Proceedings of the 1992 Winter Simulation Conference
ed. J. J. Swain, D. Goldsman, R. C. Crain, and J. R. Wilson

EFFECTS OF TIME-VARIED ARRIVAL RATES: AN INVESTIGATION IN EMERGENCY AMBULANCE SERVICE SYSTEMS

Zhiwei Zhu

Department of Management
The University of Southwestern Louisiana
Lafayette, Louisiana 70504, U.S.A

Mark A. McKnew

Department of Management
Clemson University
Clemson, South Carolina 19634, U.S.A

Jim Lee

Engineering Management Program
The University of Southwestern Louisiana
Lafayette, Louisiana 70504, U.S.A.

ABSTRACT

This paper addresses the problem of modeling time-varied arrival rates in Emergency Ambulance Service (EAS) systems. The approach is to break down a 24-hour period into several equal-length time durations such that the arrival rate during each time duration can be reasonably viewed as stationary. A discrete event simulation model was developed to describe the operations of the EAS center in Shanghai, China. Simulation experiments were performed to test the selection of duration length. The simulation output data were compared to real data collected from the Shanghai system. Statistical data analyses on the system performance indicate that the 4-hour duration model with six different arrival rates in 24 hours closely matches the operation characteristics of the Shanghai system.

1 INTRODUCTION

One of the difficulties in applying a queuing model involves the determination of a stationary arrival rate when the actual arrival rate in a real system varies substantially from hour to hour during a day. This type of time-varied arrival pattern is common for Emergency Ambulance Service (EAS) systems. Several studies (Baker, 1983; McKnew, 1985; Larson, 1975) have shown that the pattern of call arrivals for EAS fluctuates widely during the day. Generally, the calls or demand for EAS has a lull in the early morning hours, increases rapidly thereafter, and reaches its peak time at approximately 10:00 P.M. The demand then falls off gradually to the early morning low. A similar pattern of demand for EAS was observed in the study of Shanghai's EAS system in China (Zhu, 1989). The ratio in the number of calls received between the peak and valley hours during a day by the Shanghai system was as much as 10 to 1. Since queue models assume stationary arrival rates, it will be interesting to see whether the operation characteristics of a real EAS system can be accurately modelled by queuing theory.

The previous studies dealt with this problem by either taking the average arrival rate over an entire 24-hour period (Larson, 1974) or computing the mean of arrival rate within a period of time showing the highest sustained level of demand (Burwell, 1986). Taking an average arrival rate could lower the demand in a peak time in which the availability of ambulances is crucial to the performance of an EAS system. On the other hand, if the arrival rate is calculated from a truncated sample, such as a period with the highest sustained level of demand, the system performance predicted by the model may not be accurate.

This paper addresses the problem of modeling time-varied arrival rates in emergency ambulance service systems. The approach is to break down a 24-hour period into several equal-length time durations such that the arrival rate during each time duration can be reasonably viewed as stationary. A discrete event simulation model was developed to describe the operations of the EAS center in Shanghai, China. Simulation experiments were performed to test the selection of duration length. The simulation output data were compared to real data collected from the Shanghai system. Statistical data analyses on the

system performance were collected to investigate how accurately a real EAS system can be described by a queuing model with a stationary arrival rate. Specific attention will be focused on the measures of the average service time and the fraction of interdistrict calls.

2 DATA COLLECTION IN EAS SYSTEMS

2.1 The Shanghai EAS Center

The actual demand data used in this study are historical operating records from the Emergency Ambulance Service (EAS) Center in Shanghai, People's Republic of China. Shanghai is one of the most populous cities in the world. Its population of 12 million lives in 6,600 square kilometers. The entire city is composed of 12 administrative districts or 111 communities. Each district is consisted of several communities. Approximately 60 percent of the residents live in the densely populated urban districts. There are eight ambulance stations that house 15 ambulances located throughout this urban area. A station may service one or more districts. These ambulances responded to a total of 98,771 calls in 1986, resulting an average workload of 1.41 calls per hour for each station.

Large scale data recovery was difficult because the Shanghai EAS center did not store response data in a computer system. Therefore, a sample of 889 ambulance runs was randomly selected from two periods in January and August of 1987. These two months were selected because they were thought to represent the different winter and summer demand patterns. Morbidity caused by respiratory organ diseases is high in the winter due to the cold weather, while coma and shock are often the primary complaints in the hot summer days in August. Two types of information can be identified from the ambulance run records and used in developing a simulation model which simulates the Shanghai EAS system.

