

Cardiff University  
School of Mathematics

---

# MODELLING EMERGENCY MEDICAL SERVICES

By

Leanne Smith

*B.Sc. Mathematics, Cardiff University*



A thesis presented for the degree of Doctor of Philosophy  
Operational Research Group, School of Mathematics  
College of Physical Sciences

---

April 2013

# **SUMMARY**

## **MODELLING EMERGENCY MEDICAL SERVICES**

**Leanne Smith, B.Sc.**

Cardiff University, School of Mathematics

Emergency Medical Services (EMS) play a pivotal role in any healthcare organisation. Response and turnaround time targets are always of great concern for the Welsh Ambulance NHS Trust (WAST). In particular, the more rural areas in South East Wales consistently perform poorly with respect to Government set response standards, whilst delayed transfer of care to Emergency Departments (EDs) is a problem publicised extensively in recent years. Many Trusts, including WAST, are additionally moving towards clinical outcome based performance measures, allowing an alternative system-evaluation approach to the traditional response threshold led strategies, resulting in a more patient centred system.

Three main investigative parts form this thesis, culminating in a suite of operational and strategic decision support tools to aid EMS managers. Firstly, four novel allocation model methods are developed to provide vehicle allocations to existing stations whilst maximising patient survival. A detailed simulation model then evaluates clinical outcomes given a survival based (compared to response target based) allocation, determining also the impact of the fleet, its location and a variety of system changes of interest to WAST (through 'what-if?' style experimentation) on entire system performance. Additionally, a developed travel time matrix generator tool, enabling the calculation and/or prediction of journey times between all pairs of locations from route distances is utilised within the aforementioned models.

The conclusions of the experimentation and investigative processes suggest system improvements can in fact come from better allocating vehicles across the region, by reducing turnaround times at hospital facilities and, in application to South East Wales, through alternative operational policies without the need to increase resources. As an example, a comparable degree of improvement in patient survival is witnessed for a simulation scenario where the fleet capacity is increased by 10% in contrast to a scenario in which ideal turnaround times (within the target) occur.

## DECLARATION OF AUTHORSHIP

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed: .....

Date: .....

### Statement 1

This thesis is being submitted in partial fulfilment of the requirements for the degree of PhD.

Signed: .....

Date: .....

### Statement 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references. The views expressed are my own.

Signed: .....

Date: .....

### Statement 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed: .....

Date: .....

---

*This thesis is dedicated to my Dad, whose unwavering stubborn determination, perseverance and perfectionism I fortunately inherited.*

---



## **ACKNOWLEDGEMENTS**

First and foremost, I would like to thank my supervisors, Professor Paul Harper, Dr. Janet Williams and Dr. Vincent Knight, all of whom have been truly invaluable sources of knowledge, advice and support throughout my PhD, and wonderful company along the way. Thank you also to Dr. Israel Vieira for sharing his expertise. It has been an honour working with you all.

Furthermore, I am grateful to the funding offered by the EPSRC through the LANCS Initiative, and in particular to Professor Jeff Griffiths for allowing me to be part of the Cardiff LANCS Healthcare Cluster. The Welsh Ambulance Service NHS Trust provided data and insight for this project, for which I am also extremely grateful.

My Mum and brother have stood by me patiently, always encouraging me to trust my instincts and supporting my habit of never taking the easy option – I cannot thank them enough.

To all my friends and family, old and new, past and present, thank you all so much for the smiles and for sticking around during the frowns.

Finally, to my teammate, Angelico – grazie! I would not have made it to the finish line with you.

# PUBLICATIONS AND PRESENTATIONS

## Publications

Knight V.A., Harper P.R. and Smith L. (2012) Ambulance allocation for maximal survival with heterogeneous patient outcome measures. *OMEGA* 40 (6) 918-926.

Smith L., Harper P.R., Williams J.E., Knight V.A. and Vieira I.T. (2012) Modelling ambulance location and deployment. Cumberland Initiative [online].

## Awards

Early Career Researcher Poster Prize at ORAHS Conference 2011 – First Place.

St. David's Day Cutting Edge Research Poster Competition 2012 – First Place.

## Conference Contributions & Presentations

Modelling Ambulance Location and Deployment – Smith L., Harper P.R., Williams J.E., Knight V.A. and Vieira I.T.

*INFORMS*; Phoenix, Arizona, October 2012.

*EURO*; Vilnius, Lithuania, July 2012.

*LANCS Healthcare Cluster Workshop*; University of Southampton, November 2011.

Modelling Ambulance Location and Deployment in Wales – Smith L., Harper P.R., Williams J.E., Knight V.A. and Vieira I.T.

*ORAHS*; University of Twente, Holland, July 2012.

*Conference proceedings paper.*

*SCOR*; Nottingham, April 2010.

*Conference proceedings paper.*

Allocating EMS Vehicles to Maximise Survival of Heterogeneous Patients – Smith L., Harper P.R., Williams J.E. and Knight V.A.

*ORAHS*; University of Twente, Holland, July 2012.

*Conference proceedings paper.*

Allocating Welsh Emergency Medical Services to Maximise Survival – Smith L., Harper P.R. and Knight V.A.

*SCOR*; Nottingham, April 2012.

Ambulance Demand and Deployment – Smith L. and Vile J.

*HMC2 Workshop*; Cardiff School of Mathematics, April 2011.

Resource Planning and Deployment of Welsh Ambulance Services – Smith L. and Harper P.R.

*SWORDS Seminar*; Cardiff School of Mathematics, October 2011.

*OR51*; Warwick, September 2009.

*ORAHS*; Leuven, Belgium, July 2009.

*SCOR*; Lancaster, April 2009.

*PhD Symposium, LANCS Healthcare Workshop*; Cardiff Millennium Stadium, January 2009.

## **Poster Presentations**

Resource Planning and Deployment of Welsh Ambulance Services

*Making an Impact - 999 EMS Research Forum*; Cardiff, February 2013.

*Winter Simulation Conference (WSC)*; Phoenix, Arizona, December 2011.

*ORAHS*; Cardiff, July 2011.

*LANCS Initiative Board Meeting*; London, May 2010.

Modelling Ambulance Location and Deployment

*The OR Society Simulation Workshop (SW12)*; Worcestershire, March 2012.

*Speaking of Science (SOS)*; Cardiff University, April 2010.

# CONTENTS

<b>SUMMARY.....</b>	<b>I</b>
<b>DECLARATION OF AUTHORSHIP .....</b>	<b>II</b>
<b>ACKNOWLEDGEMENTS.....</b>	<b>IV</b>
<b>PUBLICATIONS AND PRESENTATIONS.....</b>	<b>V</b>
<b>1 INTRODUCTION.....</b>	<b>1</b>
1.1 Emergency Medical Services.....	1
1.1.1 Introduction .....	1
1.1.2 Associated Problems for EMS.....	1
1.1.3 Studying The Welsh Ambulance Service .....	2
1.2 Research Objectives.....	3
1.3 Research Strategy .....	4
1.4 Thesis Overview .....	6
<b>2 WELSH AMBULANCE SERVICES NHS TRUST .....</b>	<b>7</b>
2.1 EMS Responsibility.....	7
2.1.1 Patient Care.....	7
2.1.2 Star of Life.....	8
2.2 Ambulance Services of the United Kingdom.....	8
2.2.1 Introduction .....	8
2.2.2 Services.....	9
2.2.3 Challenges.....	10
2.3 The Welsh Ambulance Service.....	11
2.3.1 The Trust.....	11
2.3.2 NHS Direct.....	11
2.3.3 Emergency Service Operations.....	12
2.3.4 Resources.....	15
2.3.5 Response Targets .....	17
2.3.6 Turnaround Targets.....	19
2.3.7 Challenges.....	20
2.3.8 Interventions.....	22
2.3.9 Improvement Policies.....	24
<b>3 EMS MODELLING: A LITERATURE REVIEW.....</b>	<b>26</b>

3.1	Introduction .....	26
3.2	Medical Insight .....	27
3.2.1	Introduction .....	27
3.2.2	Triage & Categorisation.....	28
3.2.3	Response, On-scene Care & Patient Outcome.....	29
3.2.4	Challenges .....	31
3.2.5	Specialist Staff & Training.....	32
3.3	Location Problems.....	33
3.3.1	Introduction .....	33
3.3.2	<i>P</i> -Median and <i>P</i> -Center Problems.....	34
3.3.3	Covering Problems.....	35
3.4	Solution Approaches .....	38
3.4.1	Mathematical Programming.....	38
3.4.2	Queueing Theory .....	38
3.4.3	Multi-Objective Modelling.....	42
3.5	Location Analysis for Emergency Services.....	43
3.5.1	Introduction .....	43
3.5.2	Dispatching.....	44
3.5.3	Equity in Access .....	44
3.5.4	Travel Times & GIS.....	45
3.5.5	Utilisation.....	46
3.5.6	Dynamic Modelling.....	46
3.5.7	Simulation for EMS.....	47
3.6	Around the World in 999 .....	49
3.7	Generic Modelling.....	50
3.8	Model Limitations.....	51
3.9	Summary.....	52
<b>4</b>	<b>WAST: DATA ANALYSIS.....</b>	<b>54</b>
4.1	Introduction .....	54
4.2	The Data Set .....	54
4.2.1	Statement of Accuracy .....	54
4.2.2	Influences.....	55
4.2.3	Dispatching.....	56
4.2.4	Variables and Field Headers.....	57
4.2.5	Pathway .....	58
4.3	Preliminary Analysis .....	60
4.3.1	South East Structure.....	60

4.3.2	Demographics.....	61
4.3.3	Locations .....	64
4.3.4	Resources .....	65
4.3.5	Category Computation.....	66
4.4	Demand .....	67
4.4.1	Regional .....	67
4.4.2	Inter-Zone Assistance .....	67
4.4.3	Divisional.....	68
4.4.4	District.....	70
4.4.5	Time Dependency.....	71
4.4.6	Inter-Arrivals .....	75
4.5	Station Assignment .....	75
4.6	Fleet Allocations .....	81
4.7	Response Time .....	83
4.7.1	Introduction .....	83
4.7.2	Delays.....	84
4.7.3	Travel Time.....	85
4.7.4	Response Time Data Results.....	85
4.8	On-scene Service.....	86
4.9	Transportation .....	90
4.10	Turnaround and Clear Time .....	91
4.11	Summary.....	93
<b>5</b>	<b>TRAVEL TIME ESTIMATION.....</b>	<b>94</b>
5.1	Introduction.....	94
5.2	Necessity of Estimation Methods.....	95
5.3	Travel Distance Estimation.....	95
5.4	Computing Shortest Distance .....	97
5.5	Travel Time Estimation .....	99
5.5.1	Introduction .....	99
5.5.2	Scaling Factors.....	99
5.5.3	Estimation via Distance.....	99
5.6	Acceleration, Deceleration and Cruising.....	101
5.7	Travel Time Estimation by Road Type.....	102
5.8	Targeting Variation.....	103
5.8.1	Introduction .....	103
5.8.2	Regional Zoning & Preferences.....	104
5.8.3	Travel Barriers .....	105

5.8.4	Starting Point Assumption .....	106
5.9	Distributional Fits.....	106
5.10	Time-Dependent Travel Times .....	107
5.11	Limitations of Models .....	109
5.12	Estimation in Wales .....	110
5.12.1	Introduction .....	110
5.12.2	Necessity of Travel Time Prediction for Modelling WAST .....	110
5.12.3	Available Travel Time Prediction Methods .....	111
5.12.4	Response Journey Modelling.....	113
5.12.5	Transportation Journey Modelling.....	114
5.12.6	Travel Matrix Generator.....	115
5.12.7	Zoning Characteristics: South East Wales.....	116
5.12.8	Time Dependency: South East Wales.....	117
5.13	Application of Estimation Method.....	118
5.13.1	Introduction .....	118
5.13.2	Response Journey Correlation .....	118
5.13.3	Transportation Journey Correlation .....	120
5.13.4	Regression Analysis for Average Travel Time Estimation .....	121
5.13.5	Tested Models.....	122
5.13.6	Method of Least Squares Fit.....	123
5.13.7	Residual Analysis.....	125
5.14	Results.....	126
5.14.1	Google Map Locations.....	126
5.14.2	Travel Time and Distance Matrices.....	127
5.15	Conclusion.....	128
<b>6</b>	<b>LOCATION ANALYSIS .....</b>	<b>130</b>
6.1	Introduction.....	130
6.2	Improving EMS Performance with Location Analysis .....	131
6.3	Coverage .....	131
6.4	Survival.....	134
6.4.1	Introduction .....	134
6.4.2	Cardiac Arrest.....	136
6.4.3	Survival Function Development .....	139
6.4.4	Challenges .....	141
6.4.5	Survival and Location Theory .....	141
6.5	Modelling Heterogeneous Patient Groups.....	143
6.5.1	Introduction: MESLMHP .....	143

6.5.2	Model Brief: MESLMHP .....	144
6.5.3	Notation & Formulation: MESLMHP .....	145
6.6	Modelling Heterogeneous Patients and a Heterogeneous Fleet.....	147
6.6.1	Model Brief: MESLMHPHF.....	147
6.6.2	Notation: MESLMHPHF .....	148
6.6.3	Formulation: MESLMHPHF .....	153
6.7	Combating the Input Utilisation Problem: A vicious circle .....	154
6.7.1	Model Brief: MESLMHP-I and MESLMHPHF-I .....	154
6.7.2	Notation: MESLMHP-I.....	156
6.7.3	Notation: MESLMHPHF-I .....	159
6.8	Application to WAST .....	161
6.8.1	Introduction .....	161
6.8.2	Granularity.....	161
6.8.3	Genetic Algorithm.....	162
6.8.4	Iterations.....	163
6.8.5	Data Input .....	163
6.8.6	Service Procedures .....	165
6.8.7	Priority Weighting.....	166
6.9	Results.....	166
6.10	Conclusion.....	172
6.10.1	Introduction .....	172
6.10.2	Model Limitations .....	173
6.10.3	Extensions.....	173
6.10.4	Survival Approach.....	174
<b>7</b>	<b>SIMULATING AN EMS SYSTEM.....</b>	<b>176</b>
7.1	Why Simulate?.....	176
7.1.1	Definition.....	176
7.1.2	Benefits .....	177
7.1.3	Overview .....	177
7.2	Strength of Simulation.....	178
7.2.1	Introduction .....	178
7.2.2	Advantages .....	179
7.2.3	Visualisation.....	180
7.2.4	Comparison with Other Techniques.....	181
7.2.5	Limitations.....	181
7.3	Simulation Type.....	182
7.4	Programming.....	183
7.4.1	Choice of Style.....	183



7.4.2	Choice of Language .....	184
7.4.3	Time Handling .....	185
7.5	Simulation Design .....	187
7.5.1	Contribution to WAST .....	187
7.5.2	Objectives .....	188
7.5.3	Conceptual Modelling .....	189
7.5.4	Entities .....	191
7.5.5	Assumptions .....	192
7.5.6	Data Analysis .....	193
7.5.7	Input .....	194
7.5.8	Google Maps API Input .....	196
7.5.9	Outputs .....	199
7.6	Program Processes .....	200
7.6.1	Process Introduction .....	200
7.6.2	Generating Demand .....	203
7.6.3	List Structures .....	205
7.6.4	Event List .....	206
7.6.5	Waiting Events .....	208
7.6.6	Dispatch Method .....	209
7.6.7	Transportation Policy .....	211
7.7	Sampling Methods .....	212
7.8	Model Validation .....	213
7.8.1	Introduction .....	213
7.8.2	Warm-up Period .....	213
7.8.3	Run-Length and Replication Analysis .....	216
7.8.4	Verification .....	217
7.8.5	Validation .....	219
7.9	Discussion and Extensions .....	221
<b>8</b>	<b>SIMULATION RESULTS .....</b>	<b>223</b>
8.1	Why Experiment? .....	223
8.2	Model Set-up .....	223
8.2.1	Introduction .....	223
8.2.2	Data .....	224
8.2.3	Run Options .....	224
8.2.4	Parameters and Variable Values .....	225
8.3	Fleet Allocations .....	228
8.4	Simulation Scenarios .....	229
8.4.1	Benchmark Scenario Results .....	229

8.4.2	Experimental Scenarios: What if?.....	235
8.4.3	Results Summary.....	237
8.4.4	Dispatch Policy Results.....	239
8.4.5	Demand Scenario Results .....	240
8.4.6	Catastrophe Scenario Results.....	241
8.4.7	Transportation Policy Results.....	242
8.4.8	Turnaround Time Results.....	243
8.4.9	Location Model Allocation Comparisons.....	244
8.4.10	Capacity Results.....	245
8.5	Conclusions .....	247
8.5.1	Allocation Insight.....	247
8.5.2	Diversion .....	249
8.5.3	Payoff.....	249
<b>9</b>	<b>CONCLUSION .....</b>	<b>250</b>
9.1	Discussion .....	250
9.1.1	Introduction .....	250
9.1.2	Objectives.....	251
9.1.3	Modelling Conclusions.....	252
9.1.4	Investigative Conclusions and Considerations.....	254
9.2	Model Limitations.....	257
9.3	Model Extensions .....	259
9.4	Implementation.....	262
9.5	Final Reflections .....	262
	<b>APPENDIX.....</b>	<b>264</b>
	<b>GLOSSARY .....</b>	<b>275</b>
	<b>REFERENCES.....</b>	<b>276</b>

## Chapter 1

# Introduction

## 1.1 Emergency Medical Services

### 1.1.1 Introduction

Emergency Medical Services (EMS) play a vital role in the secondary healthcare of any population. In the United Kingdom, EMS are components of the many Ambulance Trusts that operate across the country. They are managed locally by the individual trusts and provide pre-hospital care and treatment to emergency medical patients in designated regions. Healthcare managers continually endeavour to improve the services provided by the trusts; since ambulance services are often the first point of contact for a potential health-service patient in an emergency situation, awareness of system efficiency and operational effectiveness is imperative to the improvement of care. Servicing the public in their hour of need – in diverse situations and medical crises – presents obstacles to any emergency organisation; yet it is essential to still provide a consistently first-class service.

### 1.1.2 Associated Problems for EMS

All Emergency Medical Service systems find themselves faced with analogous problems but with the need to find exclusive solutions – deciphering the best operational procedures and service strategies to optimise system performance in their own specific region.

System design problems may range from the decision of staffing levels (as with any business or organisation) to the best way to minimise patient risk. Location of vehicles at ambulance bases or stand-by points, operational fleet capacity, response policy and treatment locality are just some of the decisions faced daily by EMS managers. Additionally, demand to ambulance services is ever increasing (National Audit Office 2011). With an ageing and growing population (Office for National Statistics 2012), in a world where good health is promoted, pursued and protected, ambulance services need to find the best ways of providing emergency care to an informed population whilst meeting their own performance targets. Much research has been conducted into

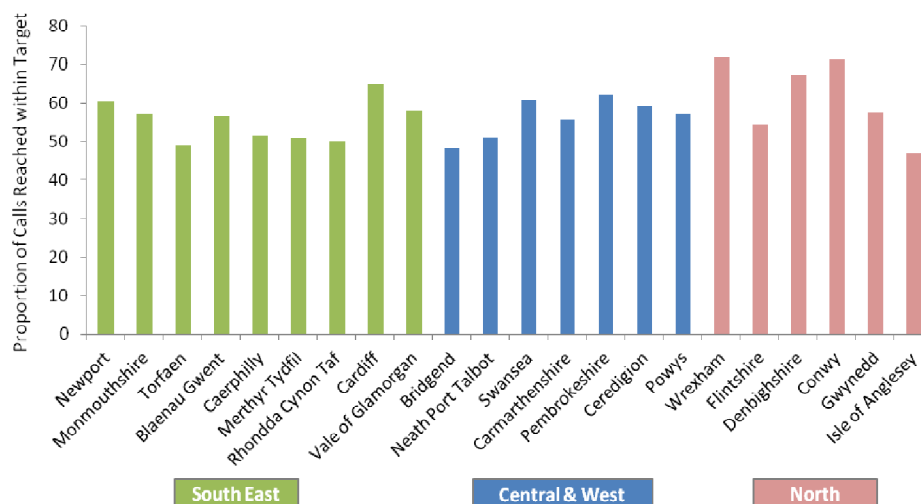
EMS systems around the world, and improving conduct within them continues to be a well investigated subject area.

Ensuing is the story of a study delving into the heart of one National Health Service (NHS) Ambulance Service in particular – the Welsh Ambulance Services NHS Trust (WAST). This thesis focuses on suggesting improvements in the service provided to the Welsh population through increasing efficiency of operations regarding the fleet, its location and deployment and service policies.

### 1.1.3 Studying The Welsh Ambulance Service

With largely sparse, sprawling populations, where even urban areas are comparatively much less dense than in England, the problems faced by WAST are likely to be accentuated by the demographics of the country. Budget cuts, increasing demand and high turnover of executive staff, has meant that WAST has struggled for many years to meet performance standards.

Welsh Government (2013) records (reproduced graphically in Figure 1.1), show that despite recent improvement there is still a necessity to increase the response time achievements of WAST, especially in problematic areas where response consistently falls below target.



**Figure 1.1** Average WAST performance for critical emergency responses throughout Wales, separated by region for the period of November 2012 – January 2013

Although advances have been made, and milestone targets are being met much more frequently and with more vigour, there is still an apparent need for further insight to the processes. Future implementation of qualitative and quantitative decision tools should assist in preventing WAST from trailing behind more urban trusts' performance and instead evolve in the field of secondary health care for Wales. A more in-depth discussion on the ambulance system in Wales and the problems it faces is given in Chapter 2.

Quite recently, support for changes in performance measures, from response time proportions to clinical outcome indicators has been vocalised. For a patient experiencing a life-threatening emergency, a good chance of survival requires a timely response, which has the added benefit of reducing patient (and bystander) anxiety, suffering and distress. It has been advocated that for non life-threatening emergencies, the single response time standard should also be replaced with clinical based measures. Since such patients suffer a variety of medical conditions, a range of responses would be better suited than a single target; however, it is acknowledged that this will only be possible *"once the evidence base and professional consensus are sufficiently developed"* (Department of Health 2005). It is for this reason that this study looks to investigate patient outcome based models for determining the impact system performance has on survival of a population.

Other issues, not discussed explicitly in this thesis but which occur recurrently throughout the academic literature and in ambulance service publications, are the situating of new facilities and evaluation of existing ones (workshops, bases and control centres), dispatching and reallocation rules of sub-fleets, staffing rules and rostering, shift lengths, cost-benefit analysis (health economics) and crisis management. Nevertheless, many of these issues will be reviewed in Chapter 3.

## **1.2 Research Objectives**

Response time performance is heavily dependent not only on the size of the fleet and resources, but also on the positions from which these vehicles respond. By simply increasing the capacity, response performance may not necessarily witness great improvement (a more detailed discussion of this peculiarity will be given in Chapter 6, section 6.2). The optimal location of the fleet is the main contributor to improved performance when targeting emergency operations. The primary objective of this study is therefore to provide an EMS, in this case WAST, with an allocation of vehicles at existing base facilities that best allows them to reach and exceed their Government set

targets under a set of system conditions. Simulation of the system, functioning with an '*optimal*' fleet allocation, will provide insight into problematic areas of the country, so that the Trust might develop superior strategies to combat any under-achievements in performance. Ultimately, the given allocation will assist WAST in meeting their performance targets whilst saving lives of the people they serve.

Using insight gained of EMS through exploration (of data and recent literature) and communication (with members of WAST), it is the intention of the research presented here to provide planning tools that offer alternative operational and strategic solutions to any ambulance service. The contribution of research is applicable to a generic EMS setting and so can easily be extended to the whole of Wales, to other trusts in the UK and to similarly structured services elsewhere. The tools designed and mathematical models developed, enable investigation on a large geographic scale as well as into the differences of modelling patient outcome over response performance – shifting the focus of the health service from a business model to be patient centred.

The five main goals of the study can be laid out, providing a frame of reference throughout the thesis, and topics for discussion in the conclusions of Chapter 9, section 9.1.2:

- Investigate if improvements to WAST's performance can be made with regards to response and turnaround phases, whilst maintaining current capacity;
- Investigate current policy impact on patient survival;
- Suggest ways in which to improve survival probability;
- Support WAST's move to clinical outcome based measures;
- Develop generic tools that may be utilised by EMS managers for future planning purposes in areas dealing with demand, fleet allocation and capacity.

### **1.3 Research Strategy**

One of the main classification areas of healthcare research as defined by Hulshof et al. (2012), is that of 'emergency care services', where either strategic, tactical or operational decision objectives form the focus of such a project. Under this heading fall the problems faced by ambulance services.

According to Hulshof et al., the three decision objectives can then be further defined for various ambulance service problems:

- Strategic planning – such as ambulance districting, ambulance coverage problems and capacity dimensioning (number of vehicles);
- Tactical planning – staff-shift planning;
- Operational planning – including ambulance dispatching, hospital facility selection, ambulance routing, vehicle relocation and prioritisation.

Previously, Operational Research (OR) techniques such as Queueing Theory, Location Analysis and Simulation, and a combination of these, have been used to study the operational and strategic planning of such Emergency Medical Services and test changes to systems (Brotcorne et al. 2003, Mason 2013). The research approach engaged throughout this thesis makes use of all three techniques but offered in a way that, if desired, allows frequent follow-up use of them by non-OR experts, including WAST analysts, planners and controllers.

EMS research has classically taken the course of reducing travel times to and from the emergency scene by locating vehicles at different places on a network (Peleg and Pliskin 2004). In OR, this is captured by the field of Location Analysis. Another useful OR practice for EMS analysis is that of Queueing Theory, where one objective would be to minimise the length of time a prioritised patient spends waiting for service. Both of these methodologies are exploited in this study in conjunction with a full-system simulation in order to answer the question of how many vehicles to locate at existing bases within a region to meet performance targets. This take on a Location Analysis problem will offer WAST the tools they require to make future decisions of resource levels to reach the variable regional demand within their performance standards.

Previous studies focus mainly on the performance driven needs of an ambulance service and not necessarily on the best result for a patient (although these often coincide, they strive for different things), whereas this study reassesses the use of the Government targets and considers the need of meeting these targets whilst ensuring best possible patient outcome. A limited amount of work has formerly been conducted in this area; however, more recently research objectives of this nature are undoubtedly becoming the focus of much EMS modelling.

Finally, the simulation model presented offers an environment in which to explore the other service phases and system aspects, including the proposed clinical outcome based objective, enabling insight to problematic policies and alternative operations.

## 1.4 Thesis Overview

The thesis begins by looking at the Welsh Ambulance Services NHS Trust (WAST) and its current operations. In Chapter 3 a presentation is made reviewing some of the more important preceding studies in the field.

In order to develop any tools for the modelling of WAST, insight to the current situation is required. Therefore, Chapter 4 sees analyses conducted on a data set provided by WAST for the purposes of this study, with the troublesome geographic conditions highlighted, enhancing understanding of the current system.

One common issue in many location problems is the accurate modelling of journey time or distance on a network. To improve the level of detail obtained compared to more traditional models, the Google Maps API is utilised and embedded within a Travel Time Matrix Generator Tool, offering a benchmark from which to predict all journey times for WAST. This solution is illustrated in Chapter 5, along with a detailed explanation of prior travel time and distance estimation techniques.

Intelligent location of an operational EMS fleet on a network is widely thought to enhance response performance. Chapter 6 presents four allocation models, demonstrated for two different performance measure standards. The resulting allocations of vehicles to existing stations in the South East of Wales are then utilised as input in the simulation modelling approach (Chapter 7) to evaluate the probability of a patient experiencing a favourable outcome from a complete service process. The results of all simulation experimentations are presented in Chapter 8 and conclusions are formulated in the final chapter.



## Chapter 2

# Welsh Ambulance Services NHS Trust

## 2.1 EMS Responsibility

### 2.1.1 Patient Care

An ambulance was customarily seen as no more than a means of transport for sick or injured people; yet, the importance of its role within today's society is one that should not be underrated. From early records of what can be thought to be the origins of the modern day ambulance – where injured patients were carried by hammock based wagons or suspensions between horses (Barkley 1978) – to the current Mercedes motor vehicles, ambulances are crucial to patient outcome in out-of-hospital medical situations.

Knights of The Order of St. John, known collectively as the 'Knights Hospitaller', provided immediate care to injured soldiers throughout the Crusades of the Middle Ages, removing them from battlefields and pioneering 'first-aid' (Nicholson 2001, The Order of St. John 2012). Nowadays, The Order of St. John is a charity providing healthcare (predominantly through ambulance services) around the world, with the mission:

*"Pro Fide, Pro Utilitate Hominum"*  
(“For the Faith, In the Service of Humanity”)

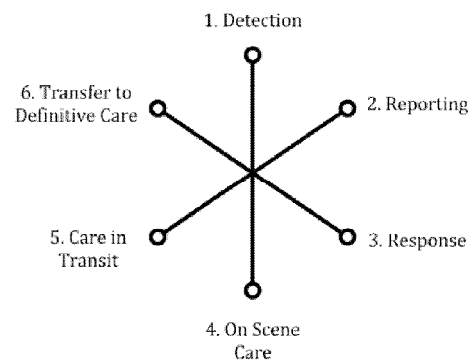
During the Napoleonic Era, Baron Dominique Jean Larrey shared compassion with wounded soldiers of the battlefield who would usually only be collected and transported to a medical centre after hostilities had ceased. Larrey introduced a tiered ambulance system in 1793, known as the “Flying Ambulances” to evacuate injured soldiers during battle to improve their chances of survival (Ortiz 1998, Skandalakis et al. 2006). Following acceptance of the necessity of a medical transport service within the military, ambulance services have evolved across the world to include purpose built EMS systems for providing emergency care alongside transportation.

### 2.1.2 Star of Life

After the evolution of civilian ambulance services in the late 19<sup>th</sup> Century, it was deemed necessary by the U.S. Department of Transport, National Highway Traffic Safety Administration, NHTSA, (NHTSA EMS 2012), for all emergency medical care services and resources to exhibit a uniform symbol allowing them to be easily identifiable. In 1977 the “Star of Life” (Figure 2.1) was approved as the symbol for resource and personnel associated with emergency medical services.



**Figure 2.1** The Star of Life



**Figure 2.2** Phases of EMS service

The Star of Life was designed by Leo Schwartz, but is an amalgamation of ancient symbols (NAEMT News 2010). The six points were based on an existing symbol of the American Medical Association and represent the six phases (Figure 2.2) an EMS goes through in response to an emergency.

The Star of Life is also displayed on fire engines in some countries since the care procedure is similar and the quest of saving lives the same; additionally, many emergency services across the world are integrated, with fire engines supporting EMS vehicles and often even enlisting paramedic staff to optimise the chances of patient survival wherever in the region an emergency arises.

## 2.2 Ambulance Services of the United Kingdom

### 2.2.1 Introduction

Traditionally, ambulance services have been used as emergency life-support; however, the proportion of genuinely life-threatening emergencies is relatively small, and so the focus of these services has shifted over time to urgent care as a whole.

Ambulance trusts in England and Wales are part of the secondary care or 'acute healthcare' service provided by the NHS, and were commissioned by local Primary Care Trusts (PCTs) in England who controlled around 80% of the English NHS budget (NHS Choices 2011c) and currently by Local Health Boards in Wales (NHS Wales 2012b). Secondary care encompasses both the elective and emergency care of its users, with the Ambulance Service making up part of the pre-hospital patient pathway.

There are some discrepancies between the operations of ambulance services of the United Kingdom. In England (excluding the Isle of Wight) there exist eleven ambulance services (NHS Choices 2011a), whereas, Wales and Scotland have only one. Elsewhere in the UK, such as with the Northern Ireland health service, a designated ambulance trust does not truly exist, instead the ambulance service comes under the 'Health and Social Care Trust'.

The Welsh Ambulance Services NHS Trust provides treatment and care to the Welsh population for pre-hospital emergencies and inter-hospital transfers. Covering urban, rural and sparse areas, the Trust operates over a diverse and often problematic region. Its responsibility to provide an efficient and tailored service in each of these areas leads to the necessity of dedicated staff and reliable strategies. The current aim of WAST is to provide equity in this service and ultimately save lives whilst maintaining a given level of performance with regards to responses and handovers. Their vision is to "deliver high quality care wherever and whenever it is needed" (WAST 2012c).

### **2.2.2 Services**

An ambulance service encompasses both Patient Care Services (PCS) and Emergency Medical Services (EMS). PCS are responsible for the safe and timely transport of scheduled and elective patients to and from medical facilities; EMS, which are the focus of this study, deal solely with unscheduled emergency care. EMS also often involves transportation of patients from the scene of an incident to a hospital facility if required, but moreover, staffed paramedics onboard EMS vehicles are able to administer medication and provide treatment directly to patients whilst in attendance or during transport.

### 2.2.3 Challenges

In this “Age of Austerity”, the NHS has a duty to make extensive efficiency savings. For the ambulance trusts, this has meant budget cuts. Financial pressure brings a necessity for better organisation of emergency services and cooperation with other health services.

Over the past few years, the Welsh Ambulance Services Trust has been a focus of local and national media attention due to its comparatively poor performance with other UK ambulance trusts. They are however not alone in pursuit of efficacy under arduous circumstances. Disruption within trusts’ structure and management (BBC News 2006a, 2010, 2012b), continual stories of long awaited responses by critical patients and shrinking budgets (BBC News 2012a, Panorama 2012), all add to the struggle of convincing the media and the public of success whilst attempting to advance UK services.

In terms of demand, the number of calls for urgent medical assistance has been increasing yearly across the UK (National Audit Office 2011) due to an ageing population resulting from the post-war baby boom. Demand also fluctuates temporally (hour, day and season), spatially, due to meteorological changes (Wong and Lai 2012) and as some suggest, with celestial movements such as the lunar cycle (Alves et al. 2003, Stomp et al. 2009). EMS systems must work to accommodate this variation in order to provide a sustained service.

Many services operate with a paramedic on board every ambulance, yet patients are often still unnecessarily taken to hospital following a response, particularly within the elderly community (Knowles et al. 2011). In 2005 a report by the Department of Health outlined problematic areas and improvement policies for English ambulance trusts. It was believed that more than one million people that end up in Accident and Emergency (A&E) departments across the country as a result of ambulance transportation could have been treated at the scene or in their homes, and that the approach should be to take healthcare *into* the community (Department of Health 2005).

Amongst visions of reducing numbers transported, guidelines were set to also enhance the speed and quality of call handling and emergency care received by patients, and to consistently improve efficiency, effectiveness and performance. The report instructs that the correct emergency response should be given “first time, in time”. One conclusion that echoes the particular objectives of the Welsh Trust and this study is that progressively “ambulance services should be designed around the needs of the patient”.

## **2.3 The Welsh Ambulance Service**

### **2.3.1 The Trust**

Having a single national ambulance service, Wales manages its EMS with a workforce of 2,500 employees and with an annual income of £72 million (WAST 2012d). As the third largest ambulance service in the UK, WAST have a duty to a population of 2.9 million people, operating resources across a large geographic area of 20,600km<sup>2</sup>. Although a small country, this area is considerably larger than most ambulance trusts would control, yet has a relatively small population, meaning much of the Trust's operational region is sparse.

Established 1<sup>st</sup> April 1998, through the merger of five predecessor ambulance services (WAO 2006), WAST operators have since handled around 300,000 emergency medical occurrences and provided more than 1.3 million non-emergency patient transportations per year through the Patient Care Services (WAST 2012b, Watkins and Price 2010).

Whilst WAST must be able to balance their operations between urban cities (of which there are officially only four) and rural communities, they must also serve the remote and mountainous regions of Wales. In addition to ground coverage, Wales has three air ambulances (funded by the Welsh Air Ambulance Charitable Trust) that provide rescue services for injured people in these areas as well as immediate emergency care (Wales Air Ambulance 2009).

### **2.3.2 NHS Direct**

In 2007, NHS Direct Wales became part of WAST (NHS Wales 2012a), providing a 24 hour telephone based medical advice service for the population. The purpose of the service is mainly triage driven, giving direction to patients as to the appropriate service and level of healthcare they require. The nurse led service assists the ambulance service when patients may not be experiencing an emergency but where an ambulance dispatch might still be required. It is unique to Wales that NHS Direct is incorporated within the Ambulance Trust. In England and Scotland, separate trusts operate the telephone service. If a call made directly to NHS Direct in Wales is deemed urgent (or more critical – see section 2.3.5) then the call is immediately passed to the ambulance service operators so that a dispatch may be arranged. Similarly, if a call arrives with the Ambulance Trust requesting medical attention for a condition that does not require the assistance of a paramedic or

transportation to hospital, then the call is passed to NHS Direct operators who may assist the patient via telephone and offer advice to deal with their condition independently. Such parallel systems offer advantages over separate operations in reducing unnecessary dispatches (Dale et al. 2003). Since the two services may coordinate and both sets of call-takers are often in the same control room, swift transfer of patients between services is easier and works with minimal disruption to the patient.

### 2.3.3 Emergency Service Operations

Current practice endeavours to increase patient care, chances of survival and of recovery, whilst exceeding strict performance targets, with various procedures in place to help achieve these tasks.

Typically, when urgent medical attention is required outside of secondary care facilities, (but not necessarily outside of the primary care structure), a bystander, third party, or the patient themselves may make a phone call to the national '999' emergency telephone line (or equivalent international number).

Operators receive emergency calls through a 'force-fed' telephone system, allowing random and equal spread of incidents to operators, ensuring fair workloads and distribution of calls. On receiving a call, operators must obtain three vital pieces of information before the call can be logged and a decision made on which vehicle to dispatch.

These three pieces of information are:

- a geographical location or postcode of the incident;
- the name of the patient and/or caller;
- description of the emergency.

The 'clock' starts measuring emergency response time when the dispatcher obtains the above pieces of information and stops measuring when the (first) EMS vehicle arrives at the scene of the incident and the crew log their arrival.

Upon receipt of the call by the emergency service telephone operators, it is assessed and categorised (as in Figure 2.5) according to the severity of the incident by an Advanced Medical Priority

Dispatch System (AMPDS). This is a unified system for the whole of the UK, whereby, the responses of the caller to questions relating to the emergency determine the type of response the incident requires. The AMPDS is one type of Emergency Medical Dispatcher, EMD, (sometimes known as Computer-Aided Dispatcher, CAD), that triages emergencies, then determines optimal vehicle dispatch choices. The AMPDS is simply a tool comprising of a visual display and a structured mathematical algorithm that determines which vehicle is closest to the incident location. It is then in the interest of the Trust and of the patient that the minimum possible delay is experienced; the emergency care providers need to reach the scene of the incident quickly in order for treatment to be of most benefit, especially in severely life-threatening situations. The chosen ambulance is given orders to dispatch to the scene by the EMS controllers.

Some CAD tools are unable to specify which ambulance is closest and so dispatchers must make an educated guess to this when vehicles are not at the stations. To improve performance, Dean (2008) suggests dynamic deployment can be used; however, to do this accurately, much more information is required by the dispatchers to determine the impact of not sending the closest available ambulance, especially “within systems that use fixed-deployment response strategies” and ones which are unable to track individual resources when they are not at their station.

Vehicles are sited based on a rotational hierarchy in Wales; when a vehicle becomes free after finishing with an incident it needs to be sited at a base or stand-by point ready to await its next call. The bases that do not already have an available vehicle positioned at them are ranked based on a weighted decision for which would be most desirable to locate at given expected demand and so the available, un-located vehicle will be sent to the highest ranked base.

An ambulance service not only has to dispatch a vehicle quickly enough from the best location to ensure a fast response, but also guarantee the correct level of care can be provided where needed; that is, a vehicle with the appropriate crew type on board must be dispatched. An assumption made throughout this thesis is that the correct crew will always be on board the chosen vehicle, since staffing levels of vehicles and scheduling is a large EMS problem in itself and not considered here.

The assignment of a hospital for transportation is almost always pre-determined by the location of the incident; however, the Trust is keen to reduce conveyance rates and treat more incidents within the community. Paramedic manned vehicles can also treat, test and administer drugs on scene, and have the ability to refer patients to social services. Special Practitioners (SPs) are a recent addition

to Welsh EMS, reducing automatic conveyance to hospital (only 4 out of 26 cases in a trial period were transported compared to the expected 26 (Watkins and Price 2010)). They are highly qualified and able to perform more procedures and administer a larger range of medicines than paramedics whilst allowing the patient to remain in their home or the community. Many incidents categorised as high priority by AMPDS are found not to be life-threatening upon attendance (improvements to AMPDS prediction abilities are recommended (Clawson et al. 2008); the SP is able to treat the patient at the scene, removing the need for transportation.

Figure 2.3 exhibits the prior discussion of operational stages for service of an emergency incident.

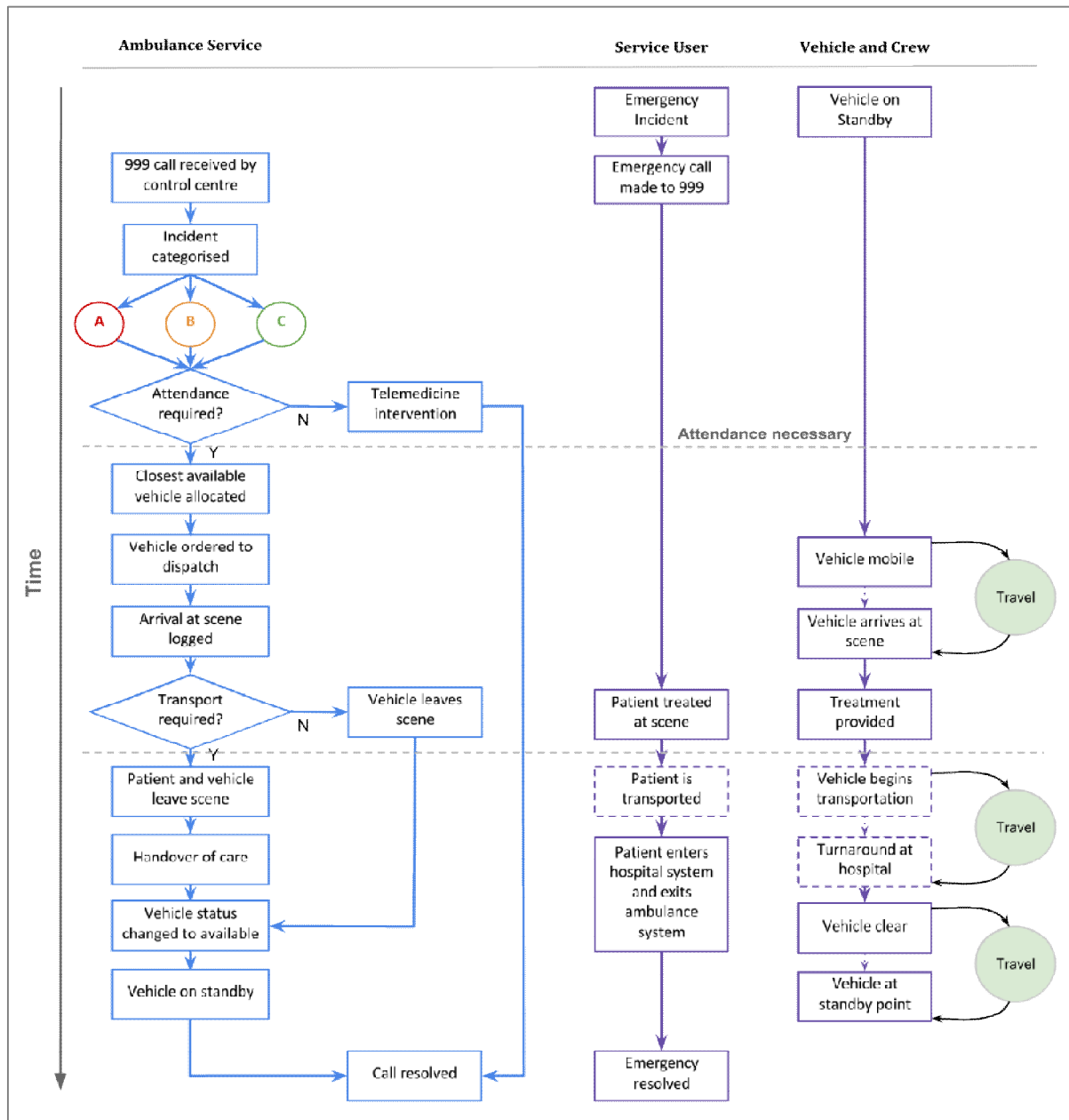
After communication with staff members at a WAST control centre in 2010, it was discovered that the service operates with locality based rosters which do not vary weekly or seasonally, but accommodate some variation by weekday based on simple average peak demand analysis. The rosters inform the number (and type) of vehicles to be on duty within each locality and the assignment to base stations. The average peak demand strategy is designed to show the system in its worst possible state based on historical data, yet the informatics team within WAST alluded to the fact that this approach might not suggest the best number of resources to deploy to meet demand.

Other issues that EMS managers may wish to investigate include:

- vehicle and crew safety and safety of patients when transporting;
- equipment selection;
- misclassification of emergencies by call-takers and automated systems;
- forecasting demand.

The service aims not only to improve patient satisfaction, care and clinical outcome through timely responses and swift handovers, but aims further to reduce cost, maximise equity and make resource utilisation (vehicles and crews) fair.





**Figure 2.3** Pathway through time for the service, service user and vehicle for a single emergency call

### 2.3.4 Resources

A wide variety of transportation modes have been used to carry sick and injured people over the centuries, and across the world. Husky dogs are not an unusual means of travel in places such as Alaska and Scandinavia, and often assist in hauling medical supplies and patient transports. During

the winters of the first world war, dog sleds pulled by teams of Huskies were used even in France (British Pathe c1915, Petwave 2012). In warmer climates, elephants made appearances in EMS teams, carrying the wounded from battlefields (National Army Museum 2012). Even hot air balloons have been used when land transportation was unsuitable (Air Ambulance Service 2012).

Nowadays, motor vehicles are common across all continents. There are two main types of vehicles used within UK ambulance trusts (Figure 2.4):

- Emergency Ambulances (EA) - traditional two-manned (either two emergency medical technicians (EMTs) or one technician and one paramedic) vehicles with the ability to transport patients if required;
- Rapid Response Vehicles (RRV) - equipped by a paramedic, these vehicles are smaller and faster than EAs, so are best used for life-threatening emergency incidents since they are in theory able to attend the scene quickly, but are unable to transport patients.



**Figure 2.4** Common EMS vehicle types in the UK (left: EA; right: RRV)

In addition to these commonly used vehicles, there also exist (WAST 2012a):

- High Dependency Units (HDU) - used primarily for transportations of lower priority patients and PCS. St. John's Ambulance Service operates with mainly these vehicle types;
- community first responders - volunteers are able to administer basic first-aid, resuscitate and use strategically placed community defibrillators;
- air ambulances - helicopters and occasionally small planes;
- bicycles - motor and pedal powered bikes are sometimes used in urban areas.

For combined EMS and PCS, there are over 700 vehicles in Wales, providing around 50% spare fleet capacity for the services. Emergency Ambulances within WAST are operational 24 hours a

day, which is not necessarily the case for the RRV fleet. For example, there are approximately 70 vehicles available for use within the South East region, but only around 40 are on shift at any one time. Many stations function with only one EA, sometimes with one additional RRV.

### **2.3.5 Response Targets**

Response times (the interval between arrival of the call and attendance of a paramedic) and turnaround times (the time spent at hospital transferring patient care) are the Key Performance Indicators (KPIs) of the Welsh Ambulance Service. It is apparent why these two are the KPIs when patient outcome is considered. In almost all emergency conditions, whether life-threatening or otherwise, a quicker response will mean less adverse effects from prolonged exposure to the condition, increasing the chances of treatment and administered medication having a more positive effect on the patient. At the hospital, if a long handover is experienced it is possible for the patient's status to deteriorate. For some conditions, the time from onset of the emergency to administration of life-saving drugs is critical to outcome (e.g. 'Golden Hour' for stroke victims) meaning the transfer of care should be minimal.

Performance standards for all UK ambulance trusts are determined by the Government. One such response performance target stipulates that a specific proportion of the population requiring an emergency ambulance response must be serviced within a set time from the receipt of an emergency call.

Emergencies are given one of a maximum of five classifications by the Trust, each with their own response time targets and monthly performance measures, summarised in Figure 2.5. The AMPDS also assigns the emergency a colour based on its urgency.

In the UK (and other countries such as the USA and Germany) targets are set to represent the difference in equity for rural and urban areas. That is, there are slightly different targets and possibly different deployment strategies depending on whether the area is rural or urban (or sparse). This represents how performance differs within and between these areas – the relative ease of serving an urban area compared to a rural one and the effect population density has on efficiency of the service (Erkut et al. 2008a, Felder and Brinkmann 2002, Fitch 2005).



**Target:** 95% to be reached within 14, 18 or 21 minutes in urban, rural and sparse regions (though these targets may be set locally throughout the UK).

### **AS2 – Urgent GP Referrals/Requests**

**(Green)**

Sometimes a patient contacts or is seen by a General Practitioner (GP) before involving the emergency medical service. The GP may decide the patient requires secondary care and so requests that the patient be transported to hospital, stating a time window for arrival.

**Target:** 95% to arrive at destination no later than 15 minutes after appointed time.

### **AS3 – Urgent Transfers**

**(Green)**

When a patient at one hospital requires transportation to another facility. In non-emergency situations the Patient Care Services (PCS) deal with these requests; however, under urgent conditions the emergency ambulance service is utilised.

