4 AMBULANCE SERVICE PLANNING: SIMULATION AND DATA VISUALISATION

Shane G. Henderson[1] and Andrew J. Mason[2]

[1]Department of Operations Research and Industrial Engineering
Cornell University
Ithaca, NY 14853

[2]Department of Engineering Science
University of Auckland
Auckland, New Zealand

SUMMARY

The ambulance-planning problem includes operational decisions such as choice of dispatching policy, strategic decisions such as where ambulances should be stationed and at what times they should operate, and tactical decisions such as station location selection. Any solution to this problem requires careful balancing of political, economic and medical objectives. Quantitative decision processes are becoming increasingly important in providing public accountability for the resource decisions that have to be made. We discuss a simulation and analysis software tool 'BARTSIM' that was developed as a decision support tool for use within the St. John Ambulance Service (Auckland Region) in New Zealand (St. Johns). The novel features incorporated within this study include

- the use of a detailed time-varying travel model for modelling travel times in the simulation,
- methods for reducing the computational overhead associated with computing time-dependent shortest paths in the travel model,
- the direct reuse of real data as recorded in a database (trace-driven simulation), and
- the development of a geographic information sub-system (GIS) within BARTSIM that provides spatial visualisation of both historical data and the results of what-if simulations.

Our experience with St. Johns, and discussions with emergency operators in Australia, North America, and Europe, suggest that emergency services do not have good tools to support their operations management at all levels (operational, strategic and tactical). Our experience has shown that a customized system such as BARTSIM can successfully combine GIS and simulation approaches to provide a quantitative decision support tool highly valued by management. Further evidence of the value of our system is provided by the recent selection of BARTSIM by the Metropolitan Ambulance Service for simulation of their operations in Melbourne, Australia. This work has led to the development of BARTSIM's successor, SIREN (Simulation for Improving Response times in Emergency Networks), which includes many enhancements to handle the greater complexities of the Melbourne operations.

KEY WORDS

Ambulance service planning, Simulation

4.1 INTRODUCTION

In 1997 we were contacted by the St. Johns Ambulance Service (Auckland region) in New Zealand, henceforth referred to as St. Johns. St. Johns wanted assistance in developing rosters for their ambulance personnel. This initial contact led to our study of ambulance service management, and to the development of a comprehensive simulation and analysis tool to assist in decision making. (We should emphasize that here the word "simulation" refers to a computer software tool, and not to the replication of realistic incident conditions where volunteers pretend to have certain injuries.) This chapter reviews some of the issues faced by St. Johns managers, and indeed ambulance service managers all over the world, and discusses the methods and tools that we developed to assist them.

The manager of an ambulance service faces a host of difficult policy questions related to operation of the service. The following list is only a sample.

- How many ambulances should be employed and where should they be stationed?

- What policies and procedures should be followed as calls for assistance are received in order to ensure rapid response to calls while obtaining quality information to allow appropriate dispatching?

- Should ambulances be used for non-urgent patient transfers in addition to the usual emergency response function?

- How should dispatching decisions be made when multiple vehicles are available for dispatch?

- How can one examine the tradeoffs associated with sharing a limited number of ambulances between a high-demand metropolitan area and a low-demand rural area? Here the issue is "fairness" in the sense of coverage, versus "efficiency" in the sense of placing ambulances where they will be in high demand.

This is a rather daunting list of problems, to which a great deal of research effort has been focused in the past. Swersey [1] provides a survey of work in emergency service planning that serves as an excellent entry point for the literature. There is a very large literature on such problems, so one might very well ask, what is the motivation for revisiting these problems?

A key difference between the ambulance-planning problem as faced before 1994 and the problem as faced today is the prevalence of data. Virtually all ambulance operations now employ some form of computer-aided dispatch (CAD) system that automatically logs the details of calls as they are received. This information is a veritable goldmine for planners! Without CAD data, ambulance studies typically relied on manual collection of data; see, for example, Swoveland et al. [2], where some of the required data was manually recorded over a period of two weeks.

A second factor that motivated much of the developments discussed in this chapter is the difference in the questions that are being asked. Much of the early development of ambulance theory focused on the questions of where and when ambulances should be operated. While this question is central to much of what we do, we are also motivated by “finer granularity” questions such as how call taking and dispatching should be performed.

To answer these and other questions at St. Johns, we developed a discrete-event simulation of ambulance operations. By manipulating the parameters of the simulation, it is possible to address, in a quantitative manner, many of the questions mentioned earlier. The flexibility of discrete-event simulation means that one can avoid simplifying assumptions that are otherwise needed to obtain performance measure predictions using other methods, such as queueing theory or Markov chain analysis. Perhaps the biggest advantage of simulation is that it is easy to explain as a decision tool to both managers and frontline personnel, so that after they understand the model, they place great store in its results. Obtaining this “buy-in” from decision makers and frontline personnel is crucial in moving from model predictions to decisions and implementation.

To reinforce these points, consider the hypercube model as surveyed in Larson and Odoni [3], and the specialization of this model to ambulance planning in Brandeau and Larson [4]. The hypercube model, while possessing great predictive power, also requires several assumptions with regard to the way that ambulances are dispatched, gives only steady-state results, and requires certain assumptions about the form of “service time” distributions, at least in the case where calls queue when all units are busy. Moreover, explaining how it works to managers is a somewhat daunting task, so that it is hard to instill a feeling of confidence in decision makers as to its predictions. In spite of these disadvantages, it seems to work very well in practice, so it remains a powerful modeling approach that, for a subset of the questions considered here, is a viable alternative to simulation.

Of course, simulation is not new to the ambulance-planning problem. Early examples are Savas [5] for ambulance operations in New York City, and Fitzsimmons [6, 7] for operations in the San Fernando Valley in Los Angeles. Swoveland et al. [2] used simulation to fit the parameters of a metamodel that predicts expected ambulance response time. The expected response time as predicted by the metamodel was then optimized using branch and bound. Simulation was used by Fujiwara et al. [8] to carefully examine a small number of alternative plans that were obtained from an optimization model developed in Daskin [9]. Lubicz and Mielczarek [10] developed a simulation model of rural ambulance operations in Poland. Ingolfsson, Erkut and Budge [11] used simulation to help in siting a "single-start station," i.e., a station from which multiple ambulances begin their shifts. In addition, the use of simulation as a tool to validate the selections of optimization models is almost universal in the literature, and continues to this day. For recent examples see Erkut et al. [12], Harewood [13] and Ingolfsson, Budge and Erkut [14]. For a recent survey of optimization methods in ambulance location problems see Brotcorne, Laporte and Semet [15]. Larson and Odoni ([3], Chapter 7) discuss general considerations related to the simulation of problems similar in form to the ambulance-planning problem.

