A simulation model to enable the optimization of ambulance fleet allocation and base station location for increased patient survival

Richard McCormack[a], Graham Coates[a,*]

[a]*School of Engineering and Computing Sciences, Durham University, South Road, Durham, DH1 3LE, UK*

Abstract

An effective emergency medical service (EMS) is a critical part of any health care system. This paper presents the optimization of EMS vehicle fleet allocation and base station location through the use of a genetic algorithm (GA) with an integrated EMS simulation model. Two tiers to the EMS model realized the different demands on two vehicle classes; ambulances and rapid response cars. Multiple patient classes were modeled and survival functions used to differentiate the required levels of service. The objective was maximization of the overall expected survival probability across patient classes. Applications of the model were undertaken using real call data from the London Ambulance Service. The simulation model was shown to effectively emulate real-life performance. Optimization of the existing resource plan resulted in significant improvements in survival probability. Optimizing a selection of one hour periods in the plan, without introducing additional resources, resulted in a notable increase in the number of cardiac arrest patients surviving per year. The introduction of an additional base station further improved survival when its location and resourcing were optimized for key periods of service. Also, the removal of a base station from the system was found to have minimal impact on survival probability when the selected station and resourcing were optimized simultaneously.

Keywords: Simulation, Optimization, Emergency medical service.

*Corresponding author. Tel.: +44 (0)191 3342479
Email address: graham.coates@durham.ac.uk (Graham Coates)

1. Introduction

An effective emergency medical service (EMS) is a critical part of any health care system. One key factor in the performance of an EMS is the speed at which emergency vehicles can respond to incidents. It is vital that at all times emergency vehicles are located so as to ensure adequate coverage and rapid response times [1]. The optimum distribution of emergency vehicles depends on the distribution of demand.

There is a significant body of literature on the effective positioning of EMS facilities and a wide variety of models have been developed to solve the problem [1–5]. Although different models are used, the overarching goal of any such model is to determine the distribution of emergency facilities and resources that best serves a given demand.

Much of the literature on EMS facility location uses models derived from the principle of set covering [5–8]. Such models aim to locate EMS resources so as to cover a set of demand nodes. At its simplest, a node is covered if an EMS resource is within a predetermined distance or response time. Early covering models ignored the stochastic nature of EMS systems and dealt with the static and deterministic location problem [9, 10]. Probabilistic models have been developed that use queuing theory to account for the fact that ambulances act as servers in a queuing system and are sometimes unavailable [11–13].

Increases in computing power and the advent of metaheuristics have facilitated the development of more complex models to solve the EMS facility location problem. For example, Aytug and Saydam [14] solved the large-scale, maximum expected coverage location problem (MEXCLP) using two forms of genetic algorithm, one being coupled with a local search algorithm, which were seen to outperform an integer programming approach and Daskin's heuristic [15] in terms of locating an optimal or near-optimal solution in a reasonable amount of time. Also, Saydam and Aytug [16] solved the MEXCLP by combining a genetic algorithm with Jarvis' and Larson's hypercube model [17, 18], the latter of which had previously only been used as a comparison tool [19, 20]. Similarly, a number of other approaches have combined a hypercube queuing model with a genetic algorithm [21–24], simulated annealing [25] and Tabu search [26–28] in order to optimize the location and allocation planning of EMS systems. Also, some metaheuristics have been used alone such as a genetic algoirithm to optimize the location of ambulances in order to contribute to higher survival rates from life-threatening

medical events [29] and simulated annealing to solve the large scale, dynamic maximal covering location problem (MCLP) [30]. Some research has focused on the stochastic and uncertain nature of the location problem of EMS vehicles. For example, Repede and Bernardo [31] introduced TIMEXCLP, the first covering model with time variation to account for the stochastic nature of variables such as travel and service times when evaluating alternative ambulance locations. On a similar theme, Ingolfsson et al. [32] included uncertainty in relation to delays and travel times when aiming to maximize expected coverage by minimizing the allocation of the number of ambulances to stations in order to provide a specified level of service.

Recent developments in the field of EMS system research include a number of areas such as the use of approximate dynamic programming [33, 34] and stochastic programming [35, 36] to solve the ambulance redeployment and/or location problem. Another important development has been the use of patient survivability as a performance metric instead of the concept of 'cover'. Erkut et al. [37] recognised the need for a new performance measure and introduced the concept of survival functions; non-linear functions that model the relationship between response time and survival rate. They incorporated a survival function, derived from medical research into cardiac arrests, within a number of existing covering models. It was found that the existing covering models produced sub-optimal EMS facility distributions when evaluated using the more realistic performance measure. Indeed, in recognizing the importance of performance measures being more closely related to patient outcomes, McLay and Mayorga [38, 39] investigated the influence of maximizing coverage for different response time thresholds on patient survival rates. Acknowledging that the findings were specific to the rural/urban geographic region considered, locating ambulances to maximize minumum response time thresholds was found to simultaneously maximized patient survival. Also, Bandara et al. [40] focused on patient survivability and its maximization via optimizing the dispatch of paramedic units to emergency calls depending on their severity. It was found that consideration of call severity, or urgency, would lead to an increase in patients' mean survival probability.

One limitation of the work presented by Erkut et al. [37] is the modelling of only one incident type. Knight et al. [41] addressed this limitation by incorporating a cardiac arrest survival function with step functions representing response time targets for other incident types. The heterogeneous objective function was a sum of the individual survival functions weighted

by their relative priorities; effectively a multi-objective optimization. This objective function was incorporated into a simple covering model.

Although covering models are popular, it is accepted in the literature that EMS simulations are more accurate [42, 43]. In fact, simulation is almost universally used in the literature to validate the selections of optimization models [44]. As well as increased realism, simulation produces a wealth of performance data, such as response time distributions and ambulance utilisation statistics, which cannot be acquired using a covering model.

1.1. Contribution of this paper

Despite recent advances in the field there are areas for further development. This paper presents work to amalgamate the recent developments in performance metrics with an accurate, simulation based model of a complex EMS system. The aim of this work is to provide useful recommendations on vehicle resourcing and ambulance base station siting from a planning perspective. The novel areas of this work are:

- The integration of heterogeneous survival functions with an EMS simulation capable of utilising real call data.
- A method for the modelling of time-variant travel times in a complex network. The incorporation of vehicle routing avoids the unsound, but common assumption that vehicles respond from their base stations.
- A multi-tiered EMS model recognising the different demands on two vehicle classes; ambulances and rapid response cars.
- Calibration and validation of the simulation model as applied to the London EMS system.
- An application case study incorporating one million calls across the entire London EMS region (2,400 km^2), demonstrating the significant improvements in patient survival achievable with optimized resource plans.
- Simultaneous optimization of base station location and the number and type of EMS vehicles located at each station (vehicle allocation). Further, the impact of introducing a new station to the London system is investigated; recommendations are made on the optimum location for the new station and how it should be resourced.

The remainder of this paper is structured as follows. In section 2, details are given of an EMS simulation model. Next, Section 3 presents the optimization heuristic and objective funcion used in conjunction with the simulation model. In section 4, the results of a number of cases considered based on real incident call data from the London Ambulance Service are presented and discussed. Finally, concluding remarks and possible areas for further work are given in Section 5.

2. Simulation model

An EMS simulation model was required in order to evaluate proposed vehicle allocations, i.e. the number and type of EMS vehicles located at each base station. Simulation was chosen over other modelling techniques due to the increased realism and accuracy it affords [44, 45]. Simulation was chosen over other modelling techniques due to the increased realism and accuracy it affords [44, 45]. Simulation gains a further advantage through the ability to directly utilise real call data (trace-driven), avoiding the simplifications required to model demand. Collating such data used to be an issue, however the majority of modern EMS providers utilise computer-aided dispatch (CAD) systems that can provide data for use in data-driven simulations. The use of real data also captures the complex, time-dependent variation in demand and resourcing, something omitted by the majority of studies.

For the purpose of the study reported in this paper, the London Ambulance Service (LAS) provided data covering all emergency calls between 1st November 2011 and 31st October 2012; almost one million calls, see Figure 1 for the geographic distribution of these calls. For each call the data contained call arrival time, incident type (initial diagnosis), location, dispatch time, number of resources dispatched, time spent at the scene, destination of patient transport (if any), time travelling to hospital, and time spent handing over the patient to hospital staff. The originating location of dispatched EMS vehicles was not included in this data. Wherever possible this data was used directly in the simulation instead of modelling, helping to minimize assumptions. Modelling was employed in cases where the required data was not available or decisions and processes depended on vehicle allocation. Figure 2 shows the EMS vehicle dispatch and service process.

The simulation model was written using C++ rather than bespoke simulation software as it afforded a superior level of control for modelling such unique processes. Further, execution speed was also a primary concern. Code

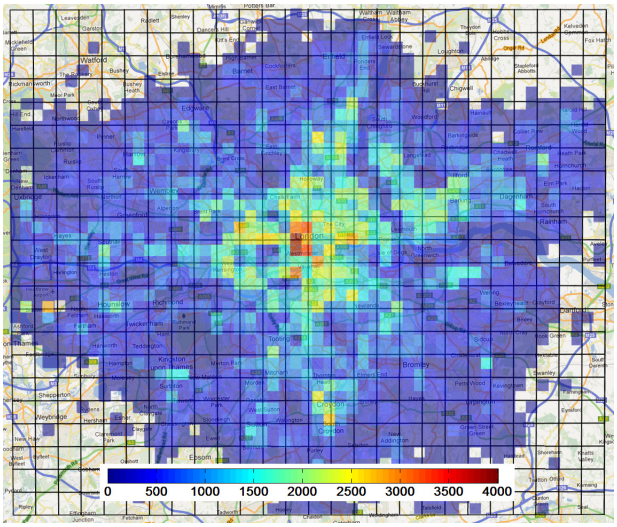

Figure 1: Geographic distribution of all calls received by the LAS for a one year period between 1st November 2011 and 31st October 2012; almost one million calls.

was designed to be as generic as possible, with minor alterations the simulation model could be applied to any large EMS system.