2.2 Data Collection

The first type of information is a series of time data, including the time an ambulance crew received a call from the dispatcher, left the station, arrived at the scene, arrived at the hospital, and returned to the home station. Mileage associated with these time periods were also recorded. There are some other relevant data which can be used for analytical purposes, including the identification number of the home stations, the caller's location, the hospital to which a patient was sent, and the type of hospital service required.

The total operation time required to serve each call is the elapsed time from ambulance dispatch time and the time the ambulance returns to the home station. This includes the travel time to scene (TTS), the time spent treating the patient at the scene (TAS), the travel time from scene to hospital (TTH), the patient unloading time at the hospital (TAH), and the time for an ambulance to return to its home station (TBS). Table 1 contains the average ambulance operation times for each of the 8 stations. It can be seen that the travel times including TTS, TTH and TBS fluctuate widely among stations. For example, TTS ranges from 7.3 to 17.8 minutes with a mean time of 9.9 minutes. Station 4 has the highest TTS because it serves a mainly rural area of Shanghai and it is isolated from other stations by a large river.

Table 1: Average Ambulance Operation Times in minutes by Station (S)

S	TTS	TAS	TTH	TAH	TBS
1	7.3†	5.7	7.3	4.2	6.9
2	12.8	5.9	9.3	3.4	8.6
3	12.3	6.6*	15.4	3.3	10.9
4	17.8*	5.3	15.8*	2.3†	12.4*
5	7.8	6.5	6.7	4.0	8.1
6	8.3	6.6*	5.9†	6.9*	6.7†
7	11.9	2.8†	12.3	3.3	9.8
8	9.6	4.4	10.5	2.4	8.3
Mean	9.9	5.5	10.4	3.7	9.0

† - smallest value in column
* - largest value in column

The second type of data collected from Shanghai is demographic data. The data provided by the Shanghai Municipal Socioeconomic Information Consultative Center were tabulated from the nation-wide population census conducted in July, 1982. The basic population reporting unit is a community, and is a primary administrative unit in Shanghai. The total population and population density for each of the 111 communities in Shanghai are useful in estimating the arrival rate of service calls of the EAS system. The detailed geographical map of Shanghai shows all sites of hospitals and ambulance stations as well as boundaries of each community and district.

Each reporting unit in the city can be treated as an independent arrival generator in the simulation model,

and therefore the arrival rate for each community has to be defined. Ideally, the proportion of calls generated by each unit should be calculated from historical data during a given period of time. Unfortunately, Shanghai's operating records do not have any information regarding the number of calls generated by each community. However, we do have residential population data for each community to use as a surrogate for measuring demand level. Therefore, the population data is used as a reasonable proxy for the number of calls generated by each community.

3 MODEL DEVELOPMENT

Based on the information discussed above, a simulation model of the Shanghai EAS system was developed using SLAM II, Simulation Language for Alternative Modeling (Pritsker, 1986). SLAM is a FORTRAN based simulation language that allows a model builder to develop models from a process interaction perspective. A discrete event modeling approach was chosen to model the EAS system because the concept of discrete event modeling fits the nature of EAS system by describing the changes that occur in the system at discrete points in time. The events defined by the changes in the state of a system are used to model the start and completion of activities of ambulances. The duration of an event is collected as event time, which will be validated and analyzed.

3.1 Discrete Events

Two FORTRAN subroutines were constructed in the simulation model to implement the logic of time-varied demand generation, EAS system operation, and collection of system performance measures. The first subroutine, ARVL, is the procedure to generate calls for service and schedule the ambulance operations. The parameter λ served as an average arrival rate over a fixed time period and was used to control the number of calls generated from a Poisson distribution based on the demand distribution of the population data. The process to identify calls from a particular community is through a 0-1 random number generator. A cumulative percentage of calls generated from each community is based on a cumulative proportion of population in each community. If a random number falls in the range of a particular cumulative proportion of population, then the caller's location is identified.

Once a call is generated, it must be assigned to a server (ambulance) from a station. The procedure involved in dispatching an ambulance to the call follows the logic of the actual dispatching policy being used by the Shanghai EAS center. When a call arrives, a list of three preferred ambulance stations, including an intradistrict and two interdistrict stations is identified. An ambulance at the intradistrict station (the most preferred station) on the list is dispatched to the scene. If none of the ambulances at the intradistrict station is available, then availability of ambulances at the second preferred station is checked. If an ambulance is still not available, then the third preferred station will be called up. Even though the probability of having all ambulances at three preferred stations busy is very small, it is possible that none of the ambulances on the list is available. The call is then put in a queue waiting for next available server, which may be either an intradistrict or interdistrict server. A call waiting in a queue for more than 5 minutes is considered to be lost or served by outside of the EAS system. The rule of first-come first-served is specified to release a call from the waiting line when an appropriate ambulance becomes free.