So what is new in this study?

First, our simulation directly reuses the data recorded in the CAD database. Real calls are fed through the simulation, rather than calls generated using the usual simulation techniques. Justification for our use of trace-driven simulation and discussion of some of the key issues can be found in Section 4.3. Such an approach resolves many difficulties, including accurate modeling of the complex dependence structure of the information related to calls including time of occurrence, location, need for transport and so forth. Of course, it also introduces other problems.

Second, we employ a sophisticated model, adapted from a model developed and used by the Auckland Regional Council [16] for regional planning purposes, to compute travel times. These travel times are used to determine which ambulance to dispatch to a call, the travel time for the ambulance to reach the call, and so forth. The effort we devote to this topic is justified by the great sensitivity of results to travel time assumptions, as noted both by the authors in a preliminary queueing analysis, and by a large proportion of the papers dealing with ambulance planning. For example, Carson and Batta [17] describe how the 30% savings predicted by their model turned into a 6% savings in actual tests, primarily due to the model not effectively capturing a certain travel time/distance relationship. The use of a simpler model based on the “square root law” [18, 19] or other approximations leads to rather large errors due to the highly irregular geography of Auckland; it is basically an isthmus between two oceans, containing many dormant volcano vents. The complex waterways and vents provide significant barriers to travel, leading to a somewhat convoluted road network. A further complication is that travel times are heavily time-dependent. The simulation makes extensive use of the travel model, and we employ several heuristics to reduce the computational effort involved. Many of the techniques used here could be used in other applications requiring travel time calculations where the travel time is time-dependent.

Third, we employ a geographic information system (GIS) to display simulation results and to examine historical performance calculated from real data. To our surprise, none of the ambulance service providers that we have talked with have used such tools in the past, and all have been tremendously excited by their potential. This has occurred in spite of the growing number of sites where a GIS is being used to draw insights from recorded data; see Peters and Hall [20]. Of course, GISs have been used many times to obtain input for simulation models (see, e.g., [21]), but GISs are not often used for displaying discrete-event simulation output. The graphical displays produced by GIS programs allow decision makers to digest copious amounts of information that were previously given in large tables. GIS output displays are currently under-utilized in discrete-event simulation studies, perhaps because of the form of the models involved. But as the ability to link discrete-event simulation software, databases, and standard GIS packages together increases, the use of GIS output display should become more prevalent.

We have been contacted many times by individuals interested in applying BARTSIM methodology to planning problems in the other emergency services, namely fire and police departments. There are many potential applications to these areas from the work presented here, and we believe that such extensions could be tremendously helpful from the practical standpoint. However, it is important to recognize some of the vital differences in these problems from the ambulance-planning problem. These differences mean that substantial effort would be required to tailor the planning methods used here. For example, the utilization rates of fire appliances are typically on the order of a few percent, while it is not uncommon to have ambulance utilization rates, at least in metropolitan areas in New Zealand, as high as 60%. In terms of police patrol planning, an important function of police patrols is to maintain police visibility, so the problems one faces can be quite different.

The remainder of this chapter is organized as follows. In Section 4.2 we discuss some of the particulars of the St. Johns problem, and outline the process that is followed when St. Johns receives an emergency call. Section 4.3 provides an overview of the simulation model underlying BARTSIM and describes some of the data-reuse issues alluded to above. Section 4.4 describes the travel model and the heuristics used to reduce computational overhead. In Section 4.5 we introduce BARTSIM itself, outline some of its GIS-based analysis capabilities, and describe how these analysis capabilities were used to provide useful insights into several decisions faced by St. Johns. Conclusions and suggestions for future research are offered in Section 4.6.

Further details on BARTSIM can be found on the BARTSIM web site (www.esc.auckland.ac.nz/stjohn).

4.2 THE PROBLEM FACED BY ST. JOHNS

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 and other non-emergency calls, into two levels. Priority 1 calls are those for which an ambulance should respond at all possible speed, including the use of lights and sirens. Priority 2 calls are calls for which an ambulance may respond at standard traffic speeds. The performance targets that St. Johns faces are shown in Table 4.1.

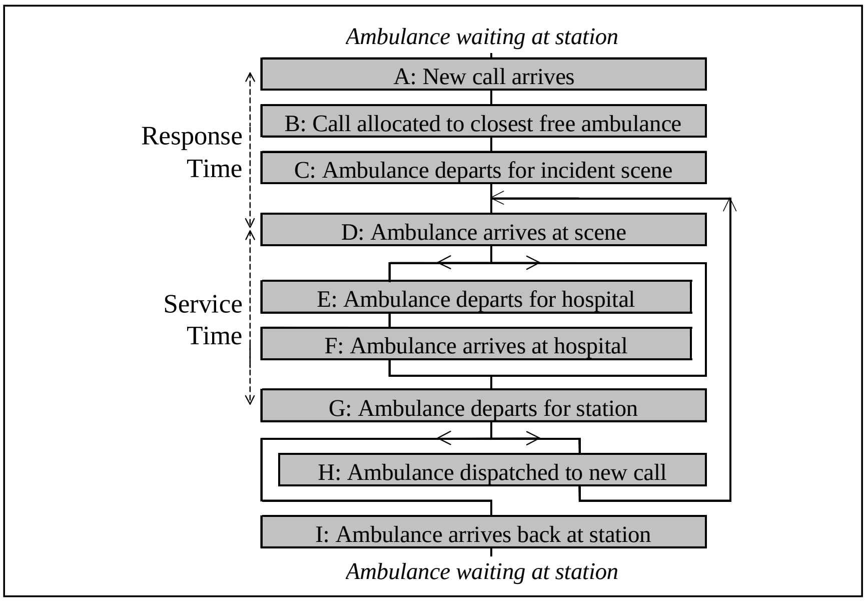
Table 4.1 Contractual service targets

	Priority 1	**Priority 2**
Metropolitan	80% in 10 minutes 95% in 20 minutes	80% in 30 minutes
Rural	80% in 16 minutes 95% in 30 minutes	80% in 45 minutes

It is interesting to note that no guidance is given in the contract as to how these figures need to be interpreted. Interpreting the targets as applying, for example, to the entire Auckland area over the entire year in aggregate will lead to far lower resource requirements than assuming, for example, that the targets must be met in each suburb during each hour of each day. One of the goals of this project has been to develop tools to assist management in exploring performance under a range of possible interpretations of the contract.

