Towards Smarter Metropolitan Emergency Response

Marcus Poulton
Birkbeck College, University of London
Malet Street, London WC1E 7HX
marcus.poulton@gmail.com

George Roussos
Birkbeck College, University of London
Malet Street, London WC1E 7HX
gr@dcs.bbk.ac.uk

ABSTRACT
A core ingredient of Smart Cities is the use of emergency services both as a lens through which to monitor their ever-changing state and as a rapid response mechanism to the needs of their population. Emergency response units in particular employ diverse ubiquitous computing technologies for sensing, resilient communication, and dispatch and depend on extensive command and control infrastructure that links into the healthcare and transportation systems. In the case of ambulance services in particular, command and control centres collate medical incident, vehicle position and status data to build a real-time picture of the City. Taking the London Ambulance Service (LAS) as our case study we develop a simulation framework and introduce an enhanced routing and dispatch method that combines concurrent assignment and redeployment of resources in a single algorithm. We provide evidence that our unified proactive relocation and dispatch model produces significant improvements in measured performance in terms of meeting citizen needs.

Author Keywords
routing, smart cities, redeployment, emergency services, simulation.

ACM Classification Keywords
I.6.5 Simulation and Modelling: Model Development

INTRODUCTION

Smart Cities are one of the most active areas of application of ubiquitous computing technologies. Notably, the information gathered and processed by emergency services in a metropolitan setting can be used as a lens for observing and reacting to human dynamics as well as the needs of individuals in the city. In this way emergency service systems and related infrastructure act as both a contributor and beneficiary to smartness and dynamic adaptation. In this paper we consider in detail the case of the ambulatory service in London to identify the costs and benefits afforded by the integration of diverse metropolitan socio-technical systems and services within a unified approach and the potential effects on the well-being of its citizens.

A well-established clinical outcome is that shorter ambulance arrival times play a criticial role [1] in the case of emergency patients involved in incidents of high severity. As a consequence, emergency response unit mobility is of key importance. The mobility characteristics of ambulances in their various forms however differ from normal civilian traffic. This is partly because ambulance crew travelling with flashing lights are exempt from traffic regulations that would otherwise impede progress to a patient. For example, ambulances are allowed to treat red traffic lights as a give way sign, are able to pass the wrong side of a keep left bollard and disobey the speed limit.

Moreover, the collection of data on the human condition by ambulatory services reveals many specific attributes that can be used to enhance social governance. For example, the temporal and spatial characteristics of acute cardiac events follow specific patterns. Driving conditions in urban road networks also have well-defined patterns that affect ambulance arrival times. Analysis of these patterns historically and in real-time can be used to govern ambulance manning levels and placement balancing strategy and tactics.

This paper explores how information captured from a variety of ubiquitous computing technolgies deployed as part of emergency response systems can be utilised to create a realistic predictive model of their performance in dense urban environments. This model reveals facts about life in the city from a healthcare perspective that has remained unobserved until now. A core ingredient of our approach is the development of an accurate and precise simulator that can be used to evaluate new ambulance dispatch algorithms. Indeed, we introduce such a novel algorithm that combines both strategic and tactical elements into a unified model and test its viability through the simulator. In the long-term we aim to refine this work by incorporating elements promoting smart governance.

In the following sections, first we provide some background on how a typical ambulance service handles emergency medical calls. This is followed by analysis of real data streams obtained from the London Ambulance Service. We then proceed to discuss the simulator developed, introduce our dispatch algorithm and assess its performance.

BACKGROUND

Incoming emergency medical calls in London are processed in one of two call centres operated by the LAS, each covering a different area of London. Typically the caller confirms the location of the patient either by passing an address or other land feature such as road junction to the call-taker. The caller is then asked a series of questions that quickly determine the type and severity of the emergency. Using this information the Command and Control system will dispatch one or more responders as and if appropriate. For life-threatening cases such as Cardiac/Respiratory Arrest also known as Category A incidents, at least two

units (vehicles with crew) being dispatched. When responders arrive at the scene they assess and provide any treatment necessary. All other non life-threatening calls are graded as Category C (there is no Category B). Approximately 75% of patients attended to are then transported to a hospital for further assessment and treatment. Once the patient has been handed over to the hospital staff, the crew are then made available for further assignments. In many cases the crew are repositioned to a location where there is a higher chance of an incident occurring within a short distance.

