A Simulation Model for Emergency Centres

Erhan Kozan and Nick Mesken

School of Mathematical Sciences, Queensland University of Technology Brisbane Qld 4001 Australia E-Mail: e.kozan,@qut.edu.au

Key words: *Simulation; Emergency services; Ambulance call centres*

EXTENDED ABSTRACT

The objective of this study is to develop and test an analytical based simulation model, which can be used to plan for emergency call centres. The simulation model can be used to analyse the effects of incoming calls on the system; analysing personnel resources at call centres and; investigate the impact of workload distribution, shift schedule adjustments and staffing profiles on response times; estimation of the impact on the resources required in relation to a reduction or increase in the number of communication centres; improving the efficiency and location of dispatch services; and improving ambulance allocation algorithm. The model with some minor alterations can be also used for different emergency centres.

The Extend model consists of five main stages:

- generate the emergency call;
- telephone routing system allocates calls to centres;
- call centres receive and handle the call and pass information to dispatcher;
- dispatcher allocates ambulance; and
- ambulance is dispatched from station.

Relationship of these stages is represented by a Flow Chart in Figure 1.

The call centre schedule could be tested to minimize costs whilst maintaining services using varied call rates.

The model can be used to evaluate both the number and type of dispatchers required as well as call centre staffing. A dispatch algorithm can be constructed easily given some realistic parameters.

The comparisons between a multiple dispatch and single dispatch system have been done essentially based around staffing numbers. A single dispatch queue enables greater flexibility with staff and given predictable call rates an efficient schedule could be devised.

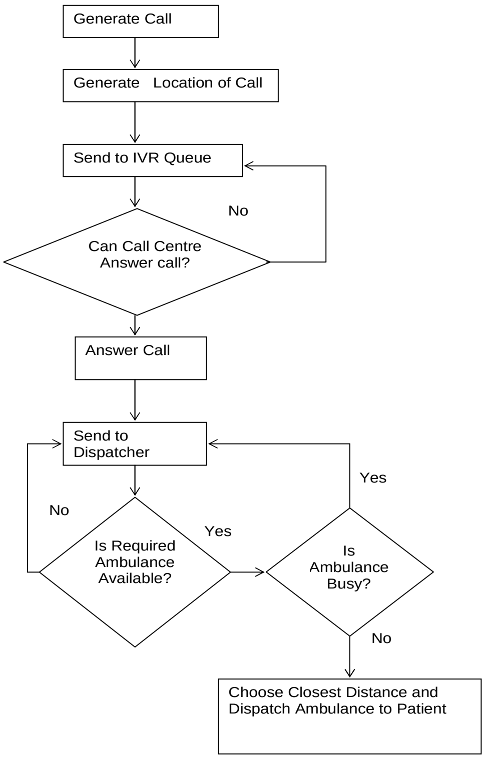
Generate Call
Generate Location of Call
Send to IVR Queue
No
Can Call Centre Answer call?
Answer Call
Send to Dispatcher
Yes
No
Yes
Is Required Ambulance Available?
Is Ambulance Busy?
No
Choose Closest Distance and Dispatch Ambulance to Patient

Figure 1. Flow Chart of the Simulation Model for Emergency Centres

Emergency management planners and health care facilities are faced with potential scenarios that may severely test their capabilities and conventional processes. This paper presents the development of a simulation-based environment that allows emergency planners and professionals to prepare more robust emergency response plans without using costly, time consuming physical drills.

1. INTRODUCTION

This paper involves the development of an integrated simulation model to evaluate the suitable capacity for emergency centres especially for ambulance centres. Results from this study will facilitate the coordination, implementation, and allocation of resources. This simulation model is explicitly designed to provide the user with a decision-making environment that enables the analysis of geographical information to be carried out in a robust.

This paper describes a simulation-based approach that allows emergency management departments to more quickly develop, test, and refine robust plans for an ever increasing list of potential threats.

The simulation model is developed for the purpose of understanding the behaviour of the emergency calls and/or evaluating various strategies for the operation of the call centres. The relationships among system's elements and the manner in which they interact determine how the overall system behaves and how well it fulfils its overall purpose (Pidd, 1996).

