Proceedings of the 1999 Winter Simulation Conference
P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, eds.

ESTIMATING AMBULANCE REQUIREMENTS IN AUCKLAND, NEW ZEALAND

Shane G. Henderson

Department of Engineering Science
University of Auckland
Private Bag 92019
Auckland, NEW ZEALAND

Andrew J. Mason

Department of Engineering Science
University of Auckland
Private Bag 92019
Auckland, NEW ZEALAND

ABSTRACT

The St. John Ambulance Service (Auckland Region) in New Zealand (St Johns) supplies ambulance services to Crown Health Enterprises. Current and future contracts specify several minimum performance targets that St Johns must achieve. As the city of Auckland grows, roads become congested, and population demographics change. These changes mean that St. Johns faces a very difficult problem: how many ambulances are needed, and where should they be placed in order to meet the service targets efficiently.

A preliminary study using queueing theory established that more ambulances were needed and suggested placements. However, the assumptions required in the queueing model were such that a more realistic modelling approach was deemed necessary to verify and refine the queueing model results. We discuss a simulation and analysis software tool BartSim developed by the authors, that is currently being used to address the issues St. Johns faces. The results obtained from BartSim are used to drive changes to the rostering process for staff at St. Johns.

1 INTRODUCTION

St Johns contracts to Crown Health Enterprises to supply emergency medical transport. The contracts stipulate that St Johns supplies a minimum level of service as specified by certain performance targets. These targets relate to response time, which is defined as the time interval between receiving a call, to the time that an ambulance first arrives at the scene. The performance targets are broken down by the location of the call (whether the call is in metropolitan Auckland, or in a rural area) and the priority of the call. St Johns classifies its emergency calls (as opposed to patient transfers etc.) into two levels. Priority one calls are those for which an ambulance should respond at all possible speed, including the use of lights and sirens. Priority two calls are calls for which an ambulance may respond at standard traffic speeds. The performance targets that St Johns faces are as follows.

Table 1: Contractual Service Targets

	Priority 1	**Priority 2**
Metro-politan	80% in 10 mins, 95% in 20 mins	80% in 30 mins
Rural	80% in 16 mins, 95% in 30 mins	no target

St Johns uses a computer-aided dispatch (CAD) system that logs information in a database on every call that they receive. The database then enables St Johns to prepare monthly reports that describe how well they meet their performance targets. Recent reports indicate that St Johns is finding it more and more difficult to meet its contractual targets. It is believed that this degradation in performance is primarily due to increasing congestion on Auckland roads.

Clearly, there is a need for a planning tool that can assist St Johns in determining how to allocate their ambulances and staff to the various stations around Auckland.

The emergency service location problem has been extensively studied, and the literature on this subject is vast. An excellent entry point is Swersey (1994).

In Section 2 we introduce the process that is followed whenever a new call arrives. We first approached the St. Johns problem through a queueing analysis, and this is discussed in Section 3. Section 4 discusses the simulation model underlying BartSim, and in Section 5 we introduce BartSim itself, and cover some of its analysis capabilities. For further details on BartSim, please visit the BartSim web site BartSim (1999).

2 THE DISPATCH PROCESS

Figure 1 shows the process that occurs when a call arrives at St Johns. Staff in the control room identify the closest available ambulance (i.e., an ambulance either idle at its base station or returning from a previous job) and dispatch this vehicle to the scene. After initial treatment at the scene, the ambulance typically transports the patient to a hospital, performs a 'handover' to hospital staff, and then returns to its base station. Where transport is not required, the ambulance returns directly to its base. In either case, the vehicle is considered available to receive calls as soon as it begins returning to base.

Figure 1 shows the two times of most interest in a call response. T8 is the contractual response time discussed earlier, while T8+T9 is the service time for a call. These times were used in the formulation of a queuing model approximation of the process.