Londons' ambulances carry extensive instrumentation that monitors their location as well as vehicle state including temperature, handbrake, door open, blue lights, siren, battery level and so forth. This information provides telemetry which is relayed to the system back end located at LAS headquarters over multiple wireless pathways including at least two 3G mobile telephony operators to ensure resilience and extended coverage as well as IEEE 802.11 when the ambulance is in the vicinity of an ambulance station. Moreover, ambulances carry a Siemens GPS unit with embedded MEMS gyros augmented with wheel sensors that measure speed. The system is capable to report positioning data accurately and provide navigational assistance even when GPS signals are weak which is critical in built environments. The information is also used by the on-board computer to provide the crew with map-based navigation, search facilities and details about the patient and the incident. Of course, similar to all UK emergency services, ambulances carry TETRA two-way transceivers which allow encrypted voice communication with the LAS Command and Control centre.

One obligation placed upon the LAS is to reach at least 75% of all Category A incidents within 8 minutes. Failure to achieve this target is met with heavy penalties. The LAS use several vehicle types to accomplish this target. The entire operational fleet consists of nearly 400 ambulance units, over 200 fast response units (FRU) and a smaller collection of bicycles and motorcycles. In London there are some 77 'standby points' or locations where vehicles and their crew will wait for work. These locations have been selected because they provide good coverage of London but also for practical reasons such as crew safety and the ability for crew to obtain refreshments. Under certain conditions considerable fricition is observed between the need to meet strategic targets and positioning tactics.

Early models that attempted to solve the coverage location problem [2] ignored road networks completely, relying instead on a set of so-called geographical atoms. Goldberg [3] used mean and variance to determine estimated travel times, an improvement on linear regression methods that preceded it. Potvin [4] used long term non-stochastic and short-term stochastic elements to produce efficient routing, thereby reducing *overall* travel time. As travel time is a key factor in survival, novel methods of traffic avoidance are investigated for example the use of crowd-sourced data has attracted considerable interest recently [5, 6].

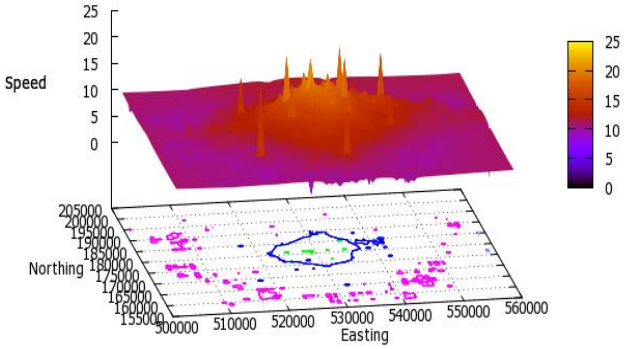

Figure 1. Spatial surface plot showing occurrence of Category A incidents in London during 2012

DATA & ANALYSIS

In this section we describe some of the key characteristic of the core data set used in this research. In particular we discuss the temporal and spatial characteristics of emergency events in London and patterns derived from the telemetry obtained from ambulances. The data used in this research was obtained from the London Ambulance dataware house. Much of this data originates from vehicle telemetry as previously described. Emergency incident data from London Ambulance was also analysed from the year 2012.

Emergency Events

Figure 1 shows how the number of life-threatening medical emergencies is distributed around London, revealing that a large proportion of these incidents occur in the centre of the city. The shape of this distribution changes throughout the day as the population swells during working hours. Figure 2 shows the total number of critical incidents and the average number of resource on duty in London, per hour, during 2012. This is at a minimum at around 4am with just over 400 critical incidents being reported during 2012. The busiest period appears to be around 6pm when just fewer than 1,100 incidents were reported.