The simulator accepts the following inputs:

- An allocation of ambulances (A) and rapid response cars (R) to base stations.
- A period of interest ($\top$). Calls that arrive within the time period $\top$ are loaded from a data file.

The trace-driven simulator processes emergency calls in accordance with a first-come first-served order. Waiting calls are stored in a global system queue and only assigned to a specific station and vehicle at the point of dispatch; the logic is described in Algorithm 1. Four types of simulation event are processed:

- New call arrival (lines 8-15). The new call (c) is assigned the required number of vehicles according to the dispatch policy. If not all of the required vehicles are available then as many as possible are dispatched and further vehicles are sent once they become available. In this study a simple myopic dispatch policy is employed, where the nearest available vehicles (shortest estimated travel times) are assigned. The nearest

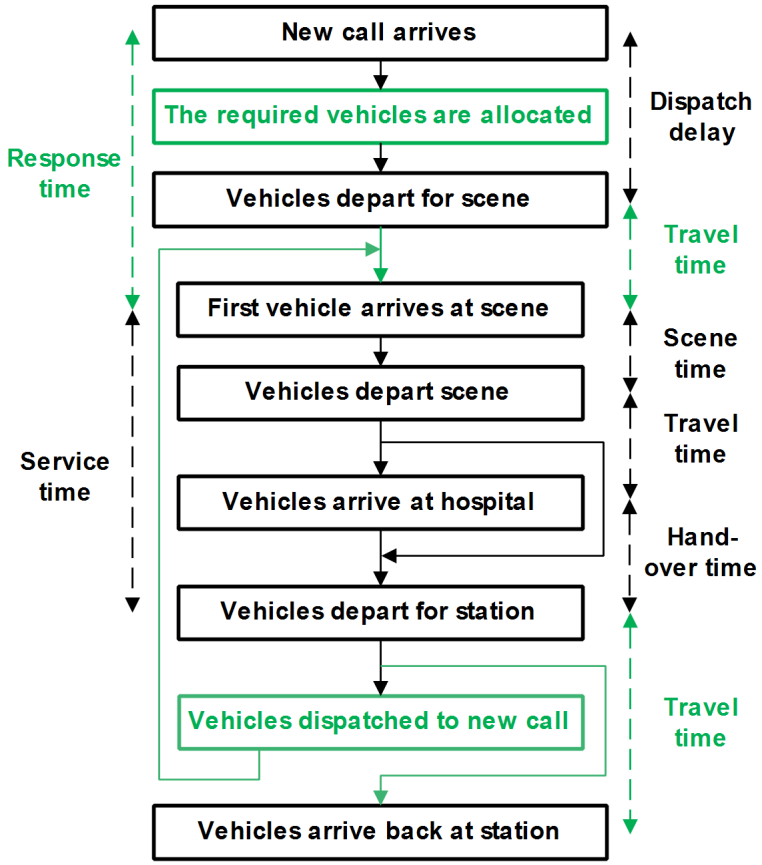

Figure 2: The EMS vehicle dispatch and service process. The details and timings of most processes could be extracted from the LAS call data. Modelling was employed where this was not viable (highlighted in green).

available vehicles may include one or more vehicles in the process of returning to base following completion of a job, these vehicles would then be re-dispatched from their current simulated location. Discussions with the LAS indicated that this is a reasonable approximation of the real dispatch policy, although an operator may decide to wait for alternative vehicles if the nearest currently available is beyond a threshold distance.

The array $V(c)$ contains the vehicles assigned to the call, the number and type of which is specified in the LAS data. If vehicles are assigned ($|V(c)| > 0$) then they are flagged as busy and a scene departure event is created. The scene departure time is dependant on the call time (t), dispatch delay (t_d), time taken for the nearest vehicle to reach the scene (t_r), and time spent at the scene (t_{sc}). These variables are unique to each call and all but the travel time is extracted from the LAS call data. Travel times are estimated using the travel time model presented in Section 2.1.

If no resources are available then the call is added to a first-in first-out queue (Q). Although this is a simple approximation to what is a complex decision for an EMS operator, it was considered a reasonable approximation by LAS personnel. This queue of waiting calls is considered whenever a vehicle finishes servicing another call.

- Scene departure (lines 16-23). The vehicles assigned to a call are leaving the scene of the incident, t_{sc} seconds after arrival of the first vehicle. If transport is required, as specified in the LAS data, one ambulance departs for hospital and a job completion event is created. At least one ambulance is sent to every call requiring transport and the first ambulance on-scene is assigned transport duties. The job completion time is dependant on the travel time to hospital (t_h), and the time spent handing over the patient to hospital staff (t_{ho}). Both of these variables are extracted directly from the LAS call data. All other vehicles are flagged as available and depart for their respective base stations.

- Job completion (lines 24-27). Service of call c is complete. Any vehicles still assigned to the call are flagged as available and depart for their respective base stations.

- Vehicle location update (lines 28-31). The locations of travelling vehi-

cles are periodically updated in line with the routing and travel time model presented in Section 2.1. The "real-time" locations of vehicles are used when deciding which vehicles should respond to calls. This is a significant departure from the existing academic literature in which it is assumed that vehicles always respond from their base stations; an assumption even made in relatively complex simulation models [43, 45]. Discussions with the LAS revealed that such an assumption is flawed for a busy EMS. In London it is not uncommon for vehicles to be away from their bases for several hours at a time.

2.1. Modelling travel times in a large and complex network

One of the principal components of the simulation model is a method for estimating travel times and routes within the London transport network. Ideally this travel time model would be capable of estimating the distance between any two precise locations, however the size and complexity of the London network make this impractical. In order to simplify the model the simulation space was discretised into a grid; a common approach to demand aggregation [14]. It was found that over 99.9% of LAS calls in the one year period considered were bounded by a 53km by 45km area approximately centred on the city centre. This area was divided into a 26 by 22 grid to form a discretised simulation space; see Figure 3. The grid resolution used was a compromise between accuracy and computational efficiency. The locations of all calls, base stations, and hospitals were snapped to the centre of the cell in which they resided. The 0.1% of calls outside these bounds were excluded from the model.

In order to effectively model travel times within London it was important to capture the time-dependent nature of the system; something rarely considered in similar studies. The most practical approach was to assume that the routes travelled by vehicles were independent of time but the speed at which vehicles travelled was not. Having time-independent routes allowed distances between nodes to be precomputed, reducing the run-time and complexity of the model. One approach to modelling distances is to use the Euclidean distance between two nodes [46]; this however, does not capture the varying complexity of a road network. One way to capture this complexity is to use geographic information systems (GIS). A tool was developed utilising the Google Maps Javascript API, which could return the travel distance, along the quickest road route, between any two nodes.

Algorithm 1 Simulator: Trace-driven simulation method

```
1: function SIMULATE(A, R, ⊤)
2:     C ← LOADCALLS(⊤)                          ▷ load call log (C) from file
3:     Q ← ∅                                     ▷ initialize empty call queue (Q)
4:     ε ← C                                     ▷ initialize event queue (ε) with calls
5:     while |ε| > 0 do
6:         remove next event (e) from ε
7:         t ← e.time()                          ▷ update current time (t)
8:         if e = new call (c) then
9:             V(c) ← DISPATCH(c, A, R)          ▷ assign vehicles
10:            if |V(c)| > 0 then
11:                flag vehicles in V(c) as busy
12:                insert scene departure event at time:
13:                t + t_d(c) + t_r(c) + t_sc(c) into ε
14:            else
15:                Q ← Q + c                     ▷ queue the call
16:        else if e = scene departure for call c then
17:            transport vehicle (if any) departs for hospital
18:            insert job completion event at time:
19:            t + t_h(c) + t_ho(c) into ε
20:            if |V(c)| > 1 then
21:                flag non-transport vehicles as available
22:                insert vehicle location update events into ε
23:                CHECKQUEUE(Q)                 ▷ answer queued calls
24:        else if e = job completion for call c then
25:            flag transport vehicle as available
26:            insert vehicle location update event into ε
27:            CHECKQUEUE(Q)                     ▷ answer queued calls
28:        else if e = location update for vehicle (v) then
29:            UPDATELOCATION(v)
30:            if not end of journey then
31:                insert vehicle location update event into ε
```

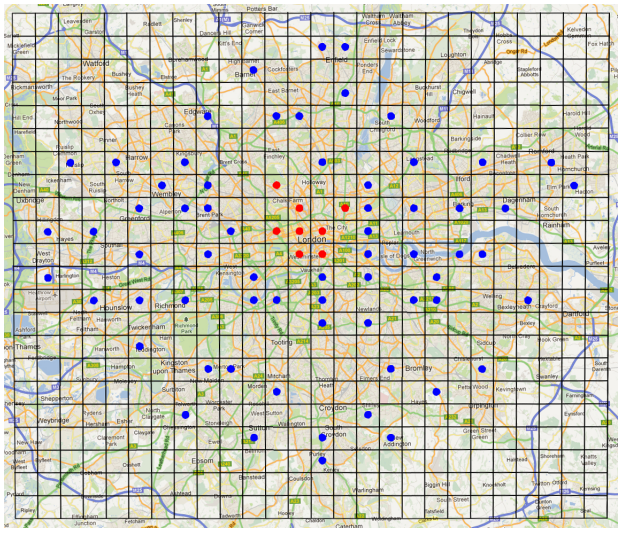

Figure 3: The LAS service area discretised into a 26 x 22 grid. Blue and red dots represent EMS vehicle base stations snapped to their model locations; red dots represent EMS vehicle base stations selected for optimization. Map data ©2013 Google.

The size of the area to be modeled posed a number of challenges. For example, a limit existed on the volume of geocoded data that could be requested from Google; a maximum of 2,500 requests per day under a free license. To produce a look-up table of distances for each node pairing in the 572 node grid would require almost 330,000 data requests, which was deemed impractical. Consequently, a number of possible solutions were considered:

- A smaller region could be modeled using a sub-set of the LAS call data. This would require a number of assumptions about which calls were answered from within the sub-region.
- The resolution of the simulation grid could be reduced. A 12 by 12 grid could have been achieved however such a coarse resolution would likely invalidate the model.
- An assumption could be made about travel routes in order to reduce the volume of GIS data required. Given that the simulation space is discretised into a grid, it could be assumed that a route between two nodes in neighbouring cells is well approximated by a straight path between

the two nodes. Thus, the overall travel route could be constructed from the straight paths between neighbouring nodes; see Figure 4.