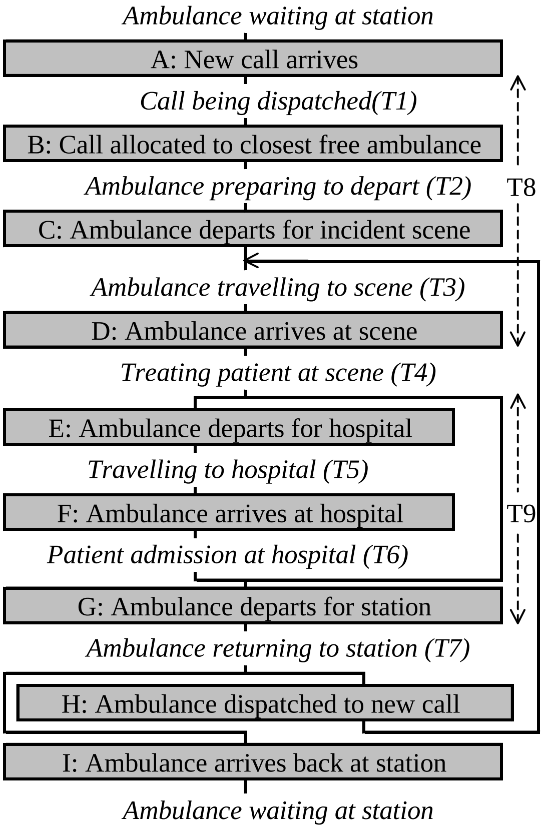

Ambulance waiting at station
A: New call arrives
Call being dispatched(T1)
B: Call allocated to closest free ambulance
Ambulance preparing to depart (T2)
T8
C: Ambulance departs for incident scene
Ambulance travelling to scene (T3)
D: Ambulance arrives at scene
Treating patient at scene (T4)
E: Ambulance departs for hospital
Travelling to hospital (T5)
F: Ambulance arrives at hospital
T9
Patient admission at hospital (T6)
G: Ambulance departs for station
Ambulance returning to station (T7)
H: Ambulance dispatched to new call
I: Ambulance arrives back at station
Ambulance waiting at station

Figure 1: The Ambulance Dispatch and Service Delivery Process

3 QUEUEING MODEL

The queueing model presented in this section was an attempt to obtain a first approximation to the required number of ambulances at each of the 16 ambulance bases currently operated by St Johns.

Consider one of the 16 ambulance bases operated by St Johns in the Auckland region. We will assume that the ambulance base serves calls within the local area, and does not send ambulances to help relieve demand in adjacent areas. We can then study the base in isolation, effectively ignoring the interactions that result from ambulances being sent to other areas. We will discuss how reasonable this "locality assumption" is later in this section.

We further assume that when no ambulance is available at the base when a call arrives, an ambulance from some other location handles the call. This is, in fact, what happens in reality.

We consider periods such as morning rush, evening rush, and weekend nights (for example) separately, because the call arrival rates and service times can vary considerably. The service time variation is primarily due to traffic.

With the above assumptions in place, it is possible to model the number of active calls (the number of calls to which an ambulance has been assigned) using an M/G/∞ queueing model. This model has been previously studied in conjunction with emergency service provisioning (Bell and Allen 1969, Fitzsimmons 1973).

The M/G/∞ model assumes a Poisson call arrival process, i.i.d. service times and an infinite number of servers. The infinite number of servers results from our assumption that calls that cannot be handled from the local base are handled by ambulances stationed elsewhere.

The arrival rate λ is easily calculated from the data available to us as the ratio of the number of calls that arrive to the total observation time. Calculating a mean service time μ^{-1} is more problematical, due to the fact that ambulances may answer calls from locations other than the base if, for example, they are returning to base after completing another callout. We took a call service time to be the time interval T8+T9 from when an ambulance was dispatched, to the time the ambulance became free, either by completing treatment of the patient at the scene with no patient transport, or by completing transport of the patient to a health facility. Note that the service time does not include a time component allowing the ambulance to return to base.

