home bases of the regular medic units. The reserve ambulances operate out of four of the 16 stations.

Calls for service are received by a single dispatching unit; minimal screening is performed. The dispatcher deploys the closest available medic unit to the call scene. Available ambulances are (1) those at home base, (2) those clearing a hospital after servicing a call, (3) those leaving a call scene when transport of a patient is not required. The no transport case includes false calls, refused transport by the patient, and incidents handled by the police. The medic unit personnel, certified cardiac rescue technicians and emergency medical technicians, have the authority to determine that a patient does not require transport or to request alternative transport equipment. If there are no available medic units, the dispatcher immediately deploys a fire suppression unit and then the first medic unit to become free. Further dispatching decisions reflect overall city coverage.

Certain situations require special response strategies. Reserve ambulances are mobilized during generally peak demand periods. This situation, referred to as a medic Red Alert, occurs when the number of non-busy ambulances falls to only two of the regular units. The reserve medic units are demobilized at the dispatcher's discretion, usually reflecting a return to overall city coverage. Multiple calls, calls requiring more than one medic unit, apply additional stress on the system.

Unlike other emergency service vehicles, which run to a scene and return to base, a medic unit has the trip to a hospital from a call scene as an additional part of its journey. The length of this leg depends on the nature of the patient's medical problem, and the hospital promixity, hospital type, and hospital alert status. A medic unit generally transports a patient to the nearest hospital but there are several exceptions. Proper medical care may require transport to a specialty referral center. Specialty centers in Baltimore include shock trauma, burn, pediatric trauma, neonatal care, eye trauma, hand, area trauma and central nervous system - spinal cord centers.

Hospitals in alert status may also prohibit transport to the nearest hospital. During a Mini-Disaster alert at a hospital, any situation in which the hospital emergency room capability is over-burdened, no patients may be received.

Proceedings of the 1982
Winter Simulation Conference
Highland * Chao * Madrigal, Editors

82CH1844-0/82/0000-0413 $00.75 © 1982 IEEE

A hospital Red Alert posture is assumed when emergency room cardiac care units are at full capacity. During a hospital Red Alert, stable cardiac patients may be transported to the second closest hospital. If the two nearest hospitals are both on Red Alert, the medic unit transports a patient to one of them, on an alternating basis. Unstable cardiac patients are transported to the closest hospital, regardless of the Red Alert status.

A SIMULATION MODEL FOR EMERGENCY MEDICAL SERVICES

Introduction

The simulation model of the Baltimore City emergency medical services incorporates the operating procedures discussed above, along with parameters derived from an extensive analysis of demand and medical data so as to capture the dynamic features of the actual system. The simulation model can aid in formulating appropriate policies by answering "what if" questions. For example, what would happen to system performance if an additional ambulance were located at a specified station? Or what would happen if an ambulance currently located at station A were moved to station B?

These questions are addressed by considering variations in total and zonal demand with respect to season, day of week, and time of day. The effectiveness of these policies is evaluated by examining changes in several primary system statistics. For example, a policy resulting in a decrease in average response time, the average time from departure to arrival at the call scene, may reflect a superior home base assignment configuration. Similarly, a policy producing a more equal distribution of work load among the medic units may be desirable; the most recent data from Baltimore City exhibited annual run totals per vehicle ranging from 2934 to 6908. Detailed information about these variations may suggest that demand could be better accommodated by varying the spatial and numerical deployment of medic units with time of day. Finally, since improvements in one statistic may coincide with a decrease in some other measure of effectiveness, the simulation can be used to elucidate the trade-offs involved in policy decisions.

The simulation program, ASSIST (Ambulance Service System with an Interactive-Option Simulation Technique), is an original simulator with incorporated subroutines for input, interactive medic unit allocation with graphics display, and post-simulation analysis. The program is coded in FORTRAN 77. The generality of packaged procedural languages such as SIMSCRIPT, SLAM II, GPSS, was supplanted by a fast-running program, streamlined for the Baltimore system. However, given similarities in emergency medical system operations in metropolitan areas, the structure and modularity of the program is sufficiently general so that, with minor modifications, it could be applied to other cities.