Analytical models cannot easily represent the complex interactions caused by random events. So, analytical based simulation model is developed to cope with analytical tools enabling data investigation and decision models enabling scenario based investigations. The simulation model is developed to describe progress in the system to address a number of specific objectives. Some of the probability distributions may not be standard probability distributions like those used in queuing theory and other mathematical models; however, simulation allows us to include these non-standard distributions into the model.

A simulation model is currently being developed for the emergency call centres using Extend V6. It contains a simple interface with predesigned blocks' being used to piece together the model. Simulation model statements of Extend are called blocks. Blocks define how the system operates. These 'blocks' consists of multiple queue types and variable adjustments which greatly simplifies the construction of a large simulation model. Each time a block is executed, the state of the system is changes. When a block is executed, an object called an entity must pass through the block. Entities typically represent items moving through the system such as calls. Similarly, a block's function normally corresponds to an operation in the real system. For example, consider the resource block: calls; when an entity executes this block, personnel resources is assigned to entity in much the same manner as personnel resources is assigned to a emergency call.

Investigations into the optimum locations of emergency vehicle sites have been conducted for a significant number of years. Ball and Lin (1992) investigated a model to pinpoint optimum locations as well as the specific types of vehicles at each station. It examined the problem from a system reliability problem with a minimum desired level of service. To do this an integer programming model was implemented with a basic simulation model used for further discussion and investigations. Sensitivity analysis was also conducted but only at a superficial level.

Marinov and Revelle (1995) also investigated the emergency vehicle site problem. However, they approached it as a maximal availability location problem and used queuing theory for the incoming calls. However, due to calls not fitting any statistical distribution the model was weak and a simulation approach proposed.

Another aspect of the emergency vehicle problem is vehicle routing, scheduling and dispatch. Hirota et al (2002) investigated a HIMS (Hierarchical Multiplex Structure) model and if it could be applied to emergency vehicles. They applied set theory to the problem and they found a quick method to find a 'good' solution but more investigation was required for an optimum solution.

An investigation into the entire emergency dispatch and service model was recently conducted in Chile by Weintraub et al (1999). A simulation model was developed of the city and its subsequent emergency services and emergency locations. They incorporated several features such as priority calls and made significant improvements to service times especially in adverse conditions when reports were highest.

Harewood (2002) recently conducted a study on ambulance deployment on the island of Barbados. Various mathematical models were examined and the multi objective with explicit costs model was thoroughly investigated. A Monte Carlo simulation was also developed to validate the optimum solution and to conduct sensitivity analysis. It was found that the main relationship is between distance vehicles have to travel and reliability of service with both offsetting each other. The social implications of the model were also examined with repercussions obvious for limiting services as per the optimum solution.

The most relevant research conducted into ambulance problems was conducted in Canada by Trudeau et al (1988). A mathematical model was developed to simulate all emergency operations. Demand forecasting, staff scheduling and operations strategy were concentrated as well as the cost versus service problem. Scheduling was then implemented using the above forecasts with segments reflecting differing demands. From this point the simulation model was created to account for the actual dispatch and service with the goal of minimising costs for a given level of service. They found several 'good' solutions yet an optimal could not be found. Some recent and related call centre literature that is associated with some aspects of simulation and call centre performance can be found in Duder and Roswenwein (2001), Eveborn and Ronnqvist (2004), Mehrotra and Profozich (1997), Miller and Bapat (1999)

2. SIMULATION MODEL

The model consists of following main stages: emergency call generating; allocation of calls to centres; handling of the calls at centres; information transfer to dispatcher; ambulance allocations and dispatching from stations. A call centre consists of many areas that need to be considered namely:

- forecasting incoming calls;
- analysing staffing/shift patterns at call centres;
- improving the efficiency and location of dispatch services;
- improving ambulance allocation algorithm;
- optimisation of patient time to hospital; and
- optimising the number of each type of ambulance at each station.