**Target:** 95% to arrive at destination no later than 15 minutes after appointed time.

In Wales, for high priority calls, the target for an initial responder is fixed and not dependent upon the density of the population; however, when a second responder is required – EA attendance – follow-up vehicle targets differ for urban and rural emergency calls. Ethical issues surround decisions to set such performance targets, many of which are discussed in the paper by Felder and Brinkman (2002). For example, if a target of 75% performance exists, and 75% of demand occurs in urban areas, it could be possible to locate enough vehicles in these regions to guarantee to meet performance targets but without equitable service to rural populations.

## **2.3.6 Turnaround Targets**

Hong and Ghani (2006) show diagrammatically the flow of an ambulance through a system, and alongside show the performance measures at each stage of process. WAST have an additional performance measure to Hong and Ghani's system – turnaround time at the hospital. The Welsh target for this phase of service is to transfer patient care to the Emergency Department (ED) within fifteen minutes for all cases and all conditions. Extending the handover time by five minutes gives the overall turnaround time target of twenty minutes for all emergency cases, allowing handover and replenishment of the vehicle ready for service of any imminent emergency call.

It is in the interest of any ambulance trust, patients and hospital staff for a patient to experience a swift handover of care. This is however, not always possible and conflicting targets for the ambulance service and the emergency departments are not conducive to small turnaround times. Without better and more consistent recording of information referring to the length of these transfers of care, advancement in this service phase is restricted (BBC News 2009a).

### 2.3.7 Challenges

Hospital handovers cost the NHS millions of pounds per year in tens of thousands of lost ambulance hours due to queueing at EDs, which have risen from 37,000 hours in 2008/2009 to around 54,000 in 2010/2011 (Hughes 2009, Jones 2011). Not only is this money wasted that could be better used within the service, distressed patients spend long periods of time waiting for transfer of care, resulting in potential deterioration in their condition and effectiveness of subsequent treatment. Furthermore, whilst waiting to handover, vehicles are 'blocked' – unavailable to attend other emergency incidents – increasing utilisation (Lowthian et al. 2011) and putting lives at risk.

Rural areas such as Monmouthshire in South East Wales often witness poor performance with regards to response time. In 2008, suggestions of closure of the existing, already unmanned, station in Monmouth (Monmouthshire Beacon 2008) caused further controversy within the community, with the public looking to WAST to provide solutions to the response time problem witnessed in the area. In one reported case, a resident waited over an hour for the attention of an EMS crew, from over 30 miles away, after making a request for help for a Category A classified emergency call.

There have been many other cases, also in the South East, where ambulances from across the border in England were dispatched to an emergency since vehicles from local Welsh stations were unable to reach the patient within a reasonable response time.

It is not exclusively South East Wales that experience problems in meeting targets. Although small yearly improvements can be seen for overall Category A eight minute response performance from 2001 to 2007, WAST are *still* struggling to meet the all Wales 65% target. Even when this is surpassed, not all LHBs make their milestone 60% target. By improving the service solely to meet

the eight minute mark, the service could be jeopardising performance with respect to other categories and outcomes.

A report into the operations of English ambulance trusts comments on how the suggestion of moving towards medical based measures for all conditions could be beneficial.

*"It is becoming increasingly inappropriate to judge responses to non-category A calls exclusively on the basis of response times rather than clinical outcomes and the care given to the patient"* - Department of Health (2005)

The report comments on the classification of ambulance service targets by population density. Such classifications are often outdated (based on Welsh data from 2004), lack clarity and present confusion in the expectation of response times. Recommendations were for a single measure to be implemented from 2006 and not the segregation by urban or rural communities (Department of Health 2005). This approach would be more equitable, and would solve the problem of population classifications requiring regular re-evaluation.

It is also suggested that 'GP urgent' performance standards should be the same as the other '999' calls – that the clock should stop with arrival at the scene and not based on a target of reaching the hospital within 15 minutes of the requested time.

Centralised ambulance service structures exist around the world with the thinking that this will improve performance through better management and highly accessible resource centres. In Wales many vehicles operate out of centrally populous areas to serve the higher demand, but this leaves the more rural areas vulnerable. When individual staff members at the South East control centre were asked what they felt were the biggest challenges faced by the service, many mentioned that more vehicles and crews were required.

Delays, outdated equipment, resource deprivations, staff sickness levels due to overworking and disrupt to management structure are all thought to be "costing lives" according to the former WAST chief executive Roger Thayne (BBC News 2009b). In 2006, WAST saw three appointments of chief executive in three months (BBC News 2006b) which would have had an obvious impact on the smooth running of any organisation. Although some six years ago (at the time of writing), this has likely had a knock-on, if not adverse, affect on the service. A workforce takes time to adapt to any newly implemented changes, especially when such high turnover in management means introduced strategies may end up being temporary.

Lower cost of living and reduced house prices are attractive to low income families and the elderly, particularly in the current economic climate. Both deprived and elderly residents have a higher probability of EMS requirement and hospital care at some point during their lives (Cadigan and Bugarin 1989), where this socio-economic demand category is also on the rise (Portz et al. 2012). The South East region of Wales is home to a large proportion of this demographic, meaning EMS resources must be prepared to serve such a population effectively.

Continual improvement to any emergency service, not just EMS, is vital in order to retain service quality. Deterioration is likely if certain national issues (some of which are discussed by Hong and Ghani (2006)) are not accounted for, including:

- increasing and ageing population;
- ageing staff;
- expanding area of coverage (although in Wales this is less of an issue since the one Trust covers the entire country);
- health consciousness - increasing demand to all NHS services.

One of the biggest present challenges to WAST is public perception. Not only does the service need to achieve its own vision of improvement but must also convince the public of its successes. By restructuring the existing service, WAST hope to satisfy service users and restore faith. Public understanding however, is another, separate issue. A large number of calls for service are deemed inappropriate as an emergency (Wrigley et al. 2002) or are thought of as misuse of the service (Knapp et al. 2009). The concept of what indicates a necessary emergency intervention needs to be tackled and clarified for the public; this, alongside a better triaging system, would relieve some of the immediate pressure on EMS systems.

### **2.3.8 Interventions**

The Welsh Government carries out regular analyses on the data collected by WAST to ensure standards are maintained and any problems with the system are highlighted. Further details regarding the analysis conducted are provided in Chapter 4.



Large variation between regions in response time performance is not always due to worse or better operational procedures. Weather, which is known to affect EMS response, varies across the country and seasonally. Demography also takes its toll; urban areas tend to perform better than rural since, although demand is higher, travel times between sites is relatively small. The Department of Health (2005) recommended that English services prioritise all emergency calls in the same way so as to minimise variation by region. In order to combat the challenges faced in Wales and suggest courses for improvement, in-depth analyses of WAST have been commissioned, returning reports of operational and strategic findings over the past few years.

One of the earliest reports was presented by the Auditor General for Wales to the National Assembly (WAO 2006). Conclusions showed, against common belief, that the Trust was not under-resourced, yet questions were raised surrounding the efficiency of use and deployment of these resources. The discussion in section 2.3.7 regarding outdated rural and urban categorisation is supported by the Audit. It was found that the postcode and address database used at the time did not conform to British Standards and so future recommendations were that the target discrepancy between rural and urban areas be disregarded.

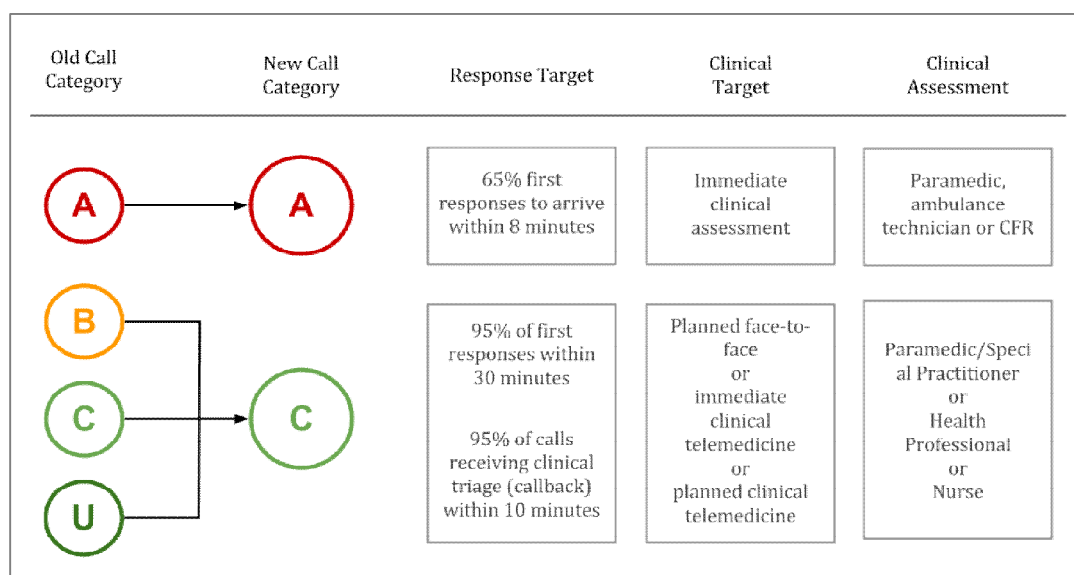
A similarly focussed investigation by Lightfoot Solutions (authorised by Health Commission Wales and WAST) found the lack of exact vehicle location details reduces the effectiveness of dispatch decisions. Through the introduction of Automatic Vehicle Location Systems (AVLS) or a mobile data system, the CAD system could more effectively track vehicles and optimise deployment. Another main finding of the report was that the reliance on overtime and planned relief for staff, although undesirable, is in fact keeping the costs to the service lower than if the service operated at the recommended staffing level (with larger delays). The efficiency review therefore recommends the contrary to the earlier Audit, offering the solution to the problem as simply increasing resources; however, the objective of this study is to show that ploughing more resources into a system is not a necessary solution and that it is possible to provide an alternative that may improve performance without incurring additional costs.

### 2.3.9 Improvement Policies

Ambulance services around the world are interested in improving their own systems, and targets as set by individual governing bodies are becoming more demanding. With the development of services comes the question of performance and whether the current measures are even appropriate.

If better triaging and assessment of incoming emergency calls could be implemented, it is possible that even continuing with current staffing levels and resource capacity, much of the pressure on the Welsh EMS system could be lessened (Lightfoot Solutions 2010). The Government offer a toolkit for commissioners to assist in understanding, identifying and combating the problems faced by EMS Trusts as laid out in sections 2.2.3 and 2.3.7 (Department of Health 2009).

The current UK system fails to record or capture by how much the response time target is exceeded. An early recommended method of measurement by the Audit Office was to use the 95<sup>th</sup> percentile, allowing the tail of the response time distribution to be captured rather than cutting off data beyond the target standard. More recently, some Trusts are moving from 'coverage' type targets to more clinical outcome based results, opening up an alternative line of questioning for researchers.



**Figure 2.6** New WAST framework for EMS performance

Since late 2011, after a claim by the National Audit Office that ambulance services put too much emphasis on response time with detrimental effects to patients (National Audit Office 2011), WAST have been instead working with patient care led strategies. Clinical quality indicators measure performance opposed to response time. Improving patient outcome can be achieved through improving efficiency, safety and care. The new performance framework is summarised in Figure 2.6 based on a document received after communication with the service detailing the change to only two category types and the planned and unscheduled methods of care (WAST March 2011).

Additional performance standards include:

- increasing percentage of patients suffering 'myocardial infarction' (heart attack) transferred to treatment centres within 150 minutes and improving patient outcome by increasing number of eligible patients receiving thrombolysis within 60 minutes;
- improving access and treatment compliance rates and percentage of patients transferred directly to a stroke team within 60 minutes.

A surge in prevalence of quantitative EMS data since 1994 (Henderson and Mason 2000), for example through CAD tools, has allowed more in depth understanding of different EMS systems and encourages investigations into better operational and strategic procedures. For example, in Ireland prior to the year 2000, data was unavailable for ambulance response times and had to be collected specifically for investigation to the system's performance (Breen et al. 2000). The discovery of unnecessarily quick responses to minor incidents and large delays experienced by some high acuity patients show priority-based classification and dispatch tools are vital for the potential improvement of the population's health. The suggestion of introducing an AVLS would also assist in this data collection in Wales, enabling more accurate data analysis and system perception. For now, the use of existing data, supported by communication and collaboration with ambulance trusts is enough to provide a valuable insight to the daily operations and larger scale tactics of an EMS system, from which the development of planning and decision support tools can aid improvement missions of such services.

## Chapter 3

# EMS Modelling: A Literature Review

### 3.1 Introduction

The National Healthcare Service (NHS) in the United Kingdom (UK) is a body that provides medical and rehabilitation care and treatment of the public. Founded by Aneurin Bevan in 1948 (NHS Choices 2011b), much work has since been and continues to be conducted in improving all NHS delivered services, ensuring better quality of care to patients and potential service users.

Pre-hospital patient care pathways often incorporate NHS ambulance service processes. The performance issue of an ambulance system is one that occurs wherever such an EMS structure exists, and is not a current issue for the UK alone. The literature review following spans a lengthy time period, with results originating from many different countries and for various services. Literature considered important to progression in the fields of health, emergency service, simulation and resource location, and those studies which encounter problems similar to an emergency service (medical or otherwise) are discussed.

Beginning with the health aspects, the effects of policies on NHS services are documented in section 3.2, with results collated from previous research and from medical documentation itself. Following an appraisal of mathematical and OR methodologies (section 3.3) and solution techniques (section 3.4) used to tackle diverse facility location and vehicle allocation problems in the past, the theoretical and implemented progression these techniques make into emergency service and emergency medical service operations is deliberated in section 3.5. The review presented here moves on to look at the difference in EMS modelling around the world and the limitations of implementing developed models.

## 3.2 Medical Insight

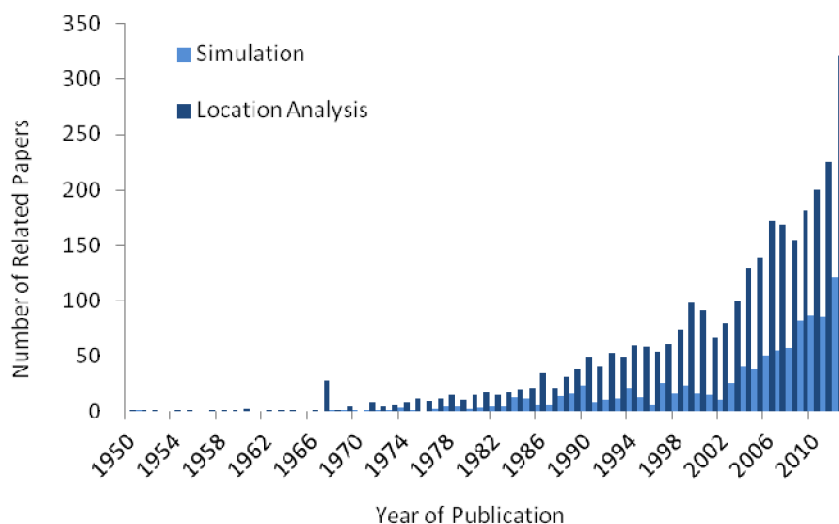
### 3.2.1 Introduction

“When someone’s heart stops beating, every second is vital”

– Mark Whitbread, Clinical Practice Manager of the London Ambulance Services NHS Trust (London Ambulance Service 2010).

The importance of immediate attendance to critical patients is indisputable. A patient’s chance of survival in many emergency conditions is initially dependent upon the speedy delivery of life-saving drugs and the administration of medical procedures such as cardiopulmonary resuscitation (CPR) and defibrillation. These procedures are almost certainly required to be conducted by trained emergency technicians or paramedics, and in many emergency situations require further care, treatment and diagnosis at a nearby hospital facility.

Response and transportation times are therefore crucial components of any EMS system, leading research to focus on better locations, optimal fleet capacities and shrewd deployment policies to improve the performance of these service phases.



**Figure 3.1** Yearly publication results from search engine ScienceDirect (SciVerse © 2013) using keyword criteria “emergency medical service, ambulance” in addition to either “location analysis” or “simulation”

A seemingly exponential growth in the number of research papers published, highlights the attempts at improving emergency secondary care provisions whilst suggesting location analysis and simulation modelling as approaches to tackling this problem in recent years (see Figure 3.1).

### 3.2.2 Triage & Categorisation

To improve the service provided by EMS Trusts, a simple, but costly, solution would be to introduce new resource elements to the system in order to manage demand and respond quickly. A more economical solution, and the core objective of many research projects, is the maximisation of existing resource efficiency (Breen et al. 2000). Savas (1969) advises that the cost-effectiveness of a solution should indeed be contemplated. Alternatives to new-resource solutions, some of which WAST have previously considered (and others which are explored throughout this thesis), include:

- substituting threshold response time targets with clinical outcome based indicators;
- maximising the use of prioritised dispatch systems;
- correctly identifying life-threatening and urgent calls – i.e. better triaging;
- utilising first responders;
- encouraging communication between operating areas;
- skill-mix of staff on ambulances – e.g. having qualified technicians on all EAs.

The Auditor General highlights call categorisation (originating from automated dispatch systems) and initial triage as issues requiring further attention in the future to improve prioritisation accuracy (WAO 2006). Furthermore, in order to enhance the smooth running of an EMS system, the Department of Health (2005) suggests making a summary of patient records available at the point of care. This would assist with triaging at the time and in the future, and in the accurate recording of service data, allowing swift handover of care at the subsequent stages of service. From an Operational Research point of view, this would also benefit future modelling endeavours.

The AMPDS (as discussed in Chapter 2, section 2.3.3) is a standard scheme, implemented by all UK Trusts. In other countries however, such a regulator does not always exist. EMS service structure, strategy and operations vary by trust, country and continent. Thakore et al. (2002) summarise why triage is necessary for all emergency calls and state how a priority-based dispatch system can reduce

overall response length. Prior to publication of the paper, not even all UK EMS Trusts operated with a prioritisation system – in these areas, all emergency calls were expected to receive an immediate response. In queueing terms, this implies the system exercises a first-in, first-out (FIFO) service policy. WAST do in essence operate with a FIFO system, complemented by the attribution of priority to patients so that separate FIFO queues exist for each emergency category; yet, in reality the queueing system is more complex than this, and the length of time the patient has spent waiting for service will also factor in their queue position.

### 3.2.3 Response, On-scene Care & Patient Outcome

*"Greater understanding of clinical best practice and technological advances [...] will make it possible to increasingly assess ambulance trusts on the quality of the care they provide, not just how quickly they get to the patient"* – Department of Health (2005).

Many medical-profession and OR researchers spend time deliberating the factors affecting patient survival and efficiency of emergency healthcare services. In some cases, the focus of a study is on EMS response time and location of resources, whereas others consider instead the training provided to responding paramedics and the effects of administering treatment prior to hospital admission. One such study (Studnek et al. 2010) assesses the impact of pre-hospital time intervals on a specific type of myocardial infarction (heart attack) patients. Time until treatment is found to be critical to patient outcome – as also known for cardiac arrest patients and stroke patients. Performance of the system is deemed acceptable if the interval from first contact to intervention is no longer than 90 minutes. By identifying the areas in pre-hospital processes for these coronary syndrome patients that require improvement, the authors claim that their model could have great clinical impact in the future.

One of the earliest references to evaluating clinical outcome given an emergency medical response and the challenge this presents to future EMS modelling was in 1984 (Hill III et al.). Despite the difficulty in obtaining data, a study of the long term survival of patients who experience an out-of-hospital cardiac arrest (OHCA) was later successfully conducted in Norway (Naess and Steen 2004). As one of the longest spanning follow-up studies, insight to patient outcome was obtained and the cost-benefits of treatment and service in relation to quality of life and length of survival was sought. From the information collected, a survival curve was found, approximately negative exponential in

shape, representing the cumulative survival of patients over time in years for an OHCA discharged from hospital alive.

Often, the probability of survival of a patient is measured until hospital discharge. For some medical conditions in particular, short-term outcome can be improved with a swift response by an appropriately trained ambulance crew. It is therefore thought, that by reducing the expected time taken to respond to the emergency the better the chance of survival of EMS service users.

Statistical significance of early paramedic intervention on cardiac patient outcome is supported by Mayer (1979) but the concept of minimising response time objectives in EMS planning is challenged. Response time is still a crucial component to survival, whether or not response times are used as performance measures, but this new way of thinking has been the main factor in the recent progression of UK EMS performance measures.

With UK pilot schemes exploring the feasibility of providing angioplasty instead of thrombolysis as a first emergency treatment (i.e. at the scene) for heart attack victims country-wide (Department of Health 2005), the shift from the traditional and accustomed eight minute response to best responses and community treatment is duly supported.

It is important to realise that some follow-on emergency treatment for such incidences may only be able to be provided at specific hospitals. Although transportation time was found to be the variable most strongly associated with achieving intervention within the target time by Studnek et al. (2010), transportation times may not be improvable if only certain facilities have the ability to intervene; therefore rural regional planning has a part to play in determining feasibility of coronary syndrome protocol.

As mentioned, for many critical conditions (stroke, cardiac arrest and heart attack), response, on-scene care, transportation and handover of care interval must all be contained in the onset to treatment time frame to maximise chances of survival. By minimising each segment of service in turn it is possible to minimise the overall emergency service length to increase chances of a favourable patient outcome.



### 3.2.4 Challenges

Progress in service improvement (by either a response time or clinical outcome based measure), is hindered by temporal and spatial demand factors, vehicle type dispatched, personnel dealing with the incident and many other conditions which the ambulance trust may not have control over.

Although, ideally, the focus of such an emergency system should be on the service it provides to its patients and not on system performance per se, adversity during service does not solely surround the patient's condition. Coats and Davies (2002) briefly mention some of the factors that contribute to difficulty experienced by EMS crews, doctors and paramedics at the scene:

- lighting (especially when incidents occur outdoors, under darkness);
- noise;
- ease of access to patients;
- weather conditions.

Weather for example, is fairly well explored when it comes to its impact on demand for an ambulance service (McLay and Mayorga 2010, Vile et al. 2012) but there exists little investigation of the affects it might also have on service (and travel) time and the increased risk response.

McLay, Boone and Brooks (2011) analyse call volume arriving to emergency services during times of extreme weather conditions (blizzards and hurricane evacuations), since it is acknowledged that the level and type of emergency (risk to patients) is higher than under normal weather conditions and the transportation network may itself be "impaired". This paper makes an unusual contribution to the literature since it is common for temporary and fast relief suggestions to be made for occurrences of extreme weather or disaster (Altay and Green 2006, Lin et al. 2012, Wright et al. 2006, Zayas-Caban et al. 2013). McLay et al. instead focus on not compromising preparedness whilst still accommodating the additional needs on the service in these periods of excessive demand.

*"Ambulance services need to become more rigorous and sophisticated in matching supply to demand, particularly given the consistent year on year increases" – Department of Health (2005).*

Adjustments to workforce plans and vehicle deployment strategies were highlighted as essential to deal with high demand in the Department of Health report. An exemplar system praised was that of the Staffordshire Ambulance Service NHS Trust, which have operated with the 'high performance

ambulance service' concept since 1994 (Turner and Nicholl 2002). The high performance concept predicts where demand is expected, so that vehicles may be positioned accordingly for time-dependent demand, reducing long-term operational expense.

A study into regular call volume predictions for EMS looks at demand varying over the time of day and day of the week (Setzler et al. 2009). Forecasting demand by only considering the expected number of calls for an individual region does not allow response times to be minimised effectively. Instead, to optimise, the forecasts should include both temporal and spatial distributions of the demand and resources (Geroliminis et al. 2009). Additionally, these aspects should be at the forefront of EMS modelling design when attempting to reduce response times and improve survival. For example, Chang and Schoenberg (2009), Ong et al. (2010) and Trowbridge et al. (2009).

### **3.2.5 Specialist Staff & Training**

In order to enhance patient survival (as WAST intend, Chapter 2 section 2.3.9), early interventions could occur from sources other than paramedic practitioners; for example, GPs, first responders, trained first aiders or bystanders. Persse et al.'s study (2003) pays homage to the tiered ambulance system whereby paramedics are not always the first line in emergency care. For some EMS incidents, paramedics are unnecessary, and by minimising their assignment to low priority calls – instead sending technicians (EMTs) – the result is of a system with more highly skilled paramedics, as these resources are able to become more specialised in a smaller range of incidents, aiding early intervention.

In Wales, Special Practitioners (SPs) are a recent addition to the EMS team. With further training and expertise, these employees help with the triaging and treatment of patients at the scene, enabling the Trust to reduce conveyance and bring care back into the community.

An assessment of pre-hospital care suggests some response protocols do not necessarily minimise mortality; however, clinical outcomes for survivors may be improved if the idea of manning all vehicles with highly skilled staff is implemented (Nicholl et al. 1998). As yet, there is a lack of commitment throughout the rest of the NHS to provide assigned pre-hospital care doctors. If this strategy were adopted in the UK, it could assist in swifter treatment of patients at the scene, or even during transportation, resulting in a higher survival probability for the most critical patients.

The appropriateness of non-paramedic first responses for pre-hospital care, is discussed in a clinical review paper (Coats and Davies 2002) in relation to road traffic accidents. Delays experienced during response may cause patients' conditions to deteriorate; yet as an alternative to investigating operational procedures to reduce these delays, improvement of professional training to support clinicians in these situations is suggested. For attending GPs at the scene of critical emergencies, patient survival could also be enhanced by the immediate administration of certain treatments (as mentioned in section 3.2.3); however, surveys suggest many lack the confidence to deal with these situations (Bloe et al. 2009). Improvement to training and inter-service collaboration is recommended to alleviate these barriers.

A range of papers have been discussed relating to the concerns and problems faced by EMS managers. The next section describes research corresponding to the use of one particular OR tool – namely Location Analysis – often used to address problems where the location of resources is fundamental to service and efficiency of a network.

### **3.3 Location Problems**

#### **3.3.1 Introduction**

Over the years, various tools such as mathematical programming, queueing theory, simulation and statistical modelling have all contributed to the development of EMS solutions, commonly through improving efficiency of resources. Such OR techniques are indispensable in solving both public and private sector problems for systems of service and delivery. One application is to location problems, which can be described as the problem of "siting facilities in some given space", where solution approaches have four main characteristics (ReVelle and Eiselt 2005):

1. customers located at nodes or on arcs;
2. facilities to be located at nodes;
3. a space in which all customers and facilities are located;
4. a metric indicating distance or time between nodes.

Location theory and its applications have played a major part in the structure of the UK's operations. With foundations in war efforts where the need to better organise, strategise and develop tactics for defence existed (Blackett 1962), this field of OR now lends itself to a wide and diverse set of

commercial and public sector problems (Daskin and Murray 2012). An extensive inspection of general location analysis papers can be found in the survey provided by ReVelle and Eiselt (2005).

One special case of the theory is the problem of locating vehicles or servers on a network to best meet a particular service user or management directed target. Spanning mainly the past five decades, the research into these areas has produced profound results from mathematical programming methods developed for a broad range of related situations.

Simulation offers an alternative environment for submitting a system to modifications and noting impacts of policy changes on overall performance. For an organisation where real-world testing is undesirable, or even impossible, simulation is a dependable and insightful tool for providing justification for operational decisions. For an EMS system, such as WAST, this technique can help convince policy makers through visual interaction that increasing efficiency in the system is possible.

The location of vehicles on a network and the service of demand are key elements to the type of problem faced by the research in this thesis and so create a starting point and become a continuous focus of the following literature review.

### **3.3.2 *P*-Median and *P*-Center Problems**

According to Smith et al. (2008) "*the major growth of location applications has occurred [...] since 1980*". Classic problems presented in the location literature however, occur much earlier and form the foundations of the bulk of modern day studies.

With the discovery of solutions to the classic *p*-Center and *p*-Median problems by Hakimi (1964, 1965), the theory around location analysis quickly became a widely deliberated topic – commercially and academically – due to an increase in demand and rise in site planning interest.

Referring to a communication network, the *p*-Median problem solution outlines the technique for optimally locating a number, *p*, 'switching' centres or points on a nodal network in such a manner as to minimise the total length of branches (wires) connecting the centre to all other nodes of the network. That is, the objective is to minimise average travel time or total average demand-weighted distance of the population to the facility.

The  $p$ -Center problem has slightly different objectives; instead of minimising the average travel time alone, it looks to optimally site  $p$  points or facilities on a network with the purpose of minimising the maximum distance from the demand to the nearest point or facility. For example, where the network can be considered to represent a highway system, police vehicles may respond to emergency incidents anywhere on the network, where large travel times are undesirable.

Almost alongside Hakimi's introduction of the  $p$ -Median problem, Maranzana published results on the location of supply points (1964). Based on Cooper's 'alternative' heuristic solution (1963) for the similar problem in the plane, the location-allocation heuristic minimises transport costs of a network through a 'centre of gravity' concept. Although unable to guarantee optimal results, it offered an important step forward at this time for location and routing problems.

### 3.3.3 Covering Problems

Over the following decade, Britain saw many further investigations in location theory drawing from the innovations of Hakimi and Maranzana. Locating major public-use facilities is prominent in the literature, but with a variety of objectives. An alternative to minimising total distance travelled by a population to facilities is to maximise the coverage of population demand with the minimum number of facilities (located at a predetermined number of candidate sites) whilst maintaining a certain level of service. This type of problem is known as 'set-covering'. Voronoi (1907) is responsible for the underlying mathematics of set-covering, with the principle of 'Voronoi Cells' being that of dividing a planar region into sub-regions using tessellations, and where distance from a discrete set of points in the plane is significant.

More than sixty years after Voronoi, the influential Location Set Covering Problem (LSCP) was first described formally (Toregas et al. 1971). The discovery of its solutions builds upon Hakimi's work, though some believe that set-covering is more generally inspired by the problems of locating emergency service facilities (Smith et al. 2008). It is viewed to be mainly a public sector problem since costs are not usually incorporated explicitly, leading to a cost-ignorant objective function; however, its popularity and use in subsequent studies (as will be shown) highlight its importance to modern location theory.

Unlike the  $p$ -Median and  $p$ -Center cases, the objective of Toregas' models is to enable coverage of an *entire* population within a pre-specified distance standard, to reduce the previously inevitable situation of some customers experiencing large travel times in the solutions. In other words, it intends to minimise the number of points required to provide a service to all other nodes on the network, where at least one service or facility is placed within the pre-specified distance standard,  $S$ , of the population it is covering.

This deterministic model makes a breakthrough into the coverage of a population; however, lacks foresight of demand. Population density, and therefore population at each node on the network is not considered by the traditional LSCP, leading to issues of unbalanced demand on servers. The numbers of servers located are often unrealistic and beyond the restrictions and resources of the problem modelled. The issues neglected by this primary insight of Toregas and of further developments in this area of location analysis are highlighted soon after by other researchers.

Work undertaken by Church and ReVelle (1974) and by White and Case (1974) essentially independently investigate the Maximal Covering Location Problem (MCLP). By maximising coverage of the population, rather than seeking to cover the entire population (as in the LSCP), the MCLP serves demand nodes within a predetermined service (time or distance) standard. The formulation considers a finite number of servers to be located optimally at candidate sites assigned to specific nodes on the network. The studies lend a popular structure for more modern models, including the ones presented in Chapter 6 of this thesis; a review of this class of covering models and their progression is given by Berman, Drezner and Krass (2010).

Limiting the real-world application of MCLP, is the absence of customer choice. Often a customer is able to choose the facility they attend or from which they receive service. The idea of 'choice' is a fairly recent interest and much literature (both for and against) exists with regards to patient choice in healthcare (Gallivan and Utley 2004, Knight et al. 2012b). In emergency services, choice is more likely to be deliberated by the operators, as opposed to the user. Silva and Serra (2008) take a different approach by incorporating directed choice within their queueing model. Control-room operators are able to deviate from an algorithm for selecting the closest available and appropriate server. This line of research is not discussed further here, despite its prevalence in modern literature, since for any emergency service, choice of whom to serve or by whom to be served would be unethical (French and Casali 2008).

Another limitation of the mentioned models (common also with many other models), can be seen by the lack of incorporation of server availability – deterministic models of this type have no link to congestion of the system.

An important extension to the MCLP is introduced in the late 1980's. The Maximum Availability Location Problem, MALP (ReVelle and Hogan 1989), can simply be thought of as a probabilistic version of the MCLP. Essentially, it strives to incorporate congestion – the chance of finding a server unavailable when required for service. Two versions are applied to Baltimore City:

- **MALP I:** makes use of the assumption of equal busy fractions across all regions in the city.
- **MALP II:** relaxes the assumption of uniform busy fractions and instead employs area specific busy fractions calculated from local estimates.

A summary of the results suggests MALP II provides better and more dispersed coverage in the solution for locating  $p$  servers than the first version. The authors were aware that solutions for location from other models do not always realise the requirements of the problem, particularly for emergency response systems. When rates of incoming emergency calls are large for a particular region, the demand on servers positioned in that region may in some cases prevent coverage. Where congestion effects occur, ReVelle and Hogan recommend MALP to ensure reliable approximation of utilisation.

So far, the *position* of resources has been considered; a method for finding the *number* of facilities required to be located is described by Neebe (1988). The intention is to keep the maximum distance travelled from demand to the facility less than some distance,  $S$ ; in particular, a range of emergency facility quantities are explored for maximum distance values (including single facility networks). Earlier studies only consider standard values of  $S$  for a specific number of servers  $p$ . Even small changes in distance or time standards may greatly affect the overall amount of facilities necessary. This is an important consideration and is vital to the optimal working of any network.

The problems discussed in this section utilise a number of OR techniques in finding solutions; the solution approaches available for location problems are now discussed, and literature demonstrating implementation is featured.

## 3.4 Solution Approaches

### 3.4.1 Mathematical Programming

Where the solution space to a problem is perhaps too large to attempt a full enumeration approach, mathematical programming, in particular, integer programming, can assist in discovering potential suitable solutions.

Often in application to EMS problems, mathematical programming approach solutions carry weaknesses (Goldberg and Paz 1991), such as:

1. deterministic travel time assumptions;
2. equal utilisation of all vehicles;
3. "a priori distribution for primary and secondary service is known";
4. independent service time of location.

Despite the limitations, some models show great realistic efforts in EMS application. For example, interruption of low-priority calls to allow resources to attend higher priority ones can be incorporated to the integer programs, for use in real-time decision support and disaster management systems (Majzoubi et al. 2012).

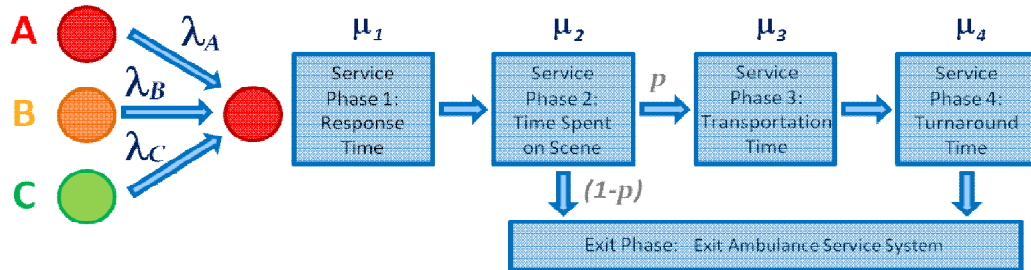
In Chapter 6, developed integer models capture the location dependence of service time by calculating all possible route travel times individually. Goldberg and Paz similarly develop models with stochastic travel times, consideration to variation in service time by location and estimated vehicle utilisation per station.

### 3.4.2 Queueing Theory

Modelling of emergency systems is a complex problem. Some efforts represent this reality solely using location analysis; however, it is understandable that queueing theory lends itself well to the situation when the structure of the system is considered (Figure 3.2). Priority queueing theory in particular fits with incoming calls for service, where emergencies are assigned a classification code for service order upon receipt. The phases that patients pass through during an emergency service



can also be represented using techniques from stochastic processes such as Markov Chains (Alanis et al. 2013) and simulation (section 3.5.7), as can server states and their locations.



**Figure 3.2** EMS system structured as a priority queueing system

The development and use of queueing theory within location analysis intends to encourage the production of fast solutions to deal with resource allocation and capacity issues on a network. Advances in location analysis seek the probabilistic location of vehicles on a network. Stochastic models allow the uncertainty of the arrivals of demand for emergency services and also probabilities of travel times and service distributions to be incorporated into the representation.

Larson's Hypercube Queueing Model (1974) (a type of 'Queueing Descriptive Model') is a cornerstone study, providing a bridge between queueing theory and location analysis. Its unique research direction proves to be invaluable in the course taken by subsequent investigations.

The Hypercube model considers server congestion rather than just coverage for siting  $N$  emergency service vehicles. A geographic region is split into 'cells', a concept explained in further detail in Chapter 5, where a (possibly non-symmetric) matrix represents the inter-cell travel times. The problem is then modelled using a continuous time Markov process to account for server availability. By tracking the state of mobile servers in a congested system, where servers in individual regions are treated independently, a solution to the steady state distribution of busy fractions is found.

The model was developed in conjunction with application to the problem of police patrol vehicle routing since its dynamic nature allows assessment of the quality of the decisions and adjusts to improve the solution.

Interestingly, the Hypercube was developed at a similar time to the Set Covering models and yet its advancements are made in the area of probabilistic, rather than deterministic, modelling. The substantial success of the Hypercube model likely comes from the combination of location analysis with queueing theory and the allowed interaction between the multiple servers on the network.

For bigger problems, involving a large number of servers, heuristics are suggested as a solution option in an approximation to the Hypercube Queueing Model (Berman et al. 1987).

In an emergency recovery tow vehicle study conducted in San Francisco Bay (Geroliminis et al. 2006), servers must be located to cover demand whilst keeping response times small. This is similar to the EMS vehicle location problem; however, this is a case of a solely urban district model, looking at highway breakdowns achieved through the Generalised Hypercube Queueing Model (GHM). Consideration is given to server availability whilst varying the number of incidents occurring in a time instant. Results are compared to the MCLP and  $p$ -Median since the GHM is a similarly structured coverage model, and finds that GHM can perform better in situations where demand rates are high.

One weakness is the exponential service time assumption of the Hypercube approximation. Jarvis (1975) develops the Mean Service Calibration (MSC), so that the Hypercube models can include location dependent service times. The MSC process follows the form:

1. set mean service time estimate of a vehicle at station  $j$  to average service time of the region;
2. use current estimates for mean service time, evaluate the model being used to gain probabilities that ambulance from  $j$  serves demand node  $i$ ;
3. use serving probabilities for each station from 2. to derive a new estimate for average service time for each vehicle;
4. if new estimates are close to current estimates of average service time, stop, else replace average service estimates by current estimates and repeat from step 2.

The next step in bridging the theory of location analysis and stochastic processes leads to the Queueing Probabilistic Location Set-Covering Problem (Marianov and ReVelle 1994). The Q-PLSCP is an extension of the PLSCP (ReVelle and Hogan 1988) where the necessity of this new probabilistic model comes from the implication of busy servers on the network. In the PLSCP, coverage is one of the objective constraints, whereas the Q-PLSCP instigates an availability

constraint of  $\alpha$  reliability that has the aim of minimising the number of servers required to cover demand. All previous research including the notion of availability assumes independent server availabilities. The Q-PLSCP however, relaxes this assumption, allowing dependence of all servers by utilising an  $M/M/s$ -loss queueing system for each region. Not only is this dependence of availability included, but it is also applied to neighbourhood-specific busy fractions, rather than the network as a whole. The results illustrate the need for the queueing aspect of the PLSCP, particularly where availability required is high, as in emergency service systems. Yet, in cases where only quite low availability is necessary, the Q-PLSCP often overestimates congestion due to the method for deriving the minimum number of servers to place within the travel standard of any node.

Following their efforts of the Q-PLSCP, the authors also attempt to solve the probabilistic MALP. Using linear programming combined with queueing theory, the Queueing Maximal Availability Location Problem, QMALP, was produced (Marianov and ReVelle 1996). The formulation follows that of the MALP, but again with the assumption of independence of busy servers relaxed. The model's objective is to maximise the (demand-weighted) population covered by the emergency service, where server availability has reliability  $\alpha$  and probabilities of servers contained in the same region being busy depend on each other. The authors claim it is the first example of queueing theory explicitly applied to MALP server busy fractions. Availability and reliability for EMS is examined further in a paper by Erkut, Ingolfsson and Budge (2008a).

The Priority Queueing Covering Location Problem, PQCLP (Silva and Serra 2008), provides an illustrative example where two priorities of emergency are considered. From the combination of maximal covering location models and queueing theory in its structure, the PQCLP also utilises heuristics to solve the problem of relating population demand with the allocation of resources.

Geroliminis et al. (2009) use queueing to locate and district emergency service vehicles with location dependent service times and non-identical server service rates. For freeway service patrol vehicles, good results were obtained for high demand periods.

Queueing theory itself may also be used as a starting point for determining a fleet capacity, before offering the findings to other more complex and realistic modelling techniques (Henderson and Mason 1999, Jenkins 2012).

Emergency systems, as has been seen, can be modelled using numerous location analysis and queueing theory solutions developed for siting resources. Such techniques endeavour to improve positioning in order to maximise response times and improve reliability of availability. The incorporation of queueing theory produces descriptive methods for solving; however, with the addition of other processes into queueing theory, the entire service operation can be addressed.

Erlang-loss models have been used to capture the situation of an EMS scenario whereby demand must be met, finding promising allocations of vehicles (Restrepo et al. 2009), with results showing good performance and accuracy even compared with the Hypercube Queueing Model.

Phase-type service distributions are described in the literature for non-Markovian queueing representations of ambulance responses to emergency calls and the situation of priority modelling. With regards to an ambulance service, the response time is only a portion of the overall service time, which intuitively leads to the possibility of using phase-type processes to model the reality of an EMS vehicle responding to and serving an emergency call.

A software package design integrating a method for modelling queueing systems with priorities and phase-type approximations of service time distributions, such as in an EMS system, is presented by Mickevicius and Valakevicius (2006). Another application of this theory is to a numerical model of quality control. In a similar vein, Valakevicius (2007) demonstrates a similar problem highlighting the use of the phase-type distribution for service phases of a priority queueing system.

### **3.4.3 Multi-Objective Modelling**

In Daskin's paper 'What you should know about location modelling' (2008), readers are informed that for multi-objective models, where only one of the objectives is solved optimally, it is likely to give poor results with respect to the other objectives. The survey paper looks at tradeoffs between two objective constraints – the average and maximum distances from located facilities on a network to demand points.

By incorporating multiple objectives and equity in busy servers, analytical models (utilising the Hypercube model) are applied to Boston EMS in order to evaluate deployment strategies and

develop a resource allocation system (Hill III et al. 1984). The novel aspect of their work is the ability to consider multiple objectives and their interactions simultaneously:

1. minimise average response time in line with an average seven minute target;
2. minimise inequity in vehicle availability.

Comparing a week's activity for the EMS logic dispatch rules with a week using rules and locations generated by the model, results showed a 30 second reduction in average response time. Adding another vehicle to the fleet would achieve the same improvement but at a quoted cost of \$150,000.

Application of location theory models and solutions to emergency services and emergency medical services has already been demonstrated, but a closer and more specific investigation at the contribution location analysis has in this field follows in the next section.

## **3.5 Location Analysis for Emergency Services**

### **3.5.1 Introduction**

In 2012, it was reported (BBC News 2012a) that, of all UK emergency calls:

- 52% are connected to the police;
- 41% are for emergency medical services;
- 6% go through to the fire brigade;
- and a mere 1% are for the coastguard and cave and mountain rescue services.

Quite often, particularly in America and Canada, the fire service is used in conjunction with the EMS. In fact, the fire service may be given instruction to dispatch to a medical emergency before an ambulance. Planning therefore can sometimes incorporate both types of fleet (Jewkes 2011, Knight et al. 2012a, Monroe 1980) and must accommodate the various objectives and roles each service plays, as well as the limitations around allocations.

Despite many simplifying assumptions inevitable when modelling reality, coverage and average response time models have been widely used for emergency service systems. The objective often being to find the fraction of emergency calls reached within the given time standard set in line with

the services' targets. Such work was conducted by Chaiken and Larson (1972) for emergency services in urban regions.

Marianov and ReVelle (1995) provide a critique of location covering models mainly applied to emergency services. An exploration into recent OR work in EMS planning and management by Ingolfsson (2013) reviews the problems associated with forecasting demand, response upon survival and utilisation on healthcare policy.

### **3.5.2 Dispatching**

Following the introduction by Larson (1974) of location application to police services, other emergency service location problems become more evident in the literature. One of the earliest appearances of ambulance service problems in the OR literature occurs through the discussion of multiple vehicle dispatches. For emergencies where more than one EMS vehicle may be required, Daskin and Haghani (1984) analyse arrival times at the scene for the first vehicle to reach the patient. In much of the emergency service research, arcs of the nodal network are thought to be representative of the road network available to the vehicles. Travel times are represented by distributions so arrival times at the scene of the incident can be modelled. Where more than one vehicle is dispatched, the important information is contained in the length of time it takes for one of the vehicles, not necessarily the first dispatched, to reach the patient. Many of the studies before this point in time assume deterministic travel times in order to handle the matter of location and also presume single server requirements. Daskin and Haghani notice the problem with the common assumption of closest single server availability in providing the best response to the emergency incident.

### **3.5.3 Equity in Access**

Equity in location analysis is a widely studied problem in its own right, especially with regards to healthcare and emergency intervention. Although not much detail on this topic is provided here it is important to highlight additional issues surrounding location research and equity of service provision.

Equitable service to each neighbourhood in a region is a prime concern and professional department managers may also specify the need for a good level of service at minimal cost (Hill III et al. 1984). The Ambulance Allocation Capacity Model (AACM) (Shiah and Chen 2007) was developed to address the capacity capabilities of the ambulance service of Taichung City, Taiwan, in particular, the inability to provide equity to the population with the current system operations. By combining probabilistic and deterministic methods used throughout location analysis, and applying them to multiple coverage decisions, equity can be incorporated to the system capacity design.

With regards to the progression seen into survival maximising research, Felder and Brinkman (2002) provide an equal access approach to EMS planning. Equal access highlights the need for an equity-efficiency trade-off. Response time is recognised to affect both the quantity and quality of saved lives by a service. Consideration of costs to the service when basing policies on equity often implies that the value of saving lives differs for urban and rural areas. To ensure patients perceive there to be equitable service, Chanta et al. (2011) provide the  $p$ -envy model to locate emergency resources.

#### **3.5.4 Travel Times & GIS**

Travel times across boundaries of sub-regions of a network have commonly been considered negligible, deeming neighbourhoods fixed and distinct and operation of servers in the different regions independent. Examples of models that make this assumption include mathematical programs such as MALP and queueing extensions, for example QMALP.

For all types of emergency service system, travel time is one of the most crucial components of a location model objective function since performance is based on the ability to respond quickly, and not just coverage. It is therefore important to realise travel may not be symmetric on a network, and that servers may be required to provide back-up coverage for neighbouring regions or nodes.

Geographical Information Systems (GIS) can be utilised by studies situating emergency service vehicles. Single line road maps (Shiah and Chen 2007) and street map databases (Azizan et al. 2012) can be used to site ambulance service vehicles within optimisation and simulation solutions.

Rural areas are known to be more problematic to EMS managers than urban districts when it comes to meeting response time targets. Some ambulance trusts operate with Global Positioning Systems

(GPS), allowing fastest routes to incidents to be found, and for vehicles to be tracked during a service, enabling subsequent analysis of operations. Gonzalez et al. (2009) found that in a trial where GPS units were introduced to an EMS provider operating in a rural area, after one year, response time to certain call types was improved compared with service from a previous year.

Within this study, travel times feature heavily in the development of new models, and so a detailed discussion is provided in Chapter 5, considering in particular the problem of rural demography.

### 3.5.5 Utilisation

Having access to the probability that each service vehicle is busy allows a modeller to determine, of these vehicles, which has the highest probability of serving a particular call from a certain demand node, and so overall performance of the system for a given set of allocations.

Calculation of these busy probabilities in earlier studies was usually taken to be the average utilisation of the system. This however, is not an accurate representation of the operations and leads to an inaccurate measure of performance and success. Instead, many researchers have attempted to find better ways of estimating such busy probabilities. Persse et al. (2003) evaluate vehicle utilisation by taking hourly levels as the calculation of:

$$\frac{(\text{number of transports} \times \text{average busy time of a vehicle serving})}{\text{total time the vehicle is on duty}}$$

This approach is similar to the one currently adopted by WAST, employed in simple demand forecasts. Although these are simple calculation estimates, often more intricate vehicle busy probabilities equations are used. A study by Goldberg and Szidarovszky (1991) develops an iterative method for solving non-linear versions of such equations.

### 3.5.6 Dynamic Modelling

Another multi-objective concept for ambulance modelling is where the response time is minimised whilst the system simultaneously aims to be best prepared to respond to future calls. A dynamic approach is necessary.



Fairly recent attempts have been made into the dynamic solution of the *location* problem, although the dynamic ambulance *allocation* problem remains relatively little studied compared to the static problem (Brailsford and Harper 2007). Dynamic models account for the relocation and reallocation aspects and the various comprehensive characteristics of any emergency service response procedure.

Automatic ambulance dispatching in Sweden was addressed in a study into dynamic ambulance relocation, creating the DYNAROC algorithm (Andersson and Varbrand 2007). The introduction of 'preparedness' to emergency service literature helps improve service modelling and reduces patient waiting lengths. Preparedness is the concept of increasing the number of operational ambulances in a region to allow coverage of the population by multiple vehicles, possibly at multiple locations, so that during a time when the primary assigned vehicles is attending a call, any future calls for service originating in a similar region will still be covered.

A model designed to forecast EMS call volumes (Setzler et al. 2009) – mentioned earlier – also addresses the differences between real-time repositioning (the fleet is relocated after one vehicle is dispatched) and dynamic deployment (using forecasts to anticipate fleet positions based on expected demand). Saydam et al. (2013) extend the dynamic coverage models to account not only for spatially and temporally dependent demand, but balancing the amount of repositioning to limit affect on crews.