ASSIST simulates the ambulance service system by stochastically imitating events occurring during ambulance service operations. That is, the model operates in discrete time intervals defined by these events. A simulation run requires the user to provide the computer program with the time period of simulation and a seed to begin random number generation. By using an interval timing mechanism, the system clock, which is updated after isolating the next event in a series of "future" events, the simulator assigns medic units to calls, generates a description of each medic unit run, places on a queue those calls that cannot be serviced immediately, implements medic unit Red Alert policy, and collects frequencies for final statistical summaries necessary for post-simulation analysis. Medic unit home base assignments are read in and stored so modifications to this input file reflect medic unit relocation, the primary policy option.

Database Source

Detailed information about each medic unit run is recorded on a run ticket prepared by the medic unit staff. 5,963 run tickets were coded into the computer with the aid of Vax-11 Datatrieve, a database management system used conjunctively with VAX-11 FMS, a forms management system. Statistical analysis was performed on samples of 1981 and some 1982 run tickets: the location of the call scene is recorded for each run and the appropriate transportation zone, defined in the next section, was assigned using a geo-based file provided by the Baltimore Regional Planning Council (RPC), a multijurisdictional planning body for the Baltimore metropolitan area. The same procedure was used to identify zones for the ambulance home bases and hospitals to which patients were transported. Times of call received, departure, arrival at call scene, arrival at the hospital, and return to station yielded a temporal picture of each run. A vast amount of medical data about each patient is entered on the ticket as well. Another source of data, cards prepared by the dispatchers, were obtained for a subset of the data; the cards recorded the time of departure from the call scene.

MODEL STRUCTURE

Geography and Movement of Medic Units

Spatial aspects are of great importance in determining appropriate responses to requests for emergency medical services. That is, the model must be able to determine the closest medic unit to a call and the closest hospital. Travel time is a crucial parameter, as it constitutes the major measure of effectiveness for the emergency medical system. Rather than imitate continuous movement of medic units, which would be extremely costly in terms of computing time and program complexity, medic unit movement is represented as movement between points. The model, therefore, is defined over a network of nodes.

The basic geographical unit in the model is the transportation zone. The RPC disaggregated the City of Baltimore into 207 discontinuous transportation zones. Travel time matrices were developed by the RPC for other RPC transportation studies. These matrices, which contain travel times between the activity centroids of all pairs of transportation zones under different traffic conditions, were used to estimate medic unit travel times in the model.

A medic unit run can be viewed in terms of movement between these geographical points and times spent at each location. Each of these intervals are called the "legs" of the run. A status vector, BUSY, associated with each medic unit defines the endpoints of the legs as follows:

For a non-transport run,

BUSY = 1: an unassigned medic unit at home base
= 2: an unassigned medic unit is dispatched
= 3: an assigned medic unit arrives at call scene
= 4: an assigned medic unit leaves the call scene.