The simulation system has the following five major subsystems namely Call Generating (CG) system, Integrated Voice Response **(**IVR) System, Call Routing (CR) System, Call Centres (CC), Dispatching (D). Several completed features of these subsystems are given below:

2.1. Call Generating (CG) System

The call generating system is operating effectively. It currently generates calls from 20 areas. Using predefined inter arrival time the call is generated along with its random location. The call length and standard deviation are also generated randomly. The type of call (problem that they are calling for) is also generated here. Currently, types are allocated randomly according to a uniform distribution but this is currently being changed to account for the different probabilities of certain calls.

Currently, this is implemented through both an Excel spreadsheet and Extend blocks. Initially, the spreadsheet contains all of the call inter-arrival times with differing times for each 15 minute block of the day. This enables call rates to be varied on demand as per reality. This spreadsheet also contains the location of the call area by using a simple Cartesian coordinate system. This is implemented through the spreadsheet as it adds greater flexibility in both including overlapping blocks and being able to specify any size area for the call region. This Cartesian plane can be altered simply to replicate any realistic region. For each call region there is a corresponding Extend module. Each module initially contains an input for both call inter-arrival times and boundaries of the region. A call is then randomly generated with both its location and length. The type of call is then allocated as per a probability algorithm and the call is then sent to the **IVRS** System queue.

2.2. Integrated Voice Response System (IVRS)

The IVRS is included in the model for further study. An IVR is basically a computer system that depending on pre programmed questions can allocate calls and services. Currently, as per reality there is no IVR system. In Extend, this equates to its simply being a FIFO queue with unrestricted output so no waiting time. This is included as with a simple addition of an activity module this system could be implemented. This can then test if such a system is feasible for an emergency vehicle system.

2.3. Call Routing (CR) System

The call routing system allocates calls to call centres. Currently this is achieved through a preference ranking system with each call area having preference for certain call centres. This calculation is done in Excel with the call location entered and allocated according to a table. A data base has been constructed with low values representing preference for an individual centre. This, accounting for busy call centres, allocates calls to the optimum centre. These preferences can be easily changed to represent reality. If all centres are busy allocation is too the original highest preference.

Initially calls enter the system and are placed in a reneging queue. This queue replicates people waiting and hanging up if waiting times are significant. As this is an emergency vehicle system time for call reneging was high as people will want an ambulance. The call is then sent to the spreadsheet to allocate a centre as discussed above. It is then placed in a FIFO queue to be answered at a centre. This queue is useful as not only does it represent waiting for a call to be answered it also enables the efficiency of the call centre system to be tested as waiting times here should be theoretically non existent. Currently it is assumed all call centres can handle all types of calls. If a centre cannot handle a specific type of call an algorithm needs to be developed to, after the call is answered, redirect it back through the system to the initial allocation phase taking into account. Also, the algorithm when all centres are full needs to be checked as should they be allocated according to preference or which centre has the shortest queue.

2.4 Call Centres (CC)

The call centres are designed to replicate the answering of calls. The calls are sent through via the call routing system as mentioned above. The mean and standard deviation of a call are calculated when the call was made is entered into a random number generator block. This block, using a lognormal distribution generates the length that the call stays in the call centre (i.e. handling time by the receptionist). This is represented with an activity block which has this time input to it. The call then waits here to be answered with a full marker sent to the call routing system to prevent allocations to this centre. Currently each centre can handle five phone calls at a time but this can be changed depending on requirements. Each centre is calculated separately so different situations can be constructed for each simply. Calls are then sent to a dispatcher.

2.5 Dispatching (D)

The dispatching system is designed to allocate the best possible ambulance to a call. The Dispatch Algorithm is basically derived from the distance required to travel with the minimum distance chosen. The dispatch algorithm mainly concerns the allocation of an ambulance (station) to the appropriate calls. The call enters into the dispatch system from the call centre (handled previously). The location and type of call is then entered into the algorithm. This forms the basis of our decision rule. The ambulance is then allocated according to the minimum distance required to travel. The call type is first entered into the system. Then the stations with the correct type of ambulance are selected to answer the call (they have an ambulance type that can handle the call and that ambulance is not busy answering another call). Then the minimum distance is selected and that ambulance is then listed as unavailable for subsequent calls until it returns. The service time is then a linear function of distance to the call. After this time has elapsed the ambulance is then available for use again. If two ambulances have a similar distance then the preference table is used to differentiate. Every call type has a preferred ambulance type is dispatched.