Owen and Daskin (1998) offer a survey paper of dynamic contributions. Brotcorne et al. (2003) provide a thorough survey of location and relocation models, spanning from early deterministic efforts, to the developments of probabilistic inputs from queueing theory, and to the advancements of dynamic modelling.

### **3.5.7 Simulation for EMS**

A popular technique used by Operational Researchers and Management Scientists for practical problem solving is simulation. Its use in an EMS environment is not novel, and many previous studies demonstrate the success that can come from suggested system set-up implementation. Due to the complexities of an EMS system, analytical modelling is often not robust enough for thorough investigation of full scenarios (Monroe 1980). Simulation provides the playground in which researchers and decision makers can witness cause and effect on a system. Procedure changes can

be suggested that may increase efficiency and performance. Models are used to determine these, capturing all the important aspects of a system and replicating its inner workings without forcing generalised assumptions where undesirable.

Using discrete-event simulation, Wu and Hwang (2009) investigate the threshold of expansion of an ambulance service fleet in order to cope with ever increasing demand and vehicle availability issues (not considering allocation). This study is one of very few that considers the effect of dispatch strategies on response time and so subsequently on availability. Initially, the closest available server is chosen for dispatch, but other options include: maintaining preparedness through repositioning vehicles after each deployment (dynamic relocation), or, sending the second closest resource in a forecasting approximation attempt.

Although entirely theoretical in development, this research was later applied to an EMS data set to see the impact operations and changes have on the system; however, due to limitations in the data and development of the model using only literature, it seems likely that its accuracy would not necessarily be sufficient for application to any other real-world service, without extensive validation. An important strength however, is the time dependent and spatial distribution inclusion for demand. The model also bases the decision for which hospital to transfer the patient on destination probabilities rather than the usual closest facility rule.

Simulation is thought to be an under-utilised (supported by Figure 3.1) but powerful tool for emergency service planning. For improved communication of OR modelling to EMS managers, Henderson and Mason develop an in-depth simulation visualisation tool, designed specifically as a tool for emergency decision process (2004). The trace-driven discrete event simulation recognises the necessity of predicted time-dependent travel times, GIS and a historical data feed. The inclusion of all these aspects, whilst being a substantial advantage of the tool, is also a direct impracticality when considering computation time. Trace-driven however, allows the model to feed directly off the intricacies of the data and avoids errors in sampling (spatially and temporally) that may occur with theoretical structures. It allows also, the small number of multiple-response incidents to be captured, which is fairly unique to this model, even if limited for more long-term planning. The authors acknowledge "fairness" versus "efficiency" (equity) issue within such a service in relation to position and capacity of demanded resources.

### 3.6 Around the World in 999

The EMS location-allocation problem has been studied world-wide. Every country, even each ambulance service, has their own operational targets, design and commitments. An overview of the differences in strategies due to regional demographics is now discussed, highlighting the problems faced by many services and the differences they must combat compared with their neighbours.

Beginning in Wales, it has already been noted that the ambulance service is a single system operating over the entire country. Although a small country, and for the most part sparsely populated, the service is regularly over-utilised resulting in poorer performance than England. Certain regions of Australia witness a similar effect. Entire populations are left with little EMS cover, meaning when emergencies do arise, patients wait long periods of time for critical responders, increasing mortality (Fitch 2005).

Efforts are currently being made into optimally locating public-access defibrillators in Canada (Chan et al. 2013). This is particularly important in countries like Canada and Australia – where large sparse expanses exist – in order to increase population survival.

Combining the previously discussed MALP and Q-PLSCP, Harwood (2002) adopts a multi-objective approach to deploying Barbados EMS vehicles. Due to the geography and demography of the island, the deployment of ambulances when another server is busy may need to suit more than one objective. It is likely desirable to maximise coverage of the population within a pre-specified distance (or time) standard with  $\alpha$  reliability, whilst for this particular scenario, it is also attractive to locate vehicles at sites which will minimise the cost of coverage. Although a typical public sector problem, with the inclusion of costs in location, the solution here becomes more like those developed for private-sector challenges.

A Chilean case study of the ambulance service introduces Key Performance Indicators (KPI) for both patients and ambulance service managers (Singer and Donoso 2008). These KPI are used in conjunction with an  $M/M/s$ -loss queueing system to assess the effects of changes of parameters in the model. Focus is given to the time dependency of the optimal geographical coverage solution for the population. Patient outcome is also considered – where long term effects on *chronic* patients is compared to *emergency* patient groups.

In Hong Kong, a lack of mainstream prioritisation standards results in all requests receiving an immediate response, despite under-performance.

Simulation modelling approaches in New Zealand have already been mentioned (section 3.5.7); however researchers here are among the leaders in simulation implementation for EMS systems. Many recent publications, both academic and consultancy led, have made valuable contributions to the progression of this OR field. Henderson and Mason (2000) developed BARTSIM in order to balance “political, economic and medical objectives” at operational, strategic and tactical levels and to answer a number of questions faced daily by ambulance service staff. Success stories exist also in the real-world implementation for a suite of models, part of the ‘Optima’ emergency service optimisation technology brand (Optima 2013).

In America, a range of location-allocation and EMS system improvement approaches are utilised. Some use tiered systems for deployment, however, others use uniform strategies (Persse et al. 2003) with varying results in survival studies. Tavakoli and Lightner (2004) develops a mathematical model that simultaneously optimises a given number of facilities and a set number of vehicles at the chosen locations.

### **3.7 Generic Modelling**

Many of these previous research attempts have underlying theoretical similarities; yet it is rare for a new study to adopt an existing model and attempt to alter it, especially in simulation, more likely a brand new model is developed. If existing research is utilised, progress could be enhanced more quickly and combined efforts are likely to lead to better long term results.

Hoping to combat the problem of unnecessary development of new models, Hillsman (1984) identifies similarities and defines desirable location problem structures in a generalised computer software package, known as the Unified Linear Model (ULM). This robust model can be adapted to suit different objective functions (built on the generalised solution for the p-Median problem) and so has the ability to derive solutions to various special cases of the original Median question, the LSCM and their extensions.

### 3.8 Model Limitations

Simulation model validation in Swoveland et al.'s (1973) paper suggested models often could have benefited from the collection of more data and further analysis, particularly with regards to demand variation (and possibly seasonality).

The suitability of performance measures of many earlier works is discussed by Erkut et al. (2008b) who claim unrealistic outcomes of early location models. Their main critique is that of the performance measure most commonly used in location literature – namely ambulance service response time proportions given a time standard. Many maximum availability or reliability location studies however, choose to look at coverage. Erkut et al. state that to their knowledge, the approach of coverage “has not been put to the test of real world relevance”. Whilst they recognise the importance of such models, and the contribution they have made to the field, many set-covering models do not fully capture the difficulties in locating EMS vehicles. Coverage of all demand points is unrealistic, particularly in rural regions, such as are seen in Wales. The ambulance service is unlikely to attempt to locate standby vehicles in order to reach all rural areas within their target, instead positioning vehicles where calls are more likely to originate based on historical occurrence, but still with reasonable accessibility to rural communities. The authors also point out that many ambulance service targets are actually system wide, not for individual sectors; therefore coverage at the target level for all neighbourhoods is inappropriate and unnecessary.

Budge et al. (2009) note the four other limiting assumptions commonly made in location problem approximation methods:

- the number of vehicles per station is generally taken to be 1;
- average workload is taken for the whole system rather than utilisation per station;
- average service time is often assumed independent of location and responding vehicle;
- server cooperation is regularly ignored, such that either all vehicles are equally likely to respond, or neighbouring stations operate completely independently.

The authors demonstrate a model where dispatch probabilities for individual servers from each vehicle station are provided. Using these and station-specific service times, an approximation for system wide busy fractions can be obtained, leading to a system modelled for ambulance allocation *and* utilisation.

### **3.9 Summary**

Further exploration of existing research and suggestions of potential research directions in queueing theory, location analysis and simulation for EMS is on the increase. Popular areas considered for future investigations (surrounding the discussions of this chapter) include specific location-allocation techniques – stochasticity and location-routing – and explicit designs of queueing theory – phase-type service distributions and priority assignments.

There are many other location problem aspects and vehicle-routing scenarios where the existing theory cannot be used realistically in application, if at all, due to the complexity of the real-world problem. Geographical representations and dependencies may not be captured thoroughly enough using computer based interpretation. Distributions of population often create theoretical problems. Some more advanced algorithms can become almost impossible to solve (at least in real-time) once all constraints of the application are considered.

From the progress already seen in the field, further success is inevitable and study into location problems persists in being a foremost focus in the OR community. Implementing academic research in public and private sector organisations is the more difficult task. Hill III et al. (1984) note that, to EMS managers, the “credibility and applicability” of the model designed is very important. Difficulties lie not only in the convincing of the integrity of a model, but often also in simultaneously pleasing managers from both civic and political backgrounds with differing motivations.

An ambulance service cannot be seen solely as a transportation service. Following the medical aspect literature review for EMS improvement problems, it is important to recognise that an EMS system should be treated as a provider of medical care in their own right within future studies, linking to the considerations of patient survival. Medical based studies already make this recognition, but more mathematical studies may ignore this fact for simplicity or make simplifying assumptions, looking exclusively at transportation and response.

The EMS is often the first point of contact for emergency patients and lengthy on scene services are not necessarily an undesirable system feature. It is possible that treatment may be administered at the scene, even where data cannot yet explicitly acknowledge the impact of this procedure. If ambulance trusts and local hospitals were able to communicate more directly, it may be that certain medicinal procedures – as in Studnek et al. (2010), where treatment is required to be administered

quickly – are the best implementation option. In these cases, a longer cycle time with the ambulance service may have a more successful patient outcome than immediate conveyance.

The implications of such policy changes and the direct affect of operational alterations are further investigated in the following chapters of this research project. With application to the South East Wales EMS system, an allocation problem is explored through mathematical programming and simulation (integrating simple queueing theory) techniques. The research results in the development of new generic models with the intention of providing insight to both academics and healthcare professionals.

## Chapter 4

# WAST: Data Analysis

### 4.1 Introduction

Wales, although occupied by a relatively small population considering its size, witnesses a substantial annual number of medical emergencies, which appear to be increasing year on year (Vile et al. 2012). Over 274,000 data records were provided for this study by WAST, covering the twelve month period of 1<sup>st</sup> January to 31<sup>st</sup> December 2009. The data set refers to approximately 175,000 unique emergency incidents originating throughout the South East region of Wales only, spanning 50 postcode districts (a *district* is represented by the first three or four characters of a UK postcode and the first two characters give the postcode *area*). Reasons for the focus on the South East of the country were discussed in Chapter 2, section 2.3.7.

Since one objective of the research presented in this thesis is to provide WAST with planning tools that may be used by the Trust in decision making, obtaining relevant real-world data is imperative to the design and ultimate implementation of any developed models. Without such detailed data, the models built may be subject to inaccuracies and may be difficult to validate. The results for the region under scrutiny need to be of use practically to the Trust, and application to other EMS systems enabled through adaption of the provided input data.

The data analysis that follows refers only to emergency calls, allowing conclusions to be made and further details obtained surrounding the emergency operations of the ambulance service in Wales, highlighting the more troublesome areas of service. The empirical information gained is used to supply details to location analysis models (Chapter 6) and a discrete event simulation (Chapter 7).

### 4.2 The Data Set

#### 4.2.1 Statement of Accuracy

Although it is possible that the accuracy in the recording of data by WAST may be imperfect (due to human and technological error in recording and logging of incident details), it is assumed that for



the purposes of this research, the level of accuracy is adequate. Since WAST uses the same data as provided for this study in their own performance analyses, any results obtained later in this thesis may be thought to be comparable with conclusions derived from this work, in accordance with the Trust's own decision making assumptions where information on these is known.

#### **4.2.2 Influences**

Throughout 2009, there was a global pandemic, commonly referred to as 'Swine Flu'. This strain of influenza virus infected over 540,000 members of the English population (Donaldson et al. 2009) and the NHS was under pressure to treat and respond to these patients as well as continue to provide as efficient a regular service as possible. NHS Direct reported an increase in the percentage of calls witnessed relating to influenza; since this peaked during summer months when typical seasonal (winter) flu calls are low, the higher call rate can be accredited to the pandemic (Public Health Wales 2010).

It is likely that certain effects of the pandemic will be evident in the data collected by WAST – a slight bias in normal operations due to an increase in demand – but this effect was minimised by the excellent additional services the NHS provided exclusively for Swine Flu outbreaks, including a collaboration scheme with NHS Direct (HPA 2009). Websites and campaigns also aimed to provide information and treatment of the condition within the community (Owen 2009); only critical cases (of which there were a few) would have required paramedic response.

Such widespread crisis situations are not uncommon scenarios for any emergency service and one WAST would possibly have to contend with similarly in the future, so no efforts are made in this study to eradicate the (likely small) contributions of the pandemic to the data set.

Generic modelling solutions provided in this thesis concern daily planning and operational procedures and not predictive or forecast modelling (which would in any case not be based on solely a year's worth of data). Any influence from a cause of increased demand during the year should not reduce the suitability of data analysis conclusions when applied to the development of modelling tools, maintaining their reliability and credibility.

### 4.2.3 Dispatching

The disparity in the total number of records provided by WAST (274,300) and the number of unique calls for service witnessed in the data set (174,665) is due to the fact that multiple vehicles are frequently dispatched to an incident. For larger incidents, perhaps several crews are necessary to deal with the scale of injuries, or in some cases assorted vehicle and crew types are required.

For a category A emergency only, in 2009, the Trust's policy was to dispatch an RRV (if available) to attend the scene as quickly as possible and for the paramedic to stabilise the patient, but with an EA dispatched simultaneously as follow-up, enabling conveyance to hospital if deemed necessary when on scene. In following this protocol, the dispatcher assigns multiple vehicles within a similar proximity to the incident in the first dispatch instance, often without knowing enough about the nature of the emergency to justify such a demand on resources. This approach however, does give high acuity patients better chances of survival through swift responses. In mathematical terms, this may be thought of as a variance reduction technique by the controllers, to give the best probability of achieving the minimum possible response time for each incident. The procedure is commonly known as a 'double dispatch' and is currently one of which WAST are trying to reduce unnecessary occurrences (WAST March 2011).

Note that, for incidents with a multiple or double dispatch, the additional vehicles do not always reach the scene; they may not be required, or they may be reallocated to another call or cancelled before on scene attendance. This is known as 'stepping down'.

Even though policy dictates that RRVs be used for the high priority patients only, since their power is their speed in response, it is seen in the data that RRVs are also dispatched to lower priority calls on some occasions. Table 4.1 shows the proportion of calls that witness an EA and RRV response (amongst other combinations), and show that on nearly half the occasions, a lone EA serves category A calls; presumably, sometimes RRVs are also dispatched but step down, and other times, an RRV may simply not be available.

**Table 4.1** Percentage of service occurrences for all vehicle combinations per category

Category:	A	B	C	AS2	AS3
1 EA Only	48.11	67.08	82.50	96.64	95.27
2 EAs	0.58	0.57	0.48	0.99	3.04
1 RRV Only	6.89	8.56	6.47	2.15	1.62
2 RRVs	0.35	0.29	0.24	0.01	0.00
1 RRV + 1 EA	43.08	22.66	10.13	0.17	0.07
3 Vehicles	0.91	0.66	0.16	0.01	0.00
Other	0.08	0.18	0.02	0.03	0.00

#### 4.2.4 Variables and Field Headers

In order to evaluate their own performance, WAST record time stamp data for many events that occur during the service of individual emergency incidents. The data set received contains 24 variables. Many of these record a time stamp for an event, i.e. the start or end time of a phase of service. Other fields specify information regarding the nature of the emergency, locations and resource details.

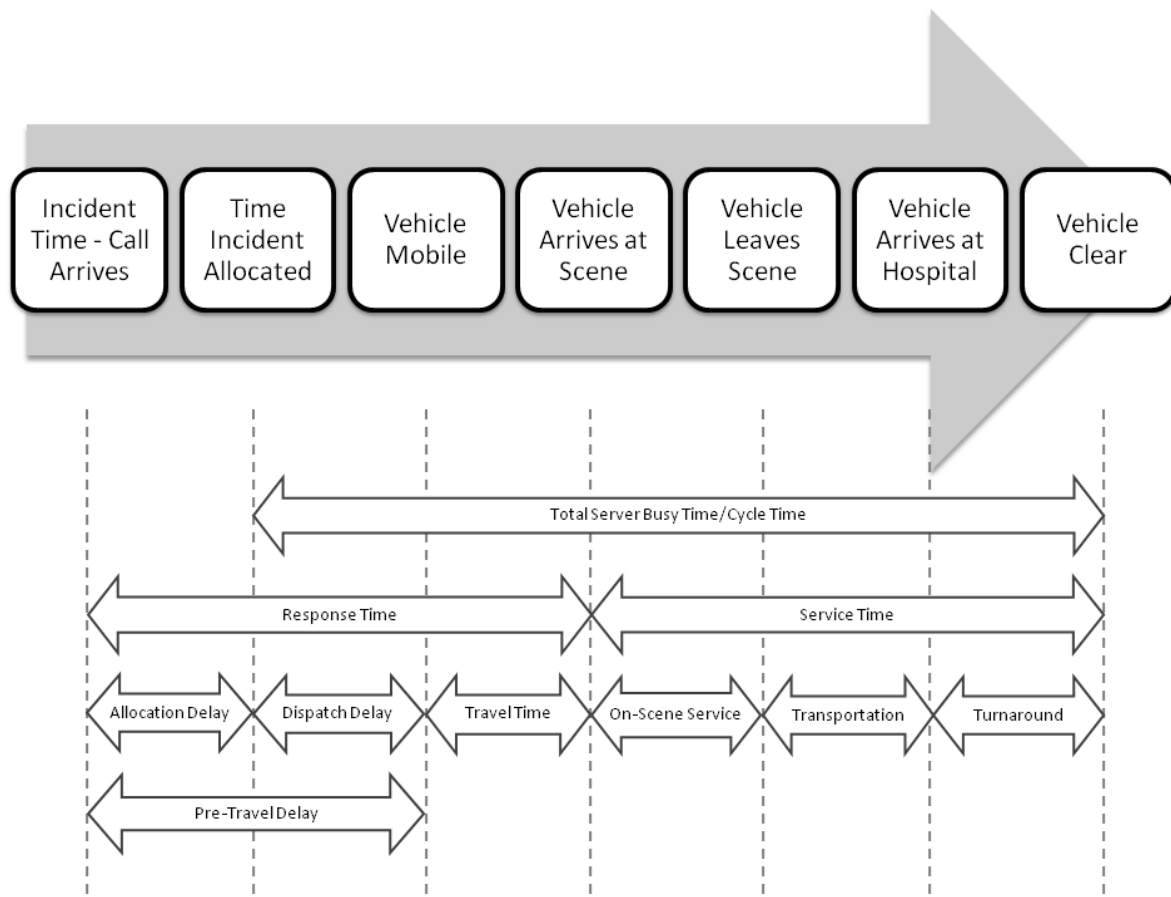
A selection of fields present in the WAST data set is given in Appendix 4.1, with comprehensive definitions of the variables' attributes.

The reasons for cancellation of service before completion by a particular vehicle (stepping down), are quite varied. They range from an error entered in the call log, to a hoax or cancelled call, or to the transfer of the patient to another emergency service (police or fire) or to another emergency medical service provider. In some cases no patient was found at the scene, or the patient took alternative transport to hospital before the EMS vehicles could attend. Due to the focus of this study and time limitations, these reasons are not investigated any further at this stage, although they provide an interest for future research opportunities.

### 4.2.5 Pathway

Although much of the EMS process is widely known, or is disclosed through communications with trusts and their publications, real-world data provide an opportunity for further insight to the Welsh operational practices. A process-flow diagram displayed in Figure 4.1 summarises potential areas of understanding gained from the data set.

Service phases associated with the seven time stamps provided in the data set are portrayed over time, from which at least ten time intervals of interest can be derived as the difference between start times of sequential stamps. Pre-travel delay and response times can be calculated and so analysed based on existing data. From the variable list in Appendix 4.1 and from Figure 4.1 it can be seen that turnaround times may be analysed in general, but explicit handover time per incident is unable to be extracted from the data of this particular study.

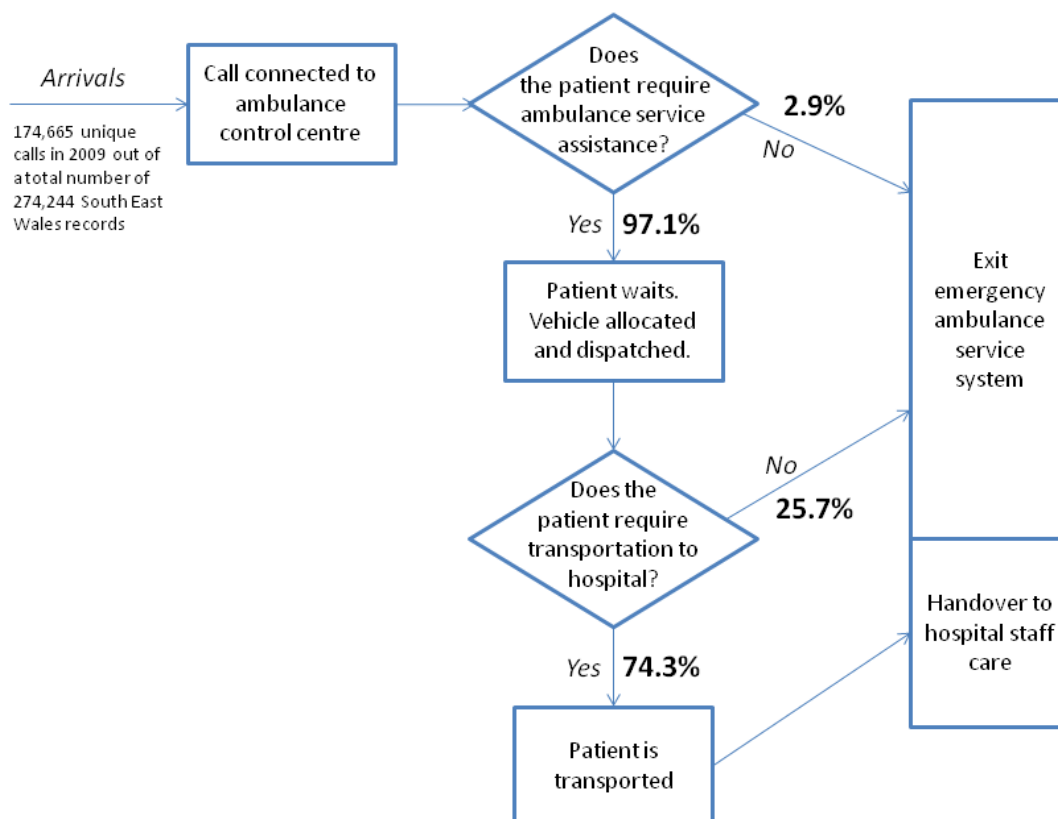


**Figure 4.1** Pathway through time of an emergency call with event time stamps and interval phases

Additional variables do also exist, and others can be computed using information from a combination of the original fields. These extras include:

- Division – based on information provided separately by WAST, detailing postcode districts located in each of the five locality divisions of the South East region;
- Category – can be determined using the AMPDS Priority and Incident Type fields (A, B, C, AS2, AS3);
- Service phase lengths – using start and end time of an event (examples given in Figure 4.1).

The data provided were fairly comprehensive but required some cleaning and organisation. Time stamps were not always chronological when compared with other time stamps of the same incident, implying error in the recording of some of the data. Checks were carried out to ensure no obvious errors or outliers exist in the data, resulting in the construction of Figure 4.2 for clarification of the number of records referring to each possible patient pathway for category A, B and C emergencies.



**Figure 4.2** Patient pathway through the system, developed during preliminary data investigation and cleaning for priority A, B and C patients

In total, only 56 records were removed (leaving a total of 274,244) due to infeasible lengths of some service phases – that is, where records were found to have negative On Scene, Transportation or Turnaround lengths. This number is relatively miniscule when compared with the original number of data records, and although tests and checks were still performed regularly during data procedures, it was not deemed a priority to continue an in depth search for further errors since such small numbers have minimal impact on the overall results obtained from any such analysis.

## 4.3 Preliminary Analysis

### 4.3.1 South East Structure

The Welsh population is divided into three regions: the North, Central & West and the South East. Operations are managed individually out of a control centre, one per region. Prior to October 2009, Wales was partitioned as 22 Local Health Boards (LHBs) within which seven different health trusts operated. The South East was home to 9 of the 22 health boards: Cardiff, Vale of Glamorgan, Merthyr, Caerphilly, Monmouthshire, Rhondda Cynon Taf, Blaenau Gwent, Torfaen and Newport. Wales' structure is now made up of a total of only seven LHBs providing all health care services (NHS Wales 2009). In the South East, where the population is largest, there exist the Aneurin Bevan, Cardiff & Vale University and Cwm Taf Health Boards (Figure 4.3).



**Figure 4.3** Local Health Boards of South East Wales (Health Maps Wales)

It is possible to define the coverage of the nine original South East health board localities of 2009 by the three new ones in the way outlined by Table 4.2; however, demand and dispatch in 2009 were based upon five geographical districts, with rosters (built using software called 'PROMIS') based on the original nine health board localities. Approximately 700 emergency calls are received by a total of six control room operators each day.

**Table 4.2** Locality assignment to South East Wales Health Boards

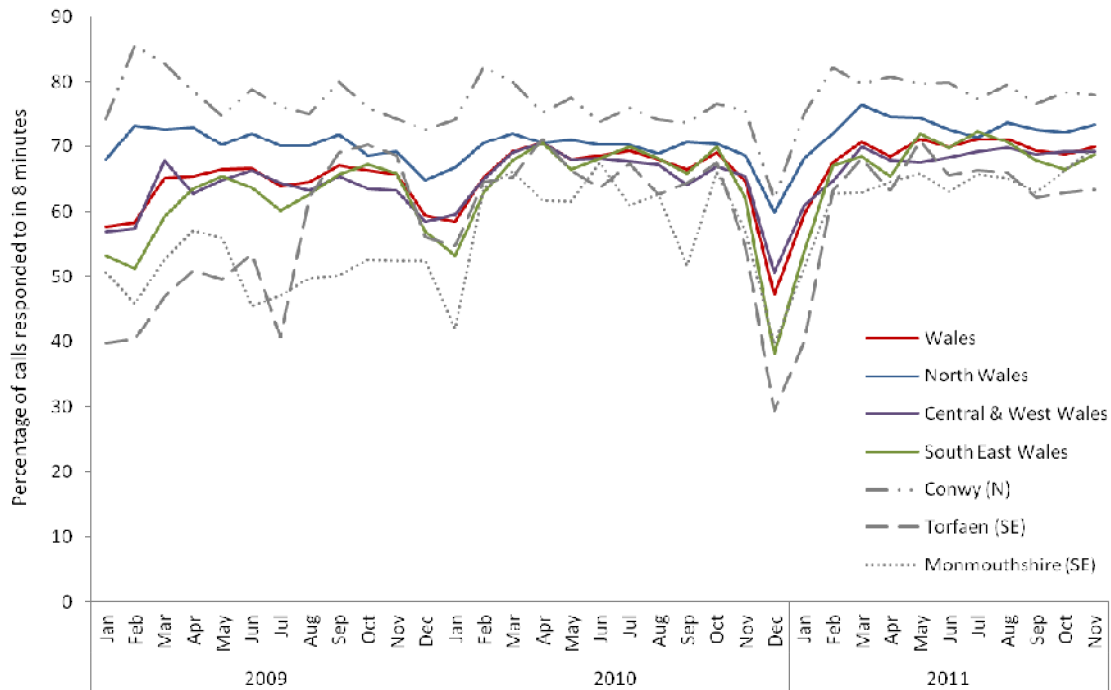
Health Board	Locality (pre October 2009)
Aneurin Bevan	Blaenau Gwent Caerphilly Monmouthshire Newport Torfaen
Cardiff & Vale	Cardiff Vale of Glamorgan
Cwm Taf	Merthyr Rhondda Cynon Taff

### 4.3.2 Demographics

The South East region has consistently struggled to meet the current response and turnaround time performance targets (Figure 4.4), particularly when compared with other areas of Wales and especially when considering that England and Scotland aim to operate to a 75% eight minute response target.

The region covers an area of approximately 2,559 km<sup>2</sup> and the Trust serves a population of 1.3 million in the three health boards (Welsh Government KAS 2009). Although mostly rural, the South East is also home to the capital city of Wales, Cardiff, and its close neighbour, the city of Newport. By national standards, the South East itself is not sparsely populated, unlike Powys (Wales Rural Observatory 2012), but its geography and position of rural valley towns compared to the higher demand areas of the urban population contribute to the large response times (and high variation) witnessed by residents of the region. Local EMS vehicles that are assigned to rural towns may end up being called to more densely populated areas at busy times, leaving fewer resources

available to serve rurally located incidents when they arise. Even if there are available vehicles, they may have been placed tactically closer to urban communities to deal with expected demand but resulting in higher response times to other areas.



**Figure 4.4** Summary of category A emergency response time performance, re-produced from statistical publications of performance by locality (Welsh Government HSA 2012)

As would be expected and supporting the earlier discussions of Chapter 2, atmospheric and seasonality conditions do appear to have an impact on performance. From Figure 4.4, many of the more severe dips in response performance, both in the South East and other areas of Wales, occur during winter months. November 2010 to January 2011 was a particularly harsh winter resulting in an upsurge in road traffic accidents, falls and influenza viruses by about 40% (BBC News 2011d). EMS systems across the UK struggled to provide their usual service due to this higher demand, terrible road conditions and congestion at hospital facilities (BBC News 2011b, c).



**Table 4.3** Station preference based on usage (frequency of emergency dispatches)

Station Code	Location	Percentage of all Dispatches	Cumulative Percentage	Frequency Order of Dispatch	EA Frequency Order	RRV Frequency Order	WASTs regular bases
1 SABW	Blackweir	15.19	15.19	1	1	1	●
2 SANP	Newport	5.90	21.09	2	4	4	●
3 SABA	Barry	5.48	26.57	3	7	3	●
4 SABG	Bassaleg	5.09	31.66	4	3	11	●
5 SACE	Cardiff East	4.94	36.61	5	2	20	●
6 SAAE	Aberdare	4.73	41.33	6	6	12	●
7 SABR	Bargoed	4.46	45.79	7	11	7	●
8 SAGI	Gelli	4.35	50.14	8	5	6	●
9 SSVF		4.23	54.37	9	70	21	
10 SAHN	Hawthorn	4.06	58.42	10	8	8	●
11 SAMR	Merthyr Tydfil	3.56	61.99	11	9	13	●
12 SAPO	Pontypool	3.37	65.36	12	16	5	●
13 SATR	Tredegar	2.84	68.21	13	21	9	●
14 SAAB	Aberbeeg	2.84	71.05	14	13	14	●
15 SABD	Blackwood	2.79	73.83	15	20	10	●
16 SACY	Caerphilly	2.77	76.60	16	14	15	●
17 SACB	Cowbridge	2.62	79.22	17	10	.	●
18 SSEF		2.34	81.55	18	.	2	
19 SACH	Parkwall	2.16	83.72	19	22	16	●
20 SAPC		1.91	85.62	20	12	18	
21 SACW	Cwmbran	1.73	87.35	21	17	.	●
22 SAAG	Abergavenny	1.69	89.05	22	19	19	●
23 SANN	Nelson	1.69	90.74	23	18	.	●
24 SAFD	Ferndale	1.69	92.43	24	15	.	●
25 SAMO	Monmouth	1.65	94.08	25	23	17	●
26 SAHQ	Headquarters	1.01	95.09	26	.	.	

### 4.3.3 Locations

#### ***Postcode Districts (Demand)***

Exactly 54 postcode districts are contained in the data set for South East Wales. Although four of these are actually locations outside of the South East boundaries (three across the border in England), representing times where service interaction occurs across zones, these occasions are very rare (only around a dozen of all records).

Via communication with WAST headquarters, it was discovered that in November 2007 the Trust stopped using an (outdated) post office dataset when referencing the location of the postcode district in which a call originates, and instead switched to a more automated gazetteer developed from Ordnance Survey address-point data – returning the central postcode for the zone in which the incident is located, and not the specific postcode. This means calls in the data set are aggregated to this smaller, limited number of (50) South East Wales locations than if all full postcode addresses of incidents were recorded, but which are more useful in WAST’s own demand analysis processes.

#### ***Station (Server Base)***

In total, 170 vehicle base stations are referenced; however for all non-unique records – that is those including multiple dispatches to an incident – 95% of all responses (not just initial responses) are serviced by only 26 common stations (Table 4.3).

Station 9 (SSVP), which, according to Table 4.3, contributes to a large proportion of services, is in fact a low preference (based on frequency of use) station for locating EAs and RRVs. On closer inspection, the majority of services by a vehicle allocated to this station are by HDUs to AS2 or AS3 calls. Details of this station were not provided by WAST since it is located outside of the South East control region. Similarly, station 18 (SSEF) is not located within the South East boundaries, even though a large proportion of services by RRVs are assigned to this station, hence further details are unavailable. It is likely ‘allocation’ to these external stations (including SAPC) of the region means these vehicles are housed at the station, but during operational hours are positioned more tactically within the region at stand-by points or other bases to await incoming emergency calls.

WAST, provided a list of fixed bases that they claim to use on a daily basis, which amount to only 22 stations. Throughout this thesis, the 22 locations are used as base station and stand-by points for analysis and further modelling. It should be noted that the fixed list provided and the 26 commonly

found preferable bases in the data, mainly account for the same locations. A comparison for interest and completeness is provided in Table 4.3.

### ***Hospital (Service Facility)***

More than 150 hospital facilities were used by the South East Wales ambulance service in 2009. Some of these listed locations however are minor or specialist facilities, accepting only a handful of patients a year. Other facilities are located geographically or administratively in other service regions, and even occasionally across the border in England (operated by the English NHS Trusts). Even so, almost 95% of incidents are handled by only eight main hospitals in the South East. Later, during discussions of conveyance rates, Table 4.10 informs of the major players in the South East hospital arena for EMS handovers at EDs.

#### **4.3.4 Resources**

There were found to be 18 different types of vehicle in use in the data set. For simplicity of the modelling to follow later, and for ease of analysis, only the main two vehicle types are considered; however for completeness in this chapter, some of the other types that are commonly used will be incorporated in discussions, where appropriate. The two major vehicle types extracted from the data, EAs and RRVs, account for over 83% of responses (Table 4.4). All other types (including HDUs) serving any type of emergency incident are grouped as 'Other' in the analysis.

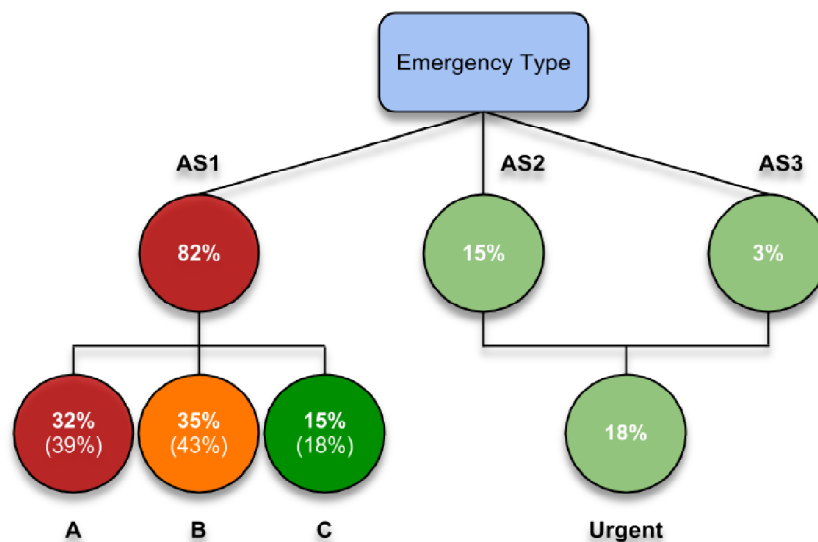
**Table 4.4** Proportion of all records and unique (initial response) services by vehicle type

Vehicle Type	Occurrences	Percentage of Responses	Unique Incidents	Percentage of First Responses
EA	166,310	60.6%	95,445	54.6%
RRV	62,729	22.9%	47,569	27.3%
Other	45,261	16.5%	31,680	18.1%
Total	274,300	100.0%	174,694	100.0%

### 4.3.5 Category Computation

Incident classification is dependent upon the severity of the emergency and type of service required (emergency – AS1, urgent – AS2 and emergency transfers – AS3). Classification therefore describes the order or priority in which to serve these received calls, as detailed in Chapter 2, Figure 2.5. Since it is expected that there will be variation between service for the different categories of calls, especially since the ambulance service targets differ dependent on the priority, the South East region structure is further split by category of calls in the data analysis, Figure 4.5.

South East AS1 (category A, B and C) calls for service make up around 82% of unique annual medical emergencies. Within this classification, the incident is given priority over other emergencies based on severity of the condition and state of the patient.



**Figure 4.5** Proportion of calls for service witnessed in 2009 per emergency type

Primarily, categorisation assists in the analysis of response time of vehicles to the scene of an emergency. Determination of this phase of service is necessary as a basis to future problem solving. Performance indicator focus is founded in the eight minute response time target that WAST must adhere to in 60% of category A cases per LHB (65% for all Wales) and similarly for the other categories.

## 4.4 Demand

### 4.4.1 Regional

The mean number of calls for service per week is 3150, averaged over all emergency types for the whole of 2009, with a standard deviation of approximately 140 calls (Table 4.5). The daily demand average is 478.6 calls with a standard deviation of 40 calls.

From Table 4.5 it seems Mondays and Thursdays are similarly variable, more so than the other weekdays. The reason for the larger spread is likely due to the long service periods often experienced with AS2 and AS3 demand, for which activity is higher on weekdays, particularly Monday mornings, due to a backlog after the weekend of GPs requesting transport for patients (Health Service Executive 2010).

**Table 4.5** Demand per weekday for the region, averaged over entire year

Weekday	Average Demand	S.D. of Demand
Sunday	442.73	30.82
Monday	456.19	34.91
Tuesday	441.50	27.97
Wednesday	439.50	26.04
Thursday	447.89	34.42
Friday	469.81	28.42
Saturday	462.17	30.59
Full Week	3150.22	139.35

### 4.4.2 Inter-Zone Assistance

Due to the nature of such emergency provisions, equitable service to the entire population is one of WAST's main objectives. Although three individual regions exist within the service area, there are times when these three regions of Wales will have to coordinate for optimal provision of medical care. In some cases, it may even be that WAST are called to incidents outside of their operational area, such as across the border of Wales with England – in counties such as Shropshire and Cheshire.

Similarly, there will be occasions where English units assist at emergencies located in Wales when resources are closest or during busy periods.

For the purposes of this analysis, based on a commonly used solution, such occurrences will be dealt with by assuming the flow of vehicles attending calls outside of the South East region of Wales is equal to the rate at which calls within the South East are responded to by non regional EMS units.

#### 4.4.3 Divisional

The Trust partitions the region into five divisions in line with conduct at the regional control centre, where telephone operators receive emergency calls and manage the vehicle dispatch tasks from five hubs in the control room. The divisions – SE1, SE2, SE3, SE4 and SE5 – represent geographic areas of the region (Figure 4.6) and the postcode districts contained within them, the breakdown of which is seen in Table 4.6. Postcode districts are designed with the aim of containing similarly sized populations but divisions will not necessarily contain equal numbers of districts, depending on area and other demographic characteristics.

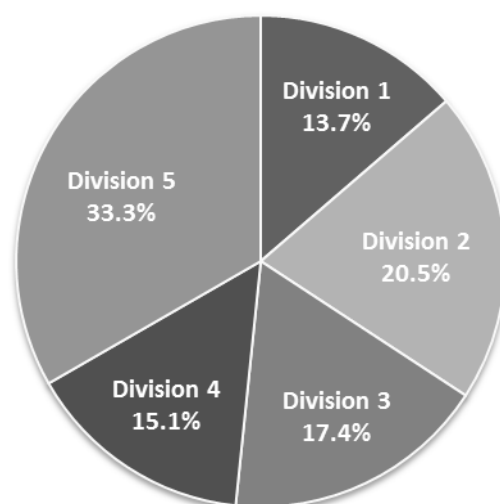


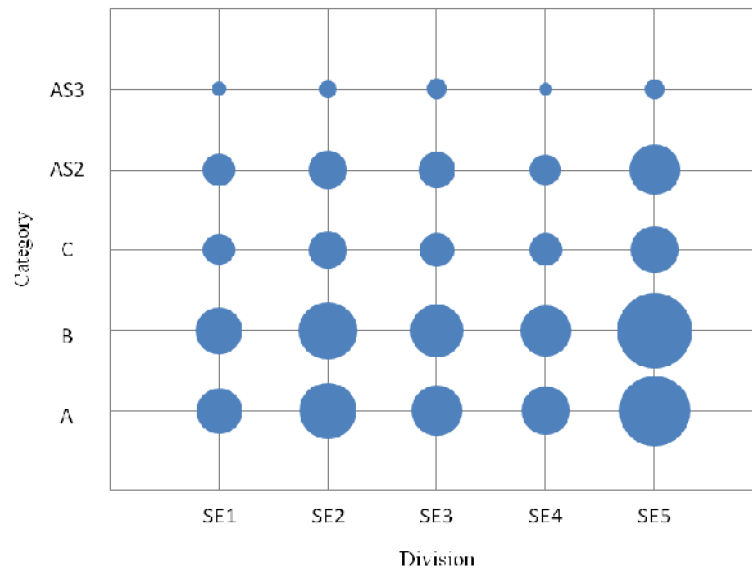
**Figure 4.6** Map of South East divisions provided by WAST (Maher and Rees 2010)

**Table 4.6** Occurrences of unique incidents in South East Wales in 2009

Division	Localities	Districts	Records
All Region		51	174,665
SE1	Monmouth, Abergavenny, Tredegar & Blackwood	8	23,911
SE2	Pontypool, Newport & Chepstow	9	35,809
SE3	Aberdare, Mountain Ash, Merthyr & Caerphilly	10	30,474
SE4	Treorchy, Pontypridd, Pontyclun & Cowbridge	13	26,327
SE5	Cardiff & Barry	11	58,144

Demand from the region can therefore also be considered at this lower divisional level. Proportions of demand arising within each are portrayed in Figure 4.7 and by category in Figure 4.8. The size of each bubble in Figure 4.8 is proportional to the number of incidents arising from within the category and division subgroup. The relative proportion shows how the demand for each of the categories within one division is spread and also the contrast of category demand with neighbouring divisions.

**Figure 4.7** Proportions of unique incidents for service by division



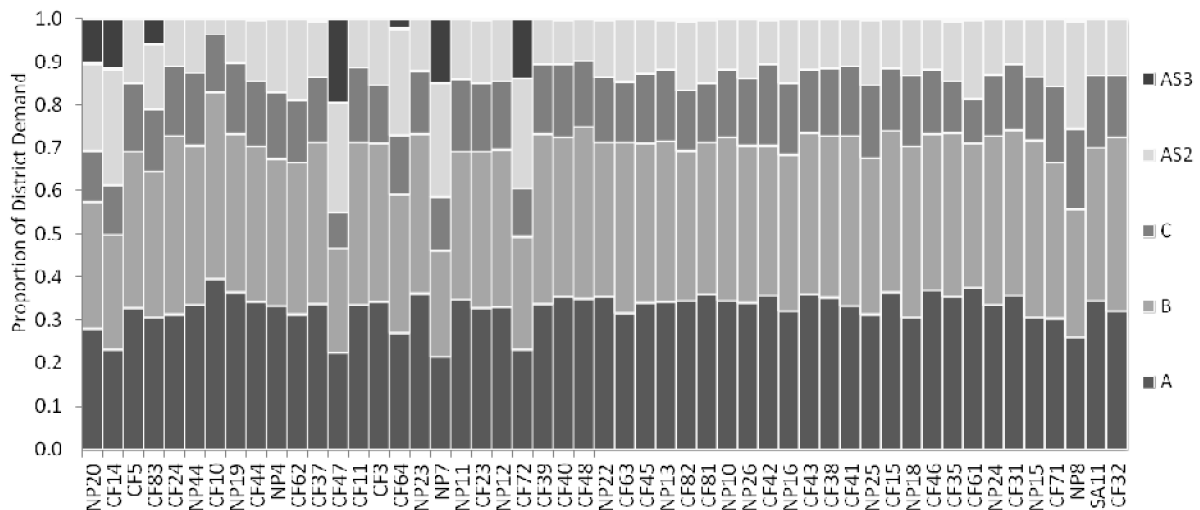
**Figure 4.8** Proportional demand by category and division for all unique incidents

As seen in Figures 4.7 and 4.8, demand varies across the South East geographically. There has been some discussion already as to the cause of this – location of urban and rural populations, deprivation and demography. The city of Cardiff is located in division 5, attributing to the largest amount of demand in the region being witnessed in SE5 – resources are pulled into the city centre, leaving more rural areas vulnerable.

#### 4.4.4 District

There are various other factors that influence the scale of demand witnessed by the service. Overall demand and demand by category can be seen to be dependent upon the geographic distribution of the population. Certain postcode districts witness larger proportions of calls for certain emergency types than others (Figure 4.9), although the majority follow a similar trend despite differences in frequency.





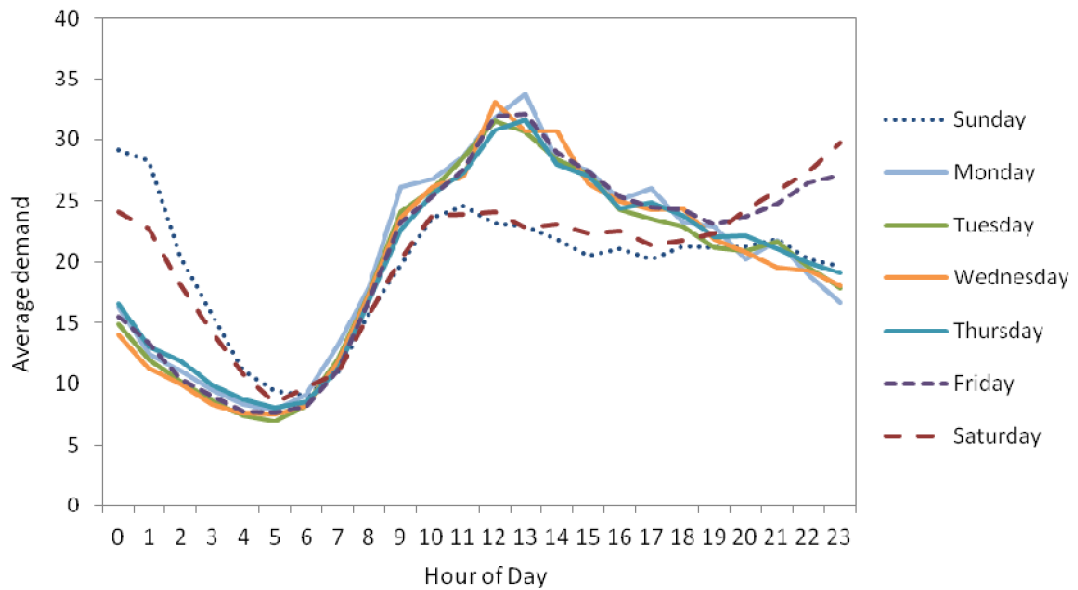
**Figure 4.9** Proportion of demand at each postcode district making up calls for service for each emergency type, in order of quantity of demand (highest to lowest: left to right)

#### 4.4.5 Time Dependency

Demand for service can also be demonstrated statistically to be dependent on time of day, day of week and month of the year, whilst also perceived to be fluctuating spatially. There is already an understanding of time dependency and seasonality affects on EMS demand (Vile et al. 2012), but the WAST data further supports this tendency.

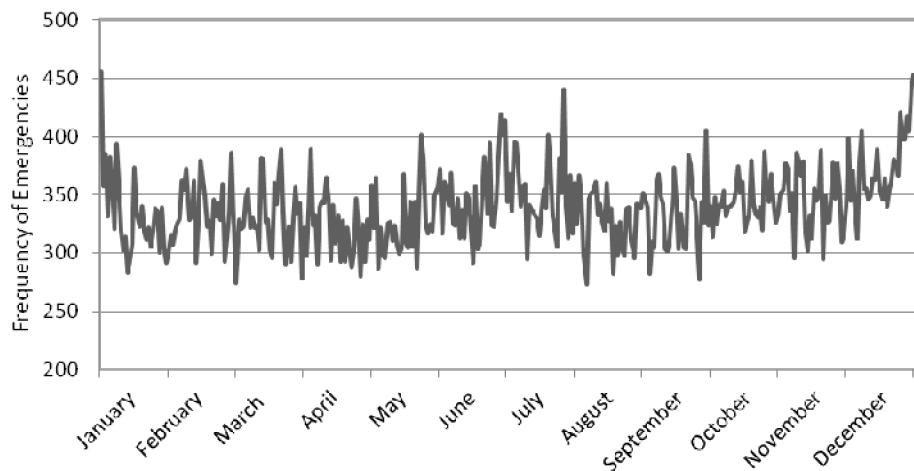
In Figure 4.10 the trend of demand for the seven individual weekdays shows a difference in the weekday versus weekend hourly profiles. Saturday and Sunday follow a similar pattern until the evening when Sunday demand becomes more like a weekday profile, and Saturday follows the Friday night trend, witnessing an increase in demand possibly due to an influence of standard social activities around these times.