From these three possible solutions, the route approximation was preferred as it avoided making potentially invalid assumptions about the LAS system. Further, the straight path assumption is not unreasonable when journeys are of a similar length to the distance between grid spaces; confirmed through analysis of the LAS data. This assumption reduced the volume of distance data required by a factor of 72; for each node, only the distance to each of its eight neighbours is required, reducing the number of data requests to 4,292. The possible solutions of modelling a smaller region and reducing the resolution of the simulation grid were not considered any further.

The accuracy of the route approximation was evaluated in order to quantify potential errors. Almost 2,400 typical journeys, of varying length, were extracted from the LAS data and their precise start and end locations were used with Google Maps to determine an accurate travel distance. These precise distances were compared to those approximated by the distance model. As anticipated, the model overestimated travel distances on average and errors increased with length of journey. A correction was built into the model to remove this average overestimation. The factor by which each distance is corrected is dependent on journey length; see Table 1.

Table 1: Average overestimation of distances by model.

Number of steps in route approximation	Average correction factor (modeled / real distance)
1	1.16
2	1.30
3	1.37
4	1.39
5	1.31
6	1.37
7	1.38
8	1.40
9	1.39
10	1.40
11	1.43

The time-dependent aspect of the travel time model is the speed at which

vehicles travel. For any given time of week it is assumed that all vehicles travel at a constant speed. A higher degree of accuracy could be achieved by considering the variable nature of travel times and the effect of different road types, however this level of detail was considered outside the scope of this study. If derived from real data, the benefit of using an average speed is that the effects and frequency of random perturbations, such as delays due to unusual traffic or road accidents, are incorporated into the model. Durations of journeys with known start and end locations were extracted from timestamps in the LAS data; these journeys corresponded to urgent transport from the scene to hospital. With these travel times, and approximated distances from the calibrated distance model, it was possible to derive vehicle speeds. For each day in the week these speeds were then aggregated and averaged based on time of day; see Figure 5. When derived from real data, the benefit of using an average speed is that the effects and frequency of random perturbations are incorporated into the model. Whilst using these average speeds does not explicitly model events such as breakdowns and heavy traffic, it does implicitly encompass the impact of these random events on travel times, and their relative probabilities.

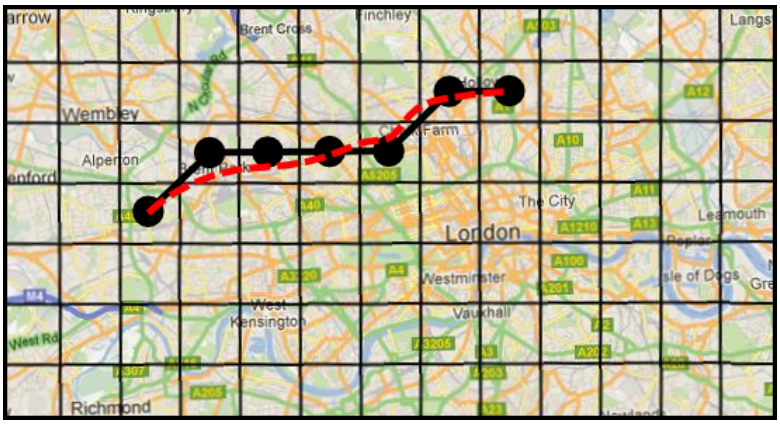

Figure 4: A fictional vehicle route is marked in red and its modeled approximation shown in black. Map data ©2013 Google.

2.2. Setting up an initial state

At the start of simulation all EMS vehicles are idle at their respective base stations, leading to unrealistically short response times until a steady

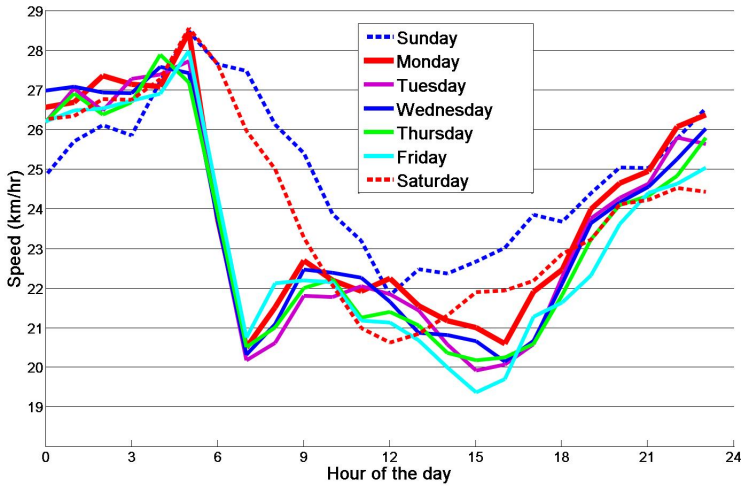

Figure 5: Mean speeds of ambulances transporting life-threatening calls to hospital.

operating state is reached. In order to prevent this affecting the evaluation of vehicle allocations, a steady operating state is created by processing a buffer of calls prior to the period of interest; no statistics are recorded for the calls in this buffer. By gradually incrementing the buffer period, it was found that processing a buffer of ninety minutes of calls or over, prior to the period of interest, produced simulation results that were independent of buffer duration. A buffer of ninety minutes was therefore used for all simulations.

2.3. Model verification, calibration, and validation

Verification and validation of the simulation were vital for model acceptance and confidence in results. The validation approach described in [43] was employed. By working closely with the LAS during model development it was ensured that the model had "face validity" (reasonable to the practitioner), "structural validity" (operated like the system), and "technical validity" (assumptions on the data were not far from reality). "Replicative validity" (predicting the past performance of the system) was achieved through model calibration. The model was used to simulate the existing distribution of EMS vehicles and the resulting survival efficiency compared to that calculated using the LAS data. The average speeds at which vehicles travelled were tuned so as to minimize the disparity between real and simulated response times, effectively replicating the current system. Figure 6 compares

actual survival efficiency to results from the calibrated simulation model. Following calibration, the root-mean-square (RMS) error between simulated and actual survival efficiency was 0.61%. The speeds derived from the LAS data were modified by 12% on average.

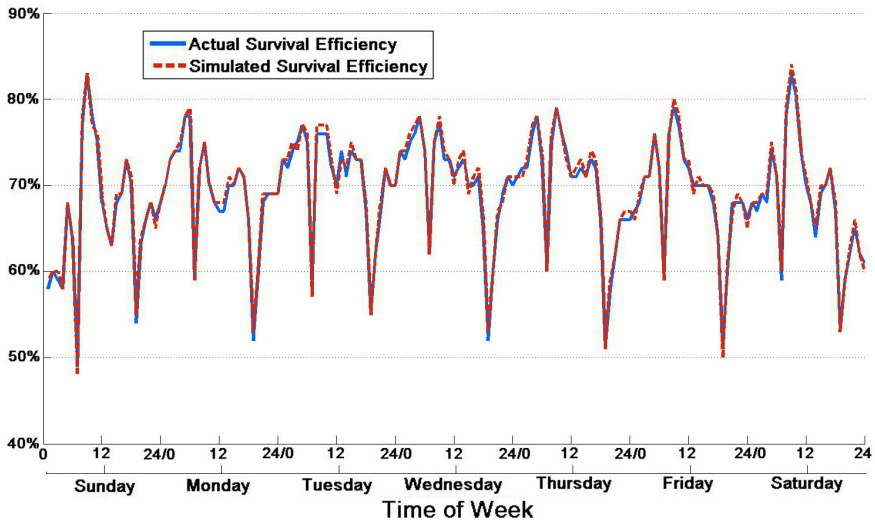

Figure 6: A comparison between simulated and actual survival efficiency following model calibration; RMS error is 0.61%.

3. Optimization

3.1. Optimization heuristic

An optimization heuristic was required for the optimization of EMS vehicle fleet allocation and base station location. The use of a GA was selected as they have been shown to robustly find good solutions for this type of facility location-allocation problem [14, 16, 47]. GAs employ a "survival of the fittest" strategy to iteratively improve the performance of a population of solutions [48].

A key aspect of any GA is the effective representation of a solution as an encoded string. Careful design of this representation guarantees feasible solutions, removing the need for continuous feasibility checks. In the optimization heuristic implemented in this study, the representation used was:

$$|\{g_1, g_2, ..., g_{2n-1}, g_{2n}\}\{g_{2n+1}\}\{g_{2n+2}, ..., g_{2n+1+m}\}|$$

where each gene (g) is a real number between 0 and 1. The first $2n$ genes encode the coordinates of n base stations with unfixed locations; n can be varied between 0 and N, where N is the total number of base stations. The next gene, g_{2n+1}, defines the ratio of ambulances to rapid response cars; this can be included in the optimization or fixed to represent existing resource levels. The final m genes encode the allocation of m vehicles to base stations; each gene is decoded to give an index between 1 and N.