Travel times to the scene are estimated from the equation derived by regressing the actual travel times on the corresponding travel distances identified from the operating records. A separate FORTRAN program was designed to read and calculate coordinates of all communities into a distance matrix based on the right angle travel metric. The distance matrix then serves as the input data for calculating the estimated travel times. The other ambulance operation times, including time at-scene, time off-scene to hospital, and the time back to station are taken from the averages calculated from the actual data. These values are different from station to station.

The second subroutine, ENDSV, performs primarily as a statistical collector on several performance measures. It starts at a point where an ambulance just released a patient at a hospital and checked with the dispatcher whether there is a call waiting in the queue. If there is a call in the queue, the ambulance is dispatched directly from the hospital. Otherwise the ambulance goes back to the home station.

3.2 Performance Measures

Statistics are collected in the simulation model based on all ambulance runs on the following system performance measures:

(1) average service time of intradistrict call by each station;

(2) average service time of interdistrict calls by each station;

(3) workload or utilization rate of ambulance for each station;

(4) fraction of calls served by interdistrict stations.

Service time is the sum of travel time to the scene

(TTS) and the treating time of patient at scene (TAS). Travel times are classified by intradistrict calls and interdistrict calls on the basis of each station's service territory. It is estimated by a function of travel distance and vehicle velocity. Workload is measured as a fraction of time that an ambulance is busy. Fraction of calls that are served by ambulances from interdistrict stations indicates a percentage of time that the first preferred ambulance (intradistrict ambulance) is not available.

3.3 Model Validation

After the model was developed, it was validated to make sure that inferences about a real system derived from the model would be correct. It is important for the simulation model to have the similarity to the real system that it represents. First, there must be an exact correspondence between the elements of the model and the items being represented. The number of service units, communities, and coordinates of their locations should be precisely described in the model. Second, the real relationship between items should be preserved. This means that the number of calls generated from each community and the number of calls served by each station should be close to the actual data. All of the above concerns were checked before the statistics about the system performance were collected and analyzed.

4 MODELING TIME-VARIED ARRIVAL RATES

4.1 Duration Length Selection

Since the call distribution of the Shanghai data shows a substantial variation during a 24-hour period, the use of a stationary arrival rate in modeling is obviously unrealistic. However, an arrival rate could be viewed as relatively constant when a 24-hour period was divided into several small time durations such as three 8-hour durations, four 6-hour durations, or six 4-hour durations. Arrival rates are found to be much more stable within each shorter time duration than a 24-hour period. The choice of the duration length should rely on the pattern of the demand distribution. A length of N ($N < 24$) is selected if the demand variation in the N hours is relatively low. To compare the selection of time duration length, a 4-hour, a 6-hour, and an 8-hour configuration along with a 24-hour model with stationary arrival rate are tested.

4.2 Starting Point Selection

Once the duration length is determined, a starting point must be selected to divide a hourly fluctuated demand data into several smaller time durations so that the hours within each duration have a similar demand pattern. To select the best starting point, hourly demand variation is measured by a coefficient of variation (CV), which is defined as the standard deviation divided by the mean of a sample. For example, in an 8-hour duration configuration, there are eight possible starting points. For each possible starting point, a CV is computed. Table 2 illustrates the coefficients of variation of possible starting points for an 8-hour duration model. It can be seen that the best starting point in this example is 7:00 A.M. where both the average and maximum CVs are minimum. Similar calculations were made on coefficients of variation of all possible starting points in order to find the best starting points for 6-hour and 4-hour durations. The best starting point was found to be 1:00 A.M. for the 6-hour duration and 2:00 A.M. for the 4-hour duration.

Table 2: Maximum and Average Coefficients of Variation (CV) of Possible Starting Points in 8-hour Duration Model

Starting Point	Maximum CV	Average CV
0:00 AM	0.37	0.26
1:00 AM	0.49	0.31
2:00 AM	0.56	0.32
3:00 AM	0.52	0.31
4:00 AM	0.43	0.28
5:00 AM	0.33	0.26
6:00 AM	0.31	0.24
7:00 AM*	0.30 †	0.22 †

† - smallest CV in column
* - best starting point.