The peak around midday of each of the five weekdays is typical of such emergency data, and similar effects are seen for emergency admissions at an ED (Knight et al. 2012b) and in other EMS studies (Monroe 1980).



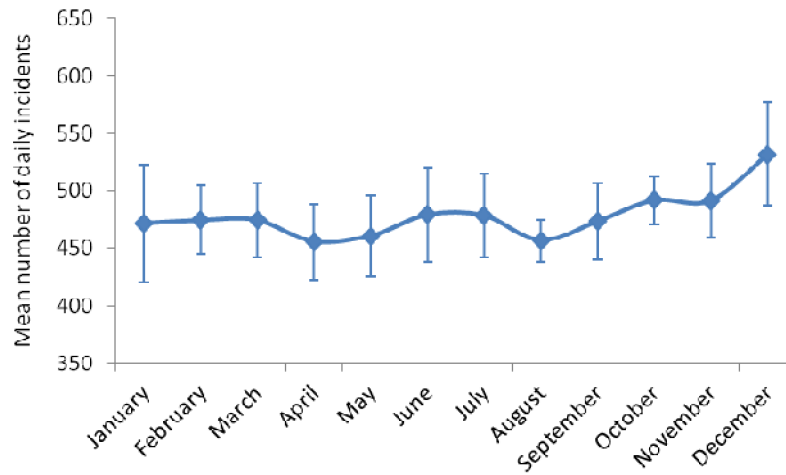
**Figure 4.10** Average demand by hour of the day per weekday

The time series of demand over the region for the whole of 2009 is depicted in Figure 4.11. The profile is as expected – heightened demand in summer and winter months, large variation across the year and daily fluctuations. This is further supported by the monthly variations of Figure 4.12.



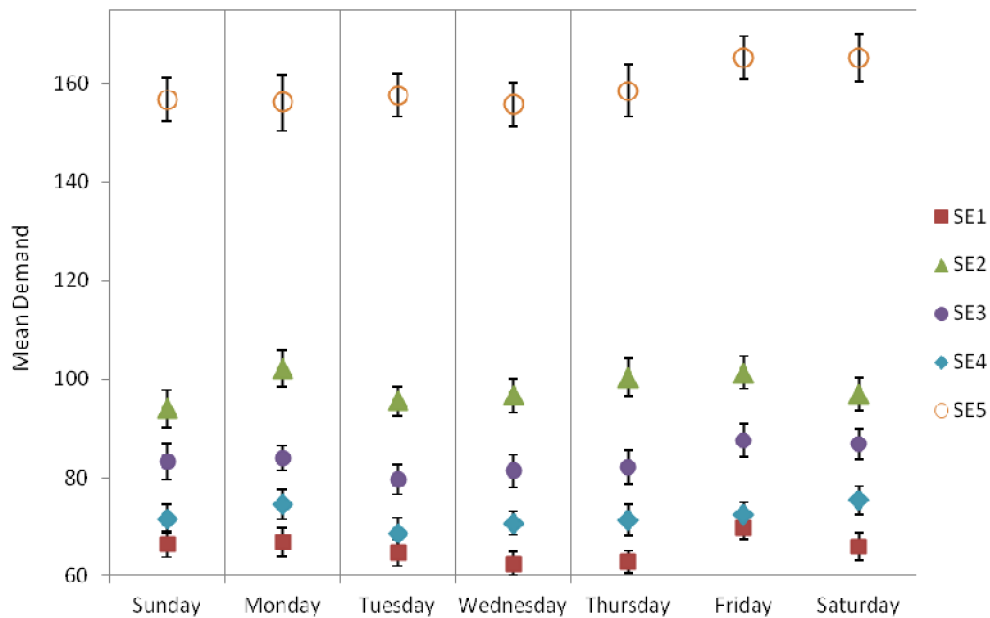
**Figure 4.11** Time series of South East demand for 2009

When looking at the average demand arising from within each division over the week (Figure 4.13), the discrepancy between demand in SE5 (Cardiff) and all other divisions is highly noticeable. The variation around the average is also larger for the more populated regions.

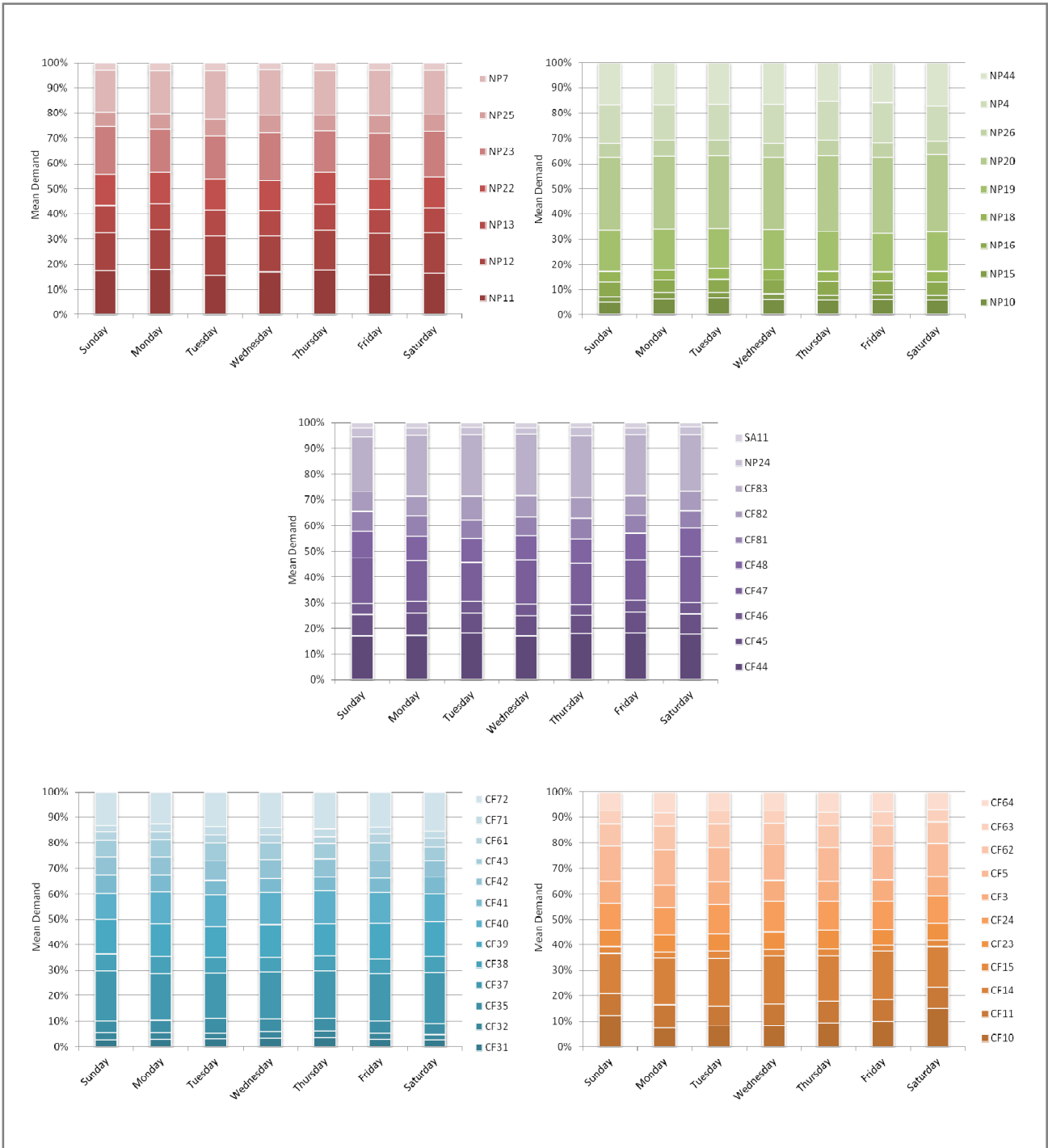


**Figure 4.12** Mean daily demand per month for 2009, with standard deviation

Additionally, the proportion of calls originating by postcode district within each of the divisions is of interest since populations will differ not only across divisions in total, leading to variation in expected demand, but also within a division, particularly where a contrast between rural and urban populations exists. In the graphs of Figure 4.14, the mean demand per postcode district is stacked by weekday to show the distribution of expected demand over the week and the split by location.



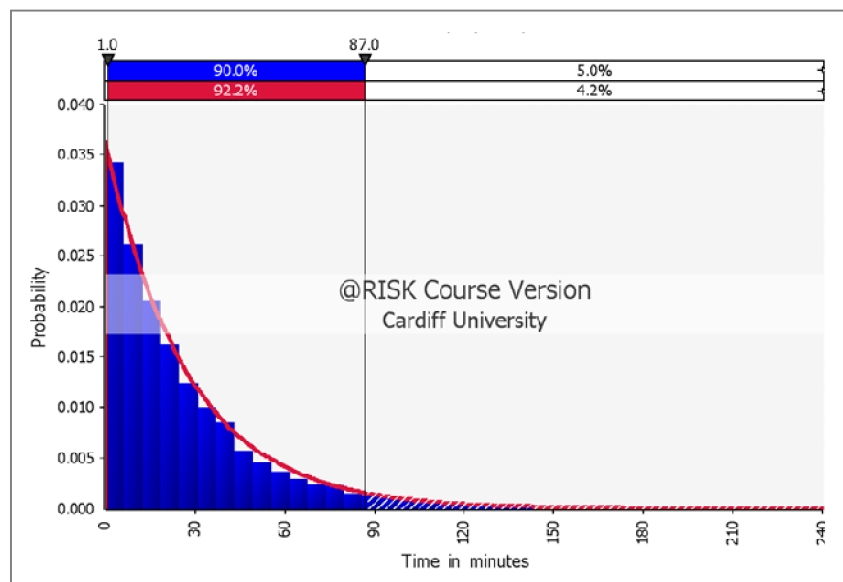
**Figure 4.13** Mean demand per division by weekday, showing 95% confidence intervals for data points



**Figure 4.14** Average proportions of daily demands for postcode districts arising in each of the five South East divisions

#### 4.4.6 Inter-Arrivals

Using the Palisade DecisionTools Suite 5.7 for Microsoft Excel (Palisade Corporation 2011), which incorporates the packages '@Risk' and 'StatTools', a distributional fit to the inter-arrival times of incidents in the South East region is found. The data, compared with a theoretical negative exponential distribution is shown in Figure 4.15, and the same pattern is witnessed for all weekdays. Although the fit is significantly different when using the Kolmogorov-Smirnoff (nor the Chi-Square) goodness-of-fit test, due to a lower frequency of zero value inter-arrival times and heavy tails of the data, the graphical representation shows a similarity from one minute intervals onwards. (Monroe (1980) similarly found the negative exponential distribution a suitable representation for emergency medical arrivals.) However, accuracy is difficult even when ignoring the zero interval range.



**Figure 4.15** Inter-arrival time distribution of historical data and theoretical statistical fit

#### 4.5 Station Assignment

The allocation of a station in servicing a particular demand node should not only be determined by the frequency at which the station is used (Table 4.3) as this will be influenced by location, but also the number and type of vehicles positioned there. Proximity to demand can indicate preference, but in some cases, the closest station is not the most preferable when calculated by comparing the set of demand node and service node pairings in the data.

A station frequency matrix derived from the 2009 South East data, for all emergency units, is provided in Table 4.7. Referring only to incidents that received an EMS unit on scene, preference is based on the number of attendances from the station to the demand node. The stations are listed in order by the overall proportion of regional calls they served (as given in Table 4.3), whilst for individual demand nodes, the preferred station is the station whose allocated vehicles attended the largest number of calls originating in that postcode district. Stations not providing any attending vehicles to calls for service from a particular demand node during 2009 will not be assigned any level of preference – the matrix entry remains blank implying an undesirable assignment to these geographical emergency incidents.

In the location analysis modelling that follows later in this thesis, such preference details are necessary for vehicle busy probability calculations. A lack of exact known vehicle allocations in the South East is the reason why direct comparison of simulation results to the real-world data for average travel times is futile (this argument is further testified in Chapter 5). Instead, preference is taken to be ordered based on proximity of a station to a demand node, Table 4.8, rather than usage prevalence in the data, as in Table 4.7.

For example, in SE5, the most preferable station for the majority of postcode districts (CF10, CF11, CF14, CF15, CF23, CF24, CF3 and CF5), based on usage in the 2009 data set (Table 4.7), is station 1 – SABW. However, when proximity and shortest travel time is used in selecting the most preferable station (Table 4.8), station 5 – SACE – instead becomes the best choice for postcode district CF23 (with SABW second choice) and CF3 (where SABW is in fact fourth closest). Station 10 is more preferable for CF15, and the final postcode district of the region (CF64) which was served most frequently by station 3 – SABA, is in fact situated closer to station 1.

**Table 4.7** Station frequency matrix based on occurrences of usage

PC	SABW	SANP	SABA	SABG	SACE	SAAE	SABR	SAGI	SSVP	SAHN	SAMIR	SAPO	SATR	SAAB	SABD	SACY	SACB	SSEF	SACH	SAPC	SACW	SAAG	SANN	SAFD	SAMO	SAHQ	Total calls	%			
SE1	NP11	14	3	19	2	15	20	10		5	20	18	6	7	4	1	16	22		9	22	8	11	17		12	13	6424	2.6		
	NP12	18	7	23	3	23	16	2	20	6	17	15	8	5	4	1	12		23	11	22	9	10	14	21	13	19	6121	2.4		
	NP13	14	8	19	6		15	11		7	19	16	5	2	1	3	17				12		10	4	18	19	9	13	3718	1.5	
	NP22	22	14		13	23	8	3	20	6	17	4	15	1	2	7	10	23	23	18	21	16	5	11	19	9	12	4606	1.8		
	NP23	17	9	21	8	21	15	12		5	18	10	7	1	2	4	19				13		11	3	16	20	6	14	6768	2.7	
	NP25	13	7	14	6	14		16		5			9	4	8	11					3		10	2			1	12	2244	0.9	
	NP7	13		7	20	8	15	17	14	21	5	18	16	6	3	4	9	21	21		12	21	11	1	18		2	10	4539	1.8	
	NP8				8		8	8		1					4	6	6						2			3	5	65	0.0		
SE2	NP10	10	2	17	1	9		15	20	4		19	7	11	8	3	16	18	20	5		6	12	20		13	14	3228	1.3		
	NP15	13	3	14	5	14				6			2	10	8	11				9		7	4			1	12	969	0.4		
	NP16	13	2	16	3	14	16	16		5	15	16	7	12	11	9	16		16	1		6	10			4	8	2839	1.1		
	NP18	11	1	15	2	14	16	16		6			4	12	9	7					3		5	10			8	13	2159	0.9	
	NP19	11	1	15	2	13	19	16	21	5	20	21	6	12	8	7	18	17		3		4	10	21	21	9	14	8600	3.4		
	NP20	10	1	18	2	14	17	15	23	6	19	25	5	13	8	7	16	20	24	4	22	3	11	21	25	12	9	14160	5.6		
	NP26	12	2	16	3	14		16		4	16		6	13	9	8					1		5	10	15		7	11	3372	1.3	
	NP4	13	2	16	3	19	17	14		7	19	21	1	8	4	9	15				11	22	6	5	18		10	12	8268	3.3	
NP44	12	2	15	3	14	19	16	21	5	18	20	1	11	7	9	17	23			6	23	4	8	21		10	13	9158	3.6		
SE3	CF44	12	13		15	19	1	3	7	10	6	2	18	13	17	16	8	19			9	19	19	4	5	19	11	8144	3.2		
	CF45	11	14	14	14	17	1	5	6	10	4	2	19	13	19	17	8	19			9			3	7		12	3812	1.5		
	CF46	11	15		15	18	3	1	7	9	5	2	20	17	14	13	6				8		18	4	10		12	2048	0.8		
	CF47	11	14	22	19	17	2	3	8	10	5	1	20	13	17	15	6					12	22	16	4	7	20	9	6218	2.5	
	CF48	12	18	21	18	21	2	3	8	10	5	1		11	15	14	6		18		9		16	4	7	17	13	4883	1.9		
	CF81	15	18		16	20	4	1	9	7	6	2	17	12	14	8	3	22				10	20	18	5	11	22	13	3687	1.5	
	CF82	13	18		14	19	4	1	7	9	6	5	17	16	15	12	2	21			21		8		20	3	11	21	10	4035	1.6
	CF83	8	15	18	6	14	5	2	9	12	3	10	20	19	20	16	1	17	25	22	22	7	23	24	4	11	25	13	10848	4.3	
NP24	15	18				4	1	8	9	6	2	18	7	12	13	5					11		14	3	10	17	16	1045	0.4		





**Table 4.8** Station preference matrix based on proximity<sup>1</sup>

PC	Station Preference Choice																							
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th	13th	14th	15th	16th	17th	18th	19th	20th	21st	22nd	23rd	
SE1	NP11	15	14	12	4	7	24	16	22	2	13	10	5	6	23	1	25	11	8	19	21	17	26	3
	NP12	15	7	14	24	16	13	12	4	10	2	22	6	11	1	25	21	8	5	23	17	3	26	19
	NP13	14	15	7	12	13	23	24	16	4	22	11	2	10	5	6	1	25	21	8	26	17	19	3
	NP22	13	7	11	15	14	24	23	16	6	12	10	4	22	25	2	21	8	26	5	1	17	19	3
	NP23	13	14	15	7	11	23	12	6	16	4	24	22	2	10	25	5	1	26	21	8	19	17	3
	NP25	26	23	19	12	2	4	13	22	14	15	7	11	5	1	16	6	10	21	17	24	3	25	8
	NP7	23	12	13	26	14	22	2	11	15	4	7	19	16	6	24	5	1	25	10	8	21	17	3
	NP8	23	13	14	11	7	12	15	26	22	6	24	2	4	19	10	16	25	21	8	5	1	17	3
SE2	NP10	4	2	5	22	1	16	15	12	14	7	24	10	3	21	13	19	23	17	25	26	8	6	11
	NP15	26	12	23	19	2	4	14	22	15	7	13	24	5	1	16	11	10	6	21	17	3	25	8
	NP16	19	26	2	4	22	23	12	5	1	16	15	14	13	7	10	21	11	17	24	3	25	8	6
	NP18	2	4	22	12	19	5	1	16	15	14	26	7	10	23	21	17	24	3	25	13	8	6	11
	NP19	2	4	22	5	12	16	19	15	1	7	14	26	10	23	21	17	24	3	25	13	8	6	11
	NP20	2	4	22	12	16	5	1	15	14	7	24	19	10	23	21	13	26	17	3	25	8	6	11
	NP26	19	2	4	22	12	5	1	16	15	14	26	7	10	23	21	17	24	3	25	13	8	6	11
	NP4	12	22	2	4	14	15	7	23	24	16	13	5	1	26	11	6	19	10	21	17	3	25	8
NP44	22	12	2	4	14	15	23	7	5	1	16	24	13	19	10	21	26	17	3	11	25	8	6	
SE3	CF44	6	11	25	24	8	13	10	15	7	16	21	14	1	12	23	5	17	4	2	3	22	26	19
	CF45	6	24	25	10	8	15	7	16	11	21	14	13	1	12	5	17	4	2	23	3	22	19	26
	CF46	24	15	7	10	16	6	25	11	21	13	8	14	12	1	4	17	5	23	2	3	22	19	26
	CF47	11	13	6	7	24	10	25	15	14	16	23	21	8	12	1	5	17	4	2	26	3	22	19
	CF48	11	13	6	7	24	10	25	15	14	16	23	21	8	12	1	5	17	4	2	26	3	22	19
	CF81	7	15	24	14	13	16	12	10	11	6	4	25	1	21	2	22	8	5	23	17	3	26	19
	CF82	24	15	7	16	10	14	6	12	11	25	4	13	1	21	5	8	2	22	17	23	3	26	19

SE3	CF83	16	10	1	24	4	15	7	5	2	21	14	22	12	25	6	8	13	11	17	3	23	19	26
	NP24	7	15	13	14	24	11	16	12	10	6	4	22	23	2	1	5	21	25	8	17	3	26	19
SE4	CF31	17	21	3	8	25	1	10	16	5	4	24	2	15	7	22	6	11	12	14	13	19	23	26
	CF32	17	21	8	25	1	3	10	16	5	4	24	2	15	7	22	6	11	12	14	13	19	23	26
	CF35	17	21	8	25	1	3	10	16	5	4	24	2	15	7	22	6	11	12	14	13	19	23	26
	CF37	10	16	24	21	1	25	8	5	17	15	7	6	4	2	11	14	3	12	22	13	23	19	26
	CF38	10	21	16	24	1	25	8	17	5	7	15	6	3	11	4	2	14	12	13	22	23	19	26
	CF39	25	8	10	21	24	16	6	17	1	15	7	11	3	5	14	4	13	12	2	22	23	19	26
	CF40	8	25	21	10	24	17	16	6	1	15	7	3	11	5	14	13	12	4	2	22	23	19	26
	CF41	8	25	21	10	6	24	16	17	11	1	15	7	3	5	13	14	4	12	2	22	23	19	26
	CF42	8	25	21	10	6	11	24	16	17	1	13	3	5	7	4	2	15	14	23	22	12	19	26
	CF43	25	8	6	10	24	21	11	16	15	7	17	1	13	5	14	3	12	4	2	23	22	19	26
	CF61	17	3	21	1	10	5	8	25	16	4	24	2	15	7	22	6	11	12	14	13	19	23	26
	CF71	17	3	21	10	1	8	25	16	5	24	4	2	15	7	6	22	11	13	12	14	19	23	26
	CF72	21	17	10	8	16	25	3	1	24	5	15	7	4	2	6	11	22	13	12	14	19	23	26
	SE5	CF10	1	5	16	10	4	2	3	21	22	17	24	15	7	25	8	12	6	14	11	19	13	23
CF11		1	5	16	10	3	21	17	4	2	24	22	15	7	25	8	6	12	11	14	13	19	23	26
CF14		1	16	5	10	4	2	21	3	15	7	17	22	24	25	8	14	12	6	13	11	19	23	26
CF15		10	1	16	5	21	24	4	15	7	2	17	25	8	3	6	22	14	11	12	13	19	23	26
CF23		5	1	4	16	2	10	22	3	21	17	12	24	15	25	14	8	7	6	19	11	23	13	26
CF24		1	5	16	10	4	2	21	3	22	17	24	15	7	25	8	12	6	14	11	19	13	23	26
CF3		5	4	2	1	16	22	10	12	15	21	3	14	7	17	24	25	19	13	23	8	6	11	26
CF5		1	5	3	16	21	17	10	4	2	24	8	22	15	7	25	12	6	11	14	13	19	23	26
CF62		3	17	1	21	5	10	16	8	25	4	24	2	15	7	22	6	11	12	14	13	19	23	26
CF63		3	1	5	17	21	10	16	8	4	25	24	2	15	7	22	6	11	12	14	13	19	23	26
CF64		1	5	3	16	10	21	17	4	2	24	22	15	7	8	25	6	12	11	14	13	19	23	26

<sup>1</sup> Stations 9, 18 and 20 of Table 4.3 do not contribute to the travel time preference matrix since their locations are unknown, leaving 23 station choices.

High station utilisation is one of the main factors leading to stations becoming less desirable to demand nodes in reality, where the station would have been the most preferable if based on proximity alone. The reasons for this may be attributed to:

- 1. Numbers of allocated vehicles:** popularity of a station based on frequency of use is dependent upon the number of vehicles allocated to the station. If there exists a station closer to a demand node than its first proximity station, it could be due to a smaller number of vehicles located at the closer station, likely leading to higher utilisation. Therefore, vehicles may actually be unable to service all demand where preferred, hence bases further from the demand node serve more often. Additionally, for the closest station, if utilisation is higher, the likelihood of a vehicle being idle at the base is small, so, vehicles could find themselves being dispatched to new calls whilst returning from other incidents, which may be less desirable than if a vehicle was sent from the next closest base.
- 2. Demand rate of neighbouring nodes:** if nearby populations have a higher demand rate, the closest station may also be closest to the neighbouring localities, in which case vehicles are likely to be called to the higher demand regions more often, meaning a less preferable station may have to serve subsequent calls from the postcode district.
- 3. Distance from hospital facilities:** some demand nodes are not situated near to a hospital facility; therefore, any time a patient requires transportation from the scene, the vehicle must travel further to and from the hospital and so is busy for a longer period of time per service. This increases utilisation, and the chance of a less preferable station being selected to serve subsequent calls originating from the same location.

## 4.6 Fleet Allocations

WAST provide allocation information, but only in relation to total fleet assignment per shift, not the operational units. This means, the numbers of vehicles quoted to be positioned at a station at any one time are markedly overestimated. Figure 4.16 shows the maximum number of EAs on shift during the week is 52 vehicles; when this is considered in addition to the number of RRVs – with a maximum of 32 (Figure 4.17) – a mean average of  $(37 + 18) / 2 = 55$  vehicles is witnessed at points during the week.

Resulting from discussions with the Trust, 40 vehicles on shift at any one time in the South East region is deemed typical. To account for the inflated fleet size suggested by the data, an approximated (operational ÷ total average fleet = 40 ÷ 55) 70% scaled version of the average WAST allocation is used as a potential benchmark if required. This overestimation likely comes from the fact that all vehicles owned by the Trust working the South East region are assigned to a base station whether or not they are currently or at all operational.

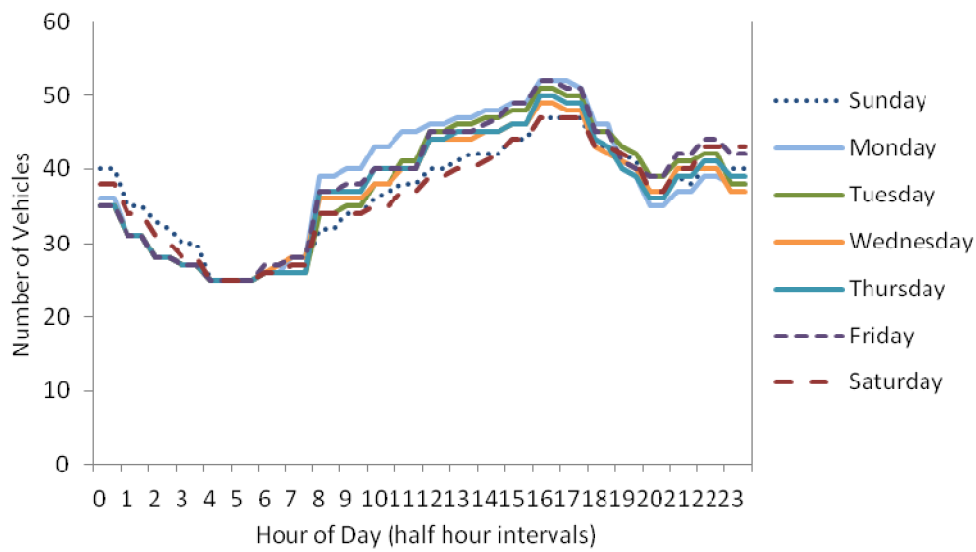


Figure 4.16 WAST EA fleet size (averaged per half hour) by day

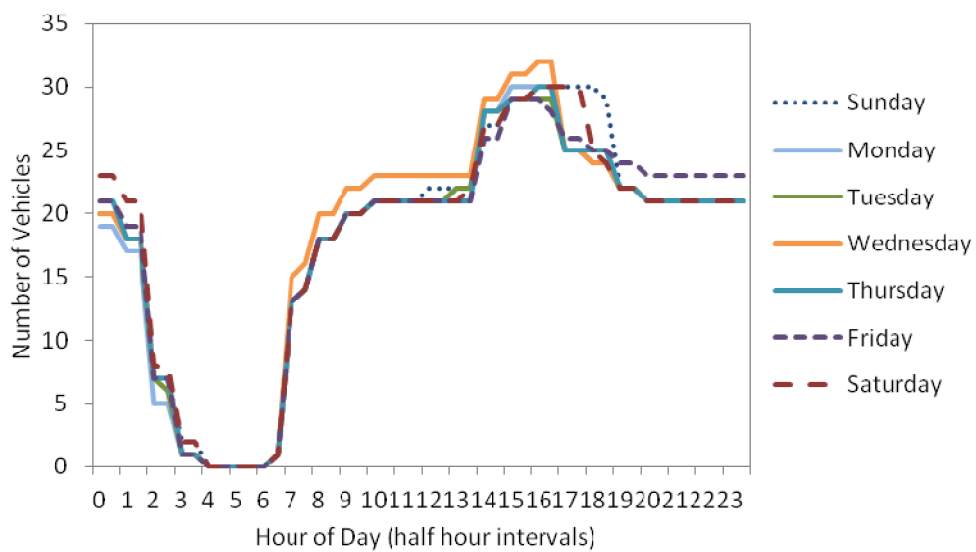


Figure 4.17 WAST RRV fleet size (averaged per half hour) by day

## 4.7 Response Time

### 4.7.1 Introduction

Opinions of the definition of response time are often inconsistent in emergency service modelling. The response time can be classified as the waiting time of the customer, i.e. the queueing time; the service time would be the period of time starting from the moment of arrival of the ambulance crew with the patient and ending whenever the objectives of the incident request are met, e.g. when the ambulance reaches the hospital, or when the patient is transferred from ambulance to hospital care, or when treatment of the patient is terminated.

In terms of a patient, response time should also be taken as the total length of time the individual spends waiting until the arrival of an EMS crew. WAST define their best response to an incident as the "*number of minutes taken for the first vehicle to arrive on-scene at an incident*" (WAST 2008).

This is still an ambiguous period of time to be defined as the standard response time. For ease of modelling, or sometimes due to limitations of available data, it is often only reasonable to calculate response time exclusively through physical travel time or time from dispatch orders by controllers. The time instant for the vehicle to become mobile after dispatch orders is sometimes ignored since it is deemed hopefully small enough to be inconsequential; however, there are times when this dispatch and the allocation delay, (here known collectively as the pre-travel delay), may account for a substantial portion of the overall response time (see Figure 4.1).

WAST calculate response time as the time after the arrival of the emergency call and after orders for dispatch have been given by the call operators until the arrival on scene (Lightfoot Solutions 2010). That is, from allocation, including the mobilisation of the vehicle and travel, to arrival with the patient. For the purposes of this study, response time will be defined as this interval – allocation to on-scene time. It is therefore possible, to independently extrapolate the two main components that make up the response time phase for analysis:

1. pre-travel delay;
2. travel time.

### 4.7.2 Delays

By definition, waiting time of a patient should incorporate any pre-travel delay that may not only be due to the crew or vehicles themselves, but also processes at the control centre. As mentioned in Chapter 2, when a call arrives with an emergency operator, the call cannot be logged, categorised or provided with service by a dispatched vehicle until three items of information have been collected:

- location of the incident;
- patient information;
- description of the emergency.

In many ambulance services, therefore an additional delay is experienced by the patient during the response service phase. This delay can be thought of as the time from the instance of the incident to the initiation of phone contact, plus the time from the call being connected to the official recording of the emergency by the operator.

The more distinct delay period known as 'pre-travel delay' accounts only for the other, measurable, time instances at which a patient is waiting and whilst a vehicle is assigned to the call but not mobile. Once an operator has logged the call and chosen an EMS unit to dispatch, there is a delay before the selected vehicle becomes mobile – the crew need to be informed that they are to be dispatched, and need to mobilise themselves and the vehicle. This may take a few minutes, but is a distinguishable phase in the WAST data set.

In previous studies, pre-travel delay has been modelled as a Lognormal distribution, in two cases with a mean of 3 minutes and a standard deviation of 1.5 minutes (Budge et al. 2010, Erkut et al. 2008b). Jewkes (2011) quoted Lognormal *response* times for a study of a Canadian EMS system; additionally, EMS *travel* times have been modelled by Lognormal distributions in the past (Budge et al. 2010).

One final area of possible delay is at the scene. Upon reaching an emergency, the crew log their arrival, yet there will still be some time associated with attempting to reach the patient after disembarking the vehicle, particularly in tower blocks, or in situations where the vehicle may have to be left some distance from the site of the emergency.

### 4.7.3 Travel Time

Taking the minimum travel time for an incident with multiple vehicle attendances, the distribution of travel from the starting locations of EMS units to an emergency scene can be found. The hypothesis is that travel time will differ between emergency types due to the urgency with which crews respond.

A comparison is made to see if there are statistically significant differences between category A, B and C travel times for EA vehicles. Since each of these groupings are non-normally distributed, a non-parametric one-way ANOVA (Kruskal-Wallis test) is used via the software package SAS Enterprise Guide (SAS 2011). The test shows significant differences with a p-value  $<0.001$ , so a post-hoc test, namely the step-down Bonferroni method, or Holm test, is conducted to determine where the differences in medians lie. The results suggest significance between all pairings, likely due to discrepancies in the frequencies of the tails of the distributions.

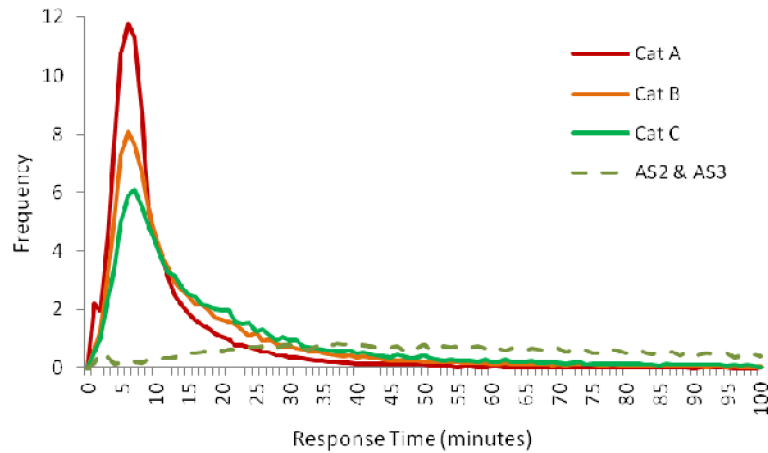
Similar results are found for category A, B and C for RRV travel times, and all post-hoc test comparisons see p-values  $<0.05$ , exposing differences at the 5% significance level.

Travel time modelling however, is a much larger problem than simply finding a distribution from the data, and will be discussed in thorough detail in the next chapter.

### 4.7.4 Response Time Data Results

For 2009, the data provides the distributions for response time based upon WAST's definition, as given in Figure 4.18, for categories A, B and C. AS2 and AS3 response times are combined since their response times are on average greatly longer due to the nature of the incident.

As is seen, the response distributions of the three high priority categories all peak around the eight minute mark, which according to the standards would appear to be satisfactory; however, the heavy tails show the large amount of variation, highlighting vast room for improvement in this service phase.



**Figure 4.18** Response time data distribution of unique incidents by category

## 4.8 On-scene Service

The length of the phase of service, whereby an EMS crew attends to the patient(s) at the incident scene, will depend on the requirements of the patient and the nature of their emergency. This is where the category of the call can determine the expected length of time a vehicle will spend servicing an emergency; yet it is important to realise that the initial telephone triaging of the condition may have been incorrect and so prediction of the needs of the patient will not always be adequate or may have instead overestimated requirements.

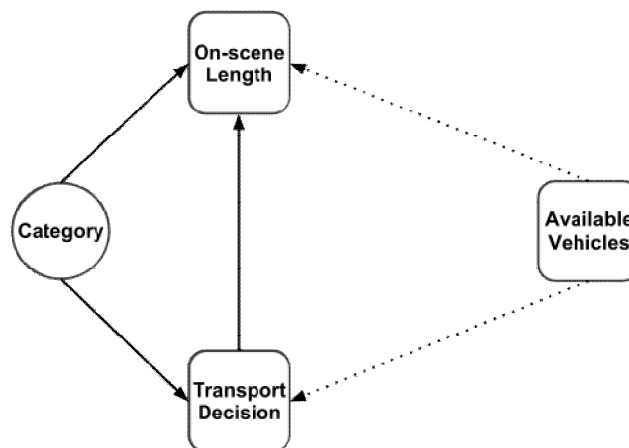
In some studies, even where priority is used to determine the dependent travel time to scene, the time spent with the patient on-scene is often assumed to be independent of nature and even deterministic (Singer and Donoso 2008). However, in Norway, it was found that a physician-manned vehicle is able to deal with only half the number of equivalent of category A incidents (known as “emergency red runs”) per hour than an ambulance attending lower priority calls (Naess and Steen 2004).

A possible triangular relationship (Figure 4.19, where a directional arrow suggests an influence) may exist between the nature of the incident, length of time on-scene and decision to transport. The distinct effect each has on the others is uncertain but it is unlikely any of these are independent from the rest. As a guideline, the nature of the incident will indicate the decision to transport and



both these may determine the expected length of time the crew spend on-scene. Vehicle utilisation also plays some part due to the crew and vehicle type available for dispatch.

It seems, from Table 4.9 that the on-scene length does differ depending on whether or not the patient ends up being transported to hospital. Crews spend less time on-scene when the patient is later transported, presumably due to more critically ill patients requiring swifter transportation. Likely, treatment and care can continue to be provided during transit as opposed to patients who do not need hospital attention but need to be stabilised at the scene.



**Figure 4.19** Relationship between category, on-scene length, transportation decision and vehicle utilisation and their influences

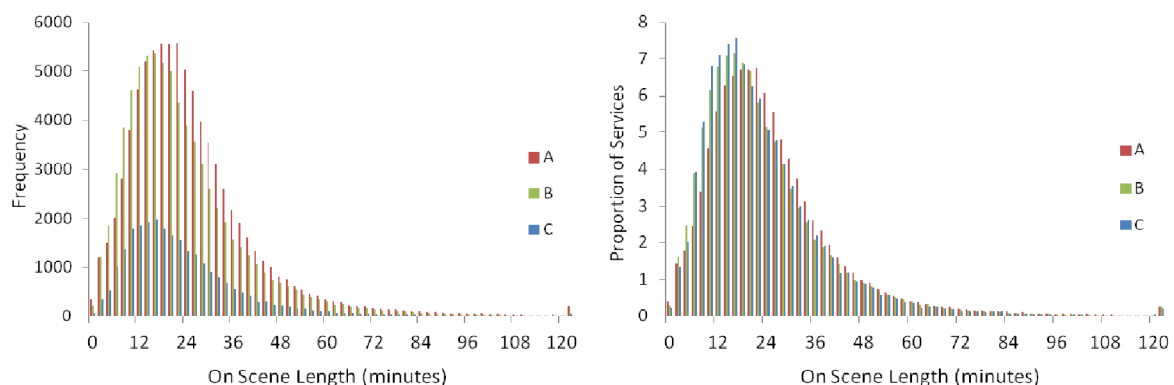
Figures 4.20, 4.21 and 4.22 provide more insight to the distributional shapes and skews for this service phase for the different emergency types and EMS units. When looking at lengths of service of absolute frequencies per category (graph A in Figure 4.20), the category C service distribution by any EMS vehicle falls much shorter than the other two. This is due to the fact that much fewer calls for this emergency type are witnessed by the Trust. Graph B of Figure 4.20 portrays the relative distributions of service per call quota for EMS attendance, showing in fact, it is category A that is slightly more negatively skewed.

Categories AS2 and AS3 have considerably different service definitions as primarily they require transportation; as expected Figure 4.21 shows a higher proportion of shorter on-scene service lengths than category A, B and C emergencies.

**Table 4.9** Summary statistics for on-scene length (in minutes) per category by vehicle type

Category	Vehicle	Proportion of all Records	Transported Patients			Non Transported Patients		
			%	Mean	SD	%	Mean	SD
A	EA	21.4%	33.18	19.51	10.22	10.80	26.87	27.16
	RRV	11.5%	1.13	21.93	14.70	35.88	26.75	71.79
B	EA	23.6%	32.71	18.28	11.12	16.64	23.28	28.11
	RRV	8.6%	1.27	20.41	15.87	23.85	27.61	65.39
C	EA	9.6%	12.82	18.68	11.75	6.63	25.97	21.74
	RRV	2.6%	0.41	19.41	14.35	4.61	32.26	65.97
AS2	HDS	6.3%	7.34	15.97	9.83	0.43	16.78	68.45
	EA	5.2%	8.61	18.26	12.08	0.69	20.21	34.69
AS3	HDS	0.8%	1.24	17.67	14.88	0.22	14.67	13.27
	EA	0.8%	1.29	14.38	13.85	0.26	16.03	20.85
Total		90.5%	100			100		

When comparing the distributions per category, the data can be broken down further to differentiate between length of on-scene service time per vehicle type (Figure 4.22). RRVs generally spend longer at the scene with larger variation, due to the fact they are initial responders. Even though they reach the scene first, usual practice suggests they will stay on site until the patient is stabilised or transported by another EMS unit in the cases of double-dispatch.



**Figure 4.20** On-scene length distributions for categories A, B and C – comparison of shape for absolute frequencies (graph A: left) and relative proportions (graph B: right)

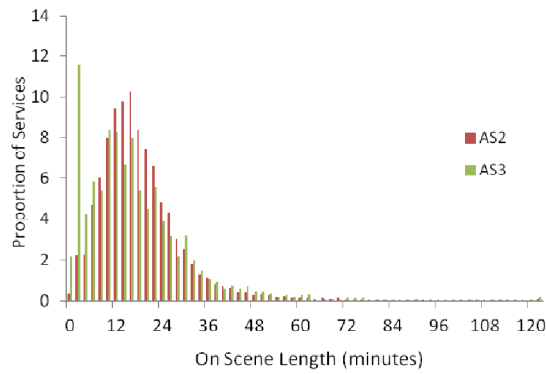


Figure 4.21 On-scene length distribution for categories AS2 and AS3

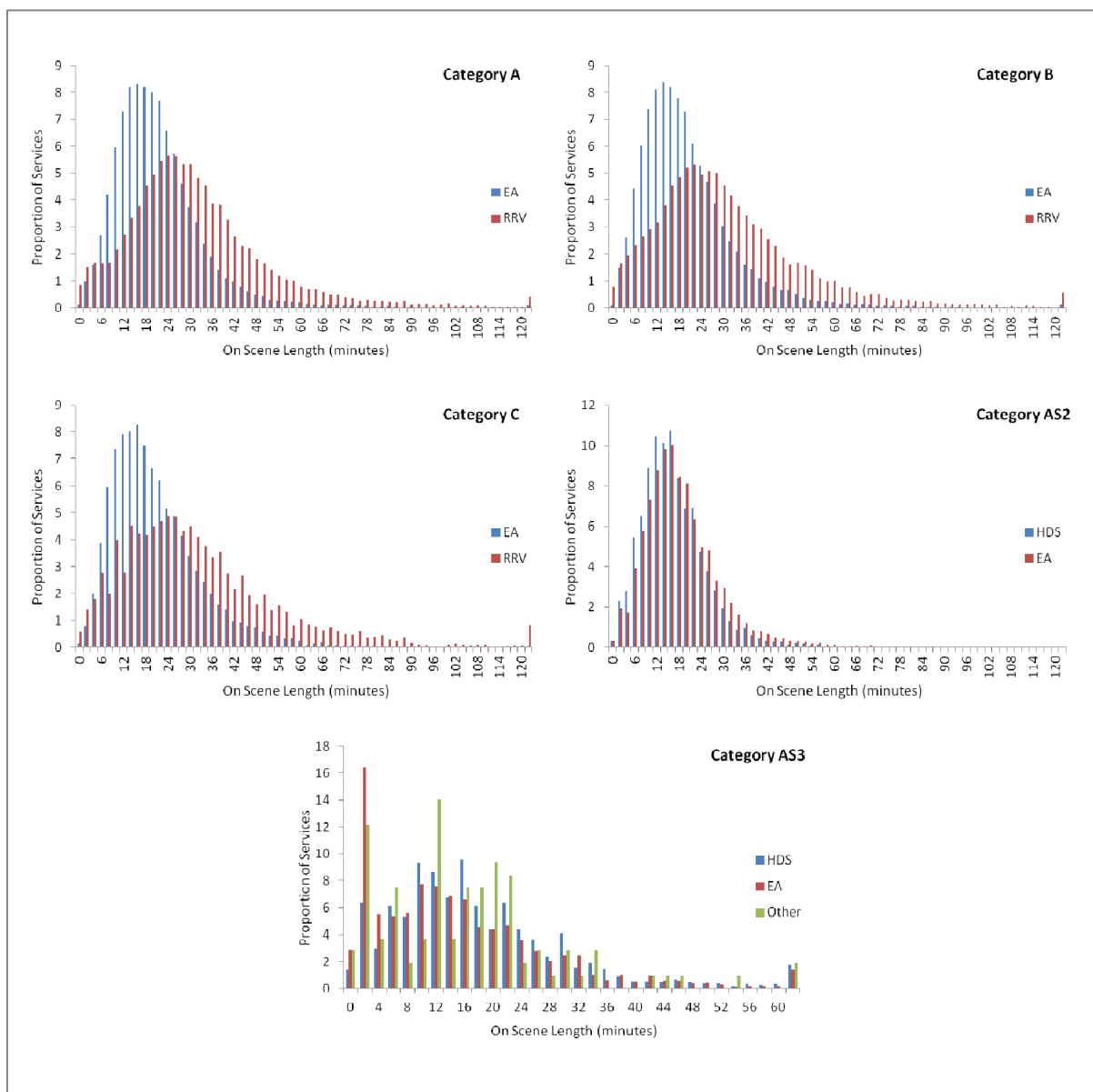


Figure 4.22 On-scene distributions by category showing differences in service by vehicle type

## 4.9 Transportation

Conveyance policy for WAST is to the nearest facility; however, there will be exceptions to this rule, where a patient requires specialist attention only provided by one or a limited number of facilities, or where a hospital's ED is experiencing a crisis or major backlog of admissions and so refuses to admit further emergency patients. In such a situation, the vehicle is diverted to the next closest suitable facility (preference for conveyance is given in Appendix 4.2). In a reply to the audit commissioned in 2006 (WAO 2006), WAST state that their aim is to transport to the best accommodating facility for the individual patients' need, rather than the closest (Audit Committee 2009); however, there is no evidence that this endeavour has been implemented in daily operations.

Taken from the data set, only seven commonly used hospitals were analysed in detail (despite 150 being witnessed), since WAST themselves make reference to only these seven facilities. When the proportion of patients transported to each of the hospitals is explored (Table 4.10), the reason for choosing only these seven for data analysis and further modelling is justified by the usage frequency.

**Table 4.10** Proportion of records referring to patient conveyance to each hospital

ID	Hospital	Unique Transports	Unique Incidents Assigned	% Unique Transports	% Unique Incident Transports	% Assigned Incident Transports
1	University Hospital of Wales, Cardiff	36045	53293	30.27	22.18	67.64
2	Royal Gwent Hospital, Newport	27901	37706	23.43	17.17	74.00
3	Royal Glamorgan Hospital, Pontyclun	15817	21572	13.28	9.73	73.32
4	Prince Charles Hospital, Merthyr	15281	21354	12.83	9.40	71.56
5	Nevill Hall Hospital, Abergavenny	12595	16136	10.58	7.75	78.06
6	University Hospital Llandough	8403	8705	7.06	5.17	96.53
7	Princess of Wales Hospital, Bridgend	3023	3723	2.54	1.86	81.20
<b>Total</b>		<b>119065</b>	<b>162489</b>	<b>100.00</b>	<b>73.28</b>	

Of the 162,489 unique incidents (out of 174,665), 93.03%, would be assigned to one of the seven listed hospitals (their closest facility) if transportation was required. Therefore, the other 143 hospitals not listed only make up 7% of records. The decision to transport actually comes after the assignment of a hospital in the care pathway, since hospital choice is made at initial triage and (usually) based on proximity rather than specialism. In some cases of course this is altered after assessment of the patient on scene. Around 73% of all emergencies in South East Wales which are assigned to one of the seven are eventually transported.

Although hospital 7, the Princess of Wales, receives only 1.86% of the overall transported incidents, which appears insignificant at first glance, it is still included in the analysis since this actually equates to approximately eight transports per day throughout the year.

When investigating transportation by category, for AS1 calls, the higher the priority the more likely the patient will require transportation (Table 4.11). Category AS2 is defined as a request for transport to hospital and is reflected in the high conveyance rate. Destinations of AS3 patients are unknown, and may not necessarily end up at a hospital – the information for this category is less insightful without further details per incident.

**Table 4.11** Proportion of transportations by emergency type

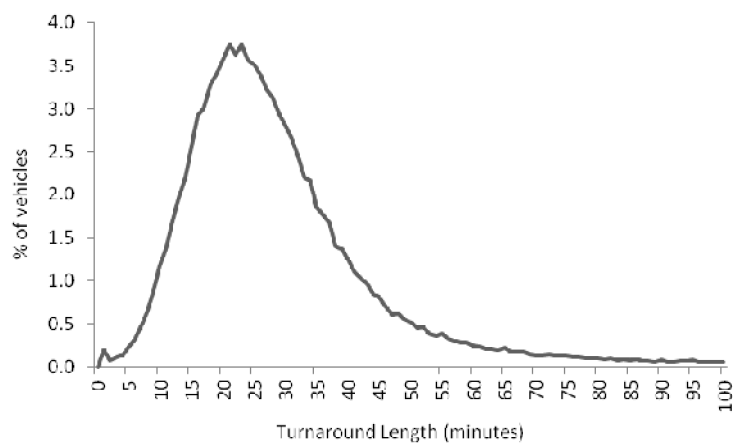
Category	% of All Transports	% of Category
A	35.55	78.99
B	35.33	71.71
C	14.15	69.26
AS2	14.97	91.33
AS3	≈0.00	45.45
<b>Total</b>	<b>100.00</b>	

#### 4.10 Turnaround and Clear Time

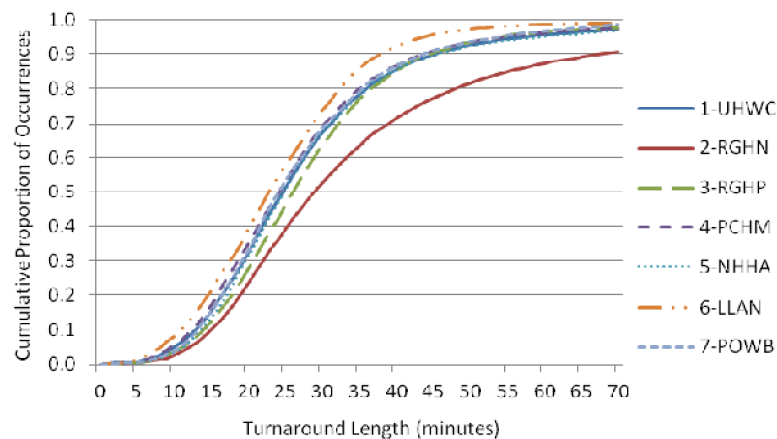
Discussions of Chapter 2 surrounding the difficulties faced by the Trust in handing over patients to hospital care within the target time are now re-evaluated. Some delay may exist for the vehicle, crew and patient upon arrival at a hospital facility, whilst the paramedics transfer patients to A&E.