An illustrative example of the chromosome representation now follows. Consider a simple problem with three emergency medical vehicles and two base stations; Station 0 has an unfixed location and Station 1 is fixed. The solution space has been divided into a 12 by 10 grid and the ratio of ambulances to rapid response cars is unfixed. One chromosome in the population is as follows:

$$|0.74, 0.80, 0.70, 0.25, 0.10, 0.74|$$

The first two genes encode a grid coordinate for Station 0 of (8,7):

$$Round(0.74 \times (12 - 1)) = 8$$
$$Round(0.80 \times (10 - 1)) = 7$$

The third gene specifies that of the three vehicles, two are ambulances:

$$Round(0.70 \times 3) = 2$$

Finally, the fourth, fifth, and sixth genes specify that Vehicles 0, 1, and 2 are assigned to Stations 0, 0, and 1 respectively:

$$Round(0.25 \times (2 - 1)) = 0$$
$$Round(0.1 \times (2 - 1)) = 0$$
$$Round(0.74 \times (2 - 1)) = 1$$

Algorithm 2 describes the GA procedure used in this study; adapted from that presented in [49]. The procedure begins with initialization of the population (P) using the complementary initialization method described in [50]. The first half of the population is randomly generated, the second half is then the "mirror image" of the first; i.e. each mirrored gene ($g_{mirrored}$) is one minus the original gene ($1 - g_{original}$). This method goes some way to ensuring reasonable diversity in the initial population. For example, if one of

the chromosomes in the first half of the population was randomly generated as:

$$|0.47, 0.08, 0.72, 0.51, 0.10, 0.67|$$

Then the "mirror" of this in the second half of the population would be:

$$|1 - 0.47, 1 - 0.08, 1 - 0.72, 1 - 0.51, 1 - 0.10, 1 - 0.67|$$

$$|0.53, 0.92, 0.28, 0.49, 0.90, 0.33|$$

At each generation, the fitness of every individual is evaluated using the simulation model and the objective function detailed in Section 3.2. Evaluations are independent and can therefore be executed in parallel. The ComputeFitness function utilises multiple threads, taking advantage of the multiple cores in modern computers. The studies presented in this paper were conducted on a quad-core computer using eight threads.

Once the fitness of each individual is known, the next generation can be populated. Individuals are selected and either copied to the next generation or crossed with another individual to produce offspring; the probability of which is set by the crossover rate (r_x). A number of different selection procedures were trialled, including cost and rank weighted roulette-wheel selection. The procedure which produced high performing individuals in the shortest time, without premature convergence, was the tournament selection procedure described in [50]; a small subset of the population is randomly picked and the fittest individual within this subset selected. Although likely, it is not guaranteed that the best individuals contribute to the next generation. Ensuring the preservation of a number of elite individuals leads to improved performance [51]. The single best individual was preserved in this implementation.

Crossover operations combine the genetic information of parents with the aim of producing fitter offspring. The crossover operator used is an adaptation of the heuristic crossover method detailed in [50], where two parents produce two offspring. There are two stages to the procedure:

1. Genes are copied from parent to child and then randomly swapped between offspring. This is known as uniform crossover.

Algorithm 2 Genetic algorithm

```
1: P ← INITIALIZE(popSize)                          ▷ initial population (P)
2: while (1) do
3:     COMPUTEFITNESS(P)
4:     P_next ← ∅                                   ▷ initialize next generation (P_next)
5:     copy elite individuals to the next generation
6:     P_next ← P_next + ELITE(P)
7:     while |P_next| < popSize do
8:         p ← SELECTION(P)                         ▷ select individual (p)
9:         rand ← RNG()                             ▷ rand ∈ (0, 1)
10:        if rand < r_x then                       ▷ crossover rate: r_x ∈ (0, 1)
11:            p_mate ← SELECTION(P)                ▷ select mate
12:            produce offspring and add to next generation
13:            P_offspring ← CROSSOVER(p, p_mate)
14:            P_next ← P_next + P_offspring
15:        else
16:            P_next ← P_next + p
17:    MUTATE(P_next)
18:    P ← P_next
19:    if HASCONVERGED(P) then
20:        return
```

2. Half of the genes in the two children (g_{child}) are randomly selected and replaced with a blend of the two parent genes (g_{parent}):

$$g_{child1} = g_{parent1} + \beta(g_{parent2} - g_{parent1}) \tag{1}$$

$$g_{child2} = g_{parent2} + \beta(g_{parent1} - g_{parent2}) \tag{2}$$

where β is a randomly generated number in the range 0 to 1.1 (inclusive). This crossover operation is effectively a linear interpolation between parent genes, where the point of interpolation is randomly selected by β. By allowing values of β greater than one we allow slight extrapolation of genetic information, helping to maintain genetic diversity. However, this extrapolation can produce infeasible genes, child gene values are therefore fixed between 0 and 1.

For example, consider two parent chromosomes copied to two offspring:

$$g_{child1} = g_{parent1} = |0.13, 0.67, 0.84, 0.33, 0.82, 0.91|$$
$$g_{child2} = g_{parent2} = |0.76, 0.97, 0.52, 0.01, 0.18, 0.39|$$

Genes 0, 3, and 4 are randomly selected and swapped between offspring (uniform crossover):

$$g_{child1} = |\mathbf{0.76}, 0.67, 0.84, \mathbf{0.01}, \mathbf{0.18}, 0.91|$$
$$g_{child2} = |\mathbf{0.13}, 0.97, 0.52, \mathbf{0.33}, \mathbf{0.82}, 0.39|$$

A β value is randomly generated, genes 2, 3, and 5 randomly selected, and equations (1) and (2) applied:

$$\beta = 1.05$$

$$g_{child1} = |0.76, 0.67, \mathbf{0.50}, \mathbf{0.00}, 0.18, \mathbf{0.15}|$$
$$g_{child2} = |0.13, 0.97, \mathbf{0.86}, \mathbf{0.35}, 0.82, \mathbf{0.94}|$$

Following crossover operations the population is subject to mutation. From the entire population, a number of randomly selected genes are replaced with randomly generated real numbers between 0 and 1. The number of mutated genes is determined by the mutation rate (r_m). The fittest chromosome is immune to mutation.

It is technically possible that following crossover and mutation operations, multiple chromosomes could translate to the same solution. Given that each problem addressed in this paper typically has in the order of 4×10^{40} possible solutions, the probability of chromosome redundancy was considered negligible.

Each iteration the GA is checked for convergence. A number of GA stopping conditions were implemented but the most effective way to ensure convergence was to run the GA for a predetermined number of generations whilst monitoring progress using real-time graphs. Each experiment was replicated 15 times with the genetic algorithm running for an average of 180 generations. Solutions remained static for a minimum of 40 generations before termination of the genetic algorithm; see Figure 7 which shows typical GA convergence.

Although a GA is a generally applicable meta-heuristic, the crossover rate, mutation rate, and population size parameters need to be tuned to suit each application. Preliminary experiments were undertaken to determine the optimum settings of these parameters. It was found that a crossover rate of 0.85 and mutation rate of 0.04 quickly produced high performing individuals, without premature convergence. The crossover rate is slightly higher than the norm but not unusually so [51]. The high mutation rate was expected, it encourages adequate exploration of what is an exponentially large and flat solution space [42]; a mutation rate of 0.03 was used for a similar problem in [14]. A population size of twenty five was selected as a compromise between performance and computation time.

3.2. Objective function

The optimization presented in this study aims to maximize the number of EMS patients that survive cardiac arrests and other life-threatening incidents. Patient survival was first used as a performance metric by Erkut et al. [37]. They used a survival function derived from a medical study that modeled the health of patients after suffering cardiac arrests [52]. Cardiac arrest incidents were used because response time is crucial to survival and the relationship has been studied extensively. Knight et al. incorporated a similar cardiac arrest survival function with step functions representing response time targets for other incident types [41]. The heterogeneous survival function was a sum of the individual survival functions weighted by their relative priorities; effectively a multi-objective optimization.

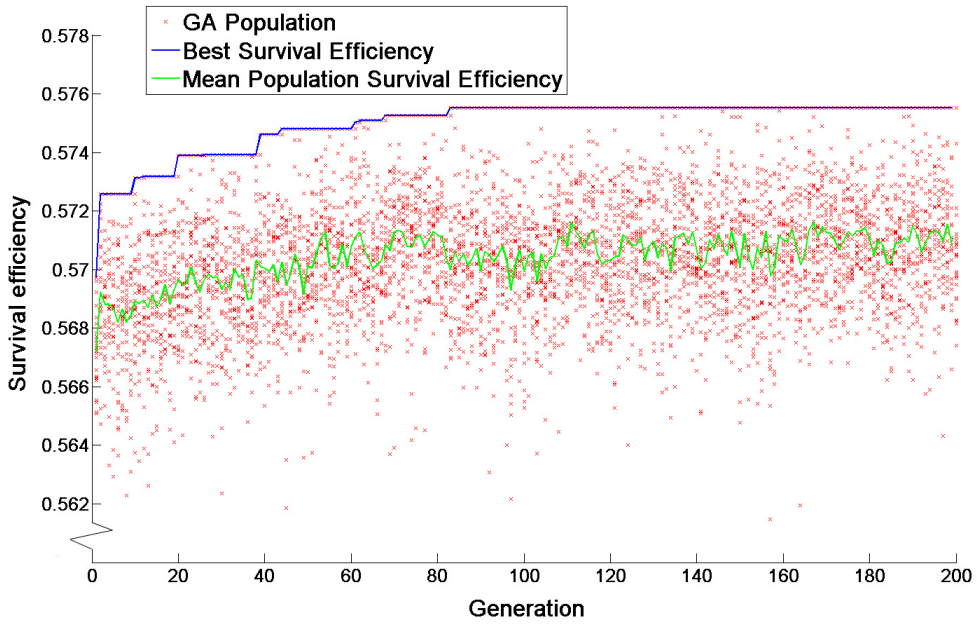

Figure 7: Typical GA convergence during experimentation.

The heterogeneous survival function used in this study is one of a number of functions that have been shown to approximate survival probability as a function of time for cardiac patients [37, 41]. This particular survival function was derived using cardiac survival data from the US and logistic regression. The survival probability (s) of two patient classes is modeled; cardiac arrests (s_c):

$$s_c = \frac{1}{1 + \exp(-0.26 + 0.139 \times T_r)} \tag{3}$$

and all other life-threatening (category A) calls (s_a):

$$s_a = \begin{cases} 1, & \text{for } T_r \leq 8 \\ 0, & \text{for } T_r > 8, \end{cases} \tag{4}$$

where T_r is the response time. The objective is maximization of the heterogeneous survival efficiency (η_s):

$$\eta_s = \frac{2\sum_{i=1}^{\gamma} s_{c_i} + \sum_{j=1}^{\delta} s_{a_j}}{2\gamma + \delta}, \tag{5}$$

where γ is the number of cardiac arrest calls and δ the number of category A calls. Cardiac arrests are weighted as twice as important as other category A calls. Non-urgent calls (category C) are excluded from the objective function as they are non life-threatening and not included in LAS performance targets.