St. Johns uses a computer-aided dispatch (CAD) system that logs, in a database, information on every call that is received. The database then enables St. Johns to prepare monthly reports that describe how well they meet their performance targets. When St. Johns first contacted us, these reports indicated that the organization was finding it more and more difficult to meet its service targets. It was (and continues to be) believed that this is primarily due to increasing congestion on Auckland roads.

Figure 4.1 The ambulance dispatch and service delivery process


Ambulance waiting at station
A: New call arrives
B: Call allocated to closest free ambulance
C: Ambulance departs for incident scene
Response Time
D: Ambulance arrives at scene
Service Time
E: Ambulance departs for hospital
F: Ambulance arrives at hospital
G: Ambulance departs for station
H: Ambulance dispatched to new call
I: Ambulance arrives back at station
Ambulance waiting at station

To fully understand these service targets it is necessary to understand the ambulance dispatch and service delivery process. Figure 4.1 shows this process, and identifies the contractual response time discussed earlier. This flowchart also helps to explain the key steps that are captured within the simulation model. When a call arrives at St. Johns, staff in the control room identify an 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. If transport is not required, the ambulance returns directly to its base from the scene. In either case, the vehicle is considered available to receive calls as soon as it begins returning to base.

4.3 THE SIMULATION MODEL

The simulation model is written using a high-level programming language without using specialist simulation software. The simulation is trace-driven, and ambulances are routed using a time-dependent travel model. 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, the logical complexity of the decisions that must be made within the model would be difficult to code in a standard package. 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 number of what-if analyses that need to be performed. Consequently, we decided to code the simulation in C, and then embed the simulation program within a custom-developed Microsoft Visual C++ application to provide a user-friendly interface. Third, this approach has allowed us to tightly couple the simulation with specialized data visualization (GIS) tools, providing integration benefits that would have been hard to achieve using any of the off-the-shelf systems that were available at the time. (Since the software development was completed, simulation software has made great strides in allowing integration with database software and code segments written in other languages.)

We were very lucky in that several years' worth of historical data was made available to us. We used this data by running trace-driven simulations: the calls that we simulate are real calls that are read in from a stored file. See p. 133 of Bratley, Fox and Schrage [22] for a discussion of issues relating to the direct reuse of historical data from the general perspective of discrete-event simulation. We confine our remarks to specifics related to the ambulance-planning problem.

The data used 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 rather complex; see, for example, Lubicz and Mielczarek [10]. For example, the location of a call is somewhat correlated with the time of day at which it is received.

Of course, if we were to use BARTSIM for long-range planning (say more 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 United Network for Organ Sharing Liver Allocation Model [23]. That model uses non-homogeneous Poisson processes to generate "arrival times"; other information about the "arrival" is obtained through a bootstrapping procedure.

An area of concern that arises in using historical data in this fashion is data validity. Indeed, many of the logged calls contain entries that are difficult to believe. For example, it is not uncommon to see durations of 1 second for the time spent at the scene of an incident. Discussions with ambulance personnel revealed that this can occur when personnel forget to notify the CAD system (through a button situated on the dashboard of an ambulance) that they have arrived at the scene. When they realize their error, they "catch up" by pushing the button multiple times. This sort of error not only corrupts the recorded time spent at the scene, but also any surrounding times, such as travel times, that are used elsewhere. Identifying such errors and devising methods for dealing with them are important research areas that we have not explored. Instead, we adopted an ad-hoc procedure where the data for a particular call is "cleaned" if it is "close" to being reasonable, or the call is deleted if the logged data is beyond repair. Of course, if too many calls require cleaning or deletion then we should be concerned, and this is the reason why more research is required in this area. Fortunately, in the St. Johns application such calls appear to occupy a very small percentage of the total calls processed, so they cannot greatly sway the overall results.

The use of trace-driven simulation allows one to deal effectively with many other issues, such as that of multiple-response calls. Multiple response calls occur, for example, because the personnel who initially respond are not legally qualified to administer needed drugs, or because the number of injured parties is large. Each response to a multiple response call is logged in the St. Johns CAD database and linked to previous entries. Within our simulation we simply replay these calls. This very simple approach could lead to potential errors when the personnel that initially respond in the simulation are qualified to assist the patient, so that further ambulance responses are not necessary. A more sophisticated simulation approach might avoid such errors by carefully analyzing the data record, but we did not do this. In any case the number of such multiple response calls is quite small.

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) meal breaks that must be held at the ambulance's base and have a certain minimum duration.

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. An important step in this project has been to establish collaborative links between St. Johns and the Auckland Regional Council, a local government body actively involved in developing strategic policy for the city of Auckland. The Auckland Regional Council made available a road network model that details both road layout information and travel times along roads (arcs) at various times of the day, including the morning and evening rush periods. The use of this data in BARTSIM is discussed in more detail in the next section.

It is possible to run the simulation and see ambulance operations unfolding on the screen. In particular, one sees ambulances traveling along the road network to and from calls. As calls arrive, they are plotted on the screen in a color indicating their priority. As calls are assigned to an ambulance, the calls change color, indicating that they are being served. This animation is extremely useful for verification and validation purposes, and for visualizing St. Johns' operations. It is also tremendously helpful in getting St. Johns personnel to accept the simulation model as a reasonable reflection of reality, and has proven invaluable in communicating our work to staff and management throughout the organization. This aspect of the simulation may seem somewhat trivial from a theoretical point of view, but has been absolutely critical in obtaining “buy in” from the decision makers. We view this selling point as a key advantage of simulation over other operations research methodologies for the ambulance-planning problem. The BARTSIM approach is intuitive and easy to understand for people with non-technical backgrounds.

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 do not report confidence intervals for our performance measures. This is mostly due to the fact that the theory of error estimation from trace-driven simulations is not well understood, so that it is not clear how to develop confidence intervals. This is an area where more research could certainly help.

A simulation model on the scale of BARTSIM requires a great deal of effort in verification and validation to ensure that the model that has been implemented is indeed what was desired, and that the model appropriately represents reality. Instead of entering into a full discussion of our efforts in this regard, which are mostly direct applications of the usual methods as outlined in Law and Kelton [24], we content ourselves with a few examples.