For a transport run,

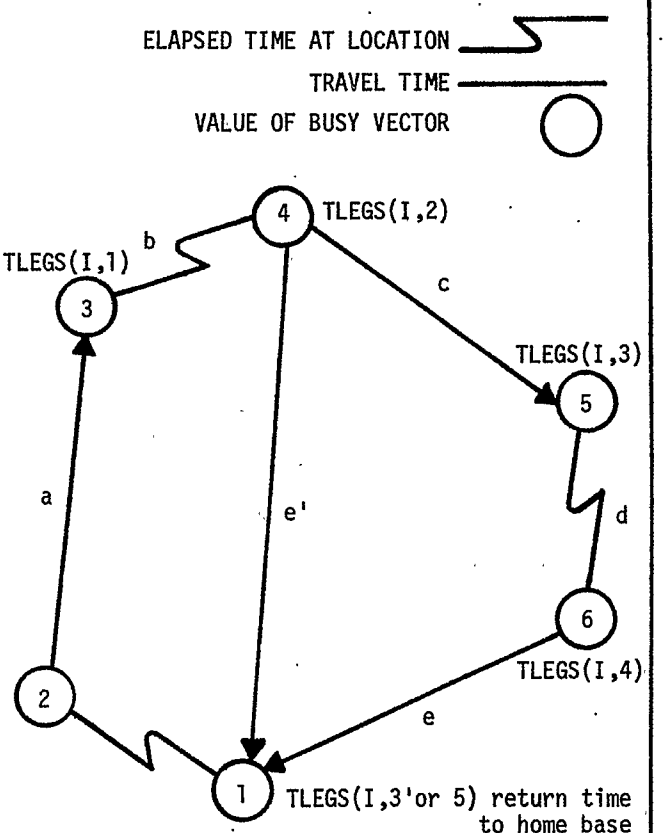
BUSY = 1-4: as above
= 5: an assigned medic unit arrives at hospital
= 6: an assigned medic unit departs the hospital.

For demobilized Red Alert medic units, BUSY = 0.

Associated with BUSY, is another array, TLEGS, which stores the time at each point. A run for medic I is defined as in diagram 1. The simulator does not account for the time between the arrival of the call at the main dispatching room and the departure of a medic unit from its fire station (time between BUSY=1 and BUSY=2).

Diagram 1

MULTIPLE LEGS OF A RUN FOR MEDIC UNIT I



Parameters determining the number of legs and the duration of each leg of a medic unit run were derived from the database. The type of call for service, stored in the status vector, TYPE, governs the number of legs, the length of time spent at the call scene, and the length of time spent at the hospital for a medic unit run. A run not requiring transport of a patient to a hospital usually requires less time than a run with the additional legs of c and d shown in diagram 1. The proportion of no transport calls relative to the total number of calls can vary by area of the city. In addition, the lengths of time spent at the call scene and at the hospital reflect the type of transport call. Calls involving suspected cardiac patients tend to exhibit longer times in both categories. An area of the city with higher frequencies of cardiac calls can monopolize more of the medic unit's time than would appear by solely examining number of calls per day by medic unit statistics. To maintain these system characteristics within the simulation, distribution functions were derived for type of call by call zone (an area of the city created by aggregating transportation zones, as discussed in the next section), length of time at

the scene by type of call, and length of time at the hospital by type of transport call. Type of call includes no transport, transport but not suspected cardiac case, and transported suspected cardiac case. A cardiac/non-cardiac distinction is made for runs involving transports.

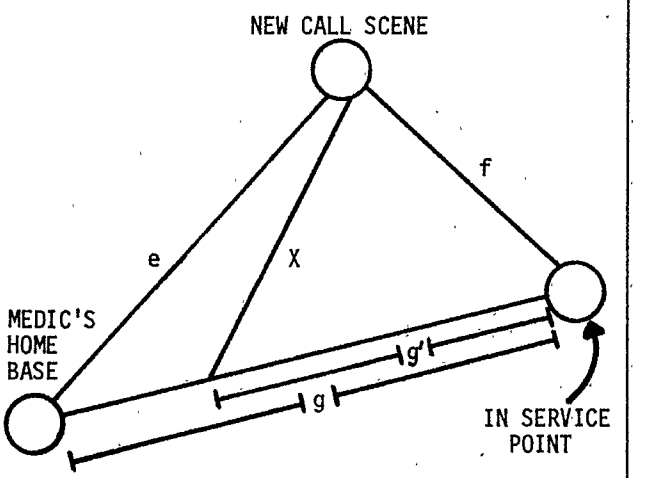
The legs a, c, and e of a medic unit run require input parameters. The first encountered travel time, a, is the most crucial as it constitutes the primary measure of system quality, the response time. To develop an accurate representation of these travel times, the RPC travel time matrices mentioned above were corrected to reflect the increased speed with which a medic unit responds to a call. An incremental time correlation was employed to provide travel times for transportation zone pairs which lacked substantial travel time data in the database. This method captures the observed variation in travel times. The correlation was performed using RPC matrices to identify all pairs of transportation zones with a particular travel time. The observed travel time was then collected for each occurrence of these pairs in the database thereby creating a distribution function of observed travel times which correspond to each RPC travel time. This procedure is followed for both peak and non-peak traffic conditions.