4. Application and evaluation

4.1. Problem definition

The ambulance fleet allocation and base station location problem is defined as follows. We are given a set of base stations $B = \{B_n\}, 1 \leq n \leq N$ to be populated with a set of EMS vehicles $V = \{V_m\}, 1 \leq m \leq M$. During simulation, vehicles respond to a set of emergency calls $C = \{C_k\}, 1 \leq k \leq K$. Each call must be processed by a subset of vehicles $V_k \subseteq V$, with each call having an independent time, location, duration, and vehicle requirement. Vehicles can process at most one call at a time.

The objective of the problem is to maximize patient survival by finding the optimal allocation of EMS vehicles to base stations and base station locations. Stations can be located at any point within the geographic limits of the model. A derivative of this problem, also addressed by this paper, fixes the location of base stations to emulate the existing EMS system.

4.1.1. Case study definition

The LAS follow a resource plan that, for each hour of the week, defines the number of ambulances and rapid response cars allocated to each of their seventy base stations. It was seen as logical to approach optimization in the same way and optimize the planned vehicle allocations for a number of these hour-long periods. Figure 8 shows the volume of calls received by the LAS over a one year period, aggregated by time of week. Probable cardiac arrest calls have been separated from other category A calls for the purposes of modelling patient health. Figure 9 outlines the current performance of the LAS as measured by response time targets and survival efficiency.

Due to the exponential size of the problem, the resource plan has only been optimized for the eight stations highlighted in red in Figure 3, selected due to their proximity to an area of high call volume. This is still far from a trivial problem since during periods of high call volume there are up to forty five vehicles allocated to these eight stations, giving 4.36×10^{40} possible combinations. Evaluating each allocation takes fifteen seconds, thus a full enumeration would take 2.07×10^{34} years. The approach detailed in this paper

is capable of optimizing the entire LAS system, the only limit on problem size is that of solution time.

Although only eight stations are being optimized, the entire LAS system was simulated. This meant that vehicles stationed at the central eight stations were free to respond to calls anywhere in the city, and vehicles at stations not being optimized were able to respond to calls within the central zone. This approach captured the cooperative aspects between optimized stations and the remaining stations in the model.

The results of key studies investigating the resourcing of these eight base stations are presented and discussed in the remainder of this section.

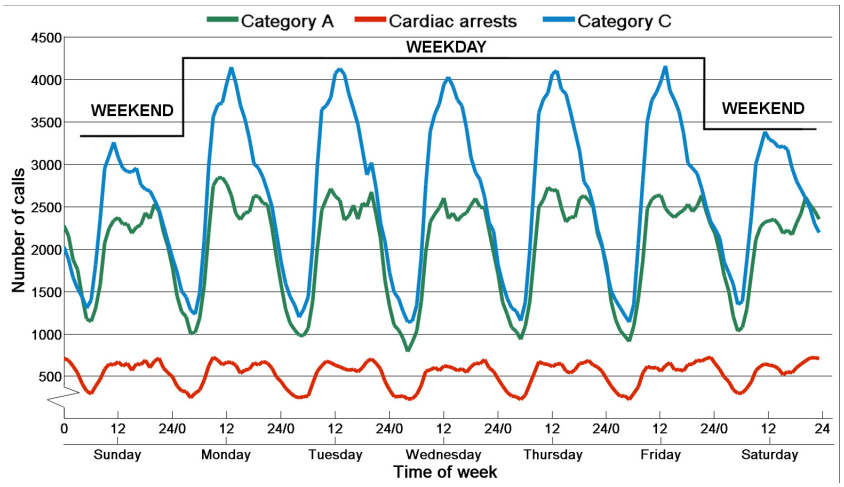

Figure 8: Call volumes for a one year period aggregated by time of week. Calls are categorised as either cardiac arrests, life-threatening (category A), or non-urgent (category C).

4.2. Results and discussion

4.2.1. Optimizing resource plans

This application investigated the number of lives that could potentially be saved by redistributing the existing resources. Periods of the week are referred to with the notation DXHY, where X is the day (0 to 6), and Y the hour (0 to 23) of interest; e.g. D0H6 refers to 6am-7am on Sunday, while D5H18 is 6pm-7pm on Friday.

Three one-hour periods, representing the range in operating conditions and current performance, were selected for optimization:

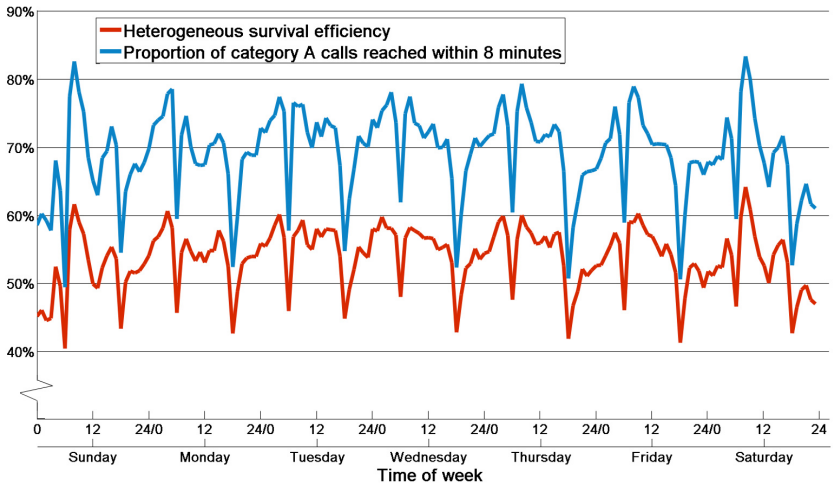

Figure 9: Current LAS performance as derived from real call data.

- D0H6 - the period with the lowest survival efficiency according to LAS data; see real η_s in Table 2.
- D2H11 - the period with the highest volume of life-threatening calls.
- D5H18 - a period with high call volume and low survival efficiency.

Multiple GA trials were conducted with all base station locations fixed and the ratio of ambulances to cars fixed to match current resourcing. For fifteen replications, computation took an average of 25.2 minutes per experiment. Table 2 presents the current and optimized resource plans and the effect on survival efficiency (η_s), number of cardiac survivors (S_c), and number of category A calls reached within eight minutes (S_{a8}). As anticipated, the largest increase in survival efficiency is observed in the period with the lowest real-life performance (D0H6). However, due to a relatively low call volume this does not translate to the greatest increase in survivors; this is observed in the period with low performance and high call volume (D5H18). Optimizing the resource plan for the D5H18 period results in three additional cardiac arrest survivors and an additional 36 category A calls reached within eight minutes.

There are no obvious trends in the reallocation of resources, indicating that the resource plan must be tuned to match the unique demand distribution in each period of the week. This emphasises the importance of incor-

Table 2: Current and optimized allocations of ambulances (A) and rapid response cars (C) for three time periods.

	D0H6		D2H11		D5H18	
Station	Current	Optimum	Current	Optimum	Current	Optimum
1	1A, 0C	2A, 2C	2A, 0C	1A, 2C	2A, 0C	4A, 0C
2	5A, 1C	5A, 1C	5A, 1C	2A, 4C	5A, 1C	3A, 6C
3	3A, 2C	0A, 1C	6A, 3C	6A, 2C	6A, 3C	2A, 1C
4	3A, 0C	2A, 2C	3A, 0C	4A, 1C	3A, 0C	5A, 2C
5	3A, 2C	3A, 0C	5A, 3C	4A, 1C	5A, 3C	3A, 0C
6	2A, 1C	4A, 1C	2A, 3C	1A, 2C	2A, 2C	5A, 3C
7	1A, 0C	1A, 1C	2A, 0C	5A, 1C	1A, 0C	3A, 1C
8	2A, 2C	3A, 0C	5A, 5C	7A, 2C	5A, 5C	4A, 1C
Real η_s	40.40%	–	54.92%	–	41.26%	–
η_s	40.21%	41.67%	53.92%	54.66%	39.96%	41.32%
$\Delta\eta_s$	–	1.46%	–	0.74%	–	1.36%
S_c	80.1	82.3	223.7	227.2	174.1	177.4
ΔS_c	–	2.2	–	3.5	–	3.3
S_{a8}	420	437	1369	1388	916	952
ΔS_{a8}	–	17	–	19	–	36

porating time dependent demand variation and travel times into any EMS model. The redistribution of resources becomes clearer when the workload of each station is examined. Table 3 presents base station utilisation under the current and the optimized resource plans, calculated as the average utilisation of the vehicles at each station. From these tables it can be seen that performance is being optimized by reallocating vehicles from low to high workload stations, i.e. balancing load. Station 3 has the lowest utilisation in the D0H6 period in which optimization reduces the resources by three ambulances and one car thus increasing utilisation from 26.98% to 30.87%.

In total, the lives of nine additional cardiac arrest patients could be saved per year in the periods examined (sum of ΔS_c in Table 2). Survival would also significantly improve for other life-threatening cases, each year an additional 72 category A calls would be reached within eight minutes (sum of ΔS_{a8} in Table 2). If these gains are representative, it could be expected that optimization of the whole resource plan would result in significantly more

Table 3: Base station utilisation under the current and optimized resource plans.

	Base station							
Period	1	2	3	4	5	6	7	8
	Utilisation under current resource plan (%)							
D0H6	50.95	47.71	26.98	39.98	41.14	60.18	65.68	46.37
D2H11	92.54	86.62	67.37	92.43	73.51	63.54	94.86	58.19
D5H18	87.78	83.58	65.58	84.00	74.18	65.75	85.91	50.63
	Utilisation under optimized resource plan (%)							
D0H6	34.06	44.80	30.87	34.26	56.88	54.62	46.03	57.10
D2H11	45.53	53.68	75.56	76.38	82.92	66.60	87.45	79.15
D5H18	86.75	58.45	76.29	65.29	90.29	68.44	76.25	74.91

lives saved per year. Further studies would be required to quantitatively assess this expectation.

Table 2 presents the best resource plans found after fifteen GA trials. Due to the stochastic nature of a GA, each trial identified a different resource plan with varying performance; this variation in GA results is presented in Table 4.

Table 4: Variation in results from fifteen GA trials.