The animation facilities of BARTSIM proved invaluable in verifying the model. By watching simulated ambulance operations over extended periods, many errors in the database of real calls were identified. As well as replaying existing calls, BARTSIM also has a facility for interactively generating calls. This was used to place calls at strategic locations for checking that the ambulance responses were as expected. Shortest paths were generated and displayed over the road network to verify the quality of the chosen routes.

The validation of a model involves ensuring that the model appropriately represents reality. In this regard, we worked very closely with a number of individuals at St. Johns. These people were closely involved in the development phase, and also assisted in performing test runs. Furthermore, we demonstrated the software and described the simulation model to groups of ambulance drivers, who provided feedback on the quality of the model. These steps also helped in the accreditation of the model, where the model is accepted and trusted by decision makers. The decision makers were so closely involved in the development and testing of the model that they felt some form of “ownership” over the system.

4.4 THE TRAVEL TIME MODEL

Auckland is built around two large harbors between two coastlines, and is dotted with dormant volcano vents. Consequently it has a highly irregular topology. Any plausible simulation of road travel cannot rely on ‘as the crow flies’ routes, or simple modifications of these to take into account a moderate number of obstacles, but must incorporate knowledge of the road network including the effects of motorways and major highways. Furthermore, the model must also incorporate the often dramatic changes in travel times that arise from varying congestion levels across the day and the week.

We obtained road data from the Auckland Regional Council detailing a network with about 2,200 nodes and 5,000 directed arcs. This Auckland Regional Transport Model (ART) is a relatively detailed transport model developed for medium term (15-25 years) project and policy planning and evaluation of regional transport strategy [16]. Traffic volumes are determined in ART using equilibrium solutions driven by origin-destination trip demands. Because the trip demands are determined using an underlying demographic model, travel times can be predicted over any planning horizon for which population forecasts are available. This ability to perform longterm planning is most useful when evaluating strategic decisions such as the location of ambulance bases.

We denote the ART road network by $G=(V,A)$, where V is the set of nodes, and A is the set of directed arcs (i,j) from node $i\in V$ to $j\in V$. By entering trip demands for different times of the day, a range of equilibrium solutions can be found, each with different travel times for the arcs. The ARC data includes the 8 a.m. morning peak travel time t_{ij}^{8}, 12 p.m. midday travel time t_{ij}^{12}, and 5 p.m. evening peak travel time t_{ij}^{17} for each arc (i,j). Weighted combinations of these times are used to estimate the travel time t_{ij}^{h} during any other hour h of the day. The weights are chosen using regression models based on actual travel times available in the St. Johns database.

We could use this model to compute dynamic shortest paths for ambulances based on time-dependent travel times whenever the simulation requires such paths. However, this would be a time-consuming computation that would greatly slow down the simulation. As a reasonable approximation, we instead pre-compute and store a range of shortest paths as follows. Of the 2,200 nodes in the network, 1,435 are used to spatially locate bends in the roads, while 765 are 'decision nodes' that define points at which a driver has a choice of direction (ignoring U-turn options). More formally, a node j belongs to the set D of decision nodes, $j\in D$, if there exists both an arc $(i, j)\in A$ and two distinct arcs from j, $(j, k_1)\in \mathrm{A}$, $(j, k_2)\in A$ with $k_1\neq j$ and $k_2\neq j$.

For each pair of decision nodes $i\in D$ and $j\in D$, we pre-compute three shortest paths, P_{ij}^{8}, P_{ij}^{12} and P_{ij}^{17} using the morning peak, midday and evening peak travel times respectively. This decision-node path information is stored in memory.

During the simulation we need to find the shortest path $S\rightarrow F$ between any arbitrary start point S and arbitrary finish point F. The shortest path process we use is heuristic, but nevertheless appears to provide a good level of accuracy.

We note that S and F need not correspond to nodes in the network. The first step in our process is to determine the spatially closest non-motorway nodes, $s\in V$ and $f\in V$, to S and F, respectively. We next determine the sets of decision nodes, $D(s)\subseteq D$ and $D(f)\subseteq D$, that are 'immediately connected' to s and f. The set of decision nodes $D(s)$ is given by $D(s)=T_s\cap D$, where $T_s\subseteq G$ is a tree with root s and with branches each constructed by adding 'outward pointing' arcs until the first decision node is reached. More formally, T_s is initialized with root $T_s=\{s\}$, and then T_s is grown by iteratively adding each arc/node pair $\{(i,j),j\}$: $i\in T_s\backslash D$, $(i,j)\in A$. Similarly $D(f)$ is determined from $D(f)=T_f\cap D$, where T_f is a tree built at f by adding all 'inward pointing arcs', i.e., adding each arc/node pair

$$\{(i,j),j\}: j\in T_f\backslash D,\ (i,j)\in A.$$

We then consider all the paths given by

$$P=\{S\rightarrow s\rightarrow d_s \overset{h}{-}\ d_f\rightarrow f\rightarrow F:\ d_s\in D(s),\ d_f\in D(f),\ h\in\{8,12,17\}\},$$

where $S\rightarrow s$ (and $f\rightarrow F$) denotes 'as the crow flies' travel from S to s (and f to F), $s\rightarrow d_s$ denotes the unique path from s to d_s in T_s,

$$d_s \overset{h}{-}\ d_f$$

denotes the pre-computed shortest path from decision node d_s to decision node d_f at hour h, $h\in\{8,12,17\}$, and $d_f\rightarrow f$ denotes the unique path from d_f to f in T_f. Each of these paths is then evaluated using the interpolated travel times for the hour in which the journey begins. The $S\rightarrow s$ and $f\rightarrow F$ travel is at some assumed off-network speed. The fastest of these paths is deemed the shortest path.

The decision node concept provides two primary benefits. First, without the use of this concept, we would need to solve an 'all shortest paths' problem on 2,200 nodes for each of the three sets of travel times. An 'all shortest paths' problem on n nodes can be solved using the Floyd-Warshall algorithm in $O(n^3)$ time (Papadimitriou and Steiglitz [25], p. 133). With the decision node concept, we solve an 'all shortest paths' problem on approximately one third (765) of the nodes, and therefore reduce the computational effort by a factor of 3^3 = 27. We also reduce the memory required to store the shortest path solutions by a factor of 3^2=9. Second, we consider several paths involving different combinations of decision nodes when deciding which route to take between any origin and destination. This means that the chosen route is a compromise between a pre-solved single fixed route, and the true shortest path as would be determined by solving a dynamic shortest path problem while the simulation is running.