The last two travel times, e and e', involve the medic unit returning to home base. The model selects travel times directly from the appropriate RPC matrix for these cases. These travel times are not crucial to the simulation, since medic units are free to service the next call upon completion of legs b or d where appropriate. In addition, data on time returned to station from the database is inconsistent, often reflecting time out of service with mechanical difficulties and other peculiarities. The workload (defined as time spent servicing a call) associated with this run is a + b + c + d. The time, e, is not included in this measure of workload since a medic unit is considered available to service another call upon completion of its run.

The special case of en-route assignment involves a triangle approximation rule, where the sides of the triangle represent expected travel times. (See Diagram 2.) It is assumed that a medic unit leaving the hospital or leaving the call scene on a non-transport call travels back to its home base. The time of departure and the system clock define how far in terms of time the medic unit has traveled, g'. The expected time for the medic unit en-route to home base to arrive at the new call scene, X, is geometrically approximated. This method was employed since the RPC matrices only supplied times and not paths between transportation zones.

Diagram 2

ETA ESTIMATION FOR EN-ROUTE ASSIGNMENT



The medic unit attribute vectors, LOCSTA, TZONE, and LOCHOS, store the locations in terms of transportation zones of the medic unit home base, the call scene, and the hospital to which a patient is transported, respectively. By using the location attribute vectors and the aforementioned status vectors, the location of a medic unit is defined at any point in time.

Geography, Temporality and the Generation of Calls for Ambulance Service

The preservation of spatial aspects of the system also involves generating calls for ambulance service in a manner that mirrors their spatial distribution in the real system. For this purpose, a larger service area unit, the call zone, was obtained by aggregating between 4 and 18 transportation zones so as to equalize the total number of calls per call zone. Call generation which mimics these distributional features is dependent upon estimates of arrival rates of calls by call zone over the entire day. Mean arrival rates of calls were therefore determined for six consecutive four hour intervals for each of the twenty-four call zones. The mean arrival rates passed a chi-square uniformity of arrival rate test establishing a basis for a Poisson process call generator in the simulation. Calls are generated using the appropriate interval by time of day to reflect the non-stationary Poisson process. To further maintain the spatial aspects of demand, a call generated for a call zone is assigned to a transportation zone contained within the call zone. This

assignment reflects the proportion of calls in a call zone which originate in each transportation zone from the data base.

Model Design

The simulation program, ASSIST, consists of nine main subroutines. (Refer to diagram 3.) The first subroutine, INITIAL, initializes the system. The system clock is set to 0. All 16 regular medic units are located at home base with BUSY set to 1. The Red Alert Reserve units are initialized as demobilized.

The subroutine, INPUT, reads in all exogenous variables, probabilistic descriptors and location variables. Another subroutine, CALLGEN is called from INPUT to generate all system calls for the simulation run period, via random number generation. Random numbers are generated as needed using the International Statistics and Mathematics Library (ISML) subroutine, GGUBS, a congruential random number generator for uniform random number over [0,1].

The third subroutine, NEXTCALL, isolates the next system call and returns the call zone from which the call arose, LASTCZ and the call time.