We were then in a position to compute the steady-state probability p_n that n ambulances (both based at the station, or based elsewhere) are busy via

$$p_n = \frac{e^{-\rho}\rho^n}{n!},$$

where $\rho=\lambda/\mu$. The number of ambulances required at the station is then computed as follows.

Suppose we are considering the requirement that 80% of priority one calls are reached within 10 minutes. We first calculate the proportion, r say, of calls that are more than T8>10 minutes travel time away from the station. (We assumed some typical values for T1 and T2.) Assuming that an ambulance responds from the ambulance base (a strong assumption), it follows that there is no way to reach these calls within the 10 minute timeframe. Now, let us assume that the only way to answer a call that is within the 10 minute radius on time is if an ambulance is available when the call comes in. If there are N ambulances stationed at the base, then the probability q_N say that an ambulance cannot immediately respond to the call is

$$q_N = p_N + p_{N+1} + ... = 1 - \sum_{k<N} p_k .$$

This quantity is easily computed in a spreadsheet, and then we choose the minimum value N so that

$$r + (1-r)\, q_N < 20\%.$$

This calculation can be repeated to determine the minimum number of ambulances required to meet the other service requirements.

The results of these calculations indicated that St Johns needed more ambulances at several of their stations to meet their contractual requirements. However, because of the many assumptions implicit in the above model, we were not confident in the actual numbers suggested by the queueing model. In particular, the assumption that calls are answered from the base is a very strong one that is not supported by the data. In reality, the proportion of calls that are answered from the base is substantially less than 50%. Furthermore, St Johns has noted that during busy periods, ambulances are redirected throughout the Auckland region and do not remain close to their base. Therefore, the locality assumption mentioned earlier is, at best, only approximately satisfied. It was therefore necessary to develop a more realistic representation of ambulance operations to better gauge the required number of ambulances at the various bases.

4 THE SIMULATION MODEL

The simulation model is written using a high-level programming language and not specialist simulation software. Real calls are fed through the simulation (rather than artificially generated calls), and ambulances are routed using a travel model adapted from data provided by the Auckland Regional Council (local government). Each of these aspects of the simulation is now discussed in more detail.

We decided not to use an "off-the-shelf" package for simulating St Johns operations for several reasons. First was the logical complexity of the decisions that had to be made within the model. For example, the dispatcher may redirect an ambulance that is responding to a Priority 2 call to a Priority 1 call. Such a decision requires detailed knowledge of travel times, ambulance locations and so forth. This decision is far easier to code using custom software in a high-level language (C) than standard simulation packages. The second reason was speed. The simulation must be very fast to facilitate the large amount of what-if analysis that needs to be done. So we decided to code the simulation in C, and we embedded the simulation program within a Microsoft Visual C++ application to provide a user-friendly interface.

We are very lucky in that an enormous amount of historical data is at our disposal (several years). We have used this data by running trace-driven simulations, i.e., the calls that we simulate are real calls that are read in from a stored file. The data we use from each call are call arrival time, call priority, call location, time spent by an ambulance at the scene, destination to which the patient was transported (if any) and time spent at the destination. The use of this historical data obviates the need to develop a statistical model for generating calls. This is a decisive advantage, as the correlation structure of calls, both temporally and spatially, is undoubtedly rather complex.

Of course, if we were to use BartSim for long-range planning (say greater than 2 years into the future), we might be more wary about using historical data, because the existing data may not be representative of conditions in the future. In such a case, one might want to use an approach similar to that used in the development of the UNOS Liver Allocation Model (Pritsker 1998). That model uses non-homogeneous Poisson processes to generate "arrival times", and other information about the "arrival" is obtained through a bootstrapping procedure. See p. 133 of Bratley, Fox and Schrage (1987) for further discussion of issues relating to the direct use of historical data.