	η_s (%)		
Period	Best	Mean	σ
D0H6	41.67	41.56	0.118
D2H11	54.66	54.61	0.035
D5H18	41.32	41.30	0.015

4.2.2. Weekday resource strategy

The LAS currently use an identical resource plan for weekdays Monday through Thursday as demand is expected to be the same. However, the demand actually observed for these days has significant differences; for example the total number of category A calls on Mondays is 8.8% higher than on Wednesdays. Figure 10 shows the current LAS performance for weekdays Monday through Thursday, as derived from the LAS data. Differences in

demand are causing survival efficiency to vary across weekdays. The difference in survival efficiency between best and worst performing days reaches a maximum of 4.9% during the midday period.

A study was conducted into whether the current strategy, referred to as plan 1, best serves weekday demand. For fifteen replications, computation took an average of 50 minutes per experiment. The periods D2H6 and D3H6 were chosen for comparison as the difference in survival efficiency (2.1%) was approximately the average of that observed in Figure 10. Resourcing was optimized for both of these time periods resulting in two potential resource plans, plan 2 optimized for D2H6 and plan 3 optimized for D3H6; see Table 5. The performance of the LAS system was evaluated using both of these plans; see Table 6. In both periods, both optimized plans outperform the current plan in terms of survival efficiency. However, of most note is that the optimal resource plan for one period is sub-optimal for the other period. That is, for the period D2H6, plan 2 outperforms plan 3 by 0.55% (a survival efficiency of 45.68% compared to 45.13%). Conversely, for the period D3H6, plan 3 outperforms plan 2 by 0.50% (a survival efficiency of 44.72% compared to 45.22%). This indicates that there is no single resource plan that is optimum for all weekdays and that the current resource strategy does not best serve weekday demand. It is instead recommended that the LAS specialise the resource plan for each weekday, accounting for the differences in demand.

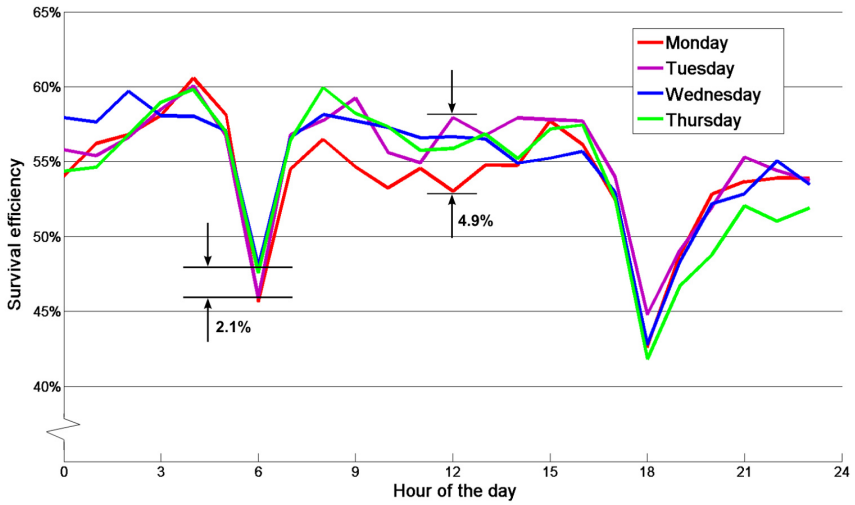

Figure 10: LAS performance for Mondays to Thursdays under the current resource plan.

Table 5: Current and optimized allocations of ambulances (A) and rapid response cars (C) for two time periods.

Station	D2H6 & D3H6 Current (Plan 1)	D2H6 Optimum (Plan 2)	D3H6 Optimum (Plan 3)
1	1A, 0C	2A, 0C	1A, 1C
2	3A, 1C	2A, 1C	3A, 2C
3	3A, 2C	2A, 1C	2A, 0C
4	2A, 0C	2A, 1C	3A, 1C
5	2A, 1C	2A, 1C	2A, 0C
6	2A, 1C	1A, 2C	1A, 1C
7	1A, 0C	2A, 0C	2A, 0C
8	2A, 0C	3A, 0C	2A, 1C

Table 6: Performance under different resource plans.

	D2H6			D3H6		
	Plan 1	Plan 2	Plan 3	Plan 1	Plan 2	Plan 3
η_s	44.36%	45.68%	45.13%	44.38%	44.72%	45.22%
$\Delta\eta_s$	–	1.32%	0.77%	–	0.34%	0.84%
S_c	74.9	74.2	74.5	84.3	84.5	84.8
ΔS_c	–	-0.7	-0.4	–	0.2	0.5
S_{a8}	442	461	453	415	419	425
ΔS_{a8}	–	19	11	–	4	10

4.2.3. System performance with varying resource levels

Discussions with the LAS revealed that it is common for resource gaps to occur, resulting in fewer than planned vehicles available for a shift. These gaps can occur for numerous reasons, including staff illness, vehicle breakdowns, or crews being questioned about incidents by police. In these situations it is vital to understand how best to redistribute the remaining resources and what effect this may have on system performance. It is also valuable, from a planning perspective, to understand what effect increasing resources may have on performance; for example net gains may be achievable by transferring resources between time periods.

The effects of varying vehicle numbers were investigated for the three periods considered in Section 4.2.1, namely D0H6, D2H11 and D5H18. Multiple

GA trials were conducted with all base station locations fixed. It is unknown which vehicles would be unavailable or introduced to the system therefore the ratio of ambulances to cars was included in the optimization. For fifteen replications, computation took an average of 49.8 minutes per experiment. Figure 11 shows the system performance as the number of vehicles is varied. There are obvious trends in both vehicle utilisation and survival efficiency, similar to those presented by Knight et al. [41], although utilisation appears linearly related to vehicle numbers across the ranges investigated.

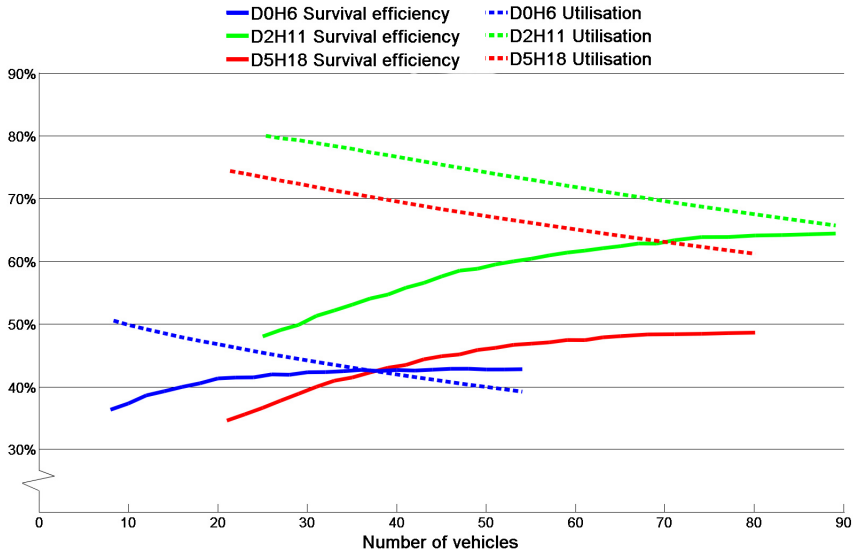

Figure 11: System performance with varying vehicle numbers. Utilisation is the average vehicle utilisation across all stations.

Increasing resources causes survival efficiency to increase, although these returns diminish. For each period, a critical resource level is reached above which there is no further increase in performance. This is the maximum performance that can be achieved with stations located in their current positions. Table 7 presents these performance ceilings and critical resource levels. With current resource levels and an optimized resource plan, the LAS could be operating at 98%, 90%, and 92% of maximum survival efficiency for the periods D0H6, D2H11, and D5H18 respectively.

Performance is close to the maximum achievable for D0H6, causing survival efficiency to be insensitive to small changes in vehicle numbers. For example, removing four vehicles from the D0H6 period results in a decrease

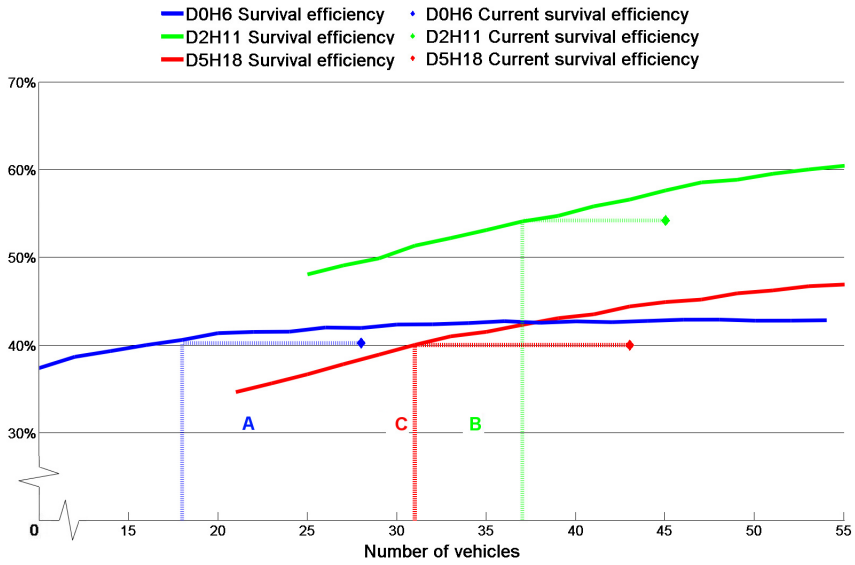

Figure 12: System performance with varying vehicle numbers. The potential to reduce the number of operating vehicles whilst retaining current performance levels is highlighted with labels A, B, and C.

in survival efficiency of 0.42%. If these vehicles were transferred for use in the D2H11 and D5H18 periods (two vehicles each), survival efficiency would increase by 0.91% and 0.49% respectively. Implementing this transfer would change the numbers of cardiac arrest survivors by -1.8, 2.8, and 2.0 for the periods D0H6, D2H11, and D5H18 respectively; a net increase of three cardiac arrest survivors per year. Another notable finding is that, adhering to the optimized resource plans, the current level of performance could be met with significantly fewer vehicles. Vehicle numbers could be reduced by 10, 8, and 12 for the D0H6, D2H11, and D5H18 periods respectively; see labels A, B, and C on Figure 12.