When an ambulance responds 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 modeling of travel speeds.

4.5 BARTSIM

BARTSIM consists of the simulation program, the travel model, and various analysis tools. The simulation and travel models have been outlined in previous sections. This section describes the analysis capabilities of BARTSIM. These capabilities may be applied to historical data as recorded by the St. Johns organization, as well as simulated data generated by the simulation component of BARTSIM. *Informed* comparisons can then be generated between alternative strategies for operating the ambulance service. These analysis capabilities have proven very useful in St. Johns' decision making, several instances of which are mentioned below.

To protect St. Johns' confidentiality, all figures presented in this section are based on simulated data, rather than actual historical data. Road travel times have been perturbed, and all performance figures subjected to random perturbation. The number of ambulances operating out of each base has also been modified, with the result that we see a lower level of performance and greater variability over the Auckland region in terms of response time than is actually the case with historical data.

We record the response time performance on every call, so that a call can be classified according 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. When a run consists of multiple weeks of real data (the runs usually consist of several months of real data), then results in the same time period in different weeks are accumulated together. Statistics are also collected on ambulance utilization.

By recording the response time performance on every call, we can generate plots such as that given in Figure 4.2. In Figure 4.2 a black dot indicates that a call was answered within the 80% time requirement, a gray dot means that the call was answered within the 95% time requirement, and a white dot indicates that neither of these response time bounds was met. (These colors have been modified from those used in the software to improve reproduction.) One can visually identify localized areas of poor performance. This is a very powerful capability that St. Johns have found extremely useful in allowing management to visually interpret data that was previously only available in aggregated database report tables. In particular, using these plots we were able to verify a belief held by some at the St. Johns organization that Silverdale (a suburb of Auckland) needed more resources, perhaps because of the strong recent growth in the region. A long-dormant station in Silverdale has since been reopened.

Figure 4.2 Response time performance in the Auckland region (data is illustrative only)

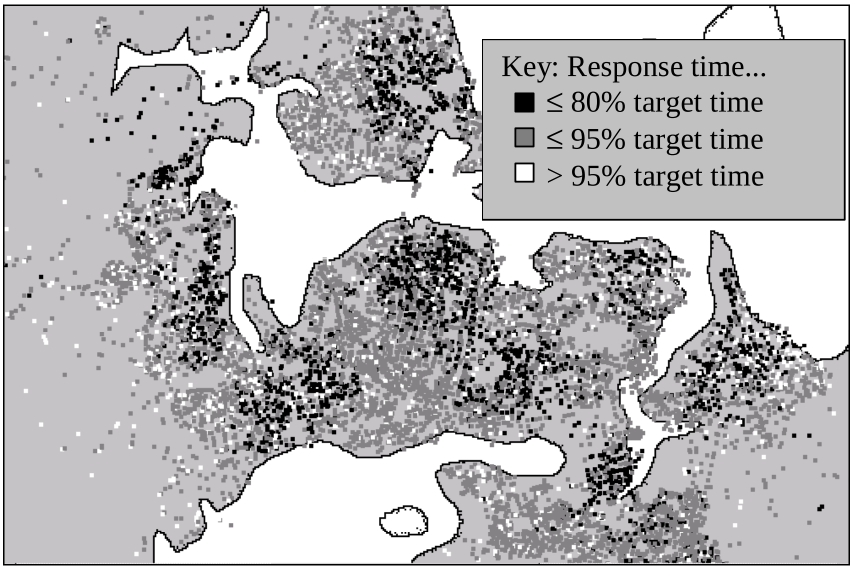
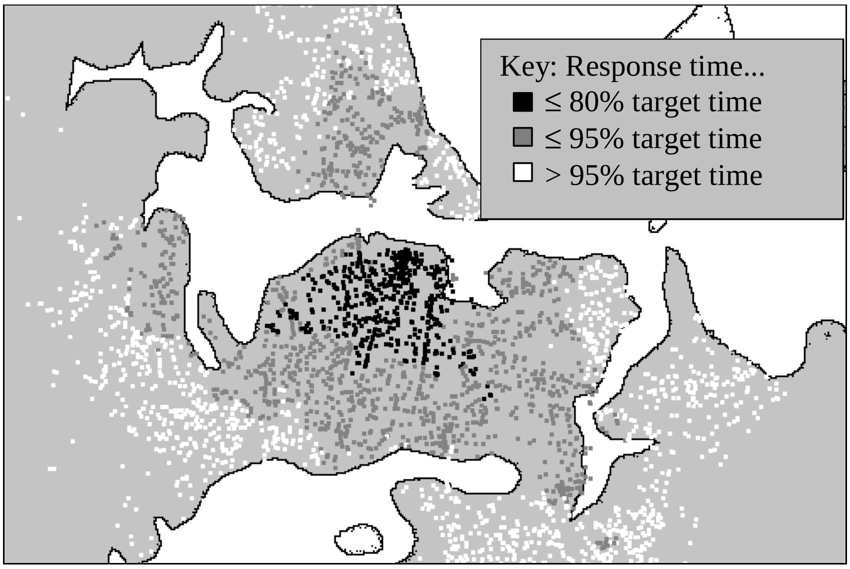


Figure 4.3 Plot of the "reach" of Pitt St. Station during the late morning/early afternoon period on weekdays (data is illustrative only)



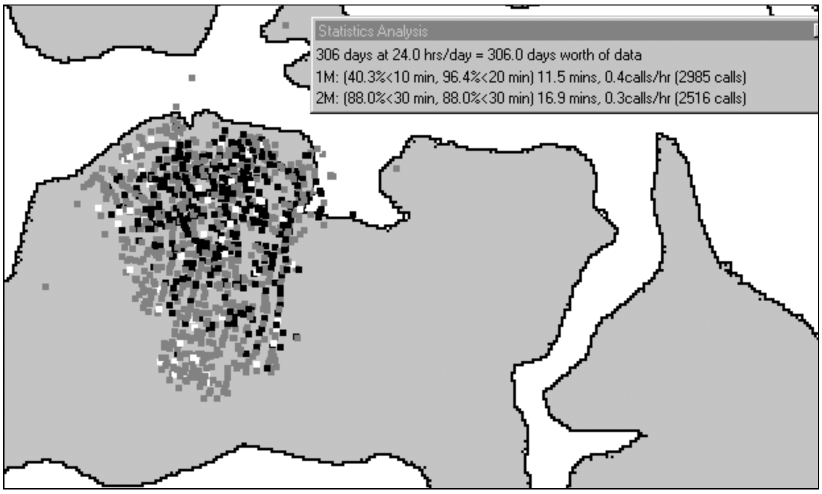
BartSim has proved to be a useful decision support tool for assisting with the allocation of ambulances to stations. During periods of low call demand, performance targets can be met by using just a few stations to cover the entire Auckland region. We can identify the “reach” of a station by producing plots like that of Figure 4.3.