A vital component of the simulation is a travel time model that computes travel times between any pair of locations in Auckland at any time. This model is based on an Auckland Regional Council road network model that supplies both road layout information and travel times along roads (arcs) at various times of the day, including the morning and evening rushes.

We could use this model to compute dynamic shortest paths for ambulances based on time-dependent travel times, but this would be a formidable computation, and would slow the simulation down to a crawl. As a reasonable approximation, we instead pre-compute shortest paths from all possible sources to all possible destinations under a given set of travel times. (Note that for this calculation, we remove nodes from the network that define only road positions and not road intersections where decisions can be made.) Time-dependent travel times are then calculated during the simulation by using time-dependent travel time data on the fixed shortest paths. The model allows entry to and exit from the road network at all non-motorway nodes; an ‘off-network’ speed is used for travel while entering or exiting the network.

When an ambulance is responding to a Priority 1 call, it travels at "lights and sirens" speed. We have captured this effect within the simulation using a multiplicative factor to decrease travel times from more standard travel speeds. This factor was fitted to data available in the database. We are currently exploring other improvements to the modelling of travel speeds.

Ambulance availability is specified in terms of when and where an ambulance is to be brought into operation, and when it is to be removed from circulation. This allows shifts to be effectively captured, along with (for example) lunch breaks that must be held at the ambulance's base and have a certain minimum duration.

5 BARTSIM

BartSim consists of the simulation program itself, the travel model, and analysis tools. The simulation and travel models have been outlined in the previous section. In this section we attempt to give some idea of how BartSim assists in determining ambulance allocations to stations and times. In particular, we focus on the analysis capabilities of BartSim.

It is possible to run BartSim, and see ambulance operation unfolding on the screen. In particular, one sees ambulances travelling along the road network to and from calls. As calls arrive, they are plotted on the screen in a colour indicating their priority. As calls are assigned to an ambulance, the calls change colour, indicating that they are being served. This animation is extremely useful for verification and validation purposes, and for visualising St Johns operations.

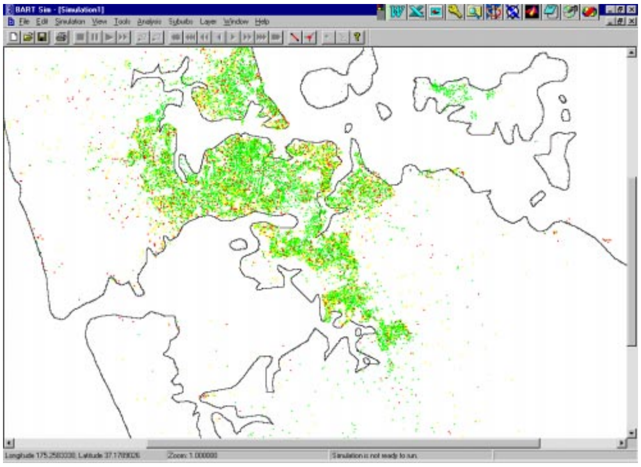
However, when one wishes to collect performance measures, the animation is an unnecessary computational overhead. In this case, animation is turned off, and the simulation proceeds without graphical feedback.

We record the response time performance on every call, so that a call can be classified as to which performance targets have been met. These "micro-statistics" may be aggregated into response time performance within every suburb of Auckland, within every half hour of the week. If a run consists of multiple weeks (the runs are usually several months in duration), then results in the same time period in different weeks are accumulated together. Statistics are also collected on ambulance utilisation.

By recording the response time performance on every call, we can generate plots such as that given in Figure 2. In Figure 2, a green dot indicates that the call was answered within the 80% time requirement, a yellow dot means that the call was answered within the 95% time requirement, and a red dot indicates that neither of these response time bounds was met. One can then visually identify localised areas of poor performance. This is a very powerful capability that we have found extremely useful.