4.2.4. Introducing a new base station

A study was carried out to determine the optimum location for a new base station, i.e. that which would result in the greatest net increase in survivors. Rather than considering all time periods, the new station location was optimized for a number of key periods, namely 6pm to 7pm for each day of the week based on the study presented in Section 4.2.1 which revealed that the greatest increase in survivors is achieved for periods with low survival

Table 7: Performance under optimized resource plans with current and critical resource levels.

	Resources for η_{sMax}		Current resources		
Period	η_{sMax} (%)	Vehicles	η_s (%)	η_s/η_{sMax}	Vehicles
D0H6	42.9	36	41.92	0.98	28
D2H11	64.3	86	57.58	0.90	45
D5H18	48.3	68	44.38	0.92	43

efficiency and high call volume.

Multiple GA trials were conducted for each of the chosen periods. Within the simulated region, the new station was free to be located anywhere whereas all other base station locations were fixed. Initial GA trials found multiple local optima with approximately equal performance, indicating that the solution space was relatively flat. To mitigate against this, the mutation rate was increased to 0.05 and the GA run-time tripled. For fifteen replications, computation took an average of 2.9 hours per experiment.

The optimum station locations for each day of the week are shown in Figure 13 in which it can be seen that there are two optimal locations, one corresponding to demand on weekdays (days 1 to 5) and the other at the weekend (days 0 and 6). The presence of these two locations is supported by the two distinct call volume patterns, weekday and weekend, observed in Figure 8. The actual distribution of life-threatening calls in these periods provides more insight as shown in Figure 14 which indicates the majority of life-threatening calls are located around the city centre on weekdays whereas these calls are more dispersed at the weekend. Given the relative proximity of the weekday and weekend optimum locations, i.e. 10 kilometres, an appropriate location for a new station could either be that which most improved performance or an average of the two locations weighted by the increase in survivors that could be achieved. To inform such a decision, performance was evaluated using the new station and compared to performance under the current resource plan; see Table 8. The new station results in an increase in survival efficiency of 3.55% during the week compared to 2.99% at the weekend. With higher call volumes during the week, this causes an increase of 31 cardiac arrest survivors per year compared to 12 for the weekend location; see delta ΔS_c in Table 8.

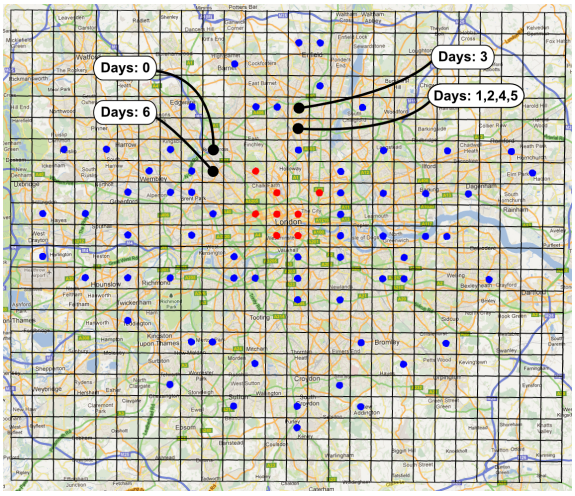

Figure 13: Black dots mark the optimum locations for a new base station, as optimized for 6pm to 7pm of each day of the week. Locations are labelled with the days with which they were found to be optimum. Map data ©2013 Google.

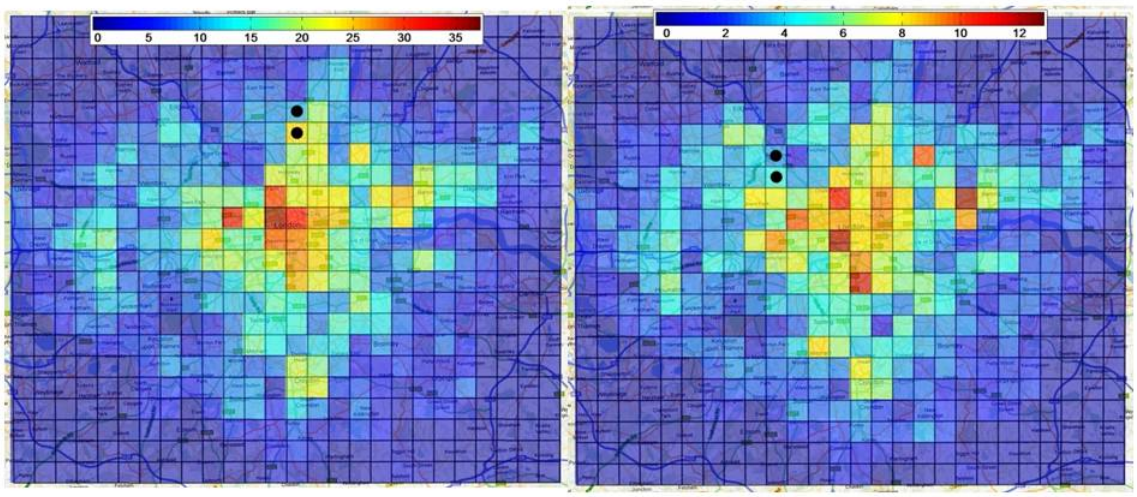

Figure 14: Distribution of life-threatening calls from 6pm to 7pm on weekdays (left) and weekends (right). The colour gradient depicts the annual volume of calls received within these periods. The optimum locations for a new base station are marked with black dots. Map data ©2013 Google.

Table 8: Performance with and without a new station.

	Weekday		Weekend	
	Current	New station	Current	New station
η_s	42.41%	45.96%	42.61%	45.60%
$\Delta\eta_s$	–	3.55%	–	2.99%
S_c	871.4	903.1	367.7	379.7
ΔS_c	–	31.7	–	12.0
S_{a8}	5022	5526	1830	1986
ΔS_{a8}	–	504	–	156

4.2.5. Removing a base station

A final study investigated the effects of removing a station with the optimum one to remove being that which resulted in the smallest net decrease in survivors. To reduce computation time, and as a proof of concept only, station removal was optimized for a single period, D1H8, rather than across all time periods. This study required the introduction of a new gene to the chromosome representation, which defined a station to be removed with any vehicles assigned to this station excluded from the simulation. GA trials were conducted with all base station locations fixed, and the ratio of ambulances to rapid response cars fixed to represent existing resource levels. For fifteen replications, computation took an average of 1.7 hours per experiment.

Station 1 was identified as the optimum station for removal, as indicated in Figure 15, with the neighbouring stations (2,3,4,6,7) seeing a net increase of one ambulance and one rapid response car to deal with the additional load; see Table 9. Further, removing station 1 and optimizing the resource plan for D1H8 resulted in a 1.23% increase in survival efficiency, only 0.04% lower than with all stations included; see $\Delta\eta_s$ in Table 9. Also, when compared to current performance, the removal of station 1 corresponds with an annual increase of 2.8 cardiac survivors in the period examined; see ΔS_c in Table 9.

5. Conclusions

This paper presents the optimization of EMS vehicle fleet allocation and base station location through the use of a GA with an integrated EMS simulation model. Novel simulation features and modelling approaches have enabled a level of realism not seen in other EMS models. In a departure

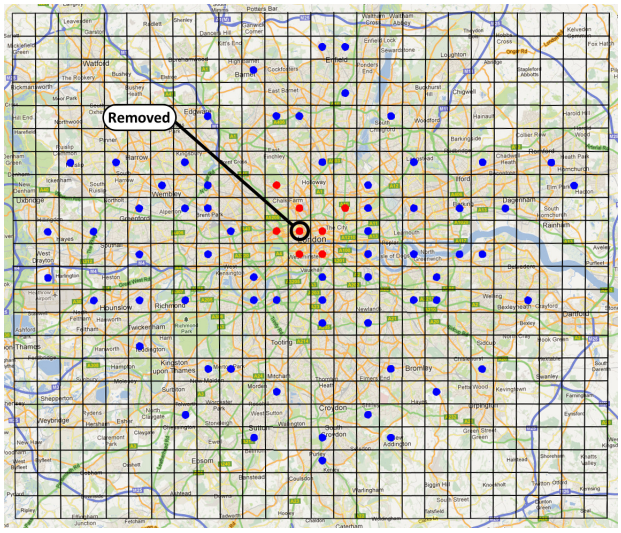

Figure 15: Removing the highlighted station was found to minimize the negative impact on performance for the period of 8am to 9am on a Monday. Map data ©2013 Google.

from existing academic literature, this model has been applied to a complex, real-life system through the use of LAS call data and geocoded locations. The aim has been to demonstrate how such a model can be used to make useful recommendations to practitioners.

A number of studies have demonstrated that significant improvements in patient survival can be achieved with optimized resource plans. In the periods examined, the lives of nine additional cardiac arrest patients could be saved per year without introducing additional resources. Indeed, the LAS indicated real value in the direct link between potential resource plans and patient survival estimates.

The LAS weekday resourcing strategy was challenged and the benefits of moving towards a customised plan assessed. Feedback from the LAS identified that the main reasons for non-optimal cover are shift-pattern and rostering requirements. Knowledge of the optimum cover is still useful however and is used to decide when and where to allocate unplanned resources. The impact of introducing a new station to the LAS system was also investigated and recommendations made on the optimum location. The LAS noted that the same techniques would be useful in identifying and evaluating potential standby points.

Future LAS strategy could see small base stations consolidated into larger

Table 9: Current and optimized allocations of ambulances and rapid response cars, including the optimum resource plan with station one removed. The performance under these plans is also presented.