In this plot, we computed the travel time from a single station to all calls. By coloring the call locations as above, we obtain a vivid picture of the area that can be covered by positioning an ambulance at a given station. Since travel time varies dramatically with the time of day, we can obtain a clearer picture of the station's reach at a given time by filtering the calls, so that we only display those arriving during a subset of the week. Figure 4.3 contains only those Priority 1 calls received in the late morning/early afternoon on weekdays. By repeating such plots for several stations, we can identify a suitable subset of stations that may be used to cover Auckland during various times.

As mentioned above, we can filter the calls so that one can "zoom in" on a particular time, or a particular area of Auckland, or both. The performance measures for the time and area of interest are then calculated, allowing one to identify response time performance for centrally located calls, for example. A sample screenshot of such an analysis is given in Figure 4.4. The small window in the upper screen area contains detailed information on contractual target performance for a case where ambulance allocation is too light, so that the targets are not met.

The plots described above are very useful for providing an overview of performance. In addition, plots such as those in Figure 4.4 allow one to provide precise numerical information on performance in a localised region. It is also desirable to be able to summarise on-time performance (relative to the contractual targets) over the entire Auckland region at once; Figure 4.5 is an example of such a plot. In this figure, the Auckland region has been broken down into rectangular regions. Within each region, we compute the percentage of Priority 1 calls reached within the required time limit (10 minutes for urban calls, 16 minutes for rural calls). To allow one to focus on regions containing significant numbers of calls, regions containing a small number of calls are suppressed in the output. Furthermore, the size (area) of the rectangles reflects the number of calls received within the region. We can also substitute other performance measures, such as the number of calls received, or the percentage of Priority 2 calls reached within the required time limit, in place of the performance measure used in this example.

Figure 4.4 Filter applied to results to identify performance in the city centre (data is illustrative only)

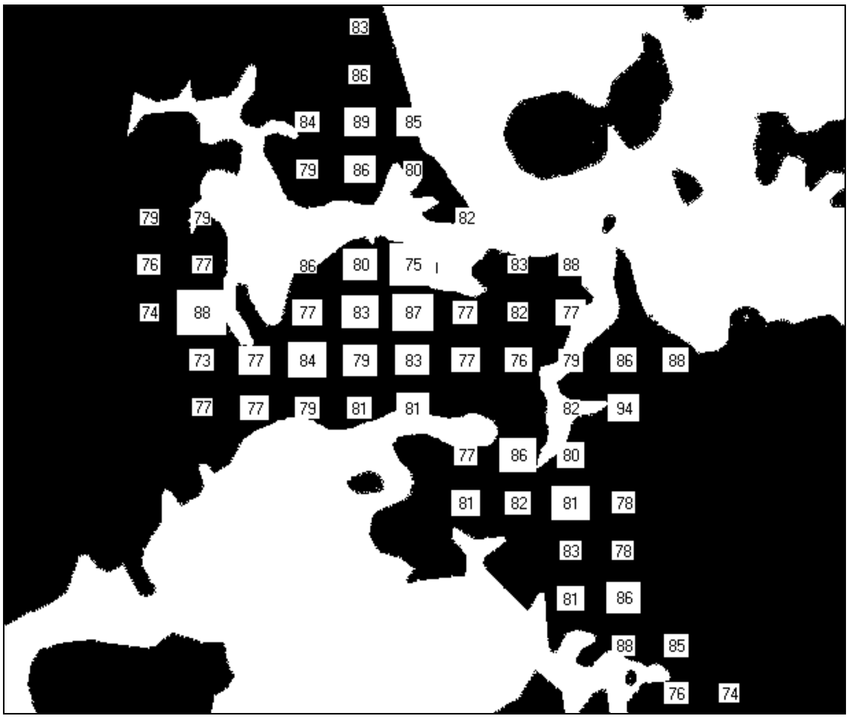


```
306 days at 24.0 hrs/day = 306.0 days worth of data
1M: (40.3%<10 min, 96.4%<20 min) 11.5 mins, 0.4calls/hr (2985 calls)
2M: (88.0%<30 min, 88.0%<30 min) 16.9 mins, 0.3calls/hr (2516 calls)
```

Figure 4.5 is perhaps the most useful of all the plots described thus far in terms of determining required ambulance allocations. We vary the ambulance allocations between bases (usually heuristically, but one could also use optimisation methods), run the simulation, and then observe the performance in terms of these plots. Using these plots, we can locate areas with both a poor overall on-time performance and a large number of calls. These areas are good candidates for extra ambulance resources. Furthermore, by filtering the calls by time and producing the same plots, we can identify times when extra ambulances are most likely to have a large impact on the performance measures.

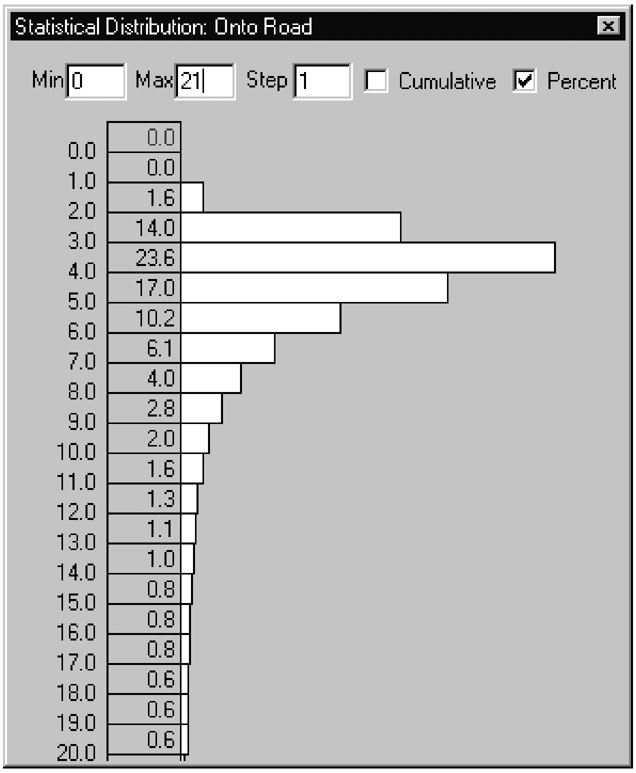
These plots revealed something unexpected when applied to historical data for the St. Johns organisation. In one small suburban area (not shown), a disproportionate (relative to neighbouring areas) number of calls were appearing. Upon investigation it was discovered that there are several accident and emergency clinics in this area, and such clinics generate many calls for St. Johns. The St. Johns organisation was apparently unaware of this situation, and is considering our recommendation that they ensure that an ambulance be relocated close to this vicinity.