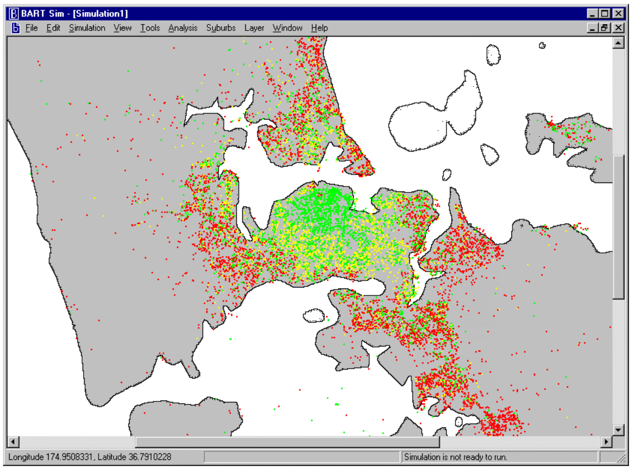
By tailoring the allocation of ambulances to stations, one can use this tool to advantage. In particular, during periods of low call demand, a few stations cover the entire Auckland region. We can identify the “reach” of these stations by producing plots like that of Figure 3. In this plot, we have placed a large number of ambulances at only one station. The number of ambulances is chosen so that no queueing for ambulances occurs, and so we identify the stations coverage area. By repeating such plots for several stations, we can identify a suitable subset of stations that may be used to cover Auckland during off-period times.

We may filter the calls, so that one can "zoom in" on a particular area of Auckland, or a particular time, or both. The performance measures for the time and area of interest are then calculated, allowing one to, for example, identify response time performance for say, centrally located calls. A sample screenshot of such an analysis is given in Figure 4. The small window in the upper-right corner contains detailed information on contractual target performance for a case where ambulance allocation is too light, so that targets are not met.


BART Sim - [Simulation1]
File Edit Simulation View Tools Analysis Suburbs Layer Window Help

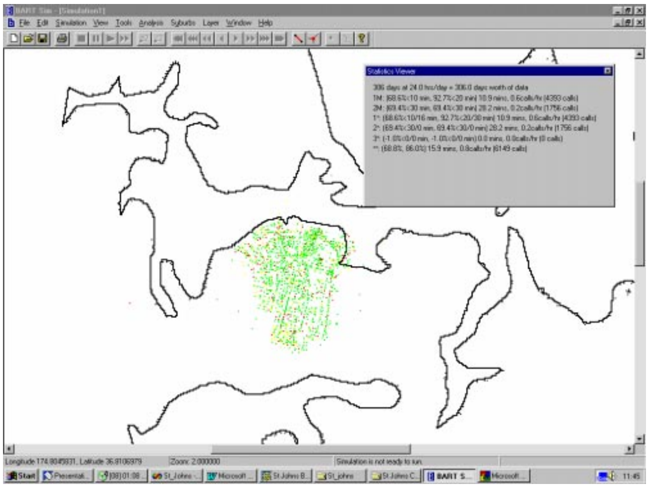
Longitude 175.2583338, Latitude 37.1789026 | Zoom: 1.000000 | Simulation is not ready to run.

Figure 2: Response Time Performance in the Auckland Region


BART Sim - [Simulation1]
File Edit Simulation View Tools Analysis Suburbs Layer Window Help

Longitude 174.9508331, Latitude 36.7910228 | Simulation is not ready to run.