During this period, the emergency vehicle is not able to be 'cleared' or ready to be dispatched to another call. Figure 4.23 shows the distribution of this delay (incorporating handover processes) across the region.

On inspection of the distribution at individual hospitals, a statistically significant difference was found for the seven hospitals when their medians were compared. This suggests separate distributions should be used for modelling each of the turnarounds at the individual facilities, although some do appear to follow similar patterns (Figure 4.24). Vandeventer et al. (2011) also find that variation in handover is strongly related to the hospital attended.



**Figure 4.23** Turnaround time data for all transported patients to the seven regional hospitals



**Figure 4.24** Turnaround time distributions for transported patients to each of the seven hospitals

## 4.11 Summary

Much of the analysis conducted in this chapter is necessary for the ensuing modelling approaches of Chapters 6 and 7. Where required, the data input utilised in the future modelling methods will be detailed, but mainly results are taken from the analysis of this chapter.

The challenges faced by WAST have been highlighted and reiterated, stating what common problems are faced by the Trust, namely:

- geographical distribution of demand;
- increasing demand;
- delays in allocation and dispatch;
- delays at hospital;
- high utilisation.

The aim of the subsequent chapters is to use the information investigated here to suggest changes to operations and strategies of an EMS system that might improve performance and reduce problematic process areas in the future.

## Chapter 5

# Travel Time Estimation

## 5.1 Introduction

Performance of any emergency service system is typically measured using response time. Whether it is a police force attending a crime scene or fire and medical services hoping to save lives, a fast response will increase chances of a positive outcome. Studies of these emergency service systems have demonstrated the importance of gaining good travel time predictors for modelling response performance, meanwhile a review by Goldberg (2004) exposes just how little work has been done.

Typically, in EMS modelling, reliable and realistic travel time estimates are required for use within developed models, to allow for detailed investigation of the operational and strategic service aspects where travel time can be thought of as a proxy for response time (Hong and Ghani 2006). Strategic policies for EMS systems mainly surround the resources and approach to improving patient outcome, for all of which, travel time is intrinsic.

Three main types of journey occur within an EMS system when serving an out-of-hospital medical emergency:

1. from an EMS vehicle base to a demand point; referred to as a 'Response' journey hereafter;
2. from a demand location to hospital facility; a 'Transportation' journey;
3. from a facility or demand location back to a vehicle base; a 'Return' journey.

This chapter presents a study of the travel times for journey types 1 and 2 recorded by the WAST in South East Wales in 2009 and searches for a suitable model to capture the behaviour over the network. Much work has been conducted in the area of travel time estimation and a review of the key studies is presented. A travel time and distance matrix generator tool has been designed which obtains journey *distances* using the embedded Google Maps Application Programming Interface (API), from which estimates for travel *times* for the journey can then be calculated. A description of the techniques employed for estimation is given. The chapter concludes by providing a re-useable results matrix of journey information for use in subsequent South East Wales EMS modelling.



## 5.2 Necessity of Estimation Methods

Travel time issues arise regularly in location analysis and many other fields of OR. These range from vehicle scheduling and routing, where network models and shortest path algorithms are used in the estimation process (Horn 2000), to cognitive human behaviour (Qi et al. 2006) and transportation (Hassan and Ferrell c2009, Sisiopiku and Roupail 1994).

Where data for travel along routes of a network does not exist, the need for estimation of distance, travel time or response time before modelling is important. Travel time data may be collected directly via observation, although this would result in regional specific information and all possible route legs within the region must be traversed (on multiple occasions to account for variation) to obtain robust estimations. This has the potential to put strain on both financial and temporal resources for any organisation. Another situation that presents a need for estimation is where certain routes of the network may not contribute any historical travel time information to a study. For example when a station facility (or hospital) is closed or not yet built, or when data points are limited for a particular region or when demand regions are to be disaggregated for accuracy, journey times on routes from these nodes to all others will not exist. Ideally, these potential or data-lacking sites would still be considered when modelling. Equally, there are many occasions where estimation is required even in the presence of ample data. Where, for example, there exists large variation in the data or uncertainty in the geographical starting and ending locations of a journey, average travel times may indeed be provided for routes, but with such uncertainty that errors may occur in any subsequent modelling.

The need for estimation within EMS is not just limited to finding journey travel time, but may also be utilised to support the probability of a vehicle reaching the incident scene within the target time (Goldberg and Paz 1991) or finding dispatch probabilities (Budge et al. 2009).

## 5.3 Travel Distance Estimation

Travel time is regularly treated as a surrogate for response time when investigating emergency service systems. In such cases accurate estimates of the expected journey time must be found between every node pair of the network being modelled. The accuracy of travel time calculations lies not only in the method of estimation selected but also in the level of detail of the data used.

There are occasions where travel times must be extracted from distance and speed, or when distances are unknown, times may be found by first using coordinates and displacement exercises to estimate distances, which are then converted to time by an appropriate method.

Two of the simplest ways to estimate travel distances between points on a network are via the use of rectilinear and Euclidean metrics. Many early emergency service studies took these approaches (Fujiwara et al. 1987, Hogg 1968) and some still calculate this way where data is unsuitable or unavailable (Silva and Serra 2008).

Euclidean distance is one of the more commonly and widely used metrics throughout practical geometry and mathematics. It assumes two fixed points given in terms of either Cartesian coordinates  $((x_1, y_1), (x_2, y_2))$  in two dimensions), or spherical coordinates (latitude and longitude) and finds the straight line difference ( $d_e$ ) of these points on a grid system using the Pythagorean formula (see Figure 5.1).

$$d_e = ((x_1 - x_2)^2 + (y_1 - y_2)^2)^{\frac{1}{2}}$$

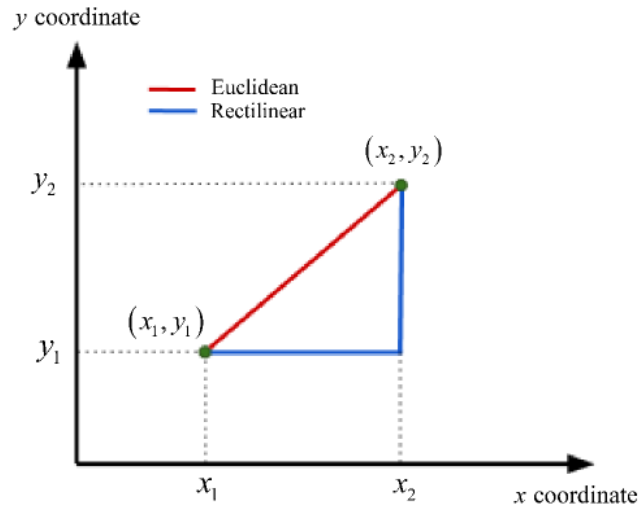
The rectilinear or rectangular metric is also used for point coordinate distance estimation within the literature, ( $d_r$ ) (Fitzsimmons 1973). Its formulation can also be seen in Figure 5.1.

$$d_r = |x_1 - x_2| + |y_1 - y_2|$$

Rectilinear distance is commonly used in studies of cities where the street network is laid out in blocks, and so also goes by the names of the 'city block', 'Manhattan' or 'Taxicab' metric. Where road direction information is known, it is possible to coincide the distance estimates to the network by altering the coordinate axis direction (Miyagawa 2009).

The Minkowski distance metric generalises the Euclidean, rectilinear and the Chebyshev (where  $p \in \mathbb{R}$  tends to infinity) distance metrics by formulating the distance from location  $i$  to  $j$  as:

$$d_{ij} = (|x_i - x_j|^p + |y_i - y_j|^p)^{\frac{1}{p}}$$



**Figure 5.1** Example of a Cartesian coordinate system

Love and Morris (1972, 1979) evaluate and compare several different distance metrics for estimating road distances in the United States. They discover the best performing estimation function to be of the type of Equation 5.1, but due to complexity in calculating parameters  $p$ ,  $s \in \mathbb{R}$  and  $k$  (weight), it is suggested that since optimal values of  $p$  and  $s$  are fairly similar, the simpler function, a weighted Minkowski formula (found to be relatively accurate), would be suitable.

$$d_{ij} = k(|x_i - x_j|^p + |y_i - y_j|^p)^{\frac{1}{s}} \quad (5.1)$$

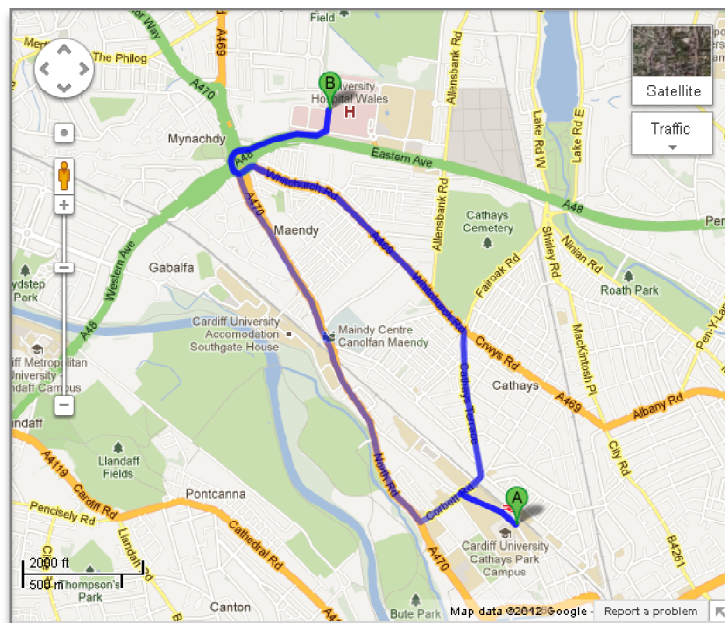
In their 1979 paper, Love and Morris compare the performance of the metrics for shortest travel distances for routes in urban and rural networks separately. Again, the more general functions are found to work best. The authors do note, however, that in urban settings it is more convenient to use the Euclidean metric which works almost as well as the general functions, and is easier to implement. (Note that where a network has “rectangular bias” the rectilinear function, rather than the Euclidean metric, is recommended.)

## 5.4 Computing Shortest Distance

In some studies, the length of all arcs or paths in a network is known, but a matrix of all possible route lengths is required. Once a method has been selected for determining the distance of an arc (as described in the section 5.3), then the shortest path can be calculated for a journey from one

node to another, that travels along one or more of the known arcs, building up a collection of route distances. A commonly used algorithm for this purpose is known as Dijkstra's algorithm (Dijkstra 1959), though this will not be described here since its computation is not required by the study.

Often with road networks, there is more than one possible journey route. For a member of the public trying to get "from A to B", or for an EMS driver going about their day job, the fastest route is usually more desirable than the shortest path. Eaton et al. (1986) and Ingolfsson et al. (2003) make use of shortest path algorithms to pre-compute travel time between nodes of their networks. Google Maps (Google ©2013) also implements a similar algorithm to provide the best possible travel option for a requested journey, rewarding shortest journey time rather than distance. New dynamic features even enable estimates under current traffic conditions. An example, showing two possible routes between an origin and a destination is given in Figure 5.2. The darker route returned is the fastest, with the fainter blue route resulting in the same distance travelled but taking slightly longer.



**Figure 5.2** Example of a network with just two nodes but two paths between the origin and destination address points, returned by Google Maps

## **5.5 Travel Time Estimation**

### **5.5.1 Introduction**

Computation of travel times can be achieved in multiple ways, with additional considerations regarding speed, road type, transport type, aggregation of points, time of day effects and geography of the network. Factors incorporated by the travel time estimation method vary according to which specific network characteristics the researcher deems important.

### **5.5.2 Scaling Factors**

An approximation of ambulance travel times may be achieved if actual travel times (based, for example, on road speed restrictions) are known. Assuming an EMS vehicle travels on average at a given rate of speed faster than a regular vehicle, if standard travel times are known (or can be accessed), then the ambulance travel time can simply be found through scaling. A study which uses this scaling approach is that of Perez (1982). There are more accurate and sophisticated methods available for travel time calculation; some studies use information stored by organisations such as the Department of Transport, or computer systems containing distance and traffic survey data (Alsalloum and Rand 2006, Love and Morris 1979) which allow more in depth analysis of travel times over a network.

In transportation and location logistic studies, it may be distance that is required to be estimated. Network specific characteristics such as density, geography of the region and traffic flow may affect the results of the estimation, but could be accounted for by a 'circuitry' factor or multiplier that corrects estimated straight-line distances. A table of calculated circuitry factors for 30 countries is provided by Ballou et al. (2002).

### **5.5.3 Estimation via Distance**

A more in-depth procedure than scaling, estimates travel time using known distances for each given route. In the most simplistic version of this approach, all that is required for computation is route

distance ( $d$ ) and overall average velocity ( $v$ ). From this, a linear relation of travel time ( $T$ ) and distance can be established.

$$T(d) = d/v \quad (5.2)$$

This may be generalised, as in Eaton et al. (1986), by expressing  $T(d)$  as the linear model:

$$T(d) = a + bd$$

However, when distance and time are treated interchangeably (Equation 5.2) then constant speed is assumed, implying travel time is proportional to distance. This has been shown to overestimate the response performance and that the estimates significantly affect results of a real-world system (Carson and Batta 1990). Other studies (detailed in the next sections) test explicitly whether this relationship is valid.

The 'square root law', coined by Kolesar and Blum, was first explored in their paper for fire engine response distances (1973). The paper demonstrates the relationship between the number of vehicles in operation and average travel time. The square root law states that "*the average [Euclidean] response distance in a region is inversely proportional to the square root of the number of [vehicle] locations*". This follows, as, if the number of stations increases, the area 'covered' by each station should become smaller, and therefore average distance should be inversely proportional to the square root of station density due to the relationship between distance and area. If the arrival rate of emergencies to a region of geographical area  $A$  square miles is  $\lambda$ , and  $n$  is the number of vehicles located within the region with service rate  $\mu$ , then the expected travel distance,  $ED_i$ , for the area  $i$  is formulated as:

$$ED_i = c_i \left( \frac{A_i}{n_i - \lambda_i/\mu_i} \right)^{\frac{1}{2}}$$

where  $c_i$  is a constant of proportionality dependent on the structure of the region and  $n > \lambda/\mu$ . Expected distance is transformed to expected response time via model 5.3.

$$ER_i = b_{0i} + b_{1i} \left( \frac{A_i}{n_i - \lambda_i/\mu_i} \right)^{\frac{1}{2}} \quad (5.3)$$

A limitation of this model is that vehicles are assumed to serve only in their designated area (Goldberg et al. 1990).

## 5.6 Acceleration, Deceleration and Cruising

If the simplistic conversion of distance to time via Equation 5.2 is applied, the resulting travel time is thought to not accurately depict reality (Goldberg et al. 1990). Speed fluctuates over each experienced trip and so is non-constant for a given route.

A seminal study of New York City fire engines by Kolesar et al. (1975) looks at the difference in travel times for short journeys, compared to longer journeys. It is found that although for relatively long journeys (distances approximately more than a mile) the relationship between travel time and distance is linear, for short journeys, travel time increases with the square-root of the distance. Each specific location pair is analysed separately and only journeys where the vehicle begins at the specified base location are considered owing to the possibility of inaccuracies that non-base locations may otherwise introduce. Due to a lack of data, the authors are not able to consider more than one type of vehicle and acknowledge that this may have brought bias to their results.

Utilising regression techniques, a continuous piecewise function, comprising of both a square-root part and a linear part, devised to map travel distance to travel time, enables the user to estimate response times of journeys whilst accounting for the change in vehicle speed. The claim is made that the speed of an EMS vehicle may be represented as accelerating for the portion of travel along smaller, rural or residential roads (where the vehicle may travel slower than when on major roadways) and decelerating when approaching the incident scene, or when leaving main roadways to travel along smaller ones. Kolesar et al. propose that on short journeys the vehicle never reaches cruising state, and spends its journey in acceleration or deceleration states. However, for longer trips, after accelerating, the vehicle is able to spend a larger proportion of time in a cruising state on main roadways before nearing the scene of the incident.

The importance of acceleration and deceleration consideration is discussed in further studies (Campbell 1992, Ingolfsson et al. 2003). Typical values of the acceleration distance are suggested by Larson & Odoni (1981). For emergency service vehicles, Kolesar et al. (1975) state that the rate of acceleration takes values ranging from 0.5 to 1.0 miles per minute<sup>2</sup> with cruising speeds of around 33 to 40 miles per hour (mph).

Kolesar et al. let  $a$  be the acceleration (assumed constant),  $d_c$  the distance required for travel before cruising state is reached,  $v_c$  cruising speed (also constant),  $D$  the route length and  $T$  the travel time to be estimated; then the travel time becomes a function of distance such that:

$$T(D) = \begin{cases} 2(D/a)^{\frac{1}{2}}, & \text{if } D \leq 2d_c \\ (v_c/a) + (D/v_c), & \text{if } D > 2d_c \end{cases}$$

Constrained non-linear regression is needed to fit to the function. For each region of the city, fits are performed iteratively to the single continuous piecewise function (and additionally to each function piece individually) via the least-squares method using average travel times for each route. It was found that the piecewise function is able to produce good estimates for all regions of New York City due to its square-root and linear components. Acceleration distance ( $d_c$ ) was found to be 0.44 miles for an acceleration value ( $a$ ) of 29.0 mph per minute and a cruising velocity ( $v_c$ ) of 39.2 mph. Since the fit was almost linear for distance values from 0.3 to 0.6 miles, the travel time values were found to be fairly stable within this range.

Many studies utilise or extend Kolesar et al.'s travel time estimation model. One such study applies the model to non-emergency vehicle travel times in urban regions using Euclidean and rectangular distances calculated via coordinates (Cook and Russell 1980). The paper states two approaches for predicting travel times: a piecewise square-root – linear function as seen in Kolesar et al. or multiple linear regression; however, Camp and DeHayes (1974) discovered that a regression model for such an estimation problem cannot be greatly improved by including independent variables in addition to distance.

More recently, the validity of Kolesar's fire engine travel time equations has been tested for use with EMS average response times. Budge et al. (2010) discovered in fact that the model is more suited to median ambulance travel times in stochastic models. They argue that since travel times are non-negative, their distributions will be skewed and so median is better predicted than mean. A parametric version of the median model is shown to work as well as a non-parametric version.

## 5.7 Travel Time Estimation by Road Type

It has already been suggested that EMS drivers and paramedics prefer to take the fastest route to the scene of an emergency rather than the route with the shortest journey time (Hong and Ghani 2006). Investigating this decision, Campbell (1992) states that "*travel times for very short trips may be very sensitive to local conditions; [...] for longer trips, local conditions will tend to average out*" and so goes on to consider mean and maximum vehicle speeds by different road types. When a journey passes along



more than one road type, the problem of estimating travel time from distance becomes more complex. The earlier models of Kolesar et al. can also be extended to incorporate this idea of variation by journey leg. Previously, using an iterative optimisation algorithm, Volz (1971) devised a 'point-to-point driving time model' to find mean response time of an EMS system. It was assumed that expected travel time already considers all the factors that may affect it (such as weather, congestion, start and end driving location, day etc.). Route distances along four different road types were obtained from maps, and through corresponding average (inverse) velocities a two-dimensional array of driving times was calculated.

Goldberg et al. (1990) suggest two ways of estimating average travel time through the use of distance data, also over four different road types, whilst capturing the variation in speeds:

- **Base-to-call:** "*Regress distances against the empirical travel times*". However, in many situations this would provide a poor fit due to large time variance per route where distance is constant.
- **Base-to-zone:** "*Regress travel distances against average travel time for each base-zone pair*". The predictions may then be used to estimate variance. This method is better in case studies with large variation since outliers will cause less disruption.

Variation around the average travel time can then be calculated from a histogram of the normalised residuals of the model. The regression was run for base-to-call and base-to-zone models, but prediction errors were found that suggest the problem lies in the structure of zones and large variance within routes, and not from the fitted models.

## 5.8 Targeting Variation

### 5.8.1 Introduction

Kolesar and Blum (1973) and Kolesar et al. (1975), also consider base-zone routes (not just all base-to-call records) for regression, and like Goldberg et al. (1990), witness large travel time variance in the data. Demand locations may refer to quite small geographic points (or sparsely populated areas), and so locations are often aggregated to a larger fixed average 'zone' instead (Figure 5.3). Since the distance is unchanging for any (and every) journey taken on a specified route, yet travel time varies, linear regression of distances cannot provide a good prediction of

expected travel time when zones cover a large area. As an alternative, and in an attempt to capture the problem of uncertainty in variance of a region, Goldberg and Paz (1991) present a method whereby the normalised residuals of the predicted travel times (from regression of average travel time data and distances per route) are used to estimate the variance of the travel time for the route.

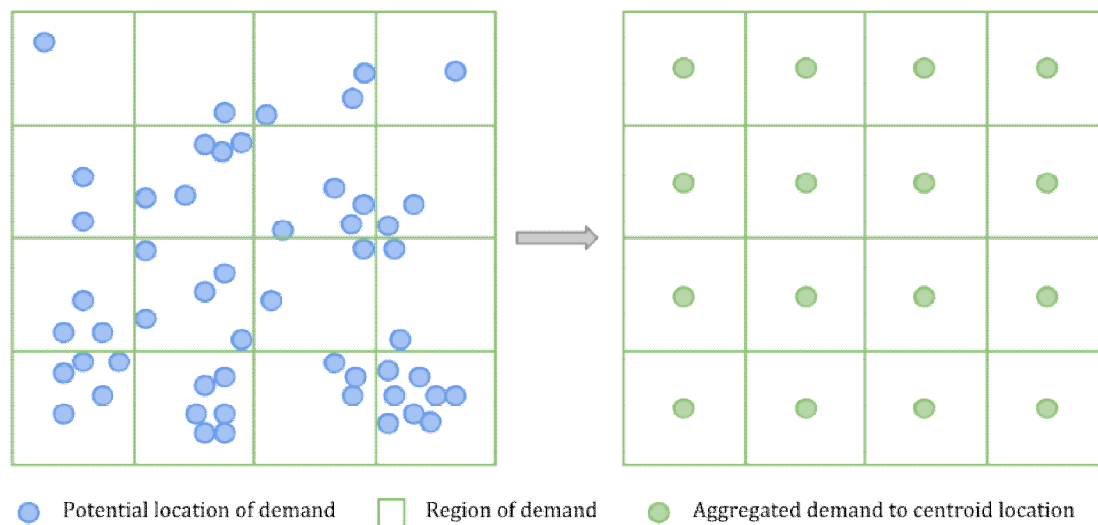
Variance may also occur due to the size of chosen zones. When an emergency arises from within a zone, its exact location is not known when modelling. This causes some error in distance and travel time estimation. However, it is assumed that the times where travel is underestimated may be balanced out somewhat by times where travel is overestimated. To reduce the effect of this error, zones should be taken to be suitable sizes, or cover appropriate population centres.

### 5.8.2 Regional Zoning & Preferences

Defining demand zones presents more problems than unexplained variation. There exists a trade-off between the level of output detail sought and computational effort required. The larger the number of defined zones, nodes or grid cells, the more accurate the solution (since the allocation of servers is dependent on the spatial distribution of demand), yet the more extensive the calculation required to find it. Hence, it is necessary to assume appropriately aggregated locations (Benveniste 1985), despite the reduction in detail that explains the variance during service.

From any node to its  $k^{\text{th}}$  closest neighbouring node, the distance is known as ' $k^{\text{th}}$  nearest distance' (Miyagawa 2009). This concept is used for ordering preferences of service nodes to demand nodes (Benveniste 1985, Hill III et al. 1984). Such a model structure is made use of by Goldberg & Paz (1991), assuming 80% of calls from a demand node are served by the closest station, and the further 20% by the second closest. The weakness in this two-station simplification is the lack of consideration for vehicle utilisation, since in reality, a vehicle from any station may serve any call if it is the best available at any given moment. Goldberg et al. (1990) build and validate their zone structure via a simulation model whose "*dispatching rules require that the zones be small enough so that there is a strict ordering of the vehicles preferred for each zone.*" This implies that each demand location will have a higher probability of being served by a particular station or specific vehicle, tending to reduce variance. In Chapter 6, a similar approach of 'dispatch preference lists' is applied to the developed allocation models. There exists literature focussing on methods for determining these zones or districts (Benveniste 1985, Carter et al. 1972, Keeney 1972, Larson and Stevenson 1972)

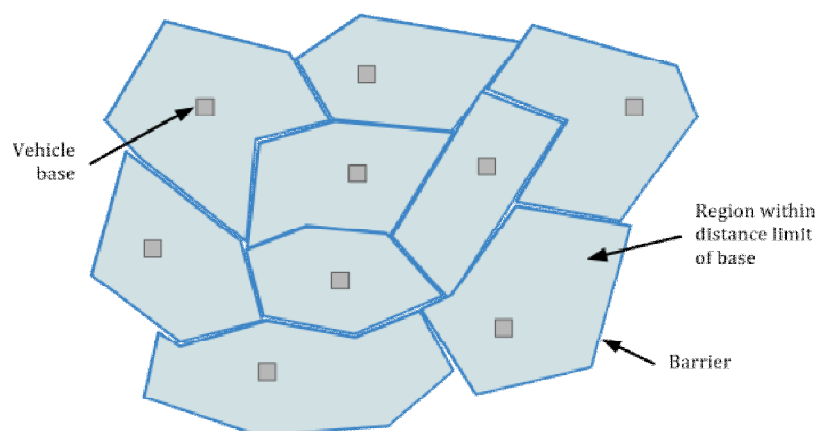
though this is not a primary concern of this study since zones (taken as the recorded postcode districts) are relatively easily deciphered from the Welsh ambulance service data.



**Figure 5.3** Representation of a region divided into sub-regions with aggregated centroid demand

### 5.8.3 Travel Barriers

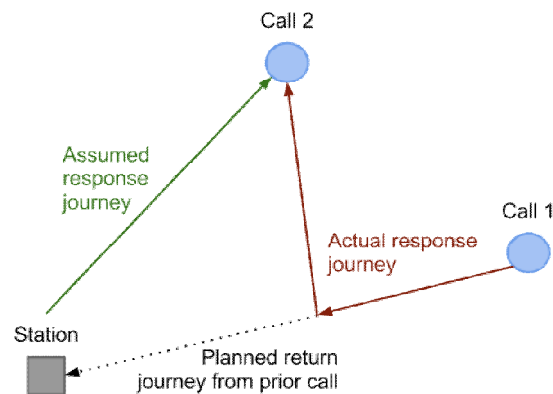
It is possible to obtain a travel matrix for a region by implementing a 'travel barrier routine'. This routine is one whereby a minimal number of routes are estimated by placing a limit on the distance a vehicle is expected to travel (Figure 5.4). It prevents extensive data collection and full route-set searches, by finding information for only a selected number of paths within regions, but lacks in detail and accuracy (Hill III et al. 1984).



**Figure 5.4** Geographical area with travel barrier routine

### 5.8.4 Starting Point Assumption

Chelst and Jarvis (1979) note that the large variation in travel time data is often due to uncertainty in the response vehicle starting point when dispatched to an incident. To combat this, an assumption is made that either all response journeys begin at the vehicle's assigned base station, or only data of trips between stations and demand nodes are considered. This is known as the 'starting point assumption' (Goldberg and Paz 1991). Although this premise is implemented in the travel time analysis following, the Welsh data provided does not necessarily adhere to the rule, as illustrated by Figure 5.5. Some discrepancy or error will still need to be accounted for in the data when vehicles begin responding from other locations or whilst en route back to their assigned base.



**Figure 5.5** Example of an incorrect starting point assumption (call 2 is assumed to be responded to by a vehicle beginning at the station but is served by a vehicle returning from an earlier call)

## 5.9 Distributional Fits

Where theoretical distributions have been successfully fitted to data, the Lognormal distribution has commonly been used (in location and simulation models) to sample travel times due to the natural skew of this type of data (Ingolfsson et al. 2008, Wu and Hwang 2009).

The Hypercube model and its approximations (mentioned in Chapter 3) have been used and extended in many studies since their development. One such study (Chelst and Jarvis 1979) uses the Hypercube to calculate cumulative distribution functions for travel time. This model would be useful for 'random variate generation' and can easily be incorporated in simulation models;

however, a large amount of computational effort is required when using the Hypercube model (Goldberg and Szidarovszky 1991). The authors claim that non-simulation models tend to estimate only average response times, but since there appear to be non-linear relationships between patient outcome and response time, travel time distributions are required. Conversely, it is noted that overall average travel times are not as sensitive to dispatch rules as travel time distributions (Chelst 1975; 1977), so have benefits where strategic policies are being investigated.

## 5.10 Time-Dependent Travel Times

There is a case for considering time-dependency in travel times even when empirical evidence does not support such a structure – Schmid and Doerner (2010) make an argument for its importance in solving ambulance location problems. Their claims are supported by operating practitioners who state that in Europe where streets are narrower than in America (where travel was not found to be time-of-day dependent), congestion has more of an effect on ambulance services.

Kolesar et al. (1975) and Budge et al. (2010) find evidence against time-dependent journey lengths. Kolesar et al. show that although differences do exist, they are not significant so may be discounted. The average travel velocity of emergency vehicles is found to only be affected slightly by the time of day and between daylight and darkness. Not as much difference as expected is found for speeds by peak and non-peak travel hours; yet, since analysis is not conducted by weekday, the authors note that this effect might be greater if weekdays and weekends are considered separately.

Where time-dependent travel times exist, it is possible to avoid building time-dependent models by taking the simpler approach of estimating route travel times per time period and running a model for the multiple time blocks separately. In such a situation, the start time period of the journey and the end time period must be regarded so that for journeys that can potentially span more than one period the appropriate speed can be assumed (Hill and Benton 1992, Horn 2000). (In EMS systems a journey is unlikely to span more than two periods assuming they refer to non-trivial blocks of a day.) It is important to employ the model under steady-state arrival conditions in order to avoid over or under-estimating coverage. A single day should be divided into time periods that contain constant demand (Goldberg and Paz 1991). Periods should be long enough to allow demand to operate under steady-state conditions so that any instant relocation assumption between blocks does not prevent the system reaching an optimal state within the time period. In busy periods utilisation

will be high, but if average utilisation and demand is used in the model, coverage will be overestimated. In quiet times, utilisation and demand averages will imply the coverage is underestimated (Goldberg and Paz 1991). Carson & Batta (1990) deal with time-dependent demand locations by breaking the 24 hour day into four 'states' and modelling the travel time between station and demand over these separately; whereas, Goldberg et al. (1990) "*assume that travel time variance is call time and location stationary*". This means that the uncertainty in travel time will not be dependent upon time of day or the node from which the demand originates.

A disadvantage of including time-dependent network travel times is the computational complexity. In addition to parameter estimation of a route, calculation of all travel times prior to modelling is demanding and requires large amount of resources (depending upon the number of time periods and locations modelled). If vehicle speeds are assumed to change over the day, travel times should also have some stochastic element that captures the uncertainty in speed at *any* time of day. The use of static models for estimation purposes is a much simpler method, and has the potential to produce good approximations in most cases for the available data.

Realising the progression of computational power, Campbell (1992) points out that "*in the future, real time dynamic route guidance information may provide impressive benefits, especially for emergency vehicles*". However, as an analytical tool, static location and discrete event simulation models will likely continue to dominate the EMS response time investigations.

Vehicle speed is also often dependent on the category of incident. Where EMS vehicles are able to increase their speed for patients in life-threatening states ('blue response'), the overall travel time may not depend on time of day since the vehicle is hindered less by traffic; but for lower priority cases, speed may be affected by congestion (time of day effect) and speed limits. Although the consideration of time-dependent travel time may be a valid one, without empirical evidence there is no justification for predicting the travel times for different blocks of time, and so for the purposes of this study, time-dependent vehicle speeds are ignored, since the data does not allow support (or disproof) of the hypotheses.

## 5.11 Limitations of Models

Firstly, none of the models found in the literature, deal with the difference in speed and overall travel time by vehicle type. The fleet for the Welsh Ambulance Trust is heterogeneous; the different vehicle types are designed specifically for different jobs and their various strengths utilised purposefully. RRVs can travel faster than EAs but are less likely to transport a patient to hospital. The heterogeneous fleet will be considered in the estimation analysis of the next section.

Secondly, all of the studies mentioned only consider travel time for the response journey of an EMS system. Since an aim of this study is to model the entire EMS system and not just the response phase, estimates of travel times for all possible routes, referring to all three journeys – response, transportation and return – must be found. In some cases, for example, the same route will be traversed by a vehicle on a response journey and a return journey. Real world distance or time will depend on the direction of travel (due to one way systems and traffic routing) and so the assumption of identical travel information in both directions of a route cannot be applied. It is therefore necessary to produce a non-symmetric travel time or distance matrix.

It is important to note that in the majority of the studies mentioned above where travel *distances* are not explicitly contained in the data, they are usually found using Euclidean or rectangular distance metrics from grid coordinates. Applications are often to American cities where road networks are generally grid plan systems of right-angled blocks of streets in urban areas, so usage of straight line distances is quite suitable. Applying these methods to non-grid plan networks in less urban areas could be misleading. Most of the studies are based on urban areas, and where rural regions are also considered, the two areas are usually treated separately. However, in many real-world applications it is desirable to have a model that may be functional for a region containing both geographies. Kolesar et al.'s work with piecewise functions was a great step towards an accurate representation of this situation, whereby it might be possible to think of regions having a square-root function for travel along rural roads and linear travel along more major roads or for longer journeys. However, since distance here may vary greatly but still only along one road type (e.g. rural) the results will unlikely be a good fit when using Kolesar's (or others') findings exactly. Instead, one aim of this project is to find a set of new functions, via the regression of non-Euclidean travel distances, returning rural and urban travel time nature of South East Wales for a heterogeneous fleet.

## 5.12 Estimation in Wales

### 5.12.1 Introduction

Being the crucial component of response time, travel time estimation methods for an EMS system must be realistic, whatever the regional structure; however, this area of South East Wales poses many geographical, strategic and operational problems when modelling, particularly given the true mix of urban and rural communities. Much of the road network in the region is of A and B class roads, with only a relatively small section of motorway running between the two cities. Predicting travel time for the region is therefore not a simple task.

### 5.12.2 Necessity of Travel Time Prediction for Modelling WAST

It is necessary to provide a method for travel time estimation in this study, primarily due to the lack of sufficiently reliable data. Although travel times for all journey types can easily be calculated from WAST's 2009 data, many anomalies exist in the records, and it is not possible to determine the starting location of the response vehicle in many cases. A low level of detail (postcode districts) recorded for emergency location means the granularity of the interpreted network is not easily increased. Additionally, the large variance for travel on given routes prevents a formal fit of a theoretical distribution to the data. Ingolfsson et al. (2008) note that the standard deviation of a route's travel time in their EMS study is on average 40% of the mean travel time for the route. Compared to service time, Benveniste et al. (1985) claim that travel times are short (Goldberg and Paz (1991) find that travel times are usually around 20% of the service time), and so ignore the variation in this part of their probabilistic system.

More succinctly, the reasons for a prediction method being developed are:

1. demand zone aggregation;
2. incorrect starting-point assumption;
3. large variance.

Origins of calls have been aggregated to postcode districts (due to the lack of detail in the data). The benefit of this is the reduction of complexity in the number of individual locations required for computation. Although the distance between a particular station and each demand node will always



be the same, it would be expected that the travel time between will vary; yet if more demand nodes are used, the accuracy in overall estimation will increase and variance decrease.

Each incident in the data set is responded to by a vehicle for which its assigned base station is known; however, it cannot be guaranteed for any incident whether the vehicle actually began its journey from the base station, or whether (as happens frequently) it began responding from its current position on the road network – either returning to base from an earlier call, or from a stand-by point (or even en route to a lower priority call in the cases of pre-emptive service). As such, when a subsequent call for service arrives with the EMS operators, and a vehicle is assigned, no matter where its current geographical position is on the network, the recorded starting point is the station. Hence, travel times may in fact be very short (if a vehicle happened to be close by and available) even when it appears that the assigned vehicle would be an undesirable choice due to distance.

Variation is likely contributed to by both the previous two points (demand node representation area and unknown starting locations) and may also come from other factors such as:

- Congestion of the network;
- Time-dependency of the traffic flow in the network (including season);
- Individual driver characteristics;
- Weather;
- Condition of roads travelled.

Any model utilising non-deterministic travel time will require the inclusion of some level of uncertainty around journey times. Instead of including all of the factors listed (and any others that may exist), this chapter aims to develop a model that can capture some aspects of variation in travel time despite the ambiguity regarding actual physical location. To do this, prediction is required not only of average travel time, but also of travel time variance; however, it will become apparent that it is difficult to get good estimates for travel time variance.

### **5.12.3 Available Travel Time Prediction Methods**

Three possible approaches to travel time estimation are considered for the application to EMS vehicles along the South East road network. Even though it is already known that response journey

time variance is large, (primarily due to uncertainty in vehicle starting location), attempts are made to apply prediction methods to this phase of service, as well as to the transportation journey phase.

### **1. Distributional Representation**

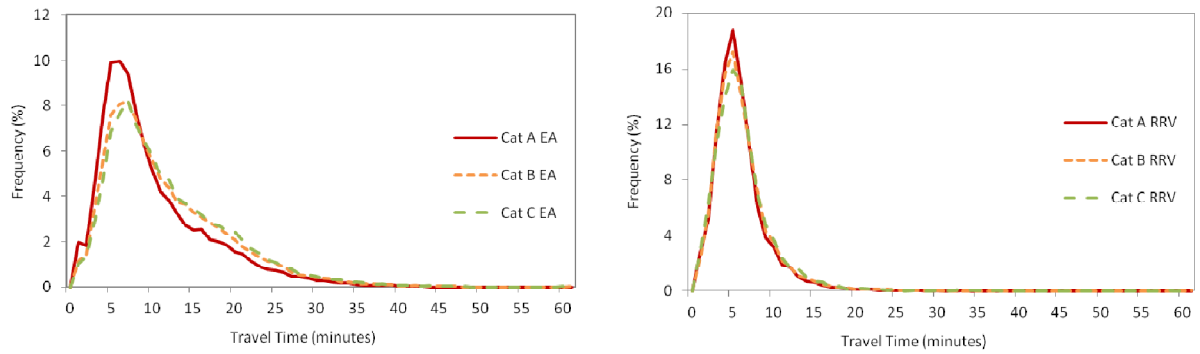
Of the common theoretical distributions available, none were found adequately suitable to represent travel time over network. Even when analysed by category, and limited to only the main demand regions, stations and hospitals, both travel time phases (response and transportation) lack significant distributional representations. The lack of any statistical fit is a consequence of the geographical structure of the region and high travel variance of the data, which produce a non-typical distribution shape and skewness, as portrayed in Figures 5.6, 5.7 and 5.8.

### **2. Average Travel Time**

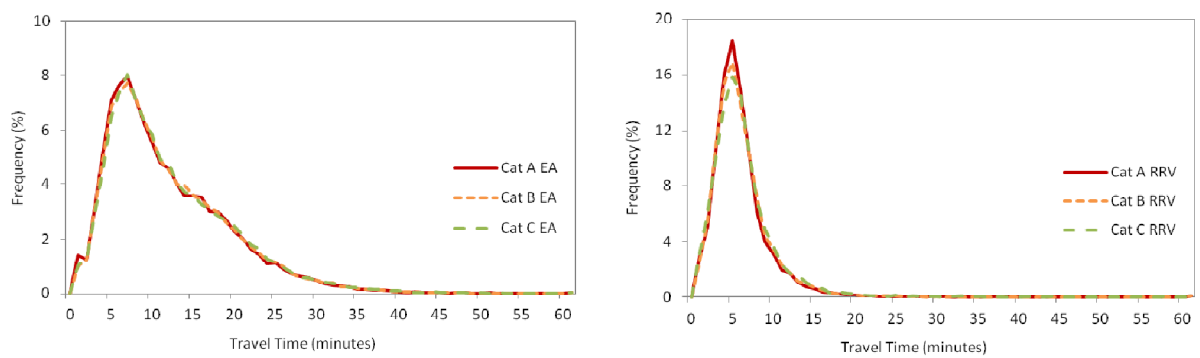
When the objective is to utilise travel times in a subsequent simulation model (where exact movements and positions of vehicles are observed), travel time should be route specific. Taking the average travel time for a route is seen to be unrealistic when it is imperative to also capture the variation in the empirical data of an EMS system such as WAST. Along a particular route in the network travel time is expected to be fairly similar over all trips, fluctuating due to some factors of uncertainty, but following a fixed pattern. This is attributable to the route's associated constant distance. Despite the ease of calculation, ambiguity in starting location of vehicles, aggregation of zones and skewness, mean that variance from the data is not a reliable measure and is not easily predicted in the case of the South East. Additionally, information may not exist for all possible routes over the network, and there is no simple way of predicting average values and variation around them for such journeys.

### **3. Travel time estimation from distance via a chosen model**

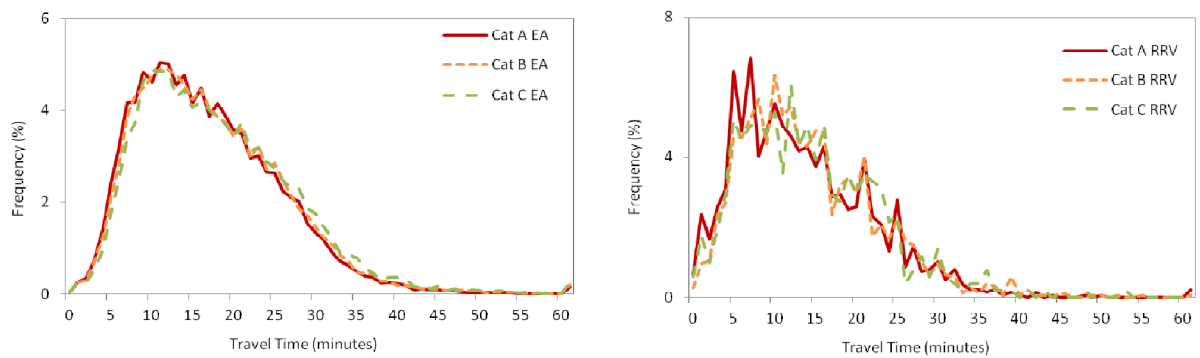
The simplest model that may be used to provide an estimate for time based on distance is that of Equation 5.2, section 5.5.3. The limitations of this constant speed assumption have already been discussed, yet there exist numerous alternative types of model whereby time can still be predicted by regressing distance. Almost all of these models are applied specifically to a particular region or city. It is therefore not sensible to adopt results directly from previous work, especially for a network that has both rural and urban characteristics. Instead, the next sections describe a model designed specifically for South East Wales where regression analysis is carried out.



**Figure 5.6** Travel time for first responding vehicle journeys



**Figure 5.7** Travel time for all response journeys



**Figure 5.8** Travel time for all transportation journeys

#### 5.12.4 Response Journey Modelling

Attempts at modelling *response* journey travel time include:

- graphical analysis;
- distribution fitting;

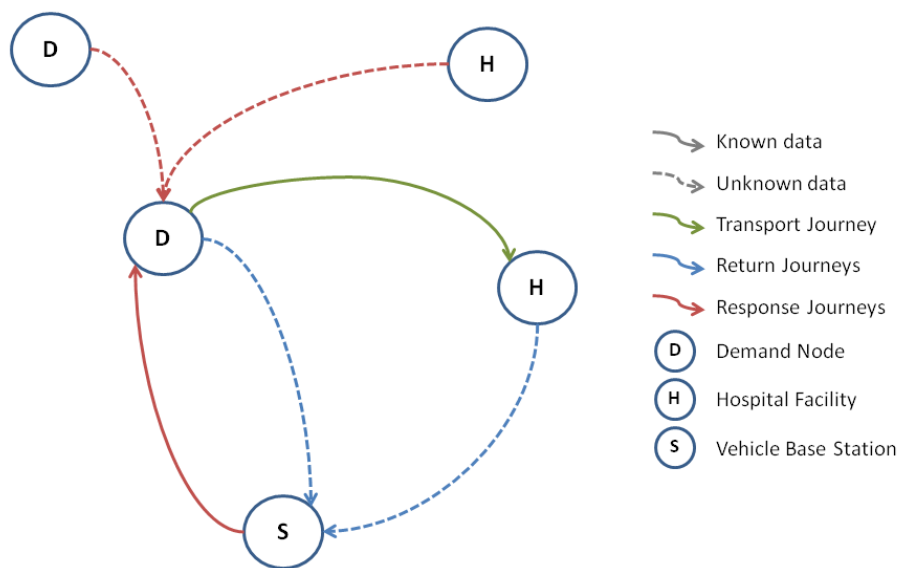
- factor scaling - comparing travel time data to Google Maps travel times and distances;
- comparison testing;
- cluster analysis to group demand locations and find factors;
- cluster analysis to group demand locations so that a significant distributional fit might be found to individual groups;

Each of these were attempted for data from the few most popular (commonly used) station choices for each demand location, since it was assumed that the variance for less utilised station-demand routes would give worse results. Preference was initially determined by the number of journeys made between the two points. This did not provide insightful results and so examination into preference was carried out with consideration to the minimum average travel time at the popular stations. Unfortunately, results were similarly unsupportive, probably as a result of smaller stations with fewer vehicles but shorter travel times being made a higher preference to more demand areas than they can realistically serve. All efforts listed failed to produce good results (statistically significant, nor visually suitable), even when classified by emergency priority and vehicle type.

### **5.12.5 Transportation Journey Modelling**

The decision was made to work with *transportation* journey data instead. (A summary of the available travel data provided in the WAST 2009 data set is given in Figure 5.9.) This transportation travel phase was expected to contain less uncertainty than response journeys since the starting (at the scene) and ending (at the hospital) locations of the vehicle are known. It is not possible to begin transportation from a location other than the scene of the incident, and there should be no cases where the patient is transferred to a location other than the recorded hospital. Exact position for the hospital facility may be used, and so in this phase, variance can be mainly attributed to aggregation of demand areas to centralised points.

The approach of modelling transportation journeys does present its own problems. There exists an issue of scaling any results for transportation journeys to response journeys when the modelling necessitates both pieces of information. How should the proportion of transportation speed that vehicles travel at when on a response journey be calculated? This problem can be trialled by spreadsheet or simulation models to see the effects of scaling.



**Figure 5.9** All possible journeys made by an EMS vehicle during an emergency service, showing whether data is collected by WAST or not for each of the journey lengths

Any incorrect assumptions can easily be rectified; updates can be made if applied to a different region or if a more accurate scaling procedure is revealed. Due to this flexibility, transportation time modelling with scalability is deemed an appropriate method to estimating response journey travel times. The third approach mentioned in the previous section 5.12.3 – ‘estimation from distance via a chosen model’ – is adopted for application to the South East region, for which travel distance must first be obtained.

### 5.12.6 Travel Matrix Generator

For modelling purposes, via both Location Analysis (Chapter 6) and simulation (Chapter 7), the key component required for response time computation is that of travel time. In order to estimate travel times, one possible way is to predict using known distances for all routes. Approximate or real distance values between all location pairs for the South East Wales EMS region must be found for inclusion in prediction models since the data set provided lacks any distance details.

For any given network, route distances may be obtained via Google Maps for each existing journey between two nodes. The Google Maps API is utilised within a purpose-built Travel Matrix

Generator Tool (Smith et al. 2011), enabling the creation of a matrix detailing journey information from an input list of geographic locations (see pseudo code, Appendix 5.1, for this process). By sending multiple requests to Google Maps, a collection of travel time and distance values between location pairs may be obtained, where a pair consists of a start location and an end location (which may or may not refer to the same geographic points). Input locations may denote full or partial addresses, full or partial postcodes or geographical references (such as latitude and longitude coordinates) of explicit or aggregated demand districts, stations and hospitals. Journey information results returned by the Google Maps API are stored in matrix form internally and in 'xml' format externally. After all requests have been made, a data table of information will be displayed within the interface where the user may select whether they wish to view travel times or travel distances for the routes. From this, the user may export the data to an external software package such as Microsoft Excel or similar, in order to analyse or work with the information further if desired.

Although it is possible to acquire travel time information for any route from Google Maps directly, due to the speed assumptions made by Google for individual roads, it is likely these will underestimate speeds travelled by EMS vehicles. Distances, which are constant for a given route, are therefore used instead. Since response time of vehicles to emergencies is the primary performance measure of ambulance services, distances must be converted into journey times for subsequent modelling and investigation of the system. Station, hospital and demand locations are supplied to the tool in return for road based distances. From these, travel times can be predicted through regression methodologies (section 5.13.4); however, the starting point assumption is still an issue with this approach and so regression is applied only to transportation journeys (distances between demand and hospital nodes).