Figure 4.5 Plot of average service quality (indicated by the numerical values) and the number of calls (indicated by the size of the white squares) for grid areas in Auckland (data is illustrative only)



BARTSIM can also produce simple histograms of various characteristics of calls, such as response time, time spent by an ambulance at the scene, and so forth. One such histogram is given in Figure 4.6, showing the time between a call being received and an ambulance being dispatched for a set of simulated metropolitan Priority 1 calls. The histogram shows very clearly that for many of the calls, a large amount of time is spent before an ambulance is dispatched to a call. Time spent in the dispatch process reduces the amount of time that an ambulance has to reach the scene of a callout if it is to meet the contractual performance targets. A plot similar to this for the historical data recorded by St. Johns was one of our most important findings for the organisation. Small decreases in these dispatch times can have (as simulations quantified) a large impact on contractual performance, so that it is worth devoting considerable effort to determining ways in which the dispatch time can be reduced. Apparent inefficiencies in the dispatch process can, when considered in view of the overall goals of the organization, actually be viewed as efficiencies, especially when the alternative expense of additional ambulance units is considered.

Figure 4.6 Distribution of the interval (in minutes) between a call being received and an ambulance responding (by radio) that it is en route (distribution is illustrative only)

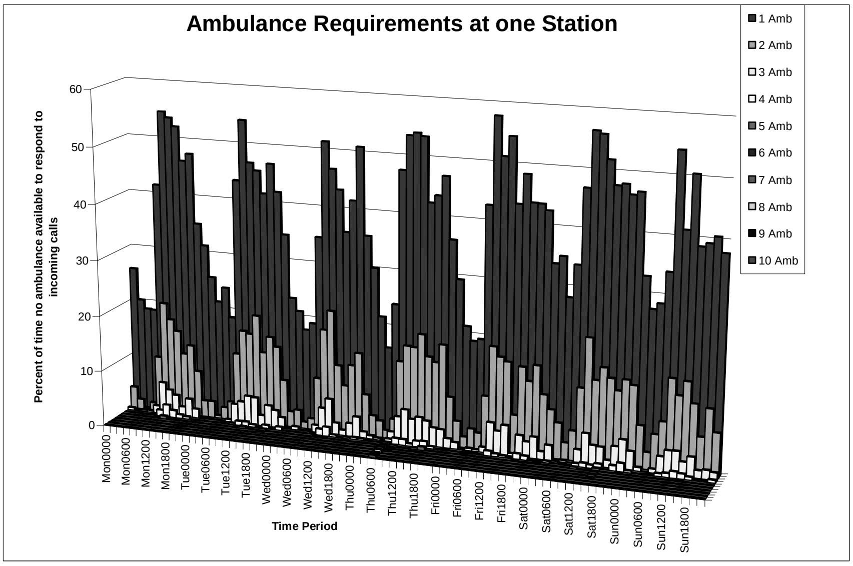


Statistical Distribution: Onto Road

Min 0 Max 21 Step 1 □ Cumulative ☑ Percent

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 4.7. 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 4.7 Ambulance utilisation/requirements at one station (data is illustrative only)



As a final example of the nontraditional uses of BARTSIM, we mention that at a certain stage St. Johns was considering the use of a dispatching strategy that was expected to have a number of effects. First, it would better match the skills of the staff with the patient's requirements at the scene, thus resulting in better care. Second, it would result in fewer Priority 1 dispatches being made because the improved data collection would allow more cases to be classified as Priority 2. Priority 2 cases have a longer target response time so the performance targets for these cases would appear to be easier to meet. However, vehicles on Priority 2 dispatches do not use lights and sirens, so the time a vehicle spends on a case increases if it is changed from Priority 1 to Priority 2. The improved case classification would come at the cost of increased dispatch times. These changes were built into the simulation using approximations for the extent of the effects, and then comparisons between the current and proposed system were drawn based on the plots discussed in this section. The analysis played a large role in determining whether the proposed system would be adopted.

4.6 CONCLUSOINS

BARTSIM has been used to evaluate several decisions considered by St. Johns, including the use of a dedicated non-emergency patient transfer service, the possible introduction of a new dispatching method, and changes to where and when ambulances should be allocated. The results of these studies have been used to shape policy at St. Johns, and we continue to work with them on these and other issues, including rostering requirements for their staff. This experience has convinced us that simulation is a powerful tool in emergency service planning that is currently underutilized. Good simulation visualization tools have proven invaluable as a communication tool for describing our work to management and staff of St. Johns. The spatial data visualization capabilities have provided management with a significantly improved understanding of their current performance and, in conjunction with the simulation model, allowed results from what-if analyses to be readily communicated and understood.

It is important in vehicle simulation models to accurately capture travel time information. We have developed heuristics that allow both accurate modeling of travel times and rapid simulation run times. In addition, we introduced the notion of a decision node, which dramatically decreases the time required to compute shortest paths in the networks. This concept may be of interest in other applications where shortest paths must be calculated in large networks.

The travel times predicted by our model are deterministic: the same time is always predicted for travel from one point to another at a given time on a given day. However, travel times can vary tremendously depending on unpredictable events such as traffic congestion, weather, and traffic accidents. It is our belief, based on some initial analysis with very simple models, that randomness in travel times can have a material effect on the predictions of a model, and this is an area that we are beginning to investigate. Some care is needed, as it is not immediately clear how to generate random travel times. In general, there will be “macro” effects, such as those described above, which affect many ambulance trips in the same way, whereas other “micro” effects, such as traffic light phasing, might be confined to a single ambulance trip.