4.3 Goodness-of-Fit Tests

Although distributions of emergency call arrivals were often assumed to be Poisson in the literature, it was necessary to test how well this assumption holds in this application. To determine the degree of fitness between the real data and an assumed theoretical distribution, the Chi-Square (χ^2) statistic was used to test the hypothesis that there is no discrepancy on frequency of calls in each one hour interval between the observed data and the Poisson distribution with an arrival rate averaged from a 24-hour period (Shannon,

1975). The hypothesis was rejected at a confidence level of 0.95. Table 3 shows that a big difference lies in both the highest and lowest frequency intervals.

The χ^2 statistics indicate that taking an average arrival rate over a 24-hour period could underestimate the demand during a peak period. Further, from a theoretical standpoint, the distribution of time intervals between calls for service should follow a negative exponential if the pattern of call arrivals is truly a Poisson distribution. The test on the time intervals between calls for service confirms that the data can not be described accurately by a Poisson distribution with an average arrival rate over a 24-hour period.

Attempts were also made to identify the data by some other distribution, such as the Erlang distribution suggested by Baker (1983). Again, the χ^2 value exceeded the tabulated critical value at a confidence level of 0.95 and suggested that the data did not follow an Erlang distribution.

Table 3: χ^2 Statistics on Comparison of Demand Distribution between Shanghai Data and Poisson Distribution with Stationary Arrival Rate

Interval (#calls/hr)	Observed Frequency	Expected Frequency
0-3	16	5.84
4	7	7.14
5	7	10.65
6	8	13.23
7	10	14.10
8	11	13.15
9	10	10.89
10	9	8.12
11	8	5.51
12-20	10	7.27

Calculated χ^2 = 24.78

Despite these rejections of the fit between the data and the theoretical distributions, it is still possible that the data may be matched by the Poisson distribution with time-varied arrival rates over a 24-hour period. It is expected that by dividing a 24-hour period into several small durations one can improve the goodness of fit by a Poisson distribution to the real data. Hourly call frequencies of each 8-hour, 6-hour and 4-hour time duration were generated from non-stationary Poisson distributions. The simulated frequencies were compared with the call frequencies from corresponding time intervals of the real data collected from the Shanghai system. The χ^2 statistics indicate that the arrival pattern of the real data can be closely matched by the Poisson distribution with different arrival rates taken from six 4-hour time durations.

Comparing the randomly generated data to the real data, it was found that taking an average arrival rate over a 24-hour period would underestimate the likelihood of the least and greatest call occurrences. The Poisson distribution using a 24-hour average has a tendency to produce average numbers of calls in each time interval, which reduces the probability of the least and most occurrences in any time interval. Given the testing results it will be interesting to see how the performance measures of an EAS system can be affected in the various time duration models.

5 SIMULATION AND RESULT ANALYSIS

To evaluate the effects of time-varied arrival rates, the Shanghai EAS system was simulated by the various time duration models with time-varied arrival rates and the best starting points determined previously. Five simulation runs were made for each time duration and each simulation run was extended beyond its time interval allowing for a period of time to establish the steady state. Several types of statistics were collected and tested. This paper reports the mean tests performed on the fraction of interdistrict calls and average service time.

Interdistrict calls and service times generated by the simulation models were compared with that of the Shanghai data. The data sets were collected to test the hypotheses specified below:

H1: There is no significant difference in the fraction of interdistrict calls between the Shanghai data and the simulation results generated from an average arrival rate over a 24-hour, 8-hour, 6-hour, or 4-hour period.

H2: There is no significant difference in the average service time between the Shanghai data with the simulation results generated from an average arrival rate over a 24-hour, 8-hour, 6-hour, or 4-hour period.

A paired *t*-test was used to determine whether or not the hypotheses can be accepted. The results reported in Table 4 show that the fraction of interdistrict calls in the real data is significantly different from the fraction generated from the 24-hour and 8-hour models. This

means that we expect to see fewer interdistrict calls than the real data when a stationary arrival rate is calculated over a 24-hour period. On the other hand, we failed to distinguish the fraction of interdistrict calls computed from the Shanghai data and the simulation models using an average arrival rate from a 6- or 4-hour time duration. The fraction of interdistrict calls of the Shanghai data can be very closely matched by six Poisson distributions with each time duration lasting four hours. No significant differences in service time were found between the real data and any type of time duration of a Poisson distribution.