The shortest distance is calculated using Cartesian Coordinate System and an ambulance is sent from the appropriate station. This is done in an Excel spreadsheet that also accounts the following factors. Initially each stations location is entered into the spreadsheet. The next set of tables is a binary table of call types against ambulance types. This table simply allocates a **1** where an ambulance can handle a type of call and a **0** when that type of ambulance cannot handle the call. The appropriate row of this table (corresponding to call type) is then sent to the top of the next table. The next table is simply the number of ambulances or each type at each station. This is then multiplied by the binary row from the previous table to give the number of ambulances at each station that are able to answer that particular call. This table is then compared with another binary table call ambulance full. This ambulance full simply has a **1** if all ambulances of that type at that station are in use and a **0** otherwise. This information is sent from Extend. This is then combined with the previous tables to see which stations have available ambulances which are available and capable of answering the call. The distance is then calculated and matched with the closest station who can handle the call. Where multiple ambulances who can handle the call are involved a preference table has been created. If distance difference is negligible and multiple ambulances are available the best ambulance is allocated. If there is a large distance differential the closest is allocated but this could be recreated if limitations of particular ambulances were known.

There are currently two differing dispatch models whilst both dispatch according to the same overall algorithm one has multiple dispatchers and one is a single dispatch. The multiple dispatch is where each call centre has an individual dispatcher. This system currently has a fault of two simultaneous calls could be allocated to the same resources with no report of it being full. This is currently being worked on. The single dispatcher currently has all call centres queues merging into one. This creates the problem of possible queuing despite resources being free. The two models are the same apart from numbers so that comparisons between the two can be made.

An exact simulation model is difficult to develop so several assumptions were made. Reality dictates that every service time has extremely high variance given travel times (Traffic, accidents, road structures, speed limits, distances etc) and treatment times as every patient is different (Same problem can have extremely differing times depending on age of patient, support around patient, unforeseen complications etc).

The key assumptions of this model reflect the difficulties in establishing a complete simulation model. Most assumptions concern distributions used to approximate and generate data. Initially in the Call Generator stage calls are assumed to arrive with exponential inter-arrival times with the call type allocated with set probabilities. The location of the call is set as uniform over a given area. The call length is also assumed to be exponentially distributed.

There are several other simplifying assumptions not related to distributions. Firstly, the calls are allocated according to a set preference based on location. All call centres can handle every type of call and if they are all busy the call is directed to the primary preference. For Dispatch it is assumed that distance is the primary driver for selection of station with a preference only used if stations are approximately the same distance from the call. The service time by the ambulance is assumed to be only dependent on distance with different speeds to calls and differing service times due to different call types not considered.

3. AN APPLICATION

3.1. Ambulance Stations

There are 270 ambulance stations and each station is designed to replicate housing ambulances and sending them on calls. A station receives the call from the dispatcher and immediately acts on it. Currently time take to service the call is simply distance but an equation to simulate different speeds will be introduced. Each ambulance has a defined location in Cartesian coordinates and is set in the spreadsheet for dispatch as mentioned previously. Each ambulance station also has an individual number of each ambulance type creating great flexibility in allocations. The distance travelled/ time taken for service also needs to include the trip to hospital and back to the station. This could also give a new starting point to receive calls from as an ambulance could be called immediately after discharging the patient. The actual time in the system is calculated using an activity block and the inputs for this simply need to be changed to account for this.

3.2. Analysis of configurations

Initially the difference between different dispatch techniques was tested. This was done with 2 differing models, one where all calls were place in a single queue to be dealt with by dispatch and the other with each call centre having its own dispatcher. Both models had identical call arrival rates and call centre service times allowing only the dispatch system to be altered.

The parameters for the followings test mainly concerned the calls. Inter-arrival times were exponentially distributed. The map of possible call locations was set on a Cartesian axis with possible locations in both x and y from 0 to 100. Call time was set at 5 minutes (exponentially distributed). These values were approximated from the small data set given from the Ambulance Service. The dispatch time was calculated in the same way in both models.