### **5.12.7 Zoning Characteristics: South East Wales**

Before finding route information between all locations via the Travel Matrix Generator Tool, if demand points are aggregated, the centre of a demand zone must be determined. For the purposes of this study, this may be done one of two ways:

1. using given Google Maps position for the *geographic* centre of a postcode district;
2. via manual estimation when the centre of a postcode district refers to a mainly uninhabited area, where the *population* centre is located elsewhere within the district.

In the second situation, when Google Maps returns the centre of a postcode district to be within a forest for example, the postcode district may simply be taken at a greater detail level (maybe referring to a main street in the nearest town) in order to locate the centre more appropriately. Communities are the areas that contribute more to the demand of the region than the rural or remote, geographically central points. Therefore, it is assumed that by taking a town centre over the district centre, the average travel times between regions are better and consistently represented when modelling. This 'manual' approach was also taken by Goldberg et al. (1990) who state:

*"The accuracy of this assumption ["that all calls from a zone occur at the average location of all calls in the zone"] depends heavily on the size of each zone. [...] The distance to the average call location does not equal the average travel distance. However, if the zone is small enough, the difference between these two values is small when compared to total trip length."*

#### **5.12.8 Time Dependency: South East Wales**

There may be time dependency contained within the travel data, but due to the consistently high variance perhaps it is not immediately apparent. Some statistical analysis was performed on the data provided by WAST by time of day and day of week, to see if any dependency for travel was significant. Very little discrepancy over segments of the day, and across weekdays, is seen, even when analysed by location. For this reason time-dependency is only considered for demand volume and not route travel times in all subsequent modelling. This is possibly explained by the fact that in Wales, there are few major roads, except around the two cities. Either congestion plays less of a role in variation than has been witnessed in other studies around the world, or perhaps, smaller rural roads mean that vehicles would not be able to travel much quicker in off-peak times than rush hours. Since the region is mainly rural, traffic would be expected to be lower than in other areas of the country, and it is possible that congestion is fairly stable within each of the South East districts.

By using a model that predicts travel time for each and every route, it is possible to apply the model to other regions, other ambulance trusts, and incorporate any time dependency quite easily if necessary compared to using regional (or district) average times or empirical data approaches.

## 5.13 Application of Estimation Method

### 5.13.1 Introduction

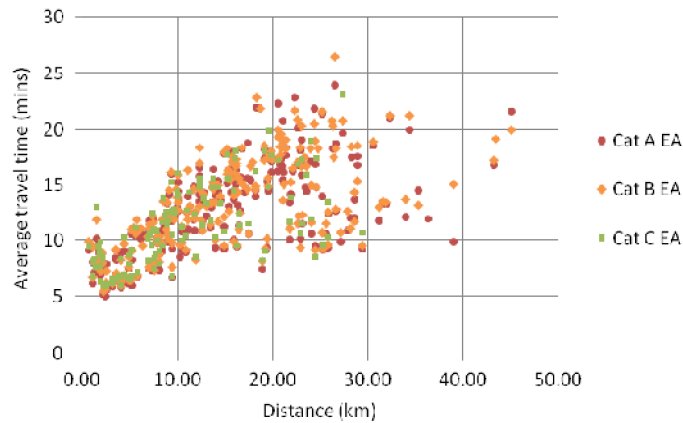
Utilising the travel distance information obtained from the Travel Matrix Generator Tool, regression techniques are used to fit models to the known routes of Figure 5.9 which are then applied to all other network routes to find suitable model parameters by emergency category and vehicle type.

In a similar fashion to Kolesar et al. (1975), approach 3 of section 5.12.3 is now applied to the WAST journey data. Travel time for a route is taken to be the average of all travel time values witnessed. For the South East region of Wales, there are 50 demand nodes (at postcode district level), 23 stations and seven hospitals. A non-symmetric matrix is required since it is not guaranteed that a route will have the same distance value in both directions. In total, there are  $(50 + 23 + 7)^2 = 6,400$  possible non-symmetric routes, with only 350 of these routes beginning at demand nodes and ending at hospital facilities (transportation journeys). Although the average travel times for routes are not an accurate representation of actual average travel time, representative regression models can be built to better estimate the expected length of a journey in the region and incorporate the variation expected due to demand aggregation, driver route decision, weather, congestion and time dependency via the inclusion of the single dependent travel variable.

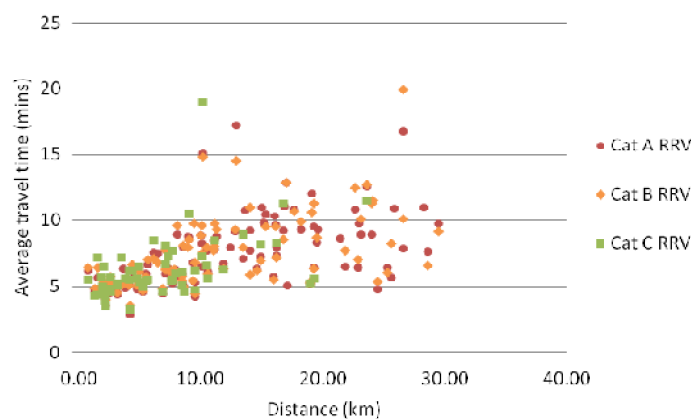
### 5.13.2 Response Journey Correlation

Initially, correlation visualisations were created for travel time and distance data in order to see the relevance and suitability of a regression technique. Response journeys are more difficult to analyse due to the inaccuracy in recording the vehicle start position, as already discussed in detail. To demonstrate this fully and for comparison with transportation journeys, correlation plots are produced (Figures 5.10 and 5.11). A trend in correlation can be seen, with coefficients of 0.64 for EAs and 0.59 for RRVs; however the spread of the data seems to increase as distance increases and there are many outlying data points. These observations are enough to tell us that linear regression analysis via the method of least squares is not directly appropriate in this case. The increase in dispersion as distance increases implies that the error data are not Normally distributed and that the homoscedasticity condition of the residuals will be violated (i.e. variance is not constant).





**Figure 5.10** Distance against average travel time for each route by category for EA response journeys (categories A, B and C only)



**Figure 5.11** Distance against average travel time for each route by category for RRV response journeys (categories A, B and C only)

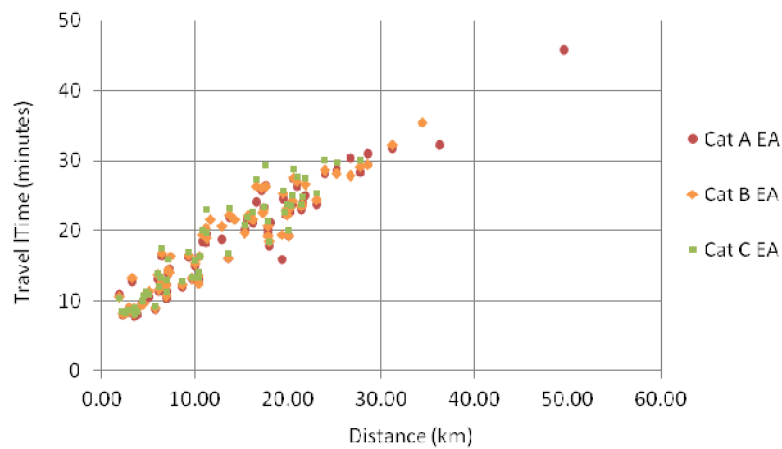
To combat this effect, regression could possibly be performed with a transformation on the dependent variable, yet this has not been attempted due to the issues that lie with the starting point assumption which a transformation would not resolve.

The average speed for a route can be calculated using the Google Maps distance and average travel time for all the journeys between a location pair. The effects of the high level of variance within the response journey data are then apparent. For example, for category A, EA vehicles only, the response journey average speed ranges from around 235 kph (146 mph) to as low as 5 kph (3 mph). It is obvious that these extremes are unviable, leading to a skew in the data, with outliers and curves as seen in the scatter-plots of Figures 5.10 and 5.11.

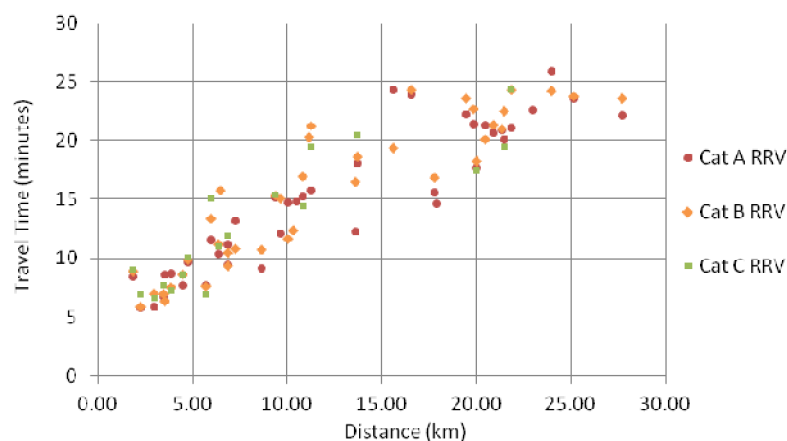
### 5.13.3 Transportation Journey Correlation

The resulting correlations for transportation journeys are portrayed graphically for each of the patient and vehicle types in Appendix 5.2. Instead, combining all data points for each of the individual categories by vehicle type produces the results shown in Figures 5.12 and 5.13.

Although more sparse for RRV journeys, there is still an obvious trend in the transportation data, supporting a strong positive correlation between distance and journey length for all emergency classifications and both vehicle types.



**Figure 5.12** Distance against average travel time for each route by category for EA transportation journeys



**Figure 5.13** Distance against average travel time for each route by category for RRV transportation journeys

Little difference is witnessed in the patterns between the emergency categories within vehicle type, yet, the difference is more pronounced across vehicle types. Despite similarities for categories within a vehicle type, grouping these may in fact hinder the estimation process. Where comparison tests prove the legitimacy of this grouping, the resulting detail level would still be decreased. Even a slight difference in speed travelled by vehicles to the various emergency categories will have an influence on system performance. It will become evident that modelling the category of incident separately allows the best insight to the service system, and will allow changes in policy to be included in subsequent analysis. In particular, category A patients are often served differently with a 'blue light' response reducing the overall travel time to these incidents.

#### 5.13.4 Regression Analysis for Average Travel Time Estimation

Regression analysis is a parameter estimation technique used where a linear relationship between variables is believed to exist (Kleinbaum et al. 2008). The value of a variable may be predicted given a combination of other variables (assumed to be without error) and a constant (or intercept). Simple linear regression is a popular choice due to the ease of investigation and aims to fit an appropriate line to data by minimising the sum of the residuals squared via calculus or numerical evaluation. This method of least squares was used successfully by Kolesar et al. (1975) and Goldberg and Paz (1991) for similar emergency service travel time problems as already mentioned. Multivariate linear regression deals with several independent variables and so it is more difficult to find the optimum model (fitted curve) unless receiving help from a computer package.

If a general linear model is assumed to be of the form:

$$\hat{y} = \alpha + \beta_1 x_1 + \dots + \beta_k x_k$$

where  $x_j$  are the independent variables (which may be raised to powers),  $\beta_j$  are the linear coefficients ( $j = 1, \dots, k$ ) and  $\hat{y}$  is the value of the model, then the difference between an observed value and a predicted one, known as a residual, is expressed as:

$$r_i = y_i - \hat{y}_i = y_i - (\alpha + \beta_1 x_{1i} + \dots + \beta_k x_{ki})$$

where  $y_i$  is the  $i^{\text{th}}$  observed value of the dependent variable,  $\hat{y}_i$  is  $i^{\text{th}}$  the predicted value from the model.

For linear regression, assumptions regarding the dependent variable are that it is:

- continuous;
- approximately Normally distributed.

Further assumptions that must be made about the model are that the random error:

- should have a Normal probability distribution;
- has a probability distribution with a mean of zero;
- has a probability distribution with constant variance for independent variables;
- associated with any two observations are independent.

The validity of linear regression can be verified through analysing graphical representations of the random error or residuals. If the scatter appears random, the residuals can be assumed to be Normally distributed, satisfying the homoscedasticity (equal variances) assumption.

### 5.13.5 Tested Models

In this study, only one independent variable is considered – distance – therefore *simple* linear regression analysis is undertaken. Four separate travel time prediction models are considered. Models 1, 2 and 4 mirror work by Kolesar et al. (1975). The additional inclusion of Model 3 was decided upon after a preliminary investigation of the other three. Model 1 is linear, the second and third have non-linear functions and the fourth is a piecewise square-root – linear function.

$$\text{Model 1: } Y = a + bX \quad (5.4)$$

$$\text{Model 2: } Y = aX^{0.5} \quad (5.5)$$

$$\text{Model 3: } Y = a + bX^{0.5} \quad (5.6)$$

$$\text{Model 4: } Y = \begin{cases} cX^{0.5}, & X \leq d \\ a + bX, & X > d \end{cases} \quad (5.7)$$

Despite some of the models representing non-linear *functions*, linear regression can be run in all cases since all *coefficients* in the models are linear in form.

Let  $Y$  be the travel time to be estimated,  $x$  the distance for the route obtained via the Travel Matrix Generator Tool,  $a$ ,  $b$  and  $c$  be appropriate regression model coefficients and  $d$  (a constant) which represents the distance at which a change in the slope of Model 4 occurs (see Kolesar et al. (1975) for further explanations regarding  $d$ ).

### 5.13.6 Method of Least Squares Fit

Parameter fits are found for these four models, predicting transportation travel time of journeys in South East Wales from Google Maps distance. Simple linear regression is conducted separately for each emergency category and vehicle type and the method of least squares is applied using the Microsoft Excel Solver add-in tool. Budge et al. (2010) use Solver and the maximum (log) likelihood approach to find the components of their proposed model. The four incident categories and two vehicle types are dealt with. Category AS2 and AS3 are only served by Emergency Ambulances (EAs) and High Dependency Units (HDUs), which are essentially equivalent, so these two emergency types are combined to form one category that receive only EA vehicle attendance. The best performing model in each case is selected, producing seven individual models in total. The comparison of the residuals and least squares for all the models can be seen in Table 5.1.

**Table 5.1** Regression analysis coefficients, minimum sum of residuals squared, correlation coefficient and R-squared values for each Model by category and vehicle type

Model	Coefficients	A EA	B EA	C EA	AS2/3 EA/HDS	A RRV	B RRV	C RRV
1	a	7.23	7.47	6.90	7.76	5.85	6.29	6.31
	b	0.80	0.82	0.90	0.94	0.73	0.76	0.75
	Min Residual Sum	341.46	362.62	325.86	488.07	222.28	250.13	113.86
2	a	5.27	5.33	5.44	6.12	4.50	4.72	4.65
	Min Residual Sum	454.14	388.07	348.55	747.35	216.08	221.93	97.79
3	a	-2.64	-1.63	-2.07	-4.77	-1.57	-1.61	-0.36
	b	5.93	5.75	6.00	7.26	4.92	5.16	4.77
	Min Residual Sum	407.15	371.34	327.59	626.75	206.77	212.06	97.51
4	b	0.88	0.90	0.94	0.93	0.74	0.70	0.80
	d	6.87	6.87	6.87	8.69	7.82	10.94	6.87
	Min Residual Sum	404.92	420.77	339.41	498.81	222.87	231.52	127.25
	Best Fit Model	1	1	1	1	3	3	3
	Correlation Coefficient	0.954	0.942	0.934	0.950	0.922	0.924	0.902
	R-Squared	0.911	0.887	0.872	0.902	0.850	0.854	0.814

For EA vehicles, the best model for all categories is the linear function of Model 1, Equation 5.4, with an R-squared value of 0.91. For RRVs however, the best fitting model is in fact the square-root function of Model 3, Equation 5.6, with R-squared value of 0.85. This result is supported by the slight curve in the correlation scatter-plot of Figure 5.13 for RRVs. The resulting models can be seen to be fairly similar for categories within a vehicle type group:

EAs per category:

$$\text{A: } \hat{Y} = 7.23 + 0.80X + \varepsilon \quad (5.8)$$

$$\text{B: } \hat{Y} = 7.47 + 0.82X + \varepsilon \quad (5.9)$$

$$\text{C: } \hat{Y} = 6.90 + 0.90X + \varepsilon \quad (5.10)$$

$$\text{Urgent: } \hat{Y} = 7.76 + 0.94X + \varepsilon \quad (5.11)$$

RRVs per category:

$$\text{A: } \hat{Y} = -1.57 + 4.92X^{0.5} + \varepsilon \quad (5.12)$$

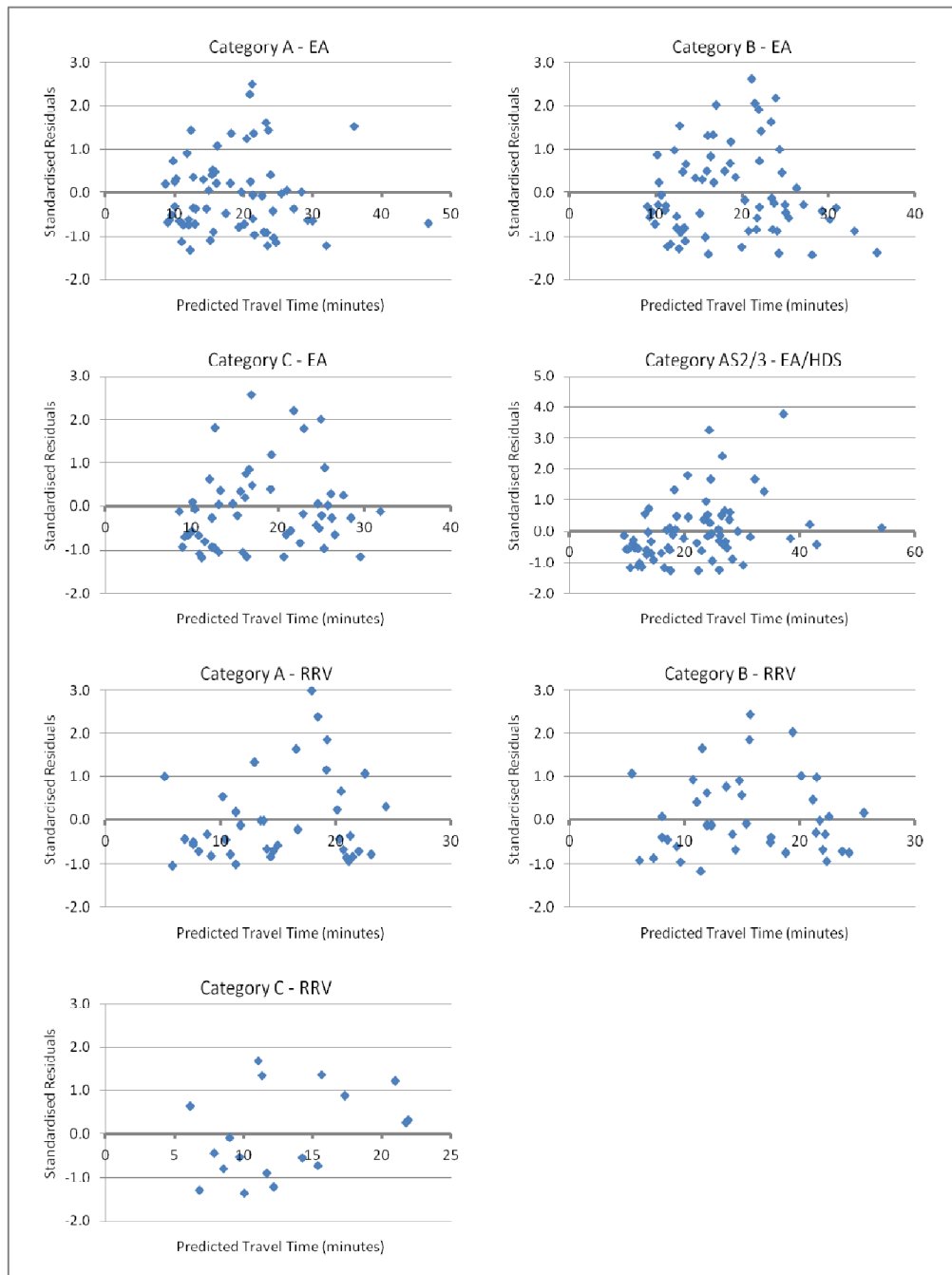
$$\text{B: } \hat{Y} = -1.61 + 5.16X^{0.5} + \varepsilon \quad (5.13)$$

$$\text{C: } \hat{Y} = -0.36 + 4.77X^{0.5} + \varepsilon \quad (5.14)$$

The error terms ( $\varepsilon$ ) in the final models account for variability in the travel time that is independent of distance. This variance, which can lead to inaccurate measurement (Budge et al. 2010) could be caused by the way the data is recorded, aggregation of zones or origin of vehicles. Other causes of expected variability in the overall travel time might be founded in weather conditions, congestion, driver preference, and vehicle condition and variety. It will be shown later, in Chapter 7, section 7.5.8, that the error term can be calculated in many ways. One possibility is to sample the error value from the associated distribution of the normalised residuals. An alternative model is to assume that the travel time itself must be sampled from an appropriate distribution, whereby the distribution mean is given by  $\hat{Y}$  and the variance could be the mean of the variance values witnessed over the routes in the data. In this vein, regression analysis using  $\hat{Y}$ , the predicted average travel time value as the independent variable could be used similarly to estimate the variation. The decision of variation prediction will also be discussed further in Chapter 7.

### 5.13.7 Residual Analysis

For regression analysis, the validity of the model must be checked before the results can be accepted. Residuals of the best fitting model in each category and vehicle case are standardised and plotted against the predicted values of travel time to check their normality and spread. Figure 5.14 shows these distributions. Since all standardised residual plots show the random error assumptions of linear regression have been met, then the suggested models in Equations 5.8 - 5.14 can be accepted.

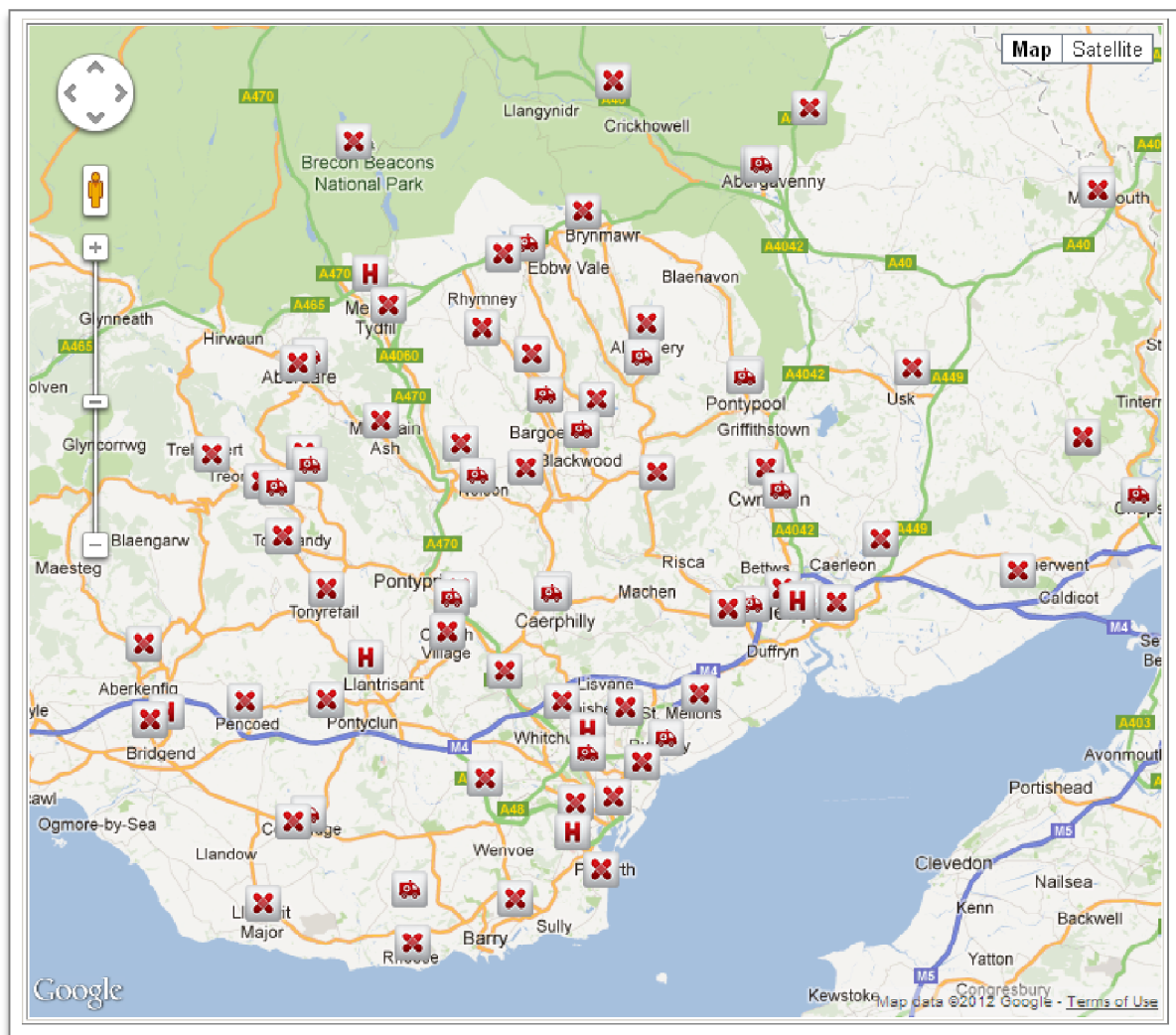


**Figure 5.14** Standardised residual plots of the chosen best fitting model for category and vehicle

## 5.14 Results

### 5.14.1 Google Map Locations

All postcode districts, station locations and hospital addresses were submitted to the Google Maps Travel Matrix Generator Tool, resulting in the travel time and distance matrices shown in section 5.14.2, and visually plotted on a static Google Map by the tool, to display the locations spatially (Figure 5.15).



**Key:** Postcode Demand District      Ambulance Base      Hospital Facility

**Figure 5.15** Google Maps API interface inbuilt to Travel Time Matrix Generator Tool, displaying all demand nodes, potential vehicle bases and hospitals in the South East Wales region



## 5.14.2 Travel Time and Distance Matrices

	CF10 (D)	CF11 (D)	CF14 (D)	CF157PE (D)	CF23 (D)	CF24 (D)	CF3 (D)	CF31 (D)	CF32 (D)	CF355NP (D)	CF375DG (D)
CF10 (D)	0	6	14	15	14	7	18	34	33	25	22
CF11 (D)	6	0	15	16	17	9	19	29	28	20	23
CF14 (D)	15	15	0	8	13	14	15	29	28	20	15
CF157PE (D)	17	17	7	0	15	16	15	20	27	10	9
CF23 (D)	14	17	13	16	0	9	9	36	35	27	23
CF24 (D)	7	10	15	16	8	0	11	37	36	28	23
CF3 (D)	17	18	15	15	9	11	0	35	34	26	22
CF31 (D)	33	30	27	25	35	36	34	0	11	13	32
CF32 (D)	32	28	26	24	34	35	32	12	0	13	31
CF355NP (D)	24	21	19	17	27	28	25	13	15	0	24
CF375DG (D)	23	23	13	9	21	22	21	33	32	24	0
CF38 (D)	26	25	17	13	25	26	24	31	30	22	10
CF39 (D)	33	29	27	23	35	36	34	25	21	19	19
CF40 (D)	40	37	31	27	39	40	38	32	28	26	23
CF41 (D)	42	41	32	28	40	41	39	32	27	31	24
CF42 (D)	53	52	43	39	51	52	50	38	33	37	35
CF434ND (D)	40	41	31	27	39	40	38	39	35	33	23
CF447RG (D)	46	46	36	32	44	45	43	52	40	46	29

**Figure 5.16** A selection of Travel Time Matrix results between all demand postcode districts, stations and hospitals for South East Wales

	CF10 (D)	CF11 (D)	CF14 (D)	CF157PE (D)	CF23 (D)	CF24 (D)	CF3 (D)	CF31 (D)	CF32 (D)	CF355NP (D)	CF375DG (D)
CF10 (D)	0	3.68	9.07	12.61	8.99	3.53	12.86	38.86	40.48	32.12	19.08
CF11 (D)	3.74	0	9.09	12.62	11.95	5.54	15.53	33.25	34.87	26.51	19.09
CF14 (D)	9.13	9.12	0	5.4	6.82	8.17	12.8	30.43	32.05	23.69	11.87
CF157PE (D)	12.7	12.69	5.35	0	12.8	11.74	15.77	30.51	32.13	23.77	6.45
CF23 (D)	0.32	12.04	6.76	13.14	0	5.91	6.06	40.22	41.04	33.47	19.61
CF24 (D)	3.78	5.76	9.63	13.16	5.75	0	9.35	38.2	39.82	31.45	19.63
CF3 (D)	10.36	15.31	12.88	16.21	6.16	7.32	0	40.89	42.51	34.14	22.69
CF31 (D)	41.02	35.68	32.8	31.5	40.24	39.19	43.1	0	6.42	9.29	37.97
CF32 (D)	40.24	34.91	32.02	30.72	39.47	38.42	42.33	6.76	0	11.5	37.2
CF355NP (D)	31.89	26.56	23.67	22.37	31.11	30.06	33.97	9.46	9.68	0	28.84
CF375DG (D)	19.16	19.15	11.81	6.82	19.25	18.2	22.28	37.02	38.64	30.28	0
CF38 (D)	21.22	23.66	13.86	8.88	21.31	20.26	24.33	27.75	29.37	21.01	5.44
CF39 (D)	31.57	26.24	23.75	19.94	31.2	30.15	33.65	19.19	15.73	12.24	12.56
CF40 (D)	36.79	31.46	26.9	23.09	34.34	33.29	37.37	24.14	20.68	17.19	15.71
CF41 (D)	36.02	35.87	28.67	24.86	36.11	35.06	39.14	27.1	23.69	21.59	17.47
CF42 (D)	41.67	41.52	34.32	30.51	41.76	40.71	44.79	28.22	24.81	28.78	23.12
CF434ND (D)	35.27	35.26	27.92	24.11	35.36	34.31	38.38	30.94	27.48	23.99	16.72
CF447RG (D)	41.12	41.1	33.76	29.96	41.21	40.16	44.23	42.14	38.68	35.18	23.23

**Figure 5.17** A selection of Travel Distance Matrix results between all demand postcode districts, stations and hospitals for South East Wales

## 5.15 Conclusion

A 2004 review paper highlights work on travel time modelling and emphasises the importance of good travel time estimates (Goldberg 2004), so that knowledge obtained from research in the field may be shared with non-OR specialists such as emergency service managers and medical directors. The problems faced and assumptions necessary when modelling are summarised without detailing the supporting mathematics. This less technical style is key in bridging the gap between research, practitioners and users of such models. Many researchers are coming to appreciate that such knowledge transfer needs to be addressed if new theoretical realisations are to be implemented in real-world applications. Hence the tool developed and described in this chapter has been created in a way that allows transfer of use to WAST themselves, and built generically so that application to numerous other problems is trivial. A guide to support users is available on the Cardiff School of Mathematics research web pages:

[www.cardiff.ac.uk/maths/research/researchgroups/opresearch/healthcare/index.html](http://www.cardiff.ac.uk/maths/research/researchgroups/opresearch/healthcare/index.html)

Unlike other similar studies found in the literature, this EMS travel time estimation analysis makes use of information known for transportation journeys (avoiding the uncertainty in starting and ending location of response journeys), fitting travel time models to the data. Vehicles are likely to travel faster when responding to patients compared to transporting and so the estimated travel times must be scaled to find response journey times if these are additionally required.

Kolesar et al.'s work implies that regression Model 4 (Equation 5.7) should provide the best fit due to the consideration of difference in speed for long and short journeys. It was in fact found that Model 4 was not superior to the simpler models when applied to the WAST system. Looking back to the data in Figure 5.12 and 5.13, the plots appear faintly non-linear. Model 4 would indeed capture a curve in data for lower distance values, yet the Welsh data curves with higher distance. It could be that Kolesar et al.'s square-root phenomenon is not explicitly present in this study since the literature often refers to distances less than a mile as 'short'. Very few of these short journeys would ever be witnessed in Wales due to its rural nature and road network structure. Another reason may be attributable to the size of demand zones chosen, for which little improvement could be made in this particular study.

Emergency Ambulances are found to travel with constant speed, going against the idea of long and short trip discrepancies. For Rapid Response Vehicles, although Model 4 did not give the smallest

Mean Squared Error (MSE), the chosen fit is that of a square-root function, suggesting speed for these smaller vehicles in attendance of the highest priority patients does indeed follow Kolesar et al.'s suggestion of acceleration with distance more closely than for the larger EA vehicles.

Categorisation of travel time data by emergency type and vehicle type may help account for the variance in the original data. It is possible, had Kolesar et al. and the other researchers mentioned in sections 5.5 - 5.10, been able to split their data similarly, the resulting models may have been applied differently to the emergency systems considered. Patient condition may have influence on the speed of vehicle and also the decision to transport. Repede and Bernardo (1994) were among the first to consider travel time as a function of the priority of call. Least squares regression fits are conducted on the WAST data separately for each emergency type and vehicle type to allow the most accurate application to the current south east Wales system. Where a system may be subject to changes in service policies, these travel time estimation models can be easily adapted, or instead serve in experimental situations (for example, simulation modelling) to aid decision makers on the implementation of such changes. It would be possible to evaluate the expected outcome of a patient of a particular category given the relative speed of a responding vehicle and also see the effect on the system should this speed be altered. Therefore, despite similar parameters being found for subgroups of the final models (Equations 5.8-5.14), all are assumed independent, as are the patient categories and responding vehicle types.

The developed chosen models and the distance matrix obtained from Google Maps will be used in the following two chapters as input to the modelling techniques explored.

## Chapter 6

# Location Analysis

### 6.1 Introduction

From allocation of relief resources in anticipation of natural disasters to the transfer of data packages in telecommunications, location theory models have been demonstrated and implemented extensively over the past few decades. Discussions of EMS system studies in Chapter 3 highlight how few of them actually model the importance of patient survival explicitly, instead capturing alternative goals such as maximising coverage or minimising average travel times. Only recently have attempts been made to encapsulate the decline of a patient's health over time taken to respond through location analysis techniques.

In the same way that the construct of coverage relates to the performance measure of meeting a fraction of calls within a target, response time success relates to a clinical outcome measures. It has been shown that real-world modelling of the latter concept is superior to the former (Bevan and Hamblin 2009, Erkut et al. 2008b, Price 2006), and since current EMS targets (of meeting calls within a time) is already a proxy for survival, progression to clinical outcome based performance measures is not far from the current situation, and one that some ambulance services are embracing.

Models of coverage are revisited but with a modern slant in this chapter. The traditional outlook of these location-allocation models is swapped for a viewpoint where performance is measured by the number of patients expecting a timely and positive outcome, referred to as 'survival'. Originating with work by Erkut et al. in 2008, the research presented here builds on standard coverage models by including a survival probability for patients of a particular type given a specific response length. The relevance and value of maximising survival over simple coverage is demonstrated, supporting the change in direction of current research and EMS Trust policies towards clinical outcome (Department of Health 2005, Turner et al. 2006, WAST March 2011). The set of models detailed aim to suggest better allocations of vehicles over a network in order to serve the population, enhancing outcome even for non life-threatened patients. The chapter concludes with a case study; produced vehicle allocations for the South East may be offered to WAST as suggested improvement on current design, or used as benchmark input to a simulation model, as demonstrated in Chapter 7.

## 6.2 Improving EMS Performance with Location Analysis

It is not practical to improve aspects of an EMS system by simply suggesting additional resources or stating processes should be shorter; such decisions should ideally be supported and quantified, and their implications verified through modelling. Even so, by merely reducing average service phase lengths of an EMS system, information regarding the utilisation of vehicles is lost, which gives limited insight to the decision's impact. For example, to improve response time performance, queueing theory (as seen in Figure 3.2, Chapter 3) may be used as a solution methodology to suggest the number of servers necessary to meet demand within set targets. Although this approach provides good comprehension of resource level impact on waiting time of patients, the response time distribution can only be improved credibly if the geographic locations of responding vehicles are considered.

When travel time of a system is a component of its KPI, making up part of the modelling objective, it is imperative to consider the full range of data – the geographic location of the resources – since average response time is not independent of vehicle position. Even when fleet sizes are increased, dramatic impact on response time should still only be seen if optimisation of the allocations is reconsidered (Jenkins 2012), despite a reduction in average response times based on simple queueing theory. For such capacity and allocation decisions, analysis may be conducted through the use of location theory to suggest fleet positioning to meet a coverage or response time threshold.

Although many location models have similar objectives of attempting to minimise some maximum distance or time travelled, or maximise the population covered by the servers, in all emergency service modelling these objectives are surrogates for the overall endeavour of saving lives, even when not stated absolutely (Goldberg 2004, Hong and Ghani 2006).

## 6.3 Coverage

Prevalent in emergency service location, capacity and deployment studies is the idea of coverage of a population. Many solutions to covering location analysis problems lie in mathematical programming techniques, particularly through integer optimisation methods. The purposes of such models are to find the locations on a network that provide the best coverage to the population, not

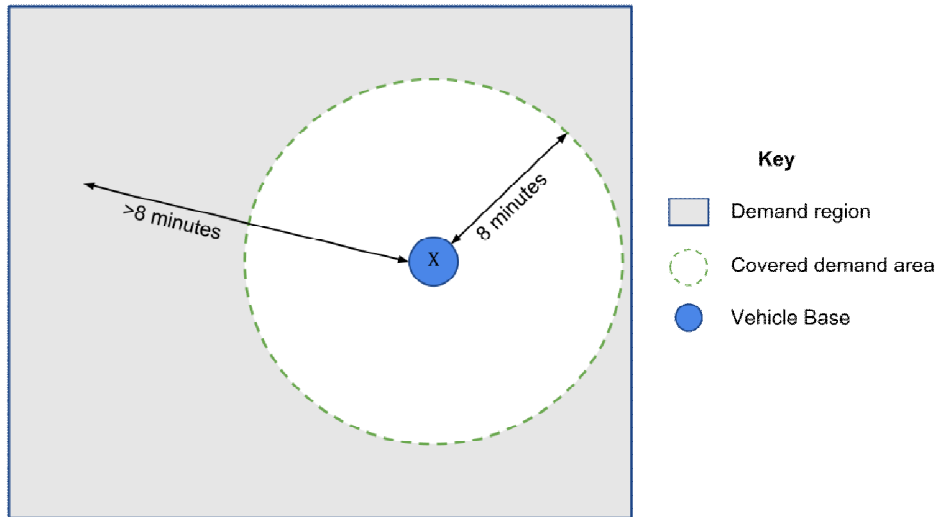
necessarily by ensuring *all* demand points are reachable within the response standard, but by maximising the amount of the population that can be reached.

Early location theory makes use of binary variables (in the objective functions and constraints) within the formulations for coverage; notably, it is often assumed that coverage takes a value of 1 if a demand node is 'covered' (can be reached within a time or distance standard) by a service node and 0 otherwise (Pan et al. 2012). Later, variables were altered to have a more continuous representation using a scale of coverage between 0 and 1 resulting in more gradual models such as Berman et al.'s which decays coverage with distance, capturing some of the sensitivity around absolute travel distances and survival (2003).

Allocation models also make use of binary and mixed-integer linear and nonlinear programming approaches, dealing with the distribution of servers across nodes of a network, where again binary variables may be used to express the decision of situating a resource at a specific node or where integer values represent the number of resources allocated to a particular node. Other models extend to the incorporation of vehicle utilisation (as was seen in Chapter 3) since coverage in earlier studies was based only on proximity and not on demand and availability of the system. Additional attempts are made to deviate from deterministic modelling, incorporating uncertainty in travel time in the network or plane with probabilities of reaching the scene in a given time standard (Budge et al. 2010). Extensions to this type of work were mentioned in the literature review and include back-up coverage and cooperative covering models.

Coverage is an agreeable objective in terms of EMS location problems due to the performance measures that are commonly in place for such services. The idea of maximising the amount of demand within a certain distance or time standard of service nodes, transfers easily into a model that maximised the percentage of calls that are reached within the target time. Erkut et al. (2008b) also point out that coverage models are easily communicated to policy-makers, service managers and the public, and that the integer programs can be solved using basic optimisation software if deterministic assumptions are made. Despite this, standard coverage models for EMS vehicle location are still subject to two types of error according to the authors.

1. **Measurement Error:** since there is no distinction between individual response times within covered areas and within non-covered areas (Figure 6.1).



**Figure 6.1** Measurement Error: covered area with response time of no more than 8 minutes

2. **Optimality Error:** in the location of resources and facilities due to measurement errors of the network even when demand magnitude is considered (see Example 6.1 later).

An additional weakness of covering models is:

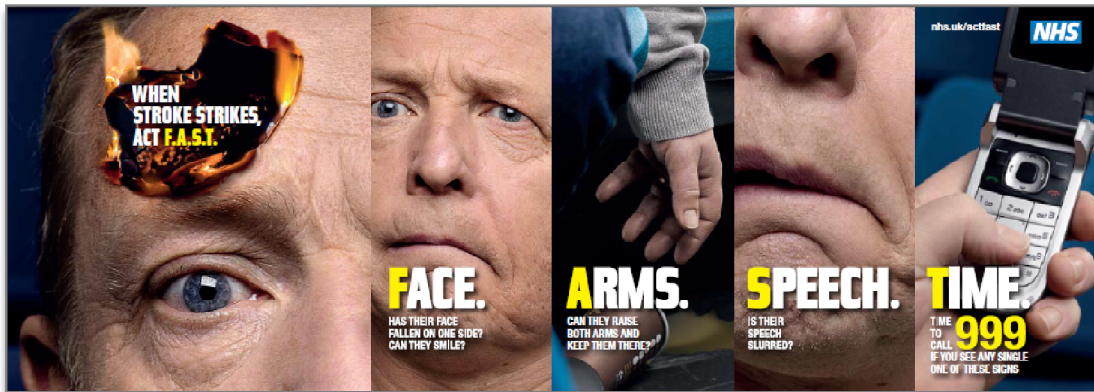
3. **Threshold Error:** if a system can be covered by a minimum of  $x$  stations, then all solutions with more than  $x$  stations will not improve the result; whereas, an increase in resources (maintaining a realistic quantity) should continually increase the probability of a positive patient outcome by reducing response time further if located sensibly.

It therefore appears a more reasonable and accountable objective function should be implemented to measure performance of an EMS system, rather than sole calculation of covered demand within a time standard.

## 6.4 Survival

### 6.4.1 Introduction

*Act F.A.S.T.* – A global acronymic slogan advertised within the UK, Ireland, Australia and the USA, used to draw public attention to the urgency of recognising a stroke and obtaining medical treatment for victims.



**Figure 6.2** F.A.S.T. Campaign (NHS 2009)

*Hard and Fast* - A recent television advertisement by the British Heart Foundation (BHF) encourages witnesses of a person experiencing cardiac arrest to administer hands-only CPR, after calling 999.



**Figure 6.3** Hands-Only CPR Campaign (British Heart Foundation 2012)

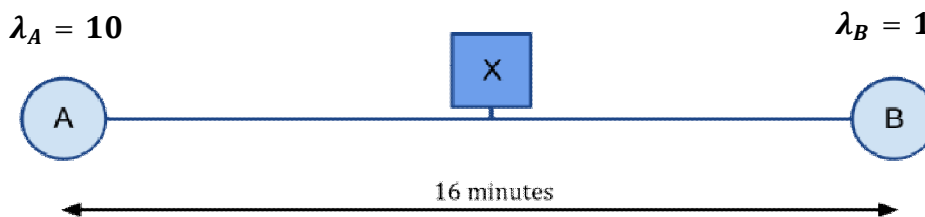
Each informative advert (Figures 6.2 and 6.3) stresses the necessity of recognition and immediate medical assistance in the case of a stroke or cardiac arrest, whether it be preliminary care from bystanders or secondary care from a local EMS. In terms of response "*faster is better, and OR models for EMS system design should take this into account*" (Erkut et al. 2008b). After the launch of the 'F.A.S.T.' campaign, England witnessed a 55% rise in emergency calls reporting possible strokes (BBC News Health 2009), showing an increase in public appreciation of rapid response.



The importance of swift attendance at the scene of an emergency in order to increase chances of survival may seem fairly obvious for such critical and life-threatening conditions, yet the quantity of research in survival based performance measures is much less than response led targets. Although the two may be thought to be identical, Erkut et al. (2008b) demonstrate (replicated here in Figure 6.4 and explained by Example 6.1) how for a network of demand nodes with populations of various sizes, response targets may lead to a poorer allocation decision with fewer 'survivors'.

### Example 6.1

In an EMS system with two demand nodes, 16 minutes road travel apart, where node *A* has a demand,  $\lambda_A$ , of 10 per unit time, and node *B* expects a demand,  $\lambda_B$ , of 1 per unit time, to operate full coverage with an eight minute response target, it makes sense to place vehicles exactly equidistant apart from both nodes at base *X* (Figure 6.4). This positioning however, does not take into account the outcome of patients, or even the relative demands at each node of the network.



**Figure 6.4** Example of coverage versus survival probability modelling constructs based on Erkut et al. (2008b)

Alternatively, if clinical outcome measures are the focus, with survival probability (representing desirable patient outcome as opposed to life or death) calculated using a simple exponent function,  $P(\text{survival}) = e^{-t}$ , (where  $t$  is travel time) then the expected number of survivors at *A* is  $10e^{-8} = 0.0034$ , and at *B* would be  $e^{-8} = 0.00034$ , giving a total number of 0.0037 possible survivors out of 11. If the vehicle were instead to be placed at *A* rather than *X*, the total number of expected survivors would be approximately 10 (91%). This simple example highlights the weakness of coverage but also the ethical issues surrounding inequitable resource distribution and of a profession where service must be all-encompassing (French and Casali 2008, Klugman 2007).

Survival can be thought of as a better option for EMS performance measurement than coverage for several reasons:

- money and lives saved are more attention grabbing than percentage met within an arbitrary time standard;
- survival is already a key message provided to the public in emergency campaigns;
- EMS targets differ around the world, whereas maximising lives saved is a common goal;
- coverage is already a proxy for survival.

Survival as a modelling objective does not refer to actual outcome of a patient within this thesis; instead the term is used to define the chance of a patient experiencing a timely response which enhances their chances of recovery. Patient outcome data were unavailable for this study; yet theoretical survival probabilities, calculated from monotonically decaying survival functions found in the literature, are used to demonstrate an attainable level of success from the response.

#### 6.4.2 Cardiac Arrest

Since the effects of a person's heart stopping can be devastating, potentially resulting in irreversible brain and heart damage or even death, easy and rapid access to Advanced Life Support (ALS) is essential to decrease mortality rates and increase quality of life for victims. Similarly to the Star of Life described in Chapter 2, cardiac arrest has its own defined 'Chain of Survival' (Figure 6.5) for the actions involved in responding to victims.



**Figure 6.5** Process of intervention for victims of cardiac arrest (ChainofSurvival.com 2012)

In an almost heroic example, one Welsh cardiac victim made a full recovery after a perfect implementation of the Chain of Survival (Newman 2012). The chain began with early bystander CPR attempts instructed by an EMS call-taker and assisted by community first responders, shortly succeeded by the arrival of a paramedic-manned RRV to provide defibrillation and followed-up of the nearest vehicle for conveyance, manned by another paramedic and a technician.

Out-of-hospital cardiac arrest (OHCA) is the most commonly researched medical condition when it comes to response time effectiveness (O'Keeffe et al. 2010). There are several studies which investigate similarly the effects on patient survival for stroke (Rajajee and Saver 2005), heart attack (myocardial infarction) (Cretin and Willemain 1979) and in-hospital cardiac arrest (IHCA) (Moretti et al. 2007); however, little is known regarding short term outcomes for other conditions given time taken from onset to intervention.

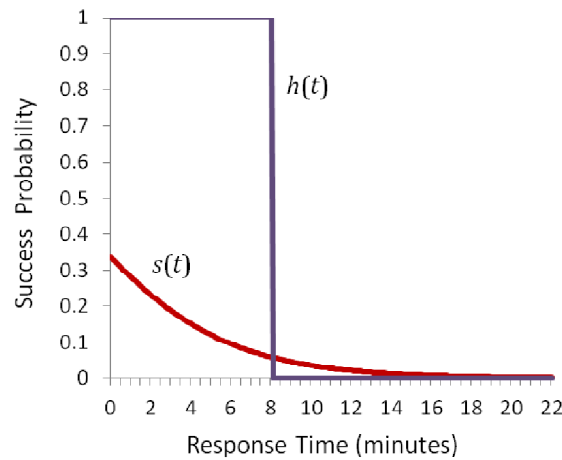
Following a large study in Canada, (OPALS 2004), pre-hospital interventions for four patient groups were evaluated:

1. cardiac arrest;
2. major trauma;
3. respiratory arrest;
4. chest pain.

Preliminary results for 2, 3 and 4 showed no real benefit from Advanced Life Support (ALS) interventions (equivalent to a paramedic response) compared with the original Basic Life Support (BLS) program. Other studies have also found that trauma patient outcome is not affected by ambulance response time (Pons and Markovchick 2002, Turner et al. 2006). For cardiac arrest however, the OPALS study advocates bystander CPR and shows early defibrillation does improve survival, in line with BLS. These findings further support Mayer's claim (1979) that there is a relationship between paramedic response time and survival rate.