Table 4: Paired *t*-tests on the Hypotheses

Hypotheses	Rejected	*t*	Significant Levels
24-Hr (H1)	Yes	3.78	.0006
24-Hr (H2)	No	1.15	.2597
8-Hr (H1)	Yes	-3.35	.0020
8-Hr (H2)	No	-1.62	.1154
6-Hr (H1)	No	-1.72	.0946
6-Hr (H2)	No	-1.30	.2009
4-Hr (H1)	No	-0.26	.7941
4-Hr (H2)	No	-0.01	.9916

We realize that an increasing number of interdistrict calls does not necessarily increase average travel time if each ambulance station is not strictly located in the geographical center of a district. It might be the case that an interdistrict service requires no more travel time than an intradistrict call since the dispatch policy states that the closest ambulance has to be dispatched to a call if the caller's own district ambulance is not available.

It is clear that taking an average arrival rate over a 24-hour period would result in a low number of interdistrict calls and underestimate the unavailability of an ambulance during a highly demand period. Fortunately, the number of interdistrict calls simulated from a 6-hour and 4-hour time duration models are statistically consistent with the Shanghai data. Consequently, it is believed that the variation of hourly demand pattern in the Shanghai data can be fully described by a Poisson distribution with the stationary arrival rates over six 4-hour time durations.

6 CONCLUSIONS

This study addressed an unresolved application issue: how well a queuing model with the assumption of a stationary arrival rate can describe the real EAS system with time-varied arrival rates over a 24-hour period. A discrete event simulation model was developed to modeling time-varied Poisson arrivals. The approach was to break down an entire 24-hour period into several small time durations such that an average arrival rate during the duration can be reasonably viewed as constant.

We compared the frequency of calls per hour generated from the simulation models, in which an arrival rate was averaged over a 24-, 8-, 6-, and 4-hour duration with the Shanghai data respectively. The χ^2 test showed that the arrival distribution of the Shanghai data was closely matched by the Poisson distribution with six different arrival rates for each of 4-hour durations. The simulation results revealed that the interdistrict calls could be underestimated if the assumption of a stationary arrival rate was made on a 24-hour time period. The model with arrival rates averaged from 4-hour duration was able to reflect the actual number of interdistrict calls.

The question of the ability of a queuing model to accurately describe a real world system with a time varied demand pattern has been examined through a simulation model. Effects of a stationary arrival rate imposed by a queuing model on some performance measures of an EAS system are examined. This study demonstrates that simulation can be a useful technique to evaluate the assumption of a queuing model.

REFERENCES

Baker, J. R. 1983. An examination of emergency medical performance and quality of care and their impact on location planning. Ph.D. dissertation, Department of Management, Clemson University, Clemson, South Carolina.

Burwell, T. H. 1987. A spatially distributed queuing model for ambulance systems. Ph.D. dissertation, Department of Management, Clemson University, Clemson, South Carolina.

Larson, R. C. 1974. Urban emergency service systems: an iterative procedure for approximating performance characteristics. R-1493-HUD, The New York City Rand Institute.

Larson, R. C. 1975. Approximating the performance of urban emergency service systems. *Operations Research* 23: 845-868.

McKnew, M. A. 1983. An approximation to the hypercube model with patrol initiated activities: an application to police. *Decision Sciences* 14: 408-418.

Pritsker, A. A. B. 1986. *Introduction to simulation and SLAM II*, 3rd ed. New York: Halsted Press.

Shannon, R. E. 1975. *Systems simulation*. Englewood Cliffs, N. J.: Prentice-Hall Inc.

Zhu, Z. 1989. A workload balancing optimization model for ambulance location: an application to Shanghai, P. R. C. Ph.D. dissertation, Department of Management, Clemson University, Clemson, South Carolina.

AUTHOR BIOGRAPHIES

ZHIWEI ZHU is Assistant Professor in Department of Management at The University of Southwestern Louisiana. He earned his Ph.D. from Clemson University in Industrial Management. His research interests lie in the area of operations research applied to service sectors and production operations management.

MARK A. McKNEW is Associate Professor of Management at Clemson University. He earned his Ph.D. from M.I.T. and his M.A. in urban planning from UCLA. His research and publication interests lie primarily in the area of operations research applied to public emergency services.

JIM LEE is Assistant Professor in Engineering Management at the University of Southwestern Louisiana. He received a B.S. degree in Industrial Engineering from Tunghai University, and M.S. and Ph.D. degrees in Industrial and Management Engineering from the University of Iowa. His research areas include simulation, manufacturing system design and planning, production and quality management, and expert systems. He has performed research projects in various areas of computer aided modeling of engineering problems.