The first test run was designed to test the effects of the different inter arrival times on the model. For this the single queue 9 dispatcher model was chosen as a benchmark with only the inter arrival times changing. As expected when calls appear more frequently (lower inter arrival time) utilization increases and average waiting time increases. As can be seen in Table 1, as the inter arrival time decreases the maximum waiting time dramatically increased. This is due to the system being flooded with not all calls being dispatched efficiently. Also, increasing the inter arrival times gives a substantial performance increase in the system. This is due to the times moving closer to the average call length reducing the elements in the system. From the above parameters it appears that an interarrival time of 4 minutes stresses the system without creating bottlenecks which is the reason for its use in subsequent tests.

Table 1. Effects of Inter arrival Times on Call Waiting Times

Interarrival Time Changes %	(minutes)	Average system Utilisation %	Average waiting time (minutes)	Max waiting time (minutes)
- 20	3.2	98	2.21	58.00
- 10	3.6	89	1.95	8.30
Benchmark	4.0	73	1.18	6.97
+ 10	4.4	50	1.13	3.06
+ 20	4.8	22	0.34	1.31

After conducting several runs the means and standard deviations of results stabilized around solid numbers. For the single dispatcher with 9 staff members (to have parity with other model) the queues had an average utilization of 73% with the maximum utilization at 82% and minimum at 68%. The average waiting time was 1.18 minutes (1 minute 11 seconds with a standard deviation of 0.2 minutes) but had a maximum waiting time of nearly 7 minutes across all runs.

The multiple dispatch model (9 dispatchers) provided similar results to the single dispatch model. Dispatchers' utilizations varied wildly between 10% and 100%. This variance was simply due to the differing call rates and preferences of each centre. Overall the rates stabilized to a utilization of 71% which was slightly less than single dispatch. This different however, appears insignificant. The main difference was in the average waiting time and its variance. The average waiting time across all 9 dispatchers was 1.37 minutes but with a standard deviation of 0.8 minutes. There was also a maximum waiting time of 11.0 minutes at one dispatcher. These large variations given the context of the model are significant and can lead to major dispatch delays.

The overall statistics given by both models indicates that there is not a major difference between running 9 dispatchers or one dispatcher queue with 9 operators. However, in the single queue case variations were lower and worst case scenario values were significantly less. Given the overall context of the model it appears the single queue system would be more reliable for emergency vehicles.

Table 2. Utilisations for Different Staff Numbers -Single Dispatcher Case-

Staff Number	Average system Utilisation %	Average waiting time (minutes)	Max waiting time (minutes)
6	84	1.4	6.73
9 *Benchmark*	73	1.18	6.97
14	62	1.04	4.46
18	48	0.63	2.87
22	31	0.52	2.35
27	8	0.26	1.78

Table 3. Utilisations for Different Staff Numbers - Multiple Dispatcher Case-

Staff Number	Average system Utilisation %	Average waiting time (minutes)	Max waiting time (minutes)
6	92	1.4	16.36
9 *Benchmark*	71	1.37	11.00
14	62	1.04	7.46
18	52	0.92	6.38
22	35	0.44	4.67
27	15	0.32	2.06

The main difference between the dispatch models begins to appear when dispatch staff numbers are 18 and above (18 in 1 centre or 2 in 9 dispatchers). As expected utilization dropped dramatically (to 48%) as did average waiting time (0.63 minutes standard deviation 0.17 minutes). However, the main change was to the maximum waiting time which was reduced to 2.87 minutes. This represents a major improvement as an emergency call centre cannot allow calls to be not dealt with. However, it is debatable whether this improvement is worth doubling the costs.

The multiple dispatch model did not fare as well when staff numbers were doubled. Utilization and average waiting times fell as expected but the maximum only moved to 6.38 minutes. This represents minimal improvement as waiting times were already low and the single dispatch case is superior for this staffing level.