An example function developed to represent survival until hospital discharge for OHCA is shown in Equation 6.1 (De Maio et al. 2003), where  $t$  is the response interval from onset to defibrillation. Graphically, this decaying survival function is represented over time in Figure 6.6 and is shown alongside the hard target step function (from Equation 6.2) for category A calls (which by definition, includes all correctly triaged cardiac arrest patients). The eight minute guideline originates from an article based on survival findings for cardiac arrest (Eisenberg et al. 1979), and has since been used

as a general target for most emergency responses (Fitch 2005), so comparison with the cardiac arrest survival approach is legitimate.



**Figure 6.6** Survival function,  $s(t)$ , estimated by De Maio (2003) compared with current category A hard target,  $h(t)$ , represented as a step function (binary coverage)

$$s(t) = (1 + e^{0.679+0.262t})^{-1} \quad (6.1)$$

$$h(t) = \begin{cases} 1 & \text{if } 0 \leq t \leq 8 \\ 0 & \text{if } t > 8 \end{cases}, t \in \mathbb{R} \geq 0 \quad (6.2)$$

When system performance using the hard target approach is translated into survival, the current target of an eight minute response greatly overestimates clinical outcome for category A patients met within the time standard (deeming all services successful), yet predicts no clinical success whatsoever if patients are responded to later. This ignorance of actual patient outcome also fails to discriminate between an instant response and one taking eight minutes. The survival function approach however, represents a varying slim, (but non-zero) chance of success given response over eight minutes, with realistic, yet less optimistic, results between nought and eight minutes, providing a more suitable platform for measures of success. Immediately, it can be seen that the current eight minute target is not optimal as there is approximately an 85 percentage point decrease in expected survival from an immediate response to one taking the target time.

Erkut et al.'s exploitation of a similar survival function for patients experiencing cardiac arrest in their models is said to be for four reasons:

- time has been shown to be critical to clinical outcome in such cases (Holmberg et al. 1998, Mayer 1979) since they are the highest priority emergency conditions;
- the current eight minute response time standard is based on cardiac medical research;
- the relationship between survival has been researched for response time to cardiac patients (but little for other conditions);
- their data contained a large proportion of cardiac arrest emergency calls.

The first three points are supported by the earlier literature discussion; the fourth point is also true of the South East Wales data set obtained from WAST for 2009, where 3% of category A calls are initially logged as cardiac arrest. Grouping heart attack, cardiac arrest and stroke incidents (which are the most prevalent high priority life-threatening conditions) accounts for more than 32% of category A demand.

The full and immediate impact on survival of a population from system changes is not clear by looking exclusively at cardiac arrests or even just category A emergencies. Potentially multiple survival curves would be required as input if every medical condition wished to be modelled accurately and separately. Eisenberg (1979) noted that neither is response time solely accountable for survival, but instead two individual time components should be considered: time from onset of cardiac arrest to cardiopulmonary resuscitation (CPR) and from onset to definitive care.

### **6.4.3 Survival Function Development**

Few pre-hospital based survival functions are found in the literature. For any emergency condition, despite the lack of research, response time is the fundamental component in terms of survival probability or even simply acceptable levels of patient care and service. It is desirable to amalgamate groups of patients with similar functions to reduce modelling time; nevertheless, group composition should maintain integrity to ensure not just the major incidents are considered.

Data analysis literature suggests a non-linear trend between survival and response. Survival curves are anticipated to plateau once a paramedic arrives with the patient, so that the probability will no

longer decrease significantly when considering survival onward to hospital (or even discharge, assuming transfer of care adheres to guidelines) (Eisenberg et al. 1990); hence the theoretical survival function is used only for response duration and not for the continuation of service or transportation (Larsen et al. 1993). In reality, survival probabilities do not reach zero as they do with some theoretical functions, but the limits can still become very low when considering such critical life-threatening conditions as cardiac arrest, with a minimum of 3% (compared to the predicted 0%) survival reported by Larsen et al. (1993) for an eight minute to CPR target.

Regression (multiple linear and logistic versions) is a commonly used tool in determining survival functions (Larsen et al. 1993, Pons et al. 2005, Valenzuela et al. 1997), since paramedic response time, although a key component, is not the sole contributor to survival. Based on the Chain of Survival, it is likely many other factors, such as bystander intervention, times between incident, call and dispatch, initial responder type, patient age (Herlitz et al. 2004), instructions given for bystander CPR (Lerner et al. 2012, Rea et al. 2001), resuscitation consistency (Valenzuela et al. 2005) and intervention type (Iwami et al. 2007, Sayre et al. 2008) will all influence patient outcome.

Valenzuela (2000) offers a survival function that was found after studying data collected from casinos, of people who suffered cardiac arrest whilst on the premises. The study supports prediction of expected survival probability from response time, but requires also information for time from attack to bystander intervention, paramedic response and other explanatory variables to be included, which in many situations are unavailable or unknown. Since this information is difficult to collect and is rare, it may be possible to utilise coefficients of explanatory variables obtained from such previous studies. Alternatively, averaging over several behavioural variables in the original regression model may provide a function that predicts survival solely on response time (or distance) (De Maio et al. 2003). Erkut et al. point out however, that the estimated survival function is dependent upon the data used in its development. The data will be influenced by the structure of the EMS system and the region from which it was derived, which may be substantially different or not even comparable to the proposed system. Calibration for application to a new region would therefore still be necessary.

#### 6.4.4 Challenges

Clinical outcome in EMS systems is affected by factors such as proximity to Automated External Defibrillators (AED) (Cardiac Science n.d.), demography, bystander willingness and ability to assist, and care provider. This leads to large variation in patient outcome for cardiac arrest (amongst other conditions), with survival rates in some locations being significantly higher than others (Field et al. 2010). It has been suggested also, that two further links should be added to the Chain of Survival seen in Figure 6.5 – ‘Early Intervention’ or ‘Recognition’ and ‘Post Cardiac Arrest Care’ – with EMS managers and operational staff ensuring any weaknesses in their specific chain are identified and monitored. The early intervention aspect refers to a link prior to the first one of the chain. A witness to the cardiac arrest could administer some form of CPR before arrival of an EMS crew in order to increase chances of survival. This perhaps is the biggest current weakness of the system in Wales, and with better awareness, triaging, telephone instruction and increased numbers of AEDs, particularly in rural areas, lives could be saved.

Survival rates are not uniform in their definition or modelling implementation (Eisenberg et al. 1990, Eisenberg et al. 1991). The most common definition is of survival to hospital discharge, assuming response to be the interval from onset to arrival of a paramedic (or beginning of ALS). For the purposes of this study, the functions utilised refer to probability of survival to hospital discharge (although the true outcome of patients is unknown from the WAST data), based upon a response time from the logging of the 999 call to arrival of an initial EMS vehicle at the scene.

#### 6.4.5 Survival and Location Theory

After many years of response based EMS performance modelling, research is moving slowly towards more clinical outcome based measures, whereby the survival and post-treatment quality of life of patients is beginning to be included in mathematical models.

Population coverage has been widely used as a proxy for patient survival in location theory, with recent inclusion of known survival curves to mathematical programming objective functions. Organisations recognise that reporting numbers of lives saved or direct costs can have more impact than reporting percentages of responses met in arbitrary time standards, placing the user at the centre of the system (Audit Committee 2009). Even though time standards have been set based on

medical research, in many cases this explanation is not general public knowledge. Improvements are more easily spotted however, when numbers of lives saved is the metric quoted across Trusts and between policies.

Survival is not the only patient focussed (as opposed to system focussed) measure that accounts for effectiveness of a service. If access were readily available to such data, it would be possible to also consider medical care costs, Quality of Life Years (QALYs), hospital length of stay and other such consequential factors. For the purposes of this study, the measurement of performance will be taken to be survival (not in terms of actual patient outcome, but in terms of probability of a positive patient outcome based on a theoretical distribution) compared with coverage.

Chelst and Jarvis (1979) state that it would be possible to make "*great savings in research time and effort if the existing models that estimate travel times could serve as the foundation on which to build these newer [outcome based] models.*"

Erkut et al. (2008b) devise a new generation of location models with the Maximum Survival Location Problem (MSLP). They account for this survival probability within an existing coverage model, locating EMS vehicles in order to maximise the survival of a population. The authors give details of the weakness of the previous coverage approaches, and demonstrate how a survival-maximising approach benefits the service and population of Edmonton, Canada. Results presented show the expected number of survivors using their proposed models, where survival (until hospital discharge) after a cardiac arrest is given by the function in Equation 6.3 based on the function devised by Valenzuela (2000).

$$s(t) = (1 + e^{0.26+0.139t})^{-1} \quad (6.3)$$

The MSLP model makes progress into the clinical outcome based location objectives, yet, as stated by Knight et al. (2012a), limitations surround the consideration of only one group of patients at a time. Many different medical conditions are dealt with daily by any EMS Trust, for which, different levels of response are required with different targets. Each critical patient would therefore have a distinct expected survival value, highlighting the need to consider response time and survival by emergency condition. A second weakness of the MSLP is that only one type of vehicle is considered for the allocation. In reality, EMS systems operate with multiple types of vehicles that each have specific roles and may be managed in different ways.



Ongoing work in Ontario, Canada, supports the findings of Erkut et al. in the importance of survival outcomes for EMS systems. After discussion with Dr. Jewkes of the University of Waterloo (2011), the new models proposed in the next sections are believed to also support current EMS endeavours and follow on suitably from earlier work in the field with a novel contribution.

Accommodating for differences in response procedures for different emergency types and sub-fleets, the Maximal Expected Survival Location Model for Heterogeneous Patients (MESLMHP) (Knight et al. 2012a) and the Maximal Expected Survival Location Model for Heterogeneous Patients with Heterogeneous Fleet (MESLMHPHF) aim to deal suitably with the allocation of a mixed fleet to existing stations whilst best serving a diverse population.

## **6.5 Modelling Heterogeneous Patient Groups**

### **6.5.1 Introduction: MESLMHP**

Where OR techniques are applied directly to real-world problems, the reliability of the modelled environment is dependent upon assumptions made and the precision of system processes captured. As witnessed in the literature, in EMS modelling, patients are often considered to be homogeneous, that is, from the same demand pool. Some studies do account for the discrepancies in patients in terms of the type of emergency for which they require service, yet the category prevalence may also depend on location. It is more unusual for a study to consider the geospatial distribution of demand in conjunction with emergency condition. Most models consider only the vehicles as entities and the output from models relates to the success of response time from the vehicle perspective. In such potential 'life-death' situations as with emergency service systems, it is important not to confuse the outcome of an individual with the output of the service.

The following new EMS allocation models aim to prioritise positive patient outcomes compared to simply maximising responses achieved within the time standards. WAST have already moved towards this style of performance measure, and some other Trusts across the UK and elsewhere are doing the same. It is not a simple transition, and of course is expected to take some time to develop fully within a system, but with modelling tools and techniques as may be supplied by OR, the possibility of a swift changeover in policy could be increased.

The first model described – the Maximal Expected Survival Location Model for Heterogeneous Patients (MESLMHP) – strives to accommodate more than one type of patient by seeking to satisfy objectives for different response measures. Inclusion of clinical outcome for various category-based targets has (at the time of writing) yet to be seen within the coverage location theory literature.

Compliance tables are one of the resulting outputs of the proposed models. Where fleet capacity can be altered by the Trust in reality, for example if additional resources were obtained, or for capacity fluctuations by shift or weekday, then an optimal allocation reference exists for operational and strategic planners to exploit.

### **6.5.2 Model Brief: MESLMHP**

Primarily, the motivation for embarking on this allocation modelling task is to improve the chances of a positive clinical outcome for EMS users. The MESLMHP aims to maximise the survival of multiple patient groups, for various emergency medical conditions, given a particular fleet. Initially, only a homogeneous fleet of specified size is considered.

Demand on an EMS comes in more than one form and not all emergencies require a response within the same target time. Some conditions, such as cardiac arrest and stroke, require immediate attention, and so a short response time target is set. Other conditions, including trauma, do still require emergency service, but with less urgency than those that are immediately life-threatening. Extending Erkut et al.'s model from a single life-threatening emergency condition (cardiac arrest) to different targets for a variety of emergency types witnessed in Wales, the proposed model's features are largely generic, allowing application to other similarly structured EMS systems and any number of emergency classifications.

Implementation of the model follows the pre-defined Welsh categories, referred to as A, B, C and Urgent in this chapter. The possibility of splitting categories further by medical condition or grouping existing categories together – for example B and C since their current hard targets are equivalent – is flexible; however, it is important to be aware when considering the formation of new groups, that for some patients, their condition might deteriorate if a long response time is experienced. Although categories B and C have the same UK target, they are modelled separately to avoid assuming equal priority. The Urgent category represents the combination of categories

AS2 and AS3 since they are similar in nature and occur less frequently in the data than the three AS1 type emergencies. By weighting demand within the mathematical programming models, service prioritisation is enabled.

Categories for survival curves may be taken to correspond to existing emergency categories, or, through clustering techniques, classes may be agreed to represent severity or a particular type of medical condition. The MESLMHP allows multiple survival functions to be implemented in the current system to see the impact on the service, whilst current hard targets may also be used to find the optimal allocation of vehicles for the current performance measures. In addition, a mixture of these two performance types (via survival and step functions) may be explored in a single system.

### 6.5.3 Notation & Formulation: MESLMHP

Let  $m$  denote the number of demand nodes,  $n$  the number of service nodes and  $k$  the number of patient types. Using the notation  $[a] = \{x \mid 1 \leq x \leq a\}$  for any  $a \in \mathbb{Z}$ , where  $m, n, k \in \mathbb{Z}$ :

- demand of type  $l \in [k]$  from demand node  $i \in [m]$  is denoted by  $\lambda_i^l$ ;
- average ambulance utilisation at a given service station  $j \in [n]$  is given by  $\pi_j$ ;
- therefore, the probability that a vehicle at station  $j$  is available to respond is given by  $(1 - \pi_j)$ .

To ensure that all stations are given an order of preference for allocation to each demand node, let  $\rho \in \mathbb{R}_{\geq 0}^{m \times n}$  denote the preference matrix; for demand node  $i$ , the  $j^{\text{th}}$  favoured choice of station,  $\rho_{ij}$ , having no available vehicle (occurring with probability  $\pi_{\rho_{ij}}^{x_j}$ , where  $x_j$  is the number of vehicles located at the station  $j$ ) implies the  $(\rho_{ij+1})^{\text{th}}$  service node will be the next station to be selected.

It is possible to have a different survival function for each patient type  $l$ ,  $s_l: \mathbb{R}_{\geq 0} \rightarrow [0,1]$ . Example survival functions were seen in Equations 6.1 and 6.3. The function used for analysis in this study is that of Equation 6.3 (Valenzuela et al. 2000).

For each demand-service node pair, the travel time is required as input to the model, represented by travel matrix  $t_{i,j} \in \mathbb{R}_{\geq 0}^{m \times n}$ . Finally, since the different groups of patients are based on the urgency of the medical condition, a weight  $w_l$  for each patient type is required in order to prioritise the demand.

All model parameters are summarised in Table 6.1, identifying also their contribution to the forthcoming models.

Beginning with the probability of survival of a patient of type  $l$  from demand node  $i$ , serviced by station  $\rho_{ij}$ , the formulation is as follows:

$$P_{i,\rho_{ij}}^l = s_l(t_{i,\rho_{ij}}) \left(1 - \pi_{\rho_{ij}}^{x_{\rho_{ij}}}\right) \prod_{r=1}^{j-1} \pi_{\rho_{ir}}^{x_{\rho_{ir}}} \quad (6.4)$$

Therefore, the objective of MESLMHP is to maximise the weighted sum over demand nodes for each of the patient groups,  $f(z)$  (with specific capacity  $z$ ), where

$$f(z) = \sum_{l=1}^k w_l \sum_{i=1}^m \lambda_i^l \sum_{j=1}^n P_{i,\rho_{ij}}^l = \sum_{l=1}^k w_l \sum_{i=1}^m \lambda_i^l \sum_{j=1}^n s_l(t_{i,\rho_{ij}}) \left(1 - \pi_{\rho_{ij}}^{x_{\rho_{ij}}}\right) \prod_{r=1}^{j-1} \pi_{\rho_{ir}}^{x_{\rho_{ir}}} \quad (6.5)$$

such that

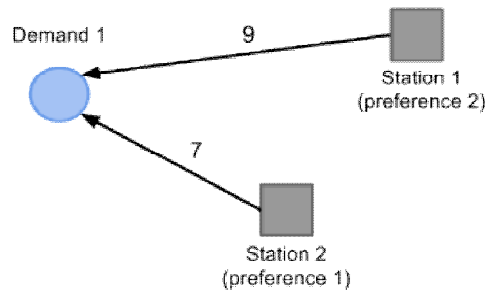
$$\sum_{j=1}^n x_j = Z \text{ and } x_j \in \mathbb{Z}_{\geq 0} \quad (6.6)$$

Constraint in Equation 6.6 ensures that the total number of vehicles,  $Z$ , are all allocated, with an integer number of vehicles allocated to a service node.

A short example is now provided to demonstrate the formulation of this modelling method for a survival based approach.

### Example 6.2

Assume a system exists, as depicted by Figure 6.7, where survival of the heterogeneous population given a response by a homogeneous fleet can be calculated using Equation 6.2.



**Figure 6.7** MESLMHP example network with one demand node and two service nodes

The travel time and preference matrices can therefore be constructed as  $t_{ij} = [9 \ 7]$  and  $\rho_{ij} = [2 \ 1]$  respectively. Assume  $x_1 = 1$ ,  $x_2 = 1$  and input utilisation for the stations to be  $\pi_j = \begin{cases} 0.5 & \text{for } j = 1 \\ 0.7 & \text{for } j = 2 \end{cases}$ , so that  $\pi_j < 1$ . Let the demand rate per hour for category A patients be  $\lambda^A = 0.9$  and for category B patients be  $\lambda^B = 0.8$ , and  $w_l = \begin{cases} 0.6 & \text{for } l = A \\ 0.4 & \text{for } l = B \end{cases}$ , then the objective function to maximise (of Equation 6.5) becomes:

$$\begin{aligned}
 f(z) &= w_A \left( \lambda_1^A \sum_{j=1}^2 s_A(t_{1,\rho_{1j}}) (1 - \pi_{\rho_{1j}}^{x_{\rho_{1j}}}) \prod_{r=1}^{j-1} \pi_{\rho_{1r}}^{x_{\rho_{1r}}} \right) \\
 &\quad + w_B \left( \lambda_1^B \sum_{j=1}^2 s_B(t_{1,\rho_{1j}}) (1 - \pi_{\rho_{1j}}^{x_{\rho_{1j}}}) \prod_{r=1}^{j-1} \pi_{\rho_{1r}}^{x_{\rho_{1r}}} \right) \\
 &= 0.6 \left( 0.9(s_A(t_{1,2})(1 - \pi_2^{x_2}) + s_A(t_{1,1})(1 - \pi_1^{x_1})\pi_2^{x_2}) \right) \\
 &\quad + 0.4 \left( 0.8(s_B(t_{1,2})(1 - \pi_2^{x_2}) + s_B(t_{1,1})(1 - \pi_1^{x_1})\pi_2^{x_2}) \right) \\
 &= 0.6(0.9(s_A(7)(1 - 0.7) + s_A(9)(1 - 0.5)0.7)) \\
 &\quad + 0.4(0.8(s_B(7)(1 - 0.7) + s_B(9)(1 - 0.5)0.7)) \\
 &= 0.6(0.9(0.23(0.3) + 0.18(0.35))) + 0.4(0.8(0.23(0.3) + 0.18(0.35))) \\
 &= 0.1135 \text{ expected proportion of survivors.}
 \end{aligned}$$

## 6.6 Modelling Heterogeneous Patients and a Heterogeneous Fleet

### 6.6.1 Model Brief: MESLMHPHF

The model described in the previous section (6.5), although considers different classes of patients, and advances the MSLP, still lacks in realism by ignoring multiple vehicle varieties. With the necessary addition to the model formulation of a fleet that is both heterogeneous in vehicle type and purpose, the resulting model may provide even more accuracy in allocations of ambulance resources in order to maximise survival of a heterogeneous population.

EMS systems often operate with a fleet containing more than one type of vehicle. Vehicles of the same variety form a sub-fleet, from which single or multiple vehicles can be dispatched to

emergencies for specific purposes. The two main vehicle types in the Wales, EAs and RRVs, are used in different combinations for the various categories of patients. An RRV is sent to the highest priority patients (since they are able to travel more quickly than EAs) with an EA as a follow-up vehicle (since RRVs do not have the capacity nor the equipment to be able to transport patients). For other patient categories, an RRV is often not dispatched at all, and instead, the EA becomes the primary responder. For this reason, it is important to model not only the patient categories, but also the difference in response operations dependent on the category.

The mathematical programming model proposed in this section considers heterogeneous patient groups as well as the two sub-fleets separately in order to optimise the allocation to best respond to all classes of patients in the correct way and with the correct vehicle and ambulance crew.

### 6.6.2 Notation: MESLMHPHF

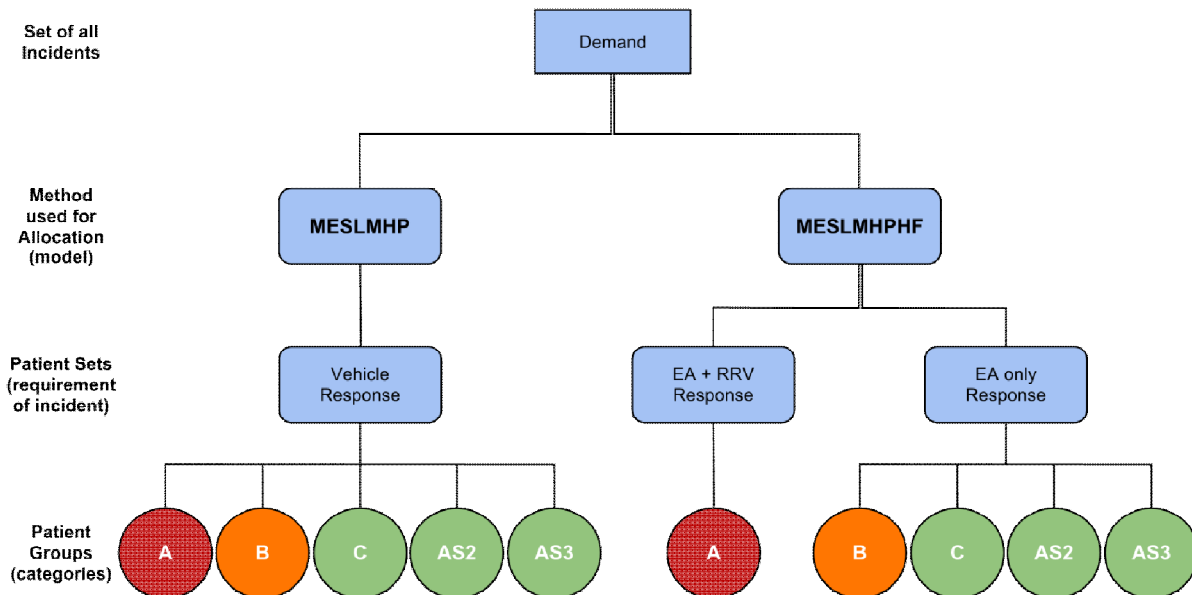
The MESLMHP formulation can be extended to incorporate different vehicle types as well as heterogeneous patient groups. The logic of this extension for two sub-fleets is now explained.

Consider two different sets of patient groups, defined as:

1. a set of patients to be responded to initially by an RRV and to be followed up by an EA service – referred to as set  $k_1$ ;
2. a set of groups that require only EA attendance – referred to as set  $k_2$ .

The first set of patients,  $k_1$ , contains the highest priority patient groups (categories) that require immediate attention. On receipt of such a call for service, an RRV will often be dispatched at the same time as an EA (if available) in the hope that the RRV will reach the patient first, begin treatment at the scene, and stabilise the patient ready for transportation (if necessary) by the following EA. If there is no RRV available, or if an EA at a closer station would be able to attend the incident faster than the best available RRV, then it is also possible for an EA to respond to these emergencies alone (i.e. for set  $k_1$ , RRV attendance is dependent on availability and the speed of response of closer vehicles).

The second set of patient categories,  $k_2$ , accounts for the lower acuity patients. They require fairly quick attendance, but are not in a life-threatening state and so an EA will be dispatched when the Trust has the available capacity to attend the scene or provide transport to hospital. In the UK, the only emergency group captured by  $k_1$  are the category A patients, as Figure 6.8 portrays.



**Figure 6.8** The structure and categorisation of emergency incidents according to the models

Before extending the MESLMHP objective function of Equation 6.5 to include appropriate terms that capture the characteristics of the two sub-fleets used in the UK, the following scenarios must first be considered. Scenarios 1 and 2 represent incidents requiring both EA & RRV attendance – patient set  $k_1$ ; scenario 3 concerns incidents that require only an EA on scene – patient set  $k_2$ . Let  $x_{j,u}$  be the number of vehicles at station  $j$  of type  $u$  and  $\pi_{j,u}$  the utilisation of these vehicles.

### Scenario 1

Assume that for an RRV from station  $j$  provides the initial response to an incident of set  $k_1$  - all vehicles at all more preferred (closer) stations must be busy. However, this ignores the fact that EAs at the preferred stations may travel slower than the current RRV, and so would not be chosen for dispatch over an RRV from station  $j$  to serve demand node  $i$ , even if available. Therefore, we must elaborate on this scenario, and state that for an RRV to respond from station  $j$ , all RRVs at all closer stations must be busy, and all (if any) *preferable* EAs at these stations must also be unavailable.

**Scenario 2**

Assume an EA at station  $j$  responds to an incident at demand node  $i$  that requires a response of type  $k_1$ , all EAs and RRVs at more preferred stations must be unavailable for this to be the case. In addition, all RRVs at less preferable stations that could travel faster to the scene than an EA from  $j$  must be busy.

**Scenario 3**

In Wales, for all patient groups other than Category A, the only vehicle type necessary on scene is that of an EA (or equivalent). Therefore, for these  $k_2$  emergencies, only the busy probabilities of EAs at all closer stations are required for calculation in the objective function, which corresponds to the formulation already described by Equation 6.5.

Considering the mathematical representation of the various response scenarios separately, we have four modelling constructs:

**1. RRV Responder (Scenario 1)**

For a  $k_1$  incident, the elements required for computation of an RRV response are the probability of an available RRV at station  $j$ , busy probabilities of all preferred RRVs, and busy probabilities of preferred EAs *if and only if* the EAs could have responded more quickly than the current RRV.

Let  $u$  be the vehicle type, so that  $u = \begin{cases} 0 & \text{if RRV} \\ 1 & \text{if EA} \end{cases}$  and  $R$  be a variable that indicates whether an EA at a more preferable station could reach the scene faster than an RRV at the considered station,

$$R = \begin{cases} 1 & \text{if } (t_{i,\rho_{ir},1} - t_{i,\rho_{ij},0}) \leq 0 \\ 0 & \text{otherwise} \end{cases}$$

where,  $t_{i,\rho_{ir},u}$  is the travel time of a vehicle of type  $u$  between demand node  $i$  and preferred station  $\rho_{i,r}$  ( $r$  is a greater preference than  $j$ ). If the travel times are equal, the best decision would be to dispatch the EA at the preferred station, so for an RRV at  $j$  to respond, the busy probability of the EA at  $r$  is still required in the calculation. The variable  $R$  takes a value of 1 if the EA at the more preferable station  $r$ , has a journey time to the scene shorter than (or equivalent to) the RRV at the current station  $j$ , and a value of 0 if the EA would be slower than the current RRV.



Multiplying these busy probabilities for all stations favoured over station  $j$ , the total probability of only more preferable vehicles being busy is:

$$\prod_{r=1}^{j-1} \pi_{\rho_{ir},0}^{x_{\rho_{ir},0}} \left( \pi_{\rho_{ir},1}^{x_{\rho_{ir},1}} \right)^R \quad (6.7)$$

( $R = 0$  creates an EA busy probability of 1 for the preferred station, implying that the EA at  $r$  would never serve demand of type  $l$  from node  $i$  over an RRV at station  $j$ ).

## 2. EA Responder (Scenario 2)

For a  $k_1$  incident, the elements required for computation of an EA response are the probability of an available EA at station  $j$ , busy probabilities of all favoured RRVs and EAs, and busy probabilities of less preferred RRVs *if and only if* the less preferable RRVs could have responded more quickly than the current EA.

Let  $Q$  be a variable that indicates whether an RRV at a less preferable station could reach the scene faster than an EA at the considered station,

$$Q = \begin{cases} 1 & \text{if } (t_{i,\rho_{iq},0} - t_{i,\rho_{ij},1}) < 0 \\ 0 & \text{otherwise} \end{cases}$$

where,  $t_{i,\rho_{iq},u}$  is the travel time of a vehicle of type  $u$  between demand node  $i$  and preferred station  $\rho_{iq}$  ( $q$  is a lower preference than  $j$ ). If the travel times are equal, the best decision would be to dispatch the EA at the current station, so the busy probability of an RRV at  $q$  is not required. The variable  $Q$  takes a value of 1 if the RRV at a less preferable station  $q$ , has a journey time to the scene shorter than the EA at the current station  $j$ . It takes a value of 0 if the EA would be the quicker vehicle to respond.

Multiplying these busy probabilities over all stations (other than current station  $j$ ), the total probability of only more preferable vehicles being busy can be written as:

$$\prod_{r=1}^{j-1} \pi_{\rho_{ir},0}^{x_{\rho_{ir},0}} \pi_{\rho_{ir},1}^{x_{\rho_{ir},1}} \prod_{q=j+1}^n \left( \pi_{\rho_{iq},0}^{x_{\rho_{iq},0}} \right)^Q \quad (6.8)$$

( $Q = 0$  creates an RRV busy probability of 1 for the less preferable station  $q$ , implying the RRV at  $q$  would never serve demand of type  $l$  over an EA at station  $j$ ).

### 3. RRV or EA Responder (Scenario 1 & 2 combined)

Merging the two mathematical formulations of constructs 1 and 2, we reach the situation, where the probability of either an RRV or an EA being the first responder to an emergency of type  $k_1$  is accounted for.

There are four extra possibilities for service of a  $k_1$  patient by vehicle  $u$  positioned at station  $\rho_{ij}$ .

1. Response by an RRV, whereby EAs at all preferred stations would in fact reach the scene quicker than the current RRV.
2. Response by an RRV, whereby EAs at only some (or none) of the preferred stations would reach the scene quicker than the current RRV.
3. Response by an EA, where all RRVs at less preferable stations would not be able to reach the scene faster than the current EA.
4. Response by an EA, where RRVs at some (or all) less preferable stations would be able to reach the scene faster than the current EA.

The total busy probability must consider the probabilities that all preferable vehicles are busy given a vehicle of type  $u$  responds to the incident.

$$\prod_{r=1}^{j-1} \pi_{\rho_{ir,0}}^{x_{\rho_{ir,0}}} \left( \pi_{\rho_{ir,1}}^{x_{\rho_{ir,1}}} \right)^{(R^{1-u})} \prod_{q=j+1}^n \left( \pi_{\rho_{iq,0}}^{x_{\rho_{iq,0}}} \right)^{Q \cdot u} \quad (6.9)$$

For an RRV response,  $u = 0$ , the second factor (EA utilisation at  $r$ ) in the first product term, will be to the power  $R$ , whereas the power  $Q \cdot u$  of the utilisation factor (of an EA at  $q$ ) in the second product term reduces to 0, giving the total product value of 1. Therefore, the final form of Equation 6.9 for an RRV response is equivalent to that of 6.7. For an EA response,  $u = 1$ , in the first product term in 6.9, the power of the second factor (EA utilisation at  $r$ ) is  $R^0 = 1$  and for the second product term, the power is simply  $Q$ , resulting in the formulation given by Equation 6.8 as required.

#### 4. EA Response Only (Scenario 3)

When the demand is for a incident type captured by set  $k_2$ , that is, only an EA is required to attend the scene, the total busy probability of all more preferable vehicles is the same as given in the MESLMHP formulation of Equation 6.5.

$$\prod_{r=1}^{j-1} \pi_{\rho_{ir}}^{x_{\rho_{ir}}}$$

Therefore, in these cases, where an EA will respond from station  $j$ , simply consider the utilisation of EAs at all stations  $r$ , where  $r$  is preferred to station  $j$  in dispatch to demand node  $i$ .

#### 6.6.3 Formulation: MESLMHPHF

The final formulation of the Maximal Expected Survival Location Model for Heterogeneous Patients with Heterogeneous Fleet (MESLMHPHF) is based on the structure of the MESLMHP given in Equation 6.5 and the discussions surrounding scenarios 1 to 3 in section 6.6.2. Additional notation is required for defining the emergency categories, following information of patient sets given in Figure 6.8.

If  $k_1$  is the set of patient groups that require both an RRV and EA response if possible and  $k_2$  is the set that require only an EA response, then let  $l$  be a patient group such that  $l \in k_1, k_2$ .

The survival probability of a patient of type  $l \in k_2$  from demand node  $i$ , serviced by a vehicle of type  $u$  found at station  $\rho_{ij}$ , can now be written as:

$$p_{i,\rho_{ij},u}^l = s_l(t_{i,\rho_{ij},u}) \left(1 - \pi_{\rho_{ij},u}^{x_{p_{ij},u}}\right) \left(\pi_{\rho_{ij},1-u}^{x_{p_{ij},1-u}}\right)^u \cdot \prod_{r=1}^{j-1} \left(\pi_{\rho_{ir},0}^{x_{\rho_{ir},0}} \left(\pi_{\rho_{ir},1}^{x_{\rho_{ir},1}}\right)^{(R^{1-u})}\right) \prod_{q=j+1}^n \left(\left(\pi_{\rho_{iq},0}^{x_{\rho_{iq},0}}\right)^{Q^u}\right) \quad (6.10)$$

The survival probability of a patient of type  $l \in k_1$ , from demand node  $i$ , serviced by an EA vehicle stationed at  $\rho_{ij}$ , can be written as:

$$p_{i,\rho_{ij},1}^l = s_l(t_{i,\rho_{ij},1}) \left(1 - \pi_{\rho_{ij},1}^{x_{p_{ij},1}}\right) \cdot \prod_{r=1}^{j-1} \pi_{\rho_{ir},1}^{x_{\rho_{ir},1}} \quad (6.11)$$

The new model aims to maximise the number of survivors from different patient groups given an allocation of a heterogeneous fleet through the addition of these two probabilities (Equations 6.10 and 6.11) summed over all stations and demand for the given patient groups and weighted accordingly.

The objective of MESLMHPHF is to maximise:

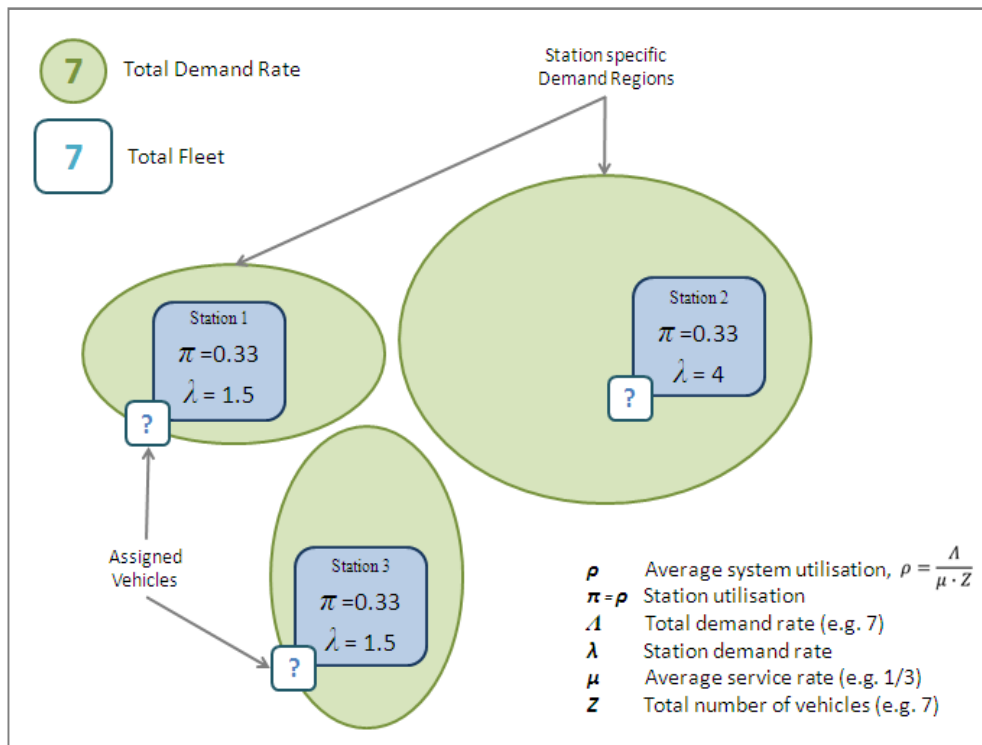
$$\begin{aligned}
 g(z) &= \sum_{l=1}^{|k_1|} w_l \sum_{i=1}^m \lambda_i^l \sum_{j=1}^n \sum_{u=0}^1 p_{i,\rho_{ij},u}^l + \sum_{l=1}^{|k_2|} w_l \sum_{i=1}^m \lambda_i^l \sum_{j=1}^n p_{i,\rho_{ij},1}^l \\
 &= \sum_{l=1}^{|k_1|} w_l \sum_{i=1}^m \lambda_i^l \sum_{j=1}^n \sum_{u=0}^1 S_l(t_{i,\rho_{ij},u}) (1 - \pi_{\rho_{ij},u}^{x_{p_{ij},u}}) (\pi_{\rho_{ij},1-u}^{x_{p_{ij},1-u}})^u \\
 &\quad \cdot \prod_{r=1}^{j-1} \left( \pi_{\rho_{ir},0}^{x_{\rho_{ir},0}} (\pi_{\rho_{ir},1}^{x_{\rho_{ir},1}})^{(R^{1-u}}) \right) \prod_{q=j+1}^n \left( (\pi_{\rho_{iq},0}^{x_{\rho_{iq},0}})^{Q \cdot u} \right) \\
 &\quad + \sum_{l=1}^{|k_2|} w_l \sum_{i=1}^m \lambda_i^l \sum_{j=1}^n S_l(t_{i,\rho_{ij},1}) \cdot (1 - \pi_{\rho_{ij},1}^{x_{p_{ij},1}}) \cdot \prod_{r=1}^{j-1} \pi_{\rho_{ir},1}^{x_{\rho_{ir},1}}
 \end{aligned} \tag{6.12}$$

## 6.7 Combating the Input Utilisation Problem: A vicious circle

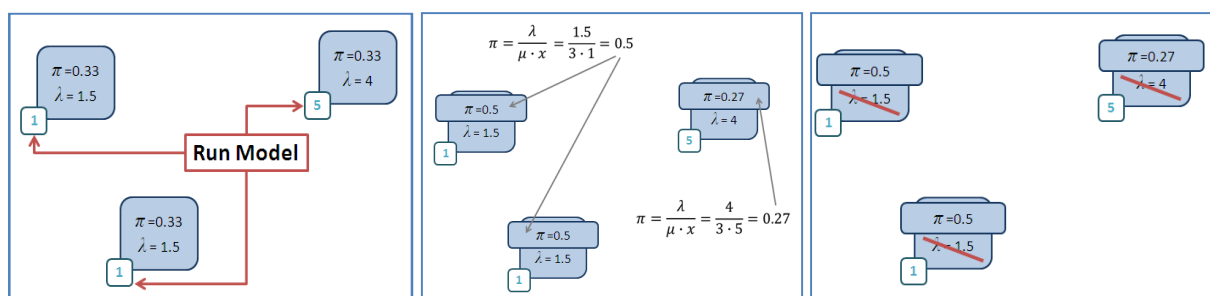
### 6.7.1 Model Brief: MESLMHP-I and MESLMHPHF-I

A structural restriction of both MESLMHP and MESLMHPHF, is that utilisation is estimated from the data and provided as input to the models (Knight et al. 2012a). The problem with this deterministic assumption of average system utilisation and constant busy probabilities is that after optimising the allocation of vehicles, in reality, the distribution of demand to stations will alter and so the utilisation at stations will in fact not be equal and will differ from the provided input utilisation  $\pi_j$ . Figures 6.9a - 6.9d demonstrate this process. Utilisation is station specific depending on the number of vehicles allocated and demand rate arriving at the station. The assumption that  $\pi_j$ , the mean utilisation of vehicles at station  $j$ , remains unchanged as allocations are optimised, limits conclusions drawn from MESLMHP and MESLMHPHF. Considering these new demand and utilisation values, it is likely that the allocations are now not optimal. A circular relationship is born.

To overcome the circular relationship between demand distribution and station utilisation, iterative versions of both the Heterogeneous Patients and Heterogeneous Patient and Fleet models – MESLMHP-I and MESLMHPHF-I are devised to take into account actual utilisation at each individual station, given the number of vehicles sited and demand to be served.



**Figure 6.9a** Example EMS system with overall average region utilisation input, demand per station region and total fleet size but with allocation unknown



**Figure 6.9b** Allocation results after optimisation model process

**Figure 6.9c** Updated utilisation given newly allocated vehicles per station

**Figure 6.9d** Incorrect demand based on new allocation and updated utilisation

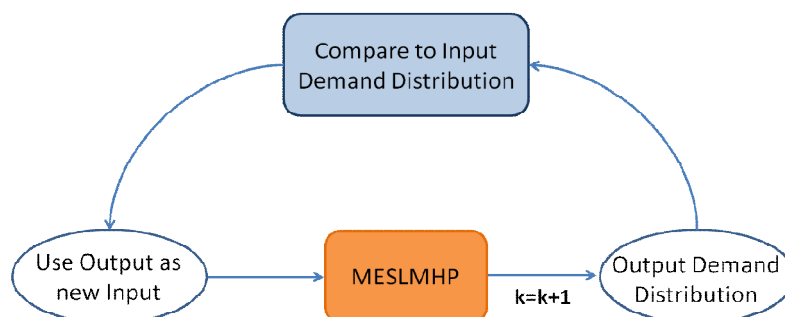
### 6.7.2 Notation: MESLMHP-I

Notation follows the convention laid out in section 6.5.3, for the MESLMHP, with the following amendments. For a summary on the parameters used in each model, and the inherent differences in the way in which utilisation is captured by each of the methods, refer to Table 6.1.

MESLMHP-I utilises the output from a MESLMHP optimisation and redistributes demand expected at each station based on the assumed optimal allocation of vehicles. The utilisation used as input to MESLMHP is now incorrect based on the new allocations and so is recalculated to use as input to the next iteration of MESLMHP. This cycle (demonstrated by Figure 6.10) continues until the chosen stopping method criteria is met:

- either the output demand distribution from one iteration is equal to the output demand distribution of the previous iteration  $\Lambda_j^{(k)} = \Lambda_j^{(k-1)}$  for all  $j \in [n]$ , where  $\Lambda_j$  is the mean demand rate at the  $j^{\text{th}}$  station – the allocation of vehicles is now suitable for the system's demand and utilisation;
- or the algorithm has run for a fixed set of iterations, with final results selected from the iteration with the smallest mean square error between the input and output demand distributions.

Insight from queueing theory into the processes at service nodes is used to model each ambulance station as an  $M_j/M_j/x_j$  queue, (random arrivals, random service rate and  $x$  servers) giving actual station utilisation of  $\pi_j = \frac{\Lambda_j}{\mu_j x_j} < 1$ .



**Figure 6.10** MESLMHP-I iteration steps to combat circular relationship between demand distribution and station utilisation in MESLMHP

After each iteration the input to the next iteration is based on the output of the previous. For a given iteration  $k$ , the input utilisation parameter is given by Equation 6.13.

$$\pi_j^{(k)} = \begin{cases} \min\left(1, \frac{\Lambda_j^{(k-1)}}{\mu_j x_j^{(k-1)}}\right); & x_j^{(k-1)} > 0 \\ \pi^* & ; x_j^{(k-1)} = 0 \end{cases} \quad (6.13)$$

where  $\Lambda_j^{(k)}$  is the approximated actual demand distribution resulting from the optimisation, depending upon the number of vehicle allocated to the base.

Difficulties lie in selecting such a method for distributing demand amongst the stations. It would be possible to use an argument similar to that employed in Equation 6.5, using the busy probabilities of vehicles and weighted demand (as demonstrated in Equation 6.14) for each station in turn.

$$\Lambda_j^{(k)} = \sum_{i,l} \lambda_i^l \left(1 - \pi_{\rho_{ij}}^{(k-1)x_{\rho_{ij}}^{(k-1)}}\right) \prod_{r=1}^{j-1} \pi_{\rho_{ir}}^{(k-1)x_{\rho_{ir}}^{(k-1)}} \quad (6.14)$$

An issue with this approach however, surrounds the total demand obtained after each iteration (Knight et al. 2012a). Due to the small probability that all vehicles and therefore all stations are busy, a proportion of the original demand is unaccounted for using this formula.

$$\begin{aligned} \sum_j \Lambda_j^{(k)} &= \sum_{i,l,j} \lambda_i^l \left(1 - \pi_{\rho_{ij}}^{(k-1)x_{\rho_{ij}}^{(k-1)}}\right) \prod_{r=1}^{j-1} \pi_{\rho_{ir}}^{(k-1)x_{\rho_{ir}}^{(k-1)}} \\ &= \sum_{i,l} \lambda_i^l \sum_j \left(1 - \pi_{\rho_{ij}}^{(k-1)x_{\rho_{ij}}^{(k-1)}}\right) \prod_{r=1}^{j-1} \pi_{\rho_{ir}}^{(k-1)x_{\rho_{ir}}^{(k-1)}} \\ &= \sum_{i,l} \lambda_i^l \left( \left(1 - \pi_{\rho_{i1}}^{(k-1)x_{\rho_{i1}}^{(k-1)}}\right) + \left(1 - \pi_{\rho_{i2}}^{(k-1)x_{\rho_{i2}}^{(k-1)}}\right) \cdot \left(\pi_{\rho_{i1}}^{(k-1)x_{\rho_{i1}}^{(k-1)}}\right) + \dots \right) \\ &= \sum_{i,l} \lambda_i^l \left(1 - \prod_{j=1}^n \pi_{\rho_{ij}}^{(k-1)x_{\rho_{ij}}^{(k-1)}}\right) < \sum_{i,l} \lambda_i^l \end{aligned}$$

As a solution to the lost demand, it is possible to include an additional term that, given the probability all vehicles are busy, splits leftover demand between stations based on the number of vehicles assigned to the base.

$$\Lambda_j^{(k)} = \sum_{i,l} \lambda_i^l \left( \left( 1 - \pi_{\rho_{ij}}^{(k-1)x_j^{(k-1)}} \right) \prod_{r=1}^{j-1} \pi_{\rho_{ir}}^{(k-1)x_j^{(k-1)}} + \frac{x_j^{(k-1)}}{Z} \prod_{j=1}^n \pi_j^{(k-1)x_j^{(k-1)}} \right) \quad (6.15)$$

Through experimentation, this technique was found to not converge due to the cyclic nature of the relationship between demand and utilisation. Appendix 6.1 shows substantial noise in the mean square error between the input and output utilisations at each iteration of an experimental scenario. These preliminary investigations were performed for a subset of the WAST data; with 18 demand nodes, 11 stations, a total number of 218 calls in the given period and a fleet capacity of 36 homogeneous vehicles. The model was run for 2000 iterations to be sure no signs of convergence existed in the allocation solution.

By setting such a fixed iteration stopping criteria, this approach may be implemented despite the convergence problem.

Alternatively, it is possible to instead approximate demand at a station more simply using the following queueing based formula:

$$\Lambda_j^{(k)} = x_j^{(k-1)} \left( \frac{\sum_{i,l} \lambda_i^l}{Z} \right)$$

If there is at least one vehicle, it is possible to calculate the next iteration input utilisation using the queueing theory formula. If there are no vehicles allocated to the station, calculating utilisation this way would not be possible, leading to a skewed view of the system where vehicles would never be placed at the base in subsequent iterations; for this case, utilisation must be calculated differently, using one of a number of methods for choosing  $\pi^*$ .

Various options were experimented with for selecting a suitable  $\pi^* < 1$ , including inferring a utilisation from the current utilisations at other busy stations, or setting to a specific estimated value. The chosen method for this case study was to calculate  $\pi^*$  as the mean utilisation of stations with at least one vehicle allocated, that is, the average utilisation of operational stations.



The MESLMHP-I algorithm can be summarised as follows:

1. Estimate  $\mu$  and  $\mu_j$  from data;
2. Assume a certain utilisation  $\pi_j^{(0)}$  for all  $j$ ;
3. Solve MESLMHP for these  $\pi_j$  and obtain allocation  $x_j$  for all  $j$ ;
4. Calculate the resulting demand distribution  $\Lambda_j^{(k)}$  for all  $j$ ;
5. Using  $M_j/M_j/x_j$ , calculate the resulting  $\pi_j$  consistent with the allocation of step 3;
6. Repeat 3, 4 and 5 until convergence criteria is met.

Initial conditions ( $k = 0$ ) for the MESLMHP-I model are as follows:

- Total number of vehicles  $Z$  distributed evenly across the stations;
- Overall mean system service rate  $\mu$  (calculated from data provided);
- Utilisation  $\pi_j^{(0)} = \sum_{i,l} \frac{\lambda_i^l}{\mu Z}$ .

### 6.7.3 Notation: MESLMHPHF-I

The iterative discussion of sections 6.7.1 and 6.7.2 can similarly be applied directly to the MESLMHPHF, adjusting only for the utilisation per vehicle type. The algorithm must now assume

$\pi_{j,u}^{(0)}$  from the data for  $u = \begin{cases} 0 & \text{if RRV} \\ 1 & \text{if EA} \end{cases}$  and calculate resulting  $\pi_{j,u}$  for each station  $j$  and vehicle type  $u$ .