Station	Current	Optimum (existing stations)	Optimum (station 1 removed)
1	2A, 2C	2A, 2C	0A, 0C
2	4A, 4C	2A, 2C	2A, 2C
3	5A, 1C	5A, 1C	4A, 1C
4	3A, 1C	5A, 3C	4A, 2C
5	4A, 3C	3A, 1C	6A, 2C
6	6A, 2C	5A, 3C	6A, 3C
7	0A, 0C	5A, 0C	7A, 2C
8	6A, 2C	3A, 3C	1A, 3C
η_s	56.45%	57.72%	57.68%
$\Delta\eta_s$	–	1.27%	1.23%
S_c	212.8	215.7	215.6
ΔS_c	–	2.9	2.8
S_{a8}	1163	1193	1192
ΔS_{a8}	–	30	29

'super stations'. An investigation was made into the impact of removing a station from the system and a recommendation made on the optimum station to remove. An investigative tool such as this could help inform decisions on future strategy.

Further work could consider rostering and shift-pattern requirements and investigate the effects of time dependant vehicle routing and stochastic variation in travel times. The effects of different dispatch methods could also be investigated. One aspect of the LAS system not considered here is how the optimum resource plan might change throughout the year. Accounting for seasonal change in demand may reveal more about how best to achieve optimum performance.

Acknowledgements

The authors express their sincere thanks to the London Ambulance Service for the provision of data.

References

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

[2] V. Marianov, C. ReVelle, Siting emergency services, In: Drezner Z (ed). Facility Location: A Survey of Applications and Methods, SPringer-Verlag, New York (1995) 199–223.

[3] C. ReVelle, Review, extension and prediction in emergency service siting models, European Journal of Operational Research 40 (1989) 58 – 69.

[4] J. B. Goldberg, Operations research models for the deployment of emergency services vehicles, EMS Management Journal 1 (2004) 20–39.

[5] X. Li, Z. Zhao, X. Zhu, T. Wyatt, Covering models and optimization techniques for emergency response facility location and planning: a review, Mathematical Methods of Operations Research 74 (2011) 281–310.

[6] M. Gendreau, G. Laporte, F. Semet, The maximal expected coverage relocation problem for emergency vehicles, The Journal of the Operational Research Society 57 (2006) pp. 22–28.

[7] C. ReVelle, H. Eiselt, Location analysis: A synthesis and survey, European Journal of Operational Research 165 (2005) 1 – 19.

[8] R. Z. Farahani, N. Asgari, N. Heidari, M. Hosseininia, M. Goh, Covering problems in facility location: A review, Computers & Industrial Engineering 62 (2012) 368 – 407.

[9] C. Toregas, R. Swain, C. ReVelle, L. Bergman, The location of emergency service facilities, Operations Research 19 (1971) 1363–1373.

[10] M. Gendreau, G. Laporte, F. Semet, Solving an ambulance location model by tabu search, Location Science 5 (1997) 75 – 88.

[11] G. Bianchi, R. L. Church, A hybrid fleet model for emergency medical service system design, Social Science & Medicine 26 (1988) 163 – 171.

[12] V. Marianov, C. ReVelle, The queueing maximal availability location problem: A model for the siting of emergency vehicles, European Journal of Operational Research 93 (1996) 110 – 120.

[13] R. D. Galvao, R. Morabito, Emergency service systems: The use of the hypercube queueing model in the solution of probabilistic location problems, International Transactions in Operational Research 15 (2008) 525–549.

[14] H. Aytug, C. Saydam, Solving large-scale maximum expected covering location problems by genetic algorithms: A comparative study, European Journal of Operational Research 141 (2002) 480–494.

[15] M. S. Daskin, A maximum expected covering location model: Formulation, properties and heuristic solution, Transportation Science 17 (1983) 48–70.

[16] C. Saydam, H. Aytug, Accurate estimation of expected coverage: revisited, Socio-Economic Planning Sciences 37 (2003) 69–80.

[17] J. P. Jarvis, Approximating the equilibrium behavior of multi-server loss systems, Management Science 31 (1985) 235–239.

[18] R. C. Larson, Approximating the performance of urban emergency service systems, Operations Research 23 (1975) 845–868.

[19] C. Saydam, J. Repede, T. Burwell, Accurate estimation of expected coverage: A comparative study, Socio-Economic Planning Sciences 28 (1994) 113 – 120.

[20] F. Y. Chiyoshi, R. D. Galvao, R. Morabito, A note on solutions to the maximal expected covering location problem, Computers & Operations Research 30 (2003) 87 – 96.

[21] A. P. Iannoni, R. Morabito, C. Saydam, A hypercube queueing model embedded into a genetic algorithm for ambulance deployment on highways, Annals of Operations Research 157 (2008) 207–224.

[22] A. P. Iannoni, R. Morabito, C. Saydam, An optimization approach for ambulance location and the districting of the response segments on highways, European Journal of Operational Research 195 (2009) 528–542.

[23] N. Geroliminis, K. Kepaptsoglou, M. G. Karlaftis, A hybrid hypercube - genetic algorithm approach for deploying many emergency response

mobile units in an urban network, European Journal of Operational Research 210 (2011) 287 – 300.

[24] H. Toro-Diaz, M. E. Mayorga, S. Chanta, L. A. McLay, Joint location and dispatching decisions for emergency medical services, Computers & Industrial Engineering 64 (2013) 917 – 928.

[25] R. D. Galvao, F. Y. Chiyoshi, R. Morabito, Towards unified formulations and extensions of two classical probabilistic location models, Computers & Operations Research 32 (2005) 15 – 33.

[26] M. Gendreau, G. Laporte, F. Semet, A dynamic model and parallel tabu search heuristic for real-time ambulance relocation, Parallel Computing 27 (2001) 1641 – 1653. Applications of parallel computing in transportation.

[27] H. K. Rajagopalan, C. Saydam, J. Xiao, A multiperiod set covering location model for dynamic redeployment of ambulances, Computers & Operations Research 35 (2008) 814 – 826. Part Special Issue: New Trends in Locational Analysis.

[28] H. Toro-Diaz, M. E. Mayorga, L. A. McLay, H. K. Rajagopalan, C. Saydam, Reducing disparities in large-scale emergency medical service systems, Journal of the Operational Research Society (2014) 1–13.

[29] S. Sasaki, A. Comber, H. Suzuki, C. Brunsdon, Using genetic algorithms to optimise current and future health planning - the example of ambulance locations, International Journal of Health Geographics 9 (2010).

[30] M. H. F. Zarandi, S. Davari, S. A. H. Sisakht, The large-scale dynamic maximal covering location problem, Mathematical and Computer Modelling 57 (2013) 710–719.

[31] J. F. Repede, J. J. Bernardo, Developing and validating a decision support system for locating emergency medical vehicles in louisville, kentucky, European Journal of Operational Research 75 (1994) 567 – 581.

[32] A. Ingolfsson, S. Budge, E. Erkut, Optimal ambulance location with random delays and travel times, Health Care Management Science 11 (2008) 262–274.

[33] M. S. Maxwell, M. Restrepo, S. G. Henderson, H. Topaloglu, Approximate dynamic programming for ambulance redeployment, INFORMS Journal on Computing 22 (2010) 266–281.

[34] V. Schmid, Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming, European Journal of Operational Research 219 (2012) 611 – 621. Feature Clusters.

[35] P. Beraldi, M. Bruni, A probabilistic model applied to emergency service vehicle location, European Journal of Operational Research 196 (2009) 323 – 331.

[36] J. Naoum-Sawaya, S. Elhedhli, A stochastic optimization model for real-time ambulance redeployment, Comput. Oper. Res. 40 (2013) 1972–1978.

[37] E. Erkut, A. Ingolfsson, G. Erdogan, Ambulance location for maximum survival, Naval Research Logistics (NRL) 55 (2008) 42–58.

[38] L. A. McLay, M. E. Mayorga, Evaluating emergency medical service performance metrics, Health Care Management Science 13 (2010) 124–136.

[39] L. A. McLay, M. E. Mayorga, Evaluating the impact of performance goals on dispatching decisions in emergency medical service, IIE Transactions on Healthcare Systems Engineering 1 (2011) 185–196.

[40] D. Bandara, M. E. Mayorga, L. A. McLay, Optimal dispatching strategies for emergency vehicles to increase patient survivability, International Journal of Operational Research 15 (2012) 195–214.

[41] V. Knight, P. Harper, L. Smith, Ambulance allocation for maximal survival with heterogeneous outcome measures, Omega 40 (2012) 918–926.

[42] J. A. Fitzsimmons, B. N. Srikar, Emergency ambulance location using the contiguous zone search routine, Journal of Operations Management 2 (1982) 225–237.

[43] J. Goldberg, R. Dietrich, J. M. Chen, M. Mitwasi, T. Valenzuela, E. Criss, A simulation model for evaluating a set of emergency vehicle base locations: Development, validation, and usage, Socio-Economic Planning Sciences 24 (1990) 125–141.

[44] S. G. Henderson, A. J. Mason, Ambulance service planning: simulation and data visualization, Handbook of Operations Research and Health Care Methods and Applications 70 (2004) 77–102.

[45] Y. Yue, L. Marla, R. Krishnan, An efficient simulation-based approach to ambulance fleet allocation and dynamic redeployment, in: AAAI Conference on Artificial Intelligence (AAAI), 2012.

[46] J. Hausner, Determining the Travel Characteristics of Emergency Service Vehicles, Technical Report, Santa Monica, CA: RAND Corporation, 1975. URL: `http://www.rand.org/pubs/reports/R1687`.

[47] C. R. Houck, J. A. Joines, M. G. Kay, Comparison of genetic algorithms, random restart and two-opt switching for solving large location-allocation problems, Computers & Operations Research 23 (1996) 587–596.

[48] J. Holland, Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence, University of Michigan Press, 1975.

[49] D. Goldberg, Genetic algorithms in search, optimization, and machine learning, Artificial Intelligence, Addison-Wesley, 1989.

[50] R. Haupt, S. Haupt, Practical Genetic Algorithms, Wiley, 2004.

[51] K. Man, K. Tang, S. Kwong, GENETIC ALGORITHMS.: Concepts and Designs, Avec disquette, Advanced Textbooks in Control and Signal Processing Series, Springer-Verlag, 1999.

[52] T. D. Valenzuela, D. J. Roe, S. Cretin, D. W. Spaite, M. P. Larsen, Estimating effectiveness of cardiac arrest interventions: A logistic regression survival model, Circulation 96 (1997) 3308–3313.