The combined simulation and data visualization tools introduced here have been of tremendous help to St. Johns, and several other ambulance companies have expressed interest in using the system within their organization. In our experience, the combination of CAD databases, GIS visualization methods and simulation leads to more informed decision making, and better utilization of resources, than the previous state of the art has supplied.

Since preparing this chapter, BARTSIM has been selected in a competitive tendering process for use in Melbourne, one of the larger cities in Australia.

As part of this work, BARTSIM has evolved into a more powerful system known as SIREN (Simulation for Improving Response times in Emergency Networks) (see http://www.optimal-decision.com). Enhancements include call generation using non-homogeneous Poisson processes, introduction of stochastic travel times, more detailed case classifications, and more sophisticated simulation logic to handle the increased operational complexity of this new problem. For example, SIREN can dispatch several vehicles to a call, one of which is left at the scene while the ambulance officers travel in the other vehicle to the hospital. Upon leaving the hospital, this vehicle then travels back to the scene where the officers return to their original vehicles. The transport model has also been enhanced to reduce the memory requirements of the pre-computed shortest paths, allowing a network with 6,000 nodes and 14,000 arcs to be handled. This network also allows shortest distance (in addition to fastest time) routes to be calculated, and includes arc-specific times for lights and sirens travel. It is pleasing to see the value that SIREN can add being recognized by another ambulance organization.

Acknowledgments

The authors would like to thank the management and ambulance officers at St. Johns for their support of this project. Tony Francis's technical assistance has been most useful, while the contributions of Shaun Camp, Marion Etches, Pauline Mellor and Scott Osmond have all proven invaluable. We are also very grateful to Russell Jones, Don Houghton, Anatole Sergejew, and the rest of the ART team at the ARC for making their transport model available to us. We wish to acknowledge the outstanding contribution of Jenny Smith in her work on an initial queuing analysis of the St. Johns problem. We also acknowledge the huge contributions made by Optimal Decision Technologies, Ltd. (www.optimal-decision.com) and Alex Currell and his team at the Metropolitan Ambulance Service in Melbourne, Australia, in the development of SIREN, BARTSIM's successor. Finally, we are most grateful to Michael Clist for programming assistance in the early stages of BARTSIM's development. This work was partially supported by New Zealand Public Good Science Fund grant UOA803, and by National Science Foundation Grant DMI-0230528.

References

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

[2] Swoveland, C., D. Uyeno, I. Vertinsky, and R. Vickson (1973). Ambulance location: A probabilistic enumeration approach. *Management Science*, 20, 686- 698.

[3] Larson, R.C. and A.R. Odoni (1981). *Urban Operations Research*. Prentice-Hall, Englewood Cliffs, NJ. Also available at http://web.mit.edu/urban_or_book/www/book/

[4] Brandeau, M.L. and R.C. Larson (1986). Extending and applying the hypercube queueing model to deploy ambulances in Boston. In A. Swersey and E. Ignall, eds. *Delivery of Urban Services*, TIMS Studies in Management Sciences 22, Elsevier. 121-153.

[5] Savas, E.S. (1969). Simulation and cost-effectiveness analysis of New York’s emergency ambulance service. *Management Science*, 15, B608-B627.

[6] Fitzsimmons, J.A. (1971). An emergency medical system simulation model. *Proceedings of the 1971 Winter Simulation Conference*. New York. 18-25.

[7] Fitzsimmons, J.A. (1973). A methodology for emergency ambulance deployment. *Management Science*, 19, 627-636.

[8] Fujiwara, O., T. Makjamroen, and K.K. Gupta (1987). Ambulance deployment analysis: A case study of Bangkok. *European Journal of Operational Research*, 31, 9-18.

[9] Daskin, M.S. (1983). A maximum expected coverage location model: Formulation, properties and heuristic solution. *Transportation Science*, 17, 48-70.

[10] Lubicz, M. and Z. Mielczarek (1987). Simulation modeling of emergency medical services. *European Journal of Operational Research*, 29, 178-185.

[11] Ingolfsson, A., E. Erkut, and S. Budge (2003). Simulation of single start station for Edmonton EMS. *Journal of the Operational Research Society*, 54, 736-746.

[12] Erkut, E., R. Fenske, S. Kabanuk, Q. Gardiner, and J. Davis (2001). Improving the emergency service delivery in St. Albert. *INFOR*, 39, 416-433.

[13] Harewood, S.I. (2002). Emergency ambulance deployment in Barbados: A multi-objective approach. *Journal of the Operational Research Society*, 53, 185-192.

[14] Ingolfsson, A., S. Budge, and E. Erkut (2003). Optimal ambulance location with random delays and travel times. Preprint. University of Alberta School of Business, Edmonton, Alberta, Canada.

[15] Brotcorne, L., G. Laporte, and F. Semet (2003). Ambulance location and relocation models. *European Journal of Operational Research*, 147, 451-463.

[16] Auckland Regional Transport (1994). Auckland Transport Models Project: Technical Working Paper 1 ‘Network Development And Inventory,’ Environment Division, Auckland Regional Council, Auckland, New Zealand.

[17] Carson, Y.M. and R. Batta (1990). Locating an ambulance on the Amherst Campus of the State University of New York at Buffalo. *Interfaces*, 20, 43-49.

[18] Kolesar, P. (1975). A model for predicting average fire engine travel times. *Operations Research*, 23, 603-614.

[19] Kolesar, P., W. Walker and H. Hausner (1975). Determining the relation between fire engine travel times and travel distances in New York City. *Operations Research*, 23, 614-627.

[20] Peters, J. and G.B. Hall (1999). Assessment of ambulance response performance using a geographic information system. *Social Science and Medicine*, 49, 1551-1566.

[21] Pidd, M., F.N. de Silva, and R.W. Eglese (1996). A simulation model for emergency evacuation. *European Journal of Operational Research*, 90, 413- 419.

[22] Bratley, P., B.L. Fox, and L.E. Schrage (1987). *A Guide to Simulation*. Springer, New York.

[23] Pritsker, A. (1998). Life and death decisions. *OR/MS Today*, August.

[24] Law, A.M. and W.D. Kelton. (2000). *Simulation Modeling and Analysis*, 3rd ed. McGraw-Hill, Boston, MA.

[25] Papadimitriou, C.H. and K. Steiglitz (1982). *Combinatorial Optimization: Algorithms and Complexity*. Prentice Hall, Englewood Cliffs, NJ.