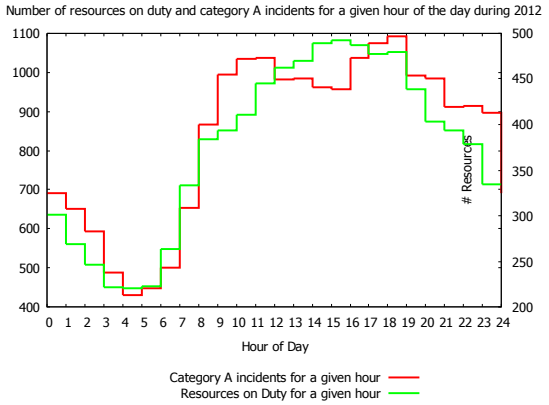

Figure 2. Occurrence of Category A incidents in 2012 by the hour along with the average number units on duty

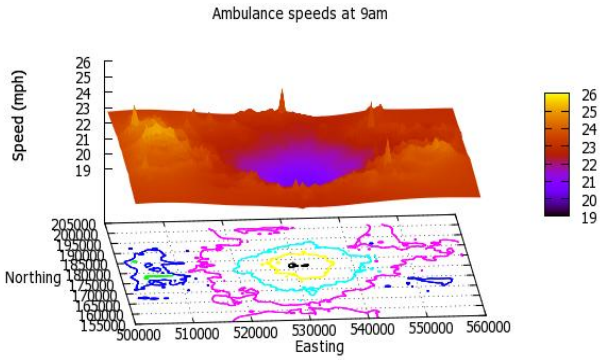

Figure 3. Surface plot of ambulance unit speeds at 9am

It can be clearly seen that the supply of resources on duty closely matches the demand. This spatial and temporal dynamic behaviour of emergency incidents adds to the complexity of where and how many resources to site at standby points.

Road Network Spatial Analysis

Our preliminary analysis aimed to determine by how much the road speeds were slower in the centre of London compared to the suburbs. The distribution shown in Figure illustrates average vehicle road speeds from 9:00-9:59 for the year 2012. When superimposed on a map of London it clearly, and obviously, shows that vehicles travelling in surrounding urban areas average higher speed than those in central London regardless of the time of day. There is also a difference between the speeds of the vehicles.

Ambulances and FRU's generated 202 million vehicle location and speed records.

Road Network Temporal Analysis

Whilst recognising that there are spatial differences in road speeds Figure 4 also shows the temporal speed difference by vehicle type. Specifically it shows vehicle type, speed and time of day data collected for the year 2012 across the whole of London. Figure 4 implies that FRUs are, as expected, faster than ambulances due to their smaller physical size and handling characteristics. Whilst this might sound obvious, there was no pre-existing data to precisely quantify this different in speed. Our analysis

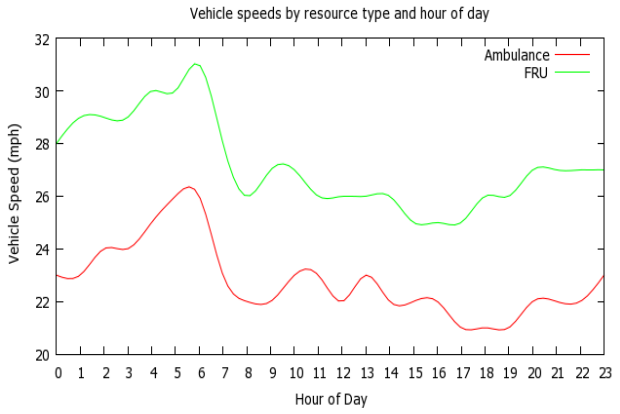

Figure 4. Vehicle speeds by hour of day

shows that on average the FRU is about 5 mph faster than ambulances in an urban environment, with the greatest variation in the morning rush hour of around 6 mph. Note that we do not take into consideration spatial variations so the difference in road speeds could vary further depending on whether vehicles are travelling in central London or in outer urban areas.

Clearly, any road speed model used for simulation would need to take into account vehicle type, spatial and temporal distribution.

SIMULATOR

We developed a discrete event simulator to model ambulance workflow so that novel dispatch algorithms could be tested. The workflow involves dispatching a resource to incidents and standby points, waiting on scene whilst dealing with the patient, optionally transporting the patient to hospital and then becoming available again for further work. Our ultimate aim was to measure the performance of the simulator in terms currently used by the LAS, i.e. the percentage of Category A calls where an ambulance arrived with 8 minutes. By improving on the dispatch model we aimed to improve the performance metric. The simulator design was based around a discrete event model with modular components acting out the various roles. In the reported experiments emulations were conducted at full scale incorporating all 400 vehicles.

At the heart of the simulator is the discrete event priority queue. The queue facilitates inter-module communication. Emergency events are replayed by passing messages to the Command and Control module. This module tracks emergency events and asks the dispatch module to recommend units for assignment. Dispatch modules typically in use in ambulance organisations today select the nearest (or quickest) available vehicle but as described later, this is not necessarily the best option.