Before assigning a medic unit to this next system call, the availability status of each medic unit must be identified. Therefore, the system must be updated to reflect any events, i.e. transitions within a medic unit run by calling the subroutine, EVENT. A search routine compares the critical times stored in TLEGS with the next call time. The vector, BUSY, and if necessary for reinitialization, TYPE, are updated, until the next event is the next system call. The queue is also tested. A call is taken off the queue when a medic unit is updated to reach the call assigned to it. When this occurs, the appropriate statistics are collected.

Call characteristics are defined in the subroutine, CHARS. The type of call and the transportation zone in which the call arises are assigned to TYPE and TZONE by assigning values to two random numbers based on the proportional distribution of these characteristics in LASTCZ.

Before handling a call it is necessary to examine the medic unit Red Alert posture. The subroutine, REDALERT, mobilizes or demobilizes these reserve units according to present Fire Department policy.

There are three versions of the subroutine, ASSIGNMU, which reflect varying policies for assigning a medic unit to a call. All three assign a closest available medic unit when available medic units exist; they differ in the handling of the case when there are no available units. Version 1 assigns the busy medic unit with the smallest estimated time of arrival (ETA), derived from its TLEGS vector. The second version assigns the first medic unit to become available regardless of its ETA. The effective difference in these policies is a change in city coverage, ultimately reflected in final response time statistics.

The third version is the interactive assignment policy. When there are no available units and an assignment is necessary, a subroutine INTERACT, is called. This version must be run on a graphics terminal, e.g. TEKTRONIX. A map of the city is displayed on the screen, showing the user the location of each mobilized medic unit, the status of the present call to which each is responding, the location of the new call, and the ETA of each medic unit to the new call. The user is prompted to assign a medic unit to the new call. If a busy medic unit is assigned, all call attributes are put on a queue. This method is employed for two reasons: 1) to keep track of back-up, and 2) to avoid incorporating queued-called statistics into the final statistical summaries should the simulation end and the calls not be taken off the queue, i.e., not actually serviced.

The interactive version was developed to incorporate the human element observed in actual medic unit dispatching, i.e., the dispatcher makes assignments based on city coverage and not simply estimated times of arrival. In addition to being a useful tool in assessing the validity of versions one and two, this option could be implemented in actual dispatcher training.

The next subroutine assigns the hospital to which a patient is transported. Presently the model assigns the closest hospital and disregards specialty referral centers and hospital alert status. Specialty referral centers were omitted due to an insufficient number of such incidents in our database. Modelling these rare events would have required an inordinately long simulation period. Hospital alert status could not be modelled because data was only available as average number of hours/day of red alert status and not by time of day. Since the modelling goal was to evaluate overall system performance under varying medic unit location alternatives, the rare events were assumed to be background noise and to have little effect on final policy outcomes.

To complete the description of a medic unit run, the TLEGS array is defined in the subroutine MURDER (Medic Unit Run Derivation). Depending upon the leg and type of call, times at each endpoint in the run are determined, as previously discussed. Statistics are collected for all unqueued calls in the subroutine COLLECT.

When the simulation period ends, final statistics are calculated in the subroutine, STATS, which

also prints the final summaries. The measures of performance derived in STATS are:

1. Response Time Measures:

   mean, variance, and distribution of response times by medic unit, call zone, and by hour.

2. Work-Load Measures:

   a. Mean and variance of # of calls/day by medic unit

   b. Mean and variance of # of calls/hour by medic unit

   c. Mean and variance of time spent servicing calls per day.

CASE STUDIES

Case studies, identified with the Fire and Planning Departments, were analyzed using the simulation model. This analysis provided valuable information about the effects of potential policy modifications to the Baltimore ambulance system in terms of changes in the city-wide and zonal effectiveness measures listed above. Additionally, the model made explicit the trade-offs inherent in the adoption of these policy options. The results to date suggest that the model can be applied to emergency medical services in other urban areas to evaluate current system effectiveness and to develop strategies for a more efficient utilization of resources to meet changing demands.

Diagram 3

FLOWCHART OF MAIN SUBROUTINES IN ASSIST