Figure 3: Plot of the "reach" of Pitt St. Station



306 days at 24.0 hrs/day = 306.0 days worth of data
1M: [68.6%<10 min, 92.7%<20 min] 10.9 mins, 0.6calls/hr [4393 calls]
2M: [69.4%<30 min, 69.4%<30 min] 28.2 mins, 0.2calls/hr [1756 calls]
1*: [68.6%<10/16 min, 92.7%<20/30 min] 10.9 mins, 0.6calls/hr [4393 calls]
2*: [69.4%<30/0 min, 69.4%<30/0 min] 28.2 mins, 0.2calls/hr [1756 calls]
3*: [-1.0%<0/0 min, -1.0%<0/0 min] 0.0 mins, 0.0calls/hr [0 calls]
**: [68.8%, 86.0%] 15.9 mins, 0.8calls/hr [6149 calls]

Figure 4: Filter Applied to Results to Identify Performance in the City Center

BartSim also produces statistics on ambulance utilisation. These statistics may be imported into a spreadsheet (we use Microsoft Excel), and analysed from there. An example of the type of graphs that can be produced is given in Figure 5. This graph depicts the underlying demand near one of the stations operated by St Johns. Each row of bars reflects the performance that can be expected over the week when a given number of ambulances are stationed at the base. In particular, each individual bar reflects, for a given number of ambulances and time of the week, the percentage of time that no ambulance is available to respond to incoming calls. This information is extremely useful for getting a first approximation to the number of ambulances required at each individual base at different times of the week. Of course, one would cover some proportion of these calls from other stations, but the plot gives an impression of the underlying demand.

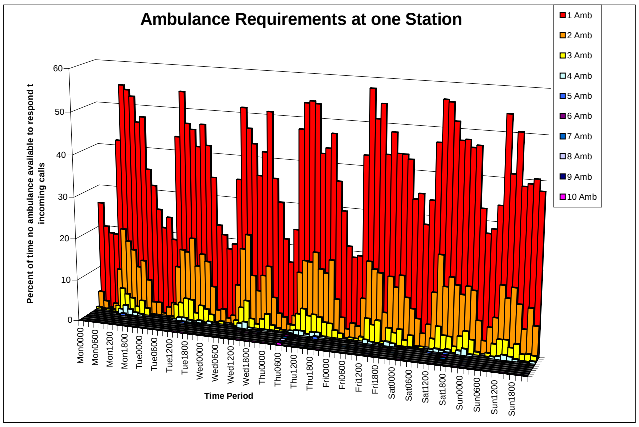


Figure 5: Ambulance Utilization/Requirements at One Station

ACKNOWLEDGMENTS

The authors are very grateful to Jenny Smith for her work on the queueing model discussed in this paper. We are also very grateful to Michael Clist for programming assistance in the early stages of BartSim's development. And we would like to thank Tony Francis of St Johns for his technical assistance throughout the process of interfacing with the CAD database. This work was partially supported by PGSF grant UOA803.

REFERENCES

BartSim (1999). http://www.esc.auckland.ac.nz/ stjohn/

Bell, C. E., D. Allen. 1969. Optimal planning of an emergency ambulance service. *Journal of Socio-Economic Planning Science* **3** 95 - 101.

Bratley, P., B. L. Fox, L. E. Schrage. 1987. *A Guide to Simulation.* Springer, New York.

Fitzsimmons, J.A. 1973. A methodology for emergency ambulance deployment. *Management Science* **19** 627 - 636.

Pritsker, A. A. B. 1998. Life & death decisions. *ORMS Today, August 1998.*

Swersey, A. J. 1994. The deployment of police, fire, and emergency medical units. In Pollock, S. M., M. H. Rothkopf, A. Barnett, eds., *Operations Research and the Public Sector*. North Holland, Amsterdam.

AUTHOR BIOGRAPHIES

SHANE G. HENDERSON joined the Industrial and Operations Engineering Department at the University of Michigan (Ann Arbor) after completing his Ph.D. in Operations Research at Stanford University. He is currently a lecturer in the Department of Engineering Science at the University of Auckland, while on leave from the University of Michigan. His research interests center around discrete-event simulation.

ANDREW J. MASON is a lecturer in the Department of Engineering Science at the University of Auckland, New Zealand. His research interests include the development of optimization systems for staff scheduling problems and the implementation of these within PC-based systems. His research activities are motivated by scheduling projects undertaken for a number of New Zealand companies and government departments.