Assignment requests are passed on to the resource module. The resource module simulates multiple vehicles by using a routing engine to plot a route to the emergency incident, following the road route at the estimated speed. Continuous distribution functions, built from historic information captured from vehicle telemetry, were used to estimate how long paramedics spent on-scene and at-hospital depending on the severity of the incident.

Routing Engine

Accurate routing estimates were the key factor in producing an accurate simulator. Analysis of the road speeds was carried out using 204 million telemetry records data captured by LAS during 2012 from the onboard GPS units travelling to an emergency. Each of the position reports were snapped to the nearest road and the road type identified. London was spatially divided up into 100x100 square cells, each of 300 meters in width and height. Speeds were averaged for each position report that occured in each cell for every hour of the day, each vehicle type and road type. This produced a 5-dimensional table

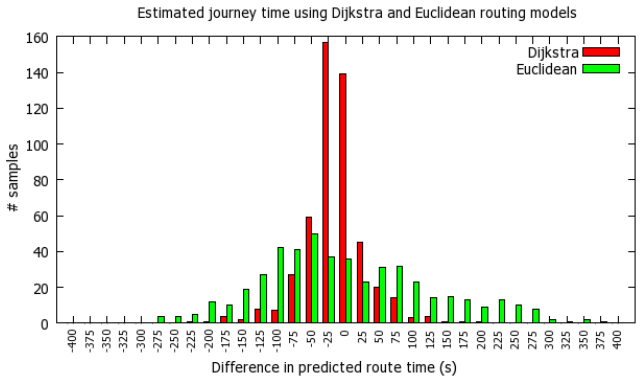

Figure 5. Accuracy of different routing engines

containing average road speeds. We built a routing engine that could calculate the quickest route between two locations using the actual road network using Dijkstra's shortest path algorithm on London's 19134 links and 15986 road nodes. The algorithm used the 5-dimensional speed data to calculate route speed and total duration.

To validate the accuracy of the routing engine we compared its performance with 500 actual journeys carried out by ambulances enroute to emergency incidents. Figure 5 shows a comparison of straight-line (Euclidean) and Dijkstra routing arrival times with actual arrival times. The Dijkstra routing engine had an accuracy of -3.085 seconds per trip, a standard deviation of 46.21 seconds with a high precision of 80% of estimated journey times within 1 minute of the actual drive time.

The routing engine is also able to calculate realistic travel-time isochrones for each vehicle. It is therefore able to calculate at any moment in time a complex heatmap, or coverage, of London that can be reached by all available ambulances. More importantly, it can therefore calculate which areas of London that cannot be reached by ambulances under the estimated traffic conditions for that time of day.

Tuning and Validation

The simulator was tuned and validated by running the simulator with historic incident data and comparing the simulated arrival time performance with actual performance. The simulator was configured to use a dispatch model that closely resembles the existing dispatch policy at LAS. This policy dispatches units that will arrive in the shortest time but does not attempt to dispatch units to standby points.

COMBINED AUTOMATIC DISPATCH MODEL

We developed the Combined Automatic Dispatch Model (CARD) to deploy resources to incidents and standby points using a static evaluation function to measure the value of the current state of deployment of ambulances around London. At any point in time there will be a number of incidents in progress, either awaiting resources to be assigned or in some other state, such as enroute, on scene or at hospital. A state with a low number of waiting incidents is preferable than a state with a larger number of waiting incidents. Additionally, where there is one incident and two resources, the "better" state is one where the assigned resource would arrive faster than the other. Resources are also better placed in locations where incidents are likely to occur. These requirements for positioning vehicles were combined using a static evaluation function (SEF) consisting of a set of five weighted basis functions. The basis functions are summed (Figure 6) to provide a single value that provides a ranking value of the current state of deployment.

$$f(x) = \sum_{i=1}^{5} b_i(x) \times w_i$$

Figure 6 - formula for weighted coverage.

The basis functions selected are directly related to the need to judge the importance of un-dispatched incidents for Category A and C incidents, the total drive time to Category A and C incidents and the overall coverage at that point in time.