The main advantage of the single dispatcher system is that during differing periods of the day different staff levels can be incorporated. For example in busy times more dispatchers can be on duty. With the parameters tested above the number of staff members in the single dispatch can be reduced as low as 6 with no major decrease in performance (Utilisation 84%, Average Waiting time 1.4 minutes and Max Waiting time 6 minutes). This can save valuable money for use other where in the system. The multiple dispatch model does not allow this flexibility with no possible decrease in staff from 9 without centres being removed from the system (defeating the purpose of having separate dispatchers with centres).

The 9 call centres can also be tested to allocate staff schedules. For this testing the single dispatcher model was used with all non call centre parameters the same as in all previous tests. The only change was to the number of operators in each call centre. Initially each centre had 5 staff each giving a total of 45.

This configuration gave an average utilization of approximately 5% with 0 calls reneged. This setup had significantly too many staff and was reduced easily. When staff numbers were reduced to 36 utilization rose to 48% with negligible reneged calls. When this was reduced even further to 27 utilization rose to 63% with approximately 0.05% calls. These results appear entirely as predicted and show that if real values for the parameters were given a call centre staffing schedule could be created more accurately. An extension of this test would be to check the difference between the 9 call centers currently in operation (as per real life) and a single large call centre. Whilst it is expected to provide similar results to the single and multiple dispatch case this needs to be empirically tested.

6. CONCLUSION

These few simple tests have highlight how an emergency vehicle system can be simulated easily involving: call generating, an Integrated Voice Response, call centres, dispatcher and ambulance stations. It is possible to evaluate both the number and type of dispatchers required as well as call centre staffing. A dispatch algorithm can be constructed easily given some realistic parameters. If more data was collected a large emergency system could be designed and analysed using this model.

The comparisons between a multiple dispatch and single dispatch system essentially are based around staffing numbers. A single dispatch queue enables greater flexibility with staff and given predictable call rates an efficient schedule could be devised. Using varied arrival rates the call centre schedule could also easily be tested to minimize costs whilst maintaining services.

Extend simulation software V6 provided a simple tool for this system and enabled simple testing. However, due to the simplicity and pre-design work on the model the simulations run slowly and inefficiently. The ease of creating and running variations of tests can easily offset this cost.

7. REFERENCES

Ball, M. and Lin, F.L. (1993), A reliability model applied to emergency service vehicle location, *Operations Research*, **41**, 18-36.

Duder, J.C. and Roswenwein, M.B, (2001), Towards zero abandonment in call centre performance, *European Journal of Operations Research*, **135**, 50-56.

Eveborn, P. and Ronnqvist M., (2004), Scheduler – A system for Staff Planning, *Annals of Operations Research*, **128**, 21-45.

Extend_6, (2002), *http://www.imaginethatinc.com.*

Harewood, S.I., (2002), Emergency Ambulance deployment in Barbados: a multi objective approach, *Journal of the Operations Research Society.* **53**, 185-192.

Hirota, K., Dong, F., Chen, K. and Takama, Y., (2002), Vehicle Routing, Scheduling and Dispatching System: Based on a HIMS Model, *LNAI*, **2275**, 76-84.

Marianov, V. and Revelle, C., (1996), The Queuing Maximal Availability Location Problem: A model for the sitting of emergency vehicles, *European Journal of Operations Research*, **93**, 110-120.

Mehrotra, V. and Profozich, D.,(1997), The best way to design your Call Centre, Telemarketing and Call Centre Solutions, *Simulation*, **16**, 28-31.

Miller, K. and Bapat, V., (1999), Case Study: Simulation of the call center environment for comparing competing call routing technologies for business case ROI projection, *Proceedings of the 1999 Winter Simulation Conference*.

Pidd M., *Tools for Thinking: Modelling in Management Science*, John Wiley & Sons, West Sussex, 1996.

Trudeau, P., Rousseau J-M., Ferland, J.A and Choquette, J.,(1988), An operations research approach for the planning and operation of an ambulance service, *INFOR*. **27**, 95-114.

Weintraub, A., Aboud, J., Fernandez, C., Laporte, G. and Ramirez, E. (1999), An emergency vehicle dispatching system for electric utility in Chile, *Journal of the Operations Research Society*, **50**, 690-696.