Table 1 - Summary of basis funcions for the static evaluation function

Basis function	Description
$\sum waiting_A$	The number of category A calls that are awaiting an ambulance to be dispatched
$\sum waiting_C$	The number of category C calls that are awaiting an ambulance to be dispatched
Weighted Coverage	The % of the London area that available resources can reach in 8 minutes multiplied by the expected incident density.
$\sum Travel\ Time_A$	The current total travel time for resources enroute to category A calls
$\sum Travel\ Time_C$	The current total travel time for resources enroute to category C calls

At any moment in time the dispatch module can evaluate, using the SEF, the current state. The dispatch algorithm can build a list of appropriate dispatch options, e.g. deploy ambulance A to incident X or deploy ambulance B to standby point Y. The best dispatch decision can then be determined by evaluating the state after each potention dispatch decision has taken place. As the SEF contains a weighted coverage element, the dispatch engine will favour deployment of a vehicle that do not leave areas of London with no available vehicles were emergency incidents are likely to occur.

The basis function weights were adjusted randomly using small perturbations over multiple simulations in order to

find suitable values. We used actual emergency incident data and actual performance figures from September 2011 as, during this period there were no outages or major events that would skew the results. The simulator was eventually able to improve performance, measured as arrival times within 8 minutes, from 74.19% to a simulated 76.84% for category A and from 57.82% to a simulated 80.28% for all other incidents.

Figure 7 shows the predicted improvement in arrival times compared with actual arrival times during that period. The histogram shows that, using CARD, the majority of category A incidents were reached in 254 seconds (4 minutes 14s) compared to 360 seconds (6 minutes) historically. These figures provide promising evidence that a combined incident and standby dispatch model can significantly improve arrival times, and therefore, the outcome of critically ill patients.

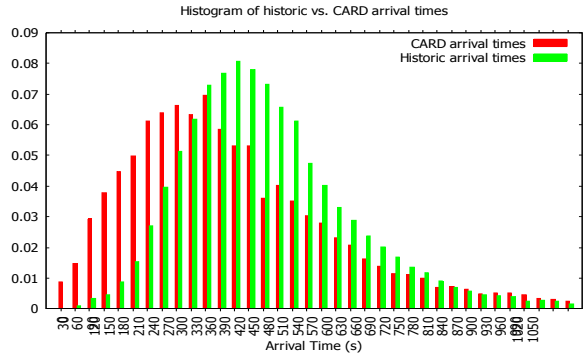

Figure 7. Historic and predicted arrival times using CARD

CONCLUSIONS

Ambulance and other emergency service fleets provide a unique perspective on the dynamics of densly populated metropolitan areas. They carry a variety of wireless communication and sensing devices that link into the complex city transportation and healthcare socio-technical systems thus revealing human and urban dynamics for example the spatial and temporal patterns of medical incidents affecting citizens. Information streams generated through the active deployment of emergency service resources can also be used to improve their performance for example by better utilising standby points to reduce arrival times and improve the prognosis especially in the case of severe incidents. Our proposals for CARD and the use of enhanced routing illustrate how such data-driven models can adapt to the changing conditions encountered in the field balancing strategic and tactical objectives. We anticipate that such benefits can be further extended to address healthcare governance concerns hence further extending the smartness of emergency response.

ACKNOWLEDGMENTS

Without the continued support of the LAS this research would not be possible.

REFERENCES

1. Valenzuela, T.D., et al., *Estimating effectiveness of cardiac arrest interventions: a logistic regression survival model.* Circulation, 1997. **96**(10): p. 3308-13.
2. Church, R. and C. ReVelle, *The maximal covering location problem.* Papers in Regional Science, 1974. **32**(1): p. 101-118.
3. Goldberg, J., et al., *Validating and applying a model for locating emergency medical vehicles in Tuczon, AZ.* European Journal of Operational Research, 1990. **49**(3): p. 308-324.
4. Potvin J, X.Y., Benyahia I, *Vehicle routing and scheduling with dynamic travel times.* Computers and Operations Research 2006. **33**: p. 1129–1137.
5. Stone, W., L. Stenneth, and J. alowibdi. *Reducing Travel Time by Incident Reporting via CrowdSourcing*. in *ICOMP'11*. 2011.
6. Janecek, A., et al., *Cellular data meet vehicular traffic theory: location area updates and cell transitions for travel time estimation*, in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. 2012, ACM: Pittsburgh, Pennsylvania. p. 361-370.