Service rates will also refer to a specific vehicle type, giving two values for  $\mu_{j,u}$  at each station.

Since service of a high priority patient requires both an EA and RRV attendance, utilisation of EA vehicles will take into account demand for all call categories, but RRV utilisation will be calculated based only on category A demand rates.

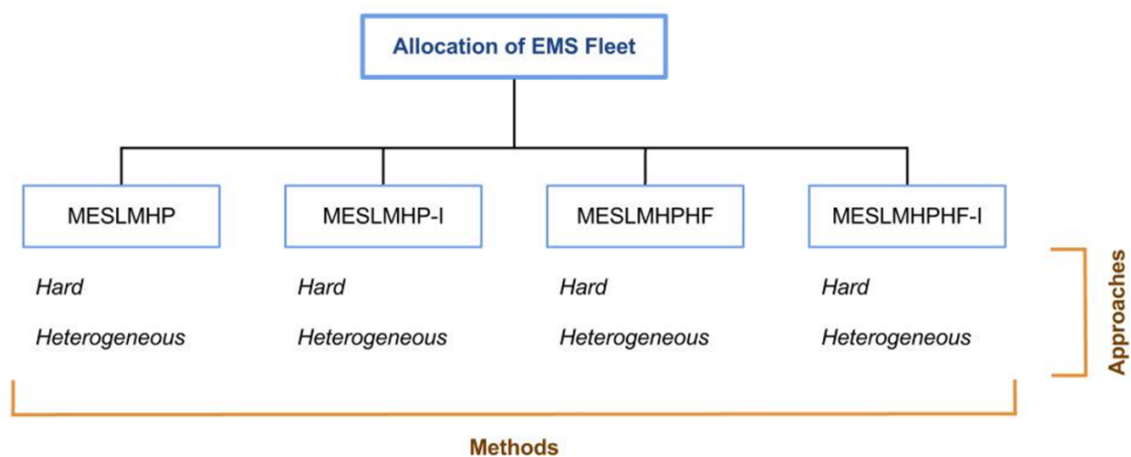
**Table 6.1** Summary of model input parameters for all four location models  
(1 – MESLMHP; 2 – MESLMHP-I; 3 – MESLMHPHF; 4 – MESLMHPHF-I)

Parameter	Description	1	2	3	4
$\lambda_i^l$	Demand at rate $\lambda$ from demand node $i \in m$ for emergency of type $l \in k$	•	•	•	•
$\Lambda_j$	Actual demand distributed to station $j \in n$		•		•
$\mu_j$	Service rate $\mu$ at station $j$	•	•		
$\mu_{j,u}$	Service rate at station $j$ for vehicle $u \in [0,1]$			•	•
$\mu$	Overall mean system service rate		•		•
$\rho_{ij}$	Station preference $j$ of demand node $i$	•	•		
$\rho_{ij,u}$	Station preference $j$ of demand node $i$ for service by a vehicle of type $u$			•	•
$\pi_j$	The utilisation of station $j$	•	•		
$\pi_{j,u}$	The utilisation of vehicle type $u$ at station $j$			•	•
$\pi^*$	Average utilisation of operational stations		•		•
$x_j$	The number of vehicles at station $j$	•	•		
$x_{j,u}$	The number of vehicles of type $u$ at station $j$			•	•
$s_l$	The survival function given for emergency of type $l$	•	•	•	•
$t_{ij}$	Predicted travel time between station $j$ and demand node $i$	•	•		
$t_{ij,u}$	Predicted travel time between station $j$ and demand node $i$ by vehicle of type $u$			•	•
$w_l$	Weight applied to category $l$ in order to prioritise emergency types	•	•	•	•
$Z$	Total number of vehicles in fleet	•	•	•	•

## 6.8 Application to WAST

### 6.8.1 Introduction

Both models and their iterative versions are now applied to the South East Wales EMS system. Two approaches (Figure 6.11) are investigated based on current and prospective system performance measures. The 'Hard' approach lends itself to the setup of the location models using the current WAST structure – step functions represent survival for all category hard target responses. The 'Heterogeneous' approach refers to one where a mixture of hard targets for lower priority patients and a survival curve (taken from the literature) for critical patients are implemented for a system based on clinical outcome.



**Figure 6.11** All methods and approaches applied to the South East Wales vehicle allocation problem

### 6.8.2 Granularity

Before experimentation between coverage and survival can begin, the structure of the network, and the degree of detail taken for demand zones must first be defined. Calls for service originating within close proximity of each other may be aggregated to form a single demand zone (as described in Chapter 5). In this case, a demand zone or node corresponds to a single postcode district since location data detail is only available at this level. For the South East region of Wales alone, accumulating demand to postcode districts produces 50 nodes for computation. In addition to demand, there are 23 regularly used vehicle bases. Overall run time will be greatly affected for

even a small increase in the number of nodes, especially since travel times and distances between all location pairs must be pre-computed (per individual regional scenario) using the Travel Matrix Generator Tool.

### 6.8.3 Genetic Algorithm

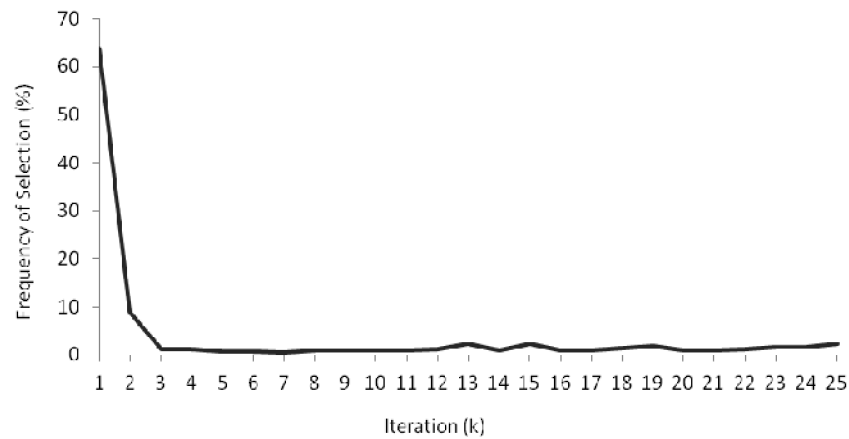
Finding an optimal solution to an integer location problem with even just 10 stations and 20 homogeneous vehicles through complete enumeration requires evaluation of over  $10^{20}$  combinations of allocations (by the uncapacitated facility location problem (Krarup and Pruzan 1990)). This is unrealistically achievable due to the computational complexity. There are two alternative solution approaches: multi-objective integer programming or Heuristic techniques. Due to the issue of scalability of patient groups and dimensionality, and the desire for a generic formulation, a heuristic method is exploited.

One particular heuristic technique that lends itself well to such problems is that of a genetic algorithm, GA, (Deb 2005, Sasaki et al. 2010). A GA is a population based heuristic that selects characteristics of parent solutions to pass on to new generations of solutions and has previously been used directly to maximise survival in EMS allocation (Erkut et al. 2008a). A benefit of solving the developed set of location models here using a GA is that the method can be applied to other, much larger EMS systems, or extended to the whole of WAST, not just the South East regional allocation strategy.

The spreadsheet models, built in Microsoft Excel, invoke the GA of the Palisade software-suite add-in, Evolver, with a population size of 50. Selection of parent solutions from the population on which to perform modifications is rank-based; as is the method of replacement of new solutions in obtaining the next generation. Uniform crossover rate (the chance that an element of a parent solution combines with another parent to generate a new solution) and adaptive mutation rate (the chance of a random swap of new solution values to diversify the gene pool) are taken to be 0.5 and 0.1 respectively. These choices are based on suggested 'rules of thumb' from various sources (Corporation 2010, Petrovic 2010, Sastry et al. 2005); they are not discussed further since WAST would not be expected to alter nor understand the operation or purpose of such parameters if application to the Trust was successful. Stopping criteria used for the optimisation of a particular scenario are 100 trials without improvement or a maximum change of 0.01% in the overall solution.

### 6.8.4 Iterations

To determine the necessary number of iterations for the iterative models, firstly the MESLMHP-I was run, distributing demand according to Equation 6.15 and stopping the model after 25 iterations. The best resulting allocation is selected as the one which gives the smallest mean square error between the input and output demand distributions. Running this version of the model for various shifts, and recording after each complete run which iteration produced the best allocation, seen in Figure 6.12, ten iterations were deemed adequate for subsequent experimentation – especially since the iterative and heterogeneous fleet models require considerably more computation time, a minimum number of iterations is desirable to balance the trade-off between optimality and run time.



**Figure 6.12** Frequency of selection of iteration as best solution to allocation problem using MESLMHP-I – 25 iteration stopping condition

### 6.8.5 Data Input

A single week's worth of data was selected from the 2009 South East Wales data set to enable more definitive representation of the system and more explicit input. Average numbers of vehicles on shift over the year and average demand are subject to large amounts of variation. By limiting the range of data, uncertainty in matching allocations with demand would be minimised and the location theory technique demonstrated for a more precise and realistic problem.

The week chosen for demonstration purposes is that of Sunday 10<sup>th</sup> May to Saturday 16<sup>th</sup> May 2009. For test and benchmark modelling, it is desirable to avoid any special calendar dates or regional

events to be sure typical demand would be adequately handled by a standard fleet. Furthermore, higher demand is witnessed during the summer school holidays and in winter due to adverse weather and so a moderate seasonal effect is obtained by selecting a spring month.

The developed models are not designed to be stochastic or dynamic; multiple versions must be run independently to capture variations in allocation patterns over time. For the chosen week, each weekday and weekend day are modelled separately, per shift. The pattern of demand is also significantly different between weekdays and weekends and over a 24-hour period. Six distinct operational shifts (shown in Table 6.2) thought to make up a typical week are extracted from a graphical representation of historical busy vehicles over time (Figure 6.13) and average arrival rate patterns (Chapter 4, Figure 4.9).

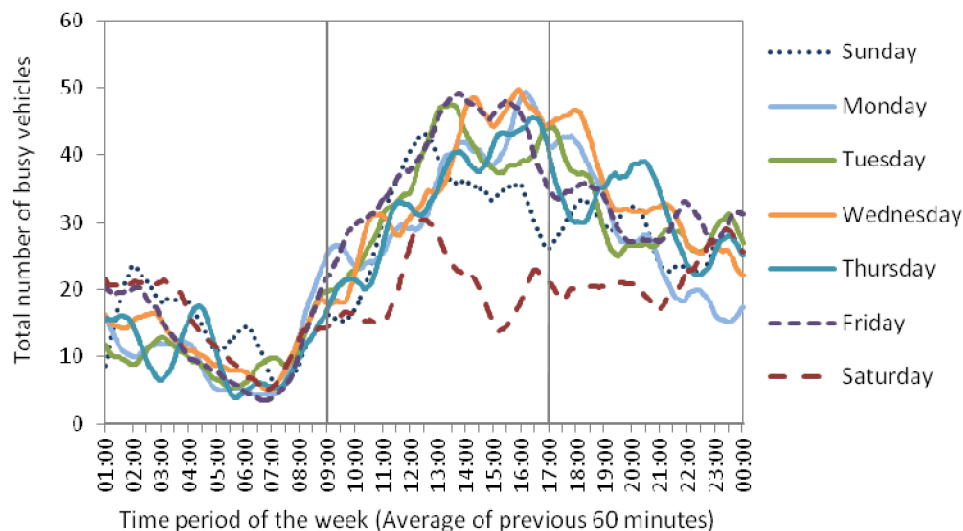
**Table 6.2** Daily data for the chosen week used in determining input values for each weekday and weekend shift for modelling

Shift	Hours of Shift		
	1am – 9am	9am – 5pm	5pm – 1am
Weekday 1	Monday-Friday		
Weekday 2	Monday-Thursday		
Weekday 3	Sunday-Thursday		
Weekend 1	Saturday-Sunday		
Weekend 2	Friday-Sunday		
Weekend 3	Friday-Saturday		

- **Arrival rates** are derived from the average number of calls arriving from a particular demand node, for each of the four categories of emergency, during each shift.
- **Service rates** per shift are procured from average cycle length found from the chosen week data. Cycle length in this case is taken to be the interval between time vehicle is allocated to the emergency call until time the vehicle becomes clear and reports itself available to attend further calls. Average cycle length from the data and cycle length for EAs and RRVs separately are converted into number of calls served per shift by each station, from which service rates  $\mu_j$  and  $\mu_{j,u}$  are obtained.

- **Fleet capacity** is the variable of the modelling process with potential for further exploration. Altering the fleet size will result in varying allocations and survival rates; for WAST this is a strategic and daily operational decision that could be expedited using such location modelling tools. Since capacity is time dependent, if WAST could access information which instructs on vehicle positioning given a total number of operational vehicles at any one point, fleet management would be a simple look-up task. All models are therefore run for various values of  $Z$  (fleet size) and results stored in compliance tables for future reference and comparison.

To give an indication of a typical fleet size as a guide for the range of  $Z$  for modelling, (and as further support for the suggestion in Chapter 4, section 4.6, that allocations provided by WAST are overestimated) the moving average (of 60 minutes) of busy vehicles at each point in the chosen week is portrayed in Figure 6.13, for all vehicles types combined. A vehicle is defined as busy from allocation to time clear.



**Figure 6.13** Number of busy vehicles (all types) over time for chosen week

### 6.8.6 Service Procedures

Since the location models purely delve into location based on demand, fleet capacity and overall average service rates (per vehicle type), no assumptions are made as to whether the patients require transportation or not. These are deterministic models, accounting for total demand rates and not individual requests for service as a simulation model might, so transportation decision is captured only by average cycle time.

For the heterogeneous fleet models (MESLMHPHF and MESLMHPHF-I), category A patients will require both an EA and RRV attendance. Categories B, C and Urgent (where Urgent is the combined AS2 and AS3) emergency patients will only require service by an EA (illustrated earlier by Figure 6.8). The calculation of utilisation is therefore dependent upon the vehicle type and suited demand.

### 6.8.7 Priority Weighting

Efforts were made to obtain importance values for the prioritisation of patient groups from WAST for weighting the categories during modelling. Figure 2.5 provides an understanding of the current priority of service structure, but WAST were unable to unequivocally state numeric values to represent differences in priorities for obvious ethical reasons. Literature on Quality of Life Years (QALYs) was investigated to see whether any numeric understanding from life years gained from a given level of response and survival probability could be applied to the developed survival coverage models. Again, no simple numeric answer exists at the time of writing. For these reasons, a selection of weights were tried and tested in a sensitivity analysis style experiment to see the effects different weight proportions had on the model outcomes.

Priority weightings are observed to be fairly stable over sensitivity analysis, such that the final decision on weights is a best guess, proportion-led outcome. Categories A, B, C and Urgent are given weights 0.6, 0.2, 0.15 and 0.05 respectively in all subsequent modelling. The larger proportion for category A reflects the life-threatening state in which patients arrive to the service. Category B and C have only a small discrepancy in weights to reflect their current target status in Wales (response time targets are equal) whilst maintaining a preference on order of service.

## 6.9 Results

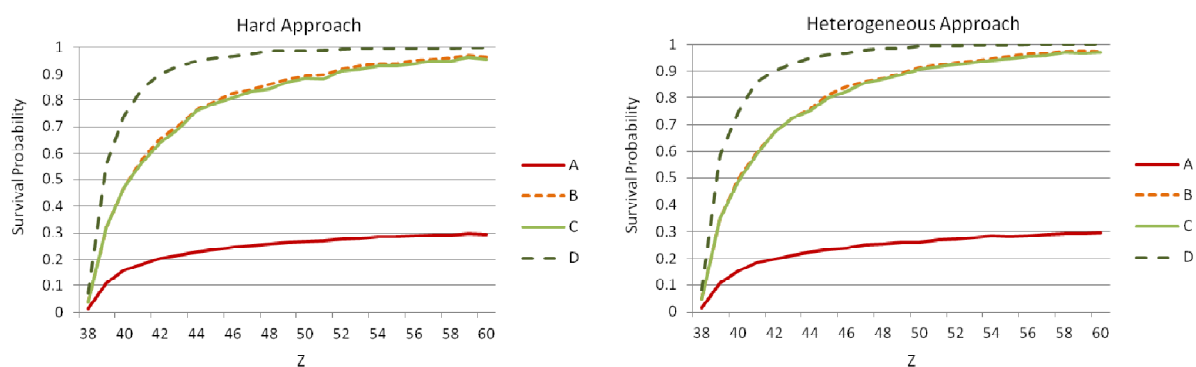
Results obtained demonstrate similar patterns across the weekday shifts and typical outcomes; therefore, only a selection are presented for brevity. The chosen results for portrayal are those of weekday shift 2 (daytime) since demand is fairly constant and consistently high during this period.

Firstly, comparisons of results across the modelling approaches for each method are illustrated.

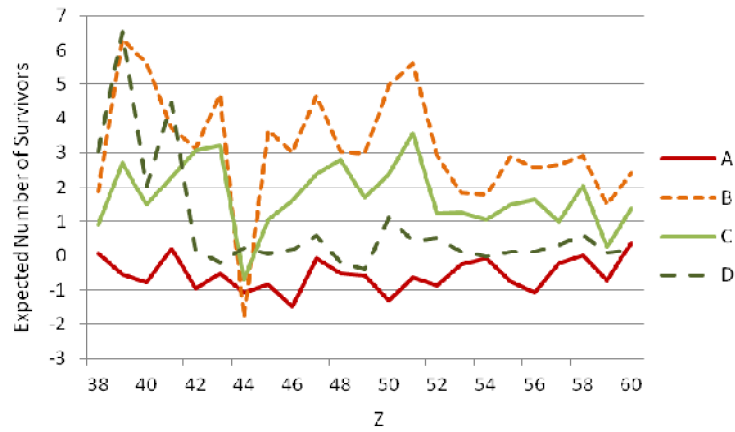


**MESLMHP**

Optimising the allocation of a homogeneous fleet to maximise survival of four patient groups via the MESLMHP produces results in accordance with Figures 6.14 and 6.15. When initial results for expected numbers of survivors of the original demand population are displayed graphically, as in 6.14, it appears there is very little difference between the overall effects of the Hard and Heterogeneous approaches in terms of total survival probability; however, on closer inspection, comparing the actual difference in total expected number of survivors, where a positive difference refers to the Heterogeneous approach maximising more survivors, it does generally perform better for category B, C and Urgent patients, but more category A patient lives are lost than with the Hard approach. This effect is due to the nature of the modelling objective. Hard targets result in vehicles being positioned in order to maximise the population that are attended within the time standard. The difference between a one minute and an eight minute response is not noted since they are both deemed successful, despite the clinical outcome of the patient likely being significantly different. When this allocation is converted into survival, using travel time as input to the survival function, the Hard target approach may have coincidentally allocated vehicles close to the high priority demand, since their hard response time target is more important than the lower priority patients. The Heterogeneous approach, although it accounts for priority through weights, aims to distribute vehicles more equitably, which may in fact lead to a lower survival probability than previously.



**Figure 6.14** Expected proportion of survivors for hard target and heterogeneous approaches from the MESLMHP method, weekday shift 2

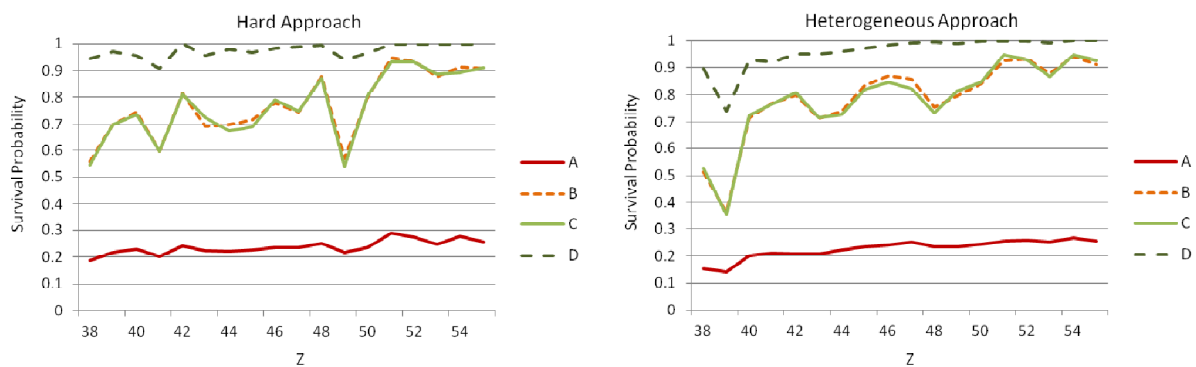


**Figure 6.15** Difference between expected number of survivors of the Heterogeneous and Hard target approaches from the MESLMHP method, weekday shift 2

### MESLMHP-I

The outcome of the Heterogeneous approach is slightly more favourable for the MESLMHP-I. The iterative method of readjusting the input utilisations based on actual demand expected at the station after an optimal allocation has been determined means the variation in number of survivors is higher. From Figure 6.16, it seems that the Heterogeneous approach gives a more stable result than the Hard approach for larger fleet sizes.

The instability in both graphs for smaller values of  $Z$  relates to the queueing theory problem embedded in the location analysis modelling structure. For small fleet sizes, the system utilisation value  $\pi$  will be close to 1 ( $\pi = \frac{\lambda}{\mu \cdot Z}$ ) meaning the length of time a patient spends waiting is likely higher than in situations with larger fleets.



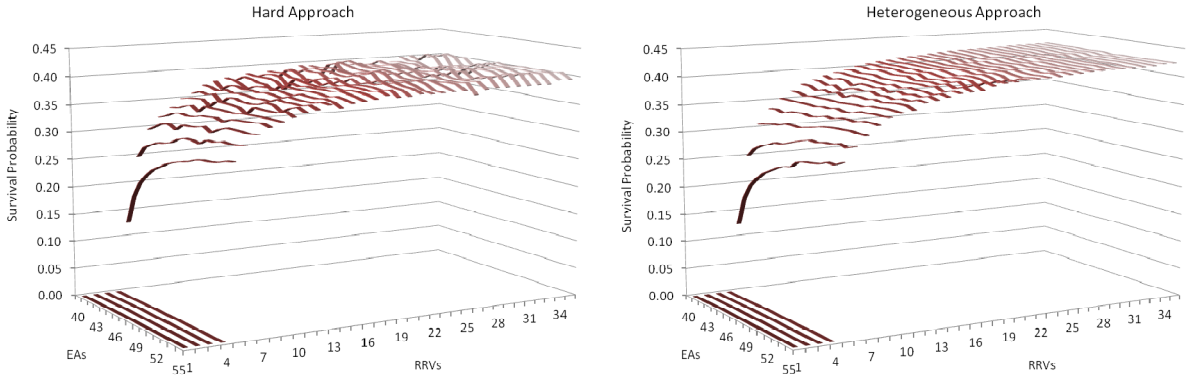
**Figure 6.16** Expected proportion of survivors for hard target and heterogeneous approaches from the MESLMHP-I method, weekday shift 2



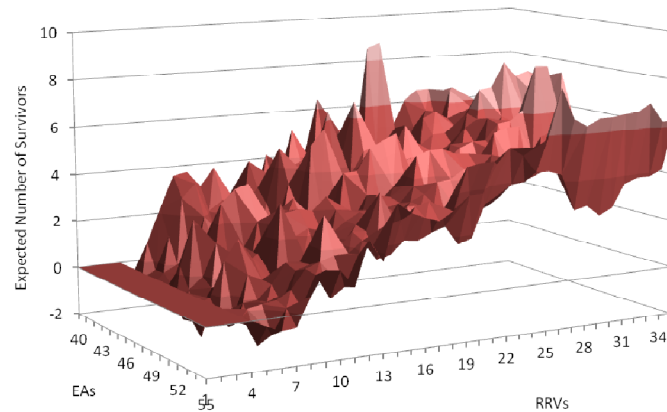
**Figure 6.17** Difference between expected number of survivors of the heterogeneous and hard target approaches from the MESLMHP-I method, weekday shift 2

**MESLMHPHF**

Looking initially only at category A patients, for a range of vehicles per type, again the comparison of Hard and Heterogeneous results in Figure 6.18 suggests little difference. Once the difference is calculated explicitly in terms of expected number of survivors however (Figure 6.19), the substantial improvement of the Heterogeneous approach on the outcome of patients is obvious.

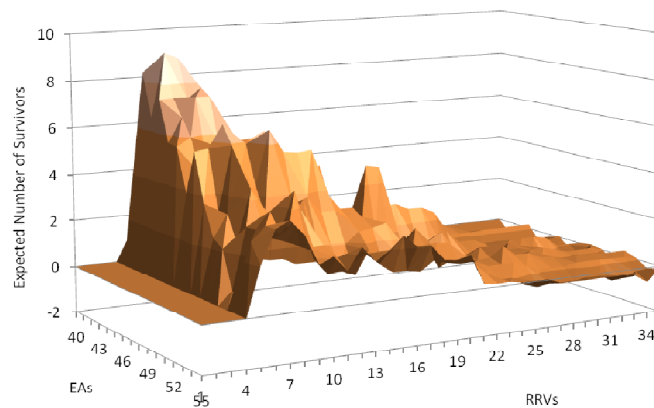


**Figure 6.18** Expected proportion of category A survivors only for hard target and heterogeneous approaches from the MESLMHPHF method, weekday shift 2



**Figure 6.19** Difference between expected number of category A survivors only of the heterogeneous and hard target approaches from the MESLMHPHF method, weekday shift 2

For category B, the difference in outcome between the two approaches is still substantially better in the Heterogeneous case (Figure 6.20); however, as the number of RRVs increase, with the ability only to serve category A patients, the difference in survivors decreases since the EAs now are more able to focus on lower priority calls than contributing to high priority patient response.

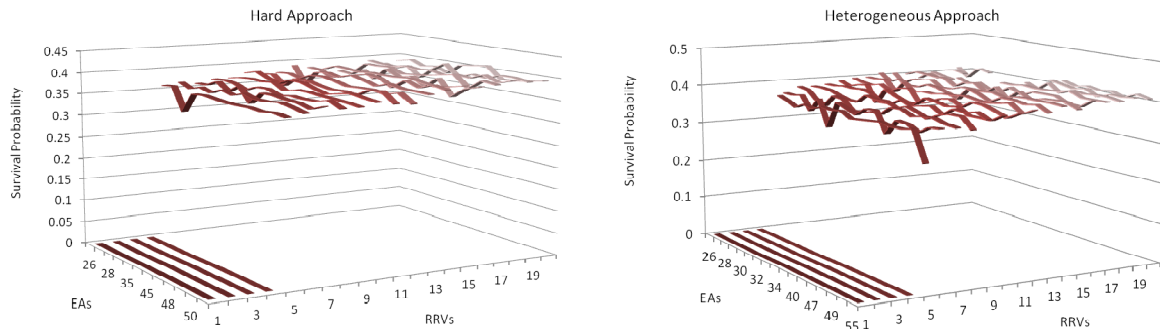


**Figure 6.20** Difference between expected number of category B survivors only of the heterogeneous and hard target approaches from the MESLMHPHF method, weekday shift 2

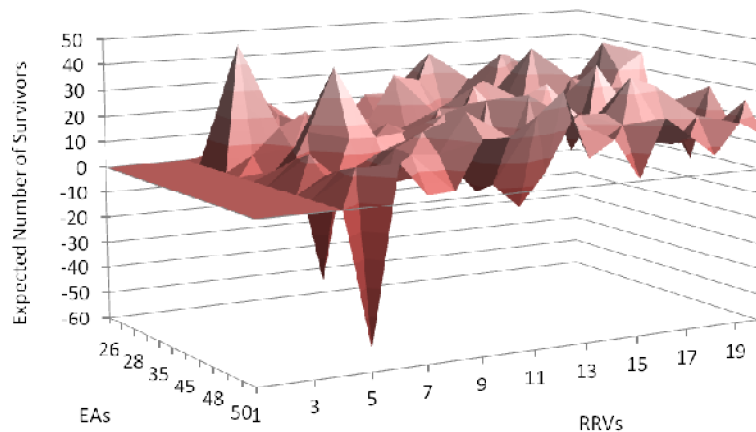
### ***MESLMHPHF-I***

Due to the exorbitant run length (up to one hour per iteration) of the iterative heterogeneous fleet model, only a select few combinations of sub-fleets were tested. The graphical results produced do

therefore not depict a full surface, making appreciation from the three-dimensional comparison curve (Figure 6.22) more difficult. A figure (removing gaps for fleet combinations that were not run) is shown, yet, scalability of this graph should be assumed with caution, the axes are not continuous or constantly distributed.



**Figure 6.21** Expected proportion of category A survivors only for hard target and heterogeneous approaches from the MESLMHPHF-I method, whole week

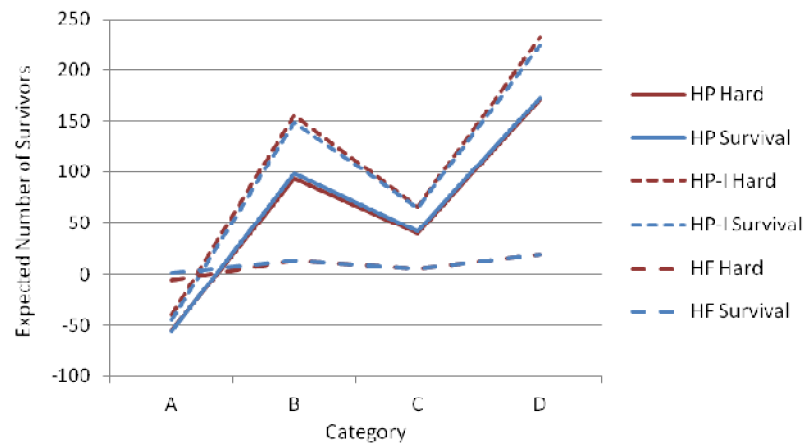


**Figure 6.22** Difference between expected number of category A survivors only of the heterogeneous and hard target approaches from the MESLMHPHF-I method, whole week

**Comparison with Actual Allocation**

WAST provided some information regarding the actual numbers of vehicles they have assigned to each station per shift across the South East region. This assignment does not guarantee the same number of vehicles are operational at any one time, but it gives a good indication of the total fleet

size available. Using this allocation and obtaining expected numbers of survivors based on their locations for the chosen weeks demand pattern, comparison with the four methods is now made.



**Figure 6.23** Difference between expected number of survivors of models compared with actual WAST allocation (weekday shift 2), for Hard and Survival function approaches per method

The difference in the expected number of survivors is depicted in Figure 6.23, where a positive difference implies a higher number of survivors. Across all categories, it is possible to see that the heterogeneous fleet model matches the current allocation performance the closest.

## 6.10 Conclusion

### 6.10.1 Introduction

From earlier results for the MESLMHP and its iterative version, it seemed that perhaps the Heterogeneous, survival maximising, approach would be less successful than originally hoped. Yet, the full complexity of the EMS system is not captured by these models and so outcomes may be misleading. Due to discrepancies in roles played by the two main operational vehicle types, the MESLMHPHF and MESLMHPHF-I offer a more accurate representation of the system and so more reliable results. Through this novel contribution, the Heterogeneous approach does prove to give better allocations in maximising positive outcome of the population. The importance of operating based on strategies that focus on the patient and not the system in fact proves to benefit both.

### 6.10.2 Model Limitations

The implemented survival curve in the developed set of location models may not be suitable for its purpose, but such curves are scarce. A single survival function, regressed from cardiac arrest data, is applied to the whole category A population. The curve utilised is also designed for response by a one-tiered system, not one that operates with BLS and ALS such as WAST. Dividing up categories in a manner accounting for differences in patient needs and condition would require the testing of various survival functions, ones suited to the system structure and possibly with better results.

Of the input data, some assumptions are made for simplicity and due to a lack of detail in the original data used for design. The travel time estimates do include residual variation, which is thought to generally capture occasions in the data where vehicles may be en-route or returning to their base when dispatched to a call and variation in speed, road conditions and human differences. It is also assumed that station preference is strictly ordered based on travel time; whereas in reality, this may not be strict – dispatch operators may use knowledge and experience to maintain a balance in coverage during dispatch and so do not always send the nearest available vehicle.

One problem with the input to the models is that the service times are taken to be the overall average cycle time of a shift. This time interval incorporates travel time from the scene to the hospital, but based on the data not the Google Maps API journey results. Since the transportation journey data is used in the regression of the Google Maps distance data to convert to travel time, it is deemed that this will not cause much error in cycle time values, since the travel time estimations are matched as best as possible to the existing historical data.

### 6.10.3 Extensions

An immediate extension to the models presented in this chapter is to an EMS system with more than two types of vehicle. WAST are also attempting to reduce the number of double-dispatches; therefore, in the MESLMHPHF models, consideration should be extended to service solely by RRVs, without EA back-up. Further work could consider the stochastic elements of such an emergency service, and capture the full extent of congestion, utilisation and time dependency through probabilistic modelling.

Capacity constraints on some, or all, stations may exist in reality, yet the allocation models assume that the entire fleet could be placed at a single vehicle base if this were the optimal allocation. This is a relatively simple constraint to implement in the current models, but has been ignored since it is assumed (and noted from preliminary tests) that subject to a fleet suitably large enough to handle the demand, optimal allocations will almost certainly be spread fairly equitably to capture demand across the network. Such capacity constraints are unlikely to be broken, and if they were, are unlikely to impact drastically on population survival, hopeful that the over-subscribed vehicles at one base could alternatively be located at the next nearest station to assist with local demand.

#### 6.10.4 Survival Approach

The development of Automated External Defibrillator programs and Community First Responders allows response time intervals to be drastically shortened; however, these programs rely heavily on the awareness of the public and the assumption of willing bystander intervention. Although this is more a medical issue than an OR modelling one, the fact that in many cases the EMS crews will be the first attendees implies work still needs to be undertaken to ensure patients receive immediate attention where needed and are not subjected to unnecessary long response delays.

Despite this, response time intervals (taken as onset to arrival of paramedics at the scene), may not actually be fully representative in terms of survival. There is a need for more information, such as:

- Was the cardiac arrest witnessed?
- Did a bystander intervene?
- How long until CPR was initiated?
- How long until defibrillation?
- Further details of BLS.
- Attendance by BLS (emergency medical technicians) or ALS (paramedic crews) or both?
- Details of other aspects of the Chain of Survival.

Discussion has shown the positive impact on patient outcome that early intervention can have for a patient experiencing an OHCA, following the Chain of Survival; however, there is still some uncertainty as to whether a BLS or ALS response would improve the rate of survival to discharge



(Chien et al. 2011). Stiell et al. (2004) found no significant difference except in admission rate to hospital; however, in their letter to the editor, Chien et al. argue that immediate outcomes are improved by BLS for some aspects of the condition but overall survival is not necessarily better for either group. They state, that rapid transportation to hospital facilities also contributes to a more positive outcome. This issue of transportation time is not captured by the location models used in this chapter and so another technique should be used as an alternative to investigate this further.

Concentration on survival as an objective should be thought not as a short term clinical outcome aim, but as an aim for long term effect improvements in an entire secondary care population through a simple shift in focus of strategy. Changes from hard targets to clinical performance measures have the benefit of accommodating the importance of the patient outcome, placing an emphasis on service of patients rather than performance of the system. Performance is still captured by the idea of having targets at all, but with survival curves, the time taken to attend the scene is a better approximation of success of the service in their underlying goal – saving lives.

## Chapter 7

# Simulating an EMS System

## 7.1 Why Simulate?

### 7.1.1 Definition

'Simulate' comes from the Latin *simulare* - to copy or represent - and, by definition (Penguin Reference 2003), means:

*"to assume the outward qualities or appearance of something".*

Before Operational Research and Management Science disciplines popularised the term 'simulation' as a type of modelling methodology, and even before technology was truly capable, the concept and structure of simulation had been in use under various guises for many years. Origins exist in military applications (Hill and McIntyre 2001) and flight simulators (Page 2000), healthcare, for medical skill training (Healthcare Simulation South Carolina 2013, Rosen 2008) and education (CreativeTeaching 2011, Zuckerman and Horn 1970). Today, simulation can be found in entertainment; conceptualised by films such as 'The Matrix' (1999), demonstrated by war games (Lenoir and Lowood 2005, Smith 2010) and computer games (Atkins 2003, Rennard 2007), and implemented in amusement park experiences (Clave 2007). Even exercise comes in simulated form, with the aid of game consoles (such as the Nintendo Wii and Microsoft's Kinect device), where 'players' *simulate* the actions of particular sports and activities without the need for sports equipment.

More commonly, in OR circles, simulating defines the act of mathematically recreating a real-world situation, system or process, often using a computer to perform experiments on the system in a controlled, safe and virtual environment. It offers an accessible environment for various, interconnected components of a system to be represented and explored with regard to a (multi-) set of objectives.

### **7.1.2 Benefits**

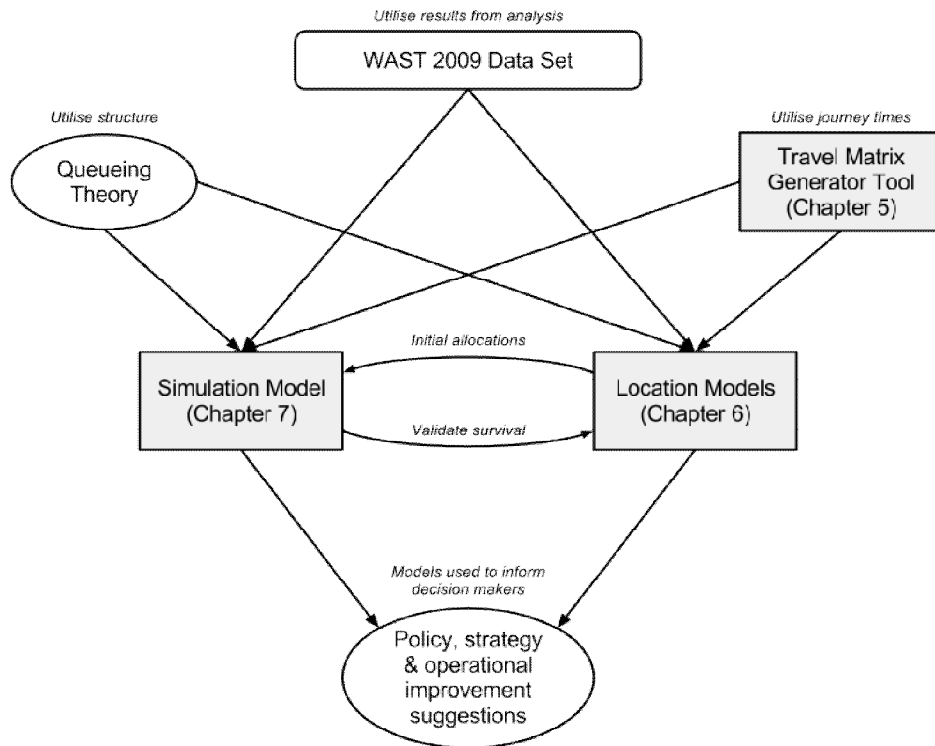
Demonstration of the power of modelling and experimentation has become ubiquitous in OR over the past few decades. (The Chapter 3 example of ambulance response time modelling prevalence in the literature, Figure 3.1, shows such an increase.)

Many organisations prefer the use of simulation experimentation to more technical methodologies, since, for the non-mathematical mindset, the understanding of simulation results can be much more intuitive than other modelling approach solutions and provides abundant interactive learning opportunities (Summers 2004). If graphics and visualisations are integrated and are to a good standard, and if the developer of the tool has the capability to communicate well the purpose, capacity and results of the model, then simulation can be a valuable asset to problem-management processes.

An advocate of this modelling aspect is Goldberg et al. (1990) who chose to use a simulation approach (with a multi-server queueing structure) in an effort to gain confidence from a client – namely the Tuscon Fire department of Arizona. The client felt more comfortable accepting output from a simulation model over a more analytical model.

### **7.1.3 Overview**

From the data analysis conclusions (Chapter 4), along with output from the Google Maps Travel Matrix Generator Tool (Chapter 5) and allocations obtained from developed location models (Chapter 6), a simulation tool is developed to allow in-depth investigation of the entire emergency ambulance service system. The interaction between the components developed in this thesis, along with the contribution of data and theory, is portrayed in Figure 7.1. The simulation tool's purpose, design and structure are detailed in the forthcoming sections, concluding with an account of hopeful future implementation in Wales.



**Figure 7.1** Interaction of the developed models and tools in potentially informing WAST's policy and operational decisions

## 7.2 Strength of Simulation

### 7.2.1 Introduction

According to Robinson (1994), the concept of simulating is built around three activities:

1. **Modelling:** the abstract representation of important features of a system, ranging from conceptual modelling to physical recreation of the current operations;
2. **Experimenting:** skill and knowledge of the system are used to gain understanding and further explore the system and its capabilities;
3. **Computing:** often used as the means to create a model and carry out the two former aspects of a simulation project.

In this study, each of these activities are considered and the description of the development process of a simulation tool is supplied in the following sections.

From spreadsheet models to purpose-built simulation packages, modern day computer simulation allows instant interaction of a proposed system model, looking at the effect of inputs (policies) on outputs (responses), where redesign and modifications are part of the interminable evolution process of simulation modelling. Manufacturing, transport, healthcare, defence, education, social networking – almost all sectors at some point or another have pioneered the use of simulation techniques, often with the expectation of reducing costs, maximising profits or improving efficiency.

### 7.2.2 Advantages

Simulation can be a powerful tool in the understanding of organisational, strategic, operational and tactical problems. Whether issues surround resource management, processes or pathways, simulation not only gives insights to situations that may be difficult to model mathematically, but allows the user to identify flaws and disparity in system knowledge and provides a canvas to replicate or test alternative ideas. It allows analysts to experience system characteristic changes in a safe environment and see the monetary and performance impact of decisions.

The benefits of simulating are numerous, hence why many emergency services (and other organisations) explore their systems in this manner (as seen in Chapter 3 section 3.5.7):

- **Understanding:** the simulated system should be easy to follow, with recognisable differences and similarities for anyone familiar with the real system, making the implications of any process modifications transparent where perhaps thought had not previously been given.
- **Safety:** in certain real-world experiments, if the consequences of system changes are unknown, implementing them can be risky to the organisation or business, or in some cases even unsafe to system users (particularly for an EMS system where patients' lives are in the balance).
- **Savings:** experimentation and system modifications are often costly. As with safety, it is often undesirable to simply change policies before testing them, since this costs time and money. Modelling prevents unnecessary changes until the optimal system state is determined.
- **Graphics:** through the inclusion of graphical displays of information and results, system operations are visibly understandable and interpretable, eliminating the need to deduce outcomes from complex formulae and abstract numeric values.

- **Scale:** experimentation can be sped up so outcomes are witnessed in a shorter period of time compared with the physical length of such activities. This is particularly useful when real changes might not take effect for months or years or when many scenarios wish to be explored.
- **Repeatability:** unlike in reality, simulated experiments can be replicated exactly to enhance the insight gained from results.
- **Flexibility:** it is possible to simulate in such a way that provides the user access to measures and estimates that would otherwise be unavailable or require years of data collection and new physical measurement of such information. Henderson and Mason (2000) state performance measures may be more accurately predicted than with more analytical modelling such as Markov chains and queueing theory.
- **Communication:** decision makers and system managers can interpret the outcomes of such a modelling approach for themselves, potentially with little assistance from the model builders. Results can be laid out simply and clearly for effective communication.

### 7.2.3 Visualisation

Simulation modelling often includes a visual aspect, whether that be specific movement of entities through the model, or (graphical) results displayed to the user. Visualisation of system structure and results is key in providing understanding of the processes contained behind the scenes in the programming code.

Designers are often not the same people as the intended users of the model and so the required formation of input and output structures should be as intuitive as possible for the eventual user, which can be aided by graphical displays and user-interface structures. An important inclusion to this simulation project is a robust user-friendly interface for the client. Results of complex logic and procedures need to be communicated simply and swiftly.

Graphical output can be demanding on computational resources and require long periods of time to develop and produce results. This study therefore takes a more conservative approach to the amount of graphical output given, finding a balance between informative and superfluous detail. Dynamic graphics are not used, but settings and simulation options are displayed as simplistically and intuitively as possible, with KPI results displayed to the user upon run completion instead.

The objective of this research is to provide WAST with a generic tool, with an interactive interface, to assist with daily planning tasks relating to locations and allocations of vehicles across any network.

#### **7.2.4 Comparison with Other Techniques**

Due to the relative inflexibility (of structure alterations, input styles and limiting Markovian assumptions) of non-simulation modelling approaches (which are mainly deterministic and static in nature) and also the potential degree of complexity of some problems, the provision of a simulation model is often more desirable. Simulation has the ability to cope with dynamic and transient effects (Pidd 2004, Robinson 1994) compared with more mathematical modelling approaches. A simulation model may also often better portray the effects on the system for more diverse scenarios and the implications of making changes to resources and control logic employed by the system. For example, in an EMS system, the analytical location-allocation model described in Chapter 6 provides static results regarding the optimal allocation of vehicles to the South East region of Wales; however, further analysis can be conducted via simulation in order to understand the stochastic performance over time of the system as a whole, given such an allocation.

#### **7.2.5 Limitations**

It is important however, to also realise the limitations of simulation when considering this line of experimentation. Often it is the case that the models developed may be over simplified for ease of representation. Real-world systems can be incredibly complex and it is usually impossible and undesirable to consider every detail when attempting to simulate their processes. Sources of inaccuracies include input and experimentation choices (Robinson 1999). The accuracy of the imitation is obviously crucial to the choices that may later be made using the model as a decision tool – model verification and validation, which deal with the accuracy and precision of representations, are discussed further in sections 7.8.4 and 7.8.5.

### 7.3 Simulation Type

Over the decades, many different types of simulation approach have been developed. The three main paradigms of simulation that assist Operational Researchers in a variety of scenarios are:

- **Discrete Event Simulation (DES):** perhaps the most commonly used approach, gaining momentum with the development of computer simulation in the late 1950's (Nance and Sargent 2002). It has since been used to model system processes in discrete time, where objects pass through the system in sequence with a time stamp to indicate the order in and time at which to process an event or change associated with the characteristic state. DES is useful when the behaviour of the system is stochastic and focuses on individual entities.
- **Agent-Based Simulation (ABS):** a slightly less common method stemming from ideas of John Von Neumann and work by Schelling (1971); multi-agent based simulation is a technique employed when there exist a large number of agents or entities, but where decisions are based on rules present for the agents interaction with each other and their behaviour within the system, rather than governed by probabilistic distributions as in DES.
- **System Dynamics (SD):** developed by Jay Forrester (1961), at the Massachusetts Institute for Technology (MIT), SD differs significantly from the other two techniques in that it can be thought of as a broader, aggregated view of an *entire* system. It is a population based view rather than entity/agent specific, looking at the flow and rate of change of populations, capturing feedback and interaction effects but for deterministic behaviour.

The approach of DES is chosen for the ambulance allocation-location problem, since its characteristics lend themselves well to the system's structure and the desired outcomes of this particular modelling task. DES compliments queueing theory, and due to the connections that queueing theory has to an EMS system, the technique is suited for application to WAST.



## 7.4 Programming

### 7.4.1 Choice of Style

A choice is made early in the planning stages of a simulation project on the style of build. For this project, it was decided the tool should be built and written in a general purpose programming language, as opposed to using a pre-developed simulation software package.

When deciding how best to conduct a simulation, the benefits of utilising a purpose built package as opposed to contriving one via a programming language must be considered. The question is whether the cost (for time and expertise) of manufacturing a new model in a coding language is outweighed by the cost of acquiring licences and understanding of existing commercial modelling software in which to develop the model (for which build time is still required and rigidity exists) (Goldberg et al. 1990).

Developing a completely new tool for this study has a higher cost in terms of development than exploiting a package, but benefits are found in the level of detail, complete flexibility of content and design and dynamic complexity captured by the final model. Additional logic can be added to models built in most pre-developed packages, but the user is bounded by the software. Through the creation of an original program, the developer gains further understanding of the system and appreciation of accuracy and assumption implications when making attempts to incorporate aspects of the real-world system. Henderson and Mason (2000) use a high-level programming language for this logical complexity reason and for two others: speed and integration with specialist GIS tools.

From the outcomes of the modelling project in this thesis, non-simulation experts should be able to make use of the tool in the future without the necessity to understand 'off-the-shelf' simulation packages or purchase expensive licences. The resulting tool is a standalone application that may also be run on more than one computer console at once, without the need for multiple copies of professional software – reducing running costs for the client and increasing speed. Some aspects incorporated to the simulation are also rarely seen in other models (such as a heterogeneous fleet, road map travel data and survival considerations).

### 7.4.2 Choice of Language

The tool developed for this project is written in the Visual C# programming language. Visual C# language is similar to C++ and Java (in that it offers class-based, run-time compilation), but utilises services of the .NET framework (instead of the computer hardware or operating system directly). It is developed in the Microsoft Visual Studio (Microsoft Corporation ©2010) Integrated Development Environment (IDE) which provides a place to create various application types using various .NET classes for execution. This .NET framework consists of class libraries of pre-written code and the Common Language Runtime (CLR) which handles memory, execution and security at runtime (Murach 2008).

The developed tool is an application that is executable directly from a user's PC, operating with a minimum specification of Microsoft Windows 2000 (Windows XP or later is required for development).

One reason for this choice over other general purpose languages is that it enables object-oriented programming (OOP). 'Objects' are manipulated whereby the focus falls on the form of the data and not the logic. Since this simulation is being driven by historical data, and the desired output should also inform through collected data (presented similarly to the data WAST collect themselves), this is of benefit to the project at hand.

*Simula* is considered to be the first object-oriented programming (OOP) language (Dahl 2002, Kindler 2007). Apparent from its name is the relationship between the language and its intent for use with simulation. As an OOP language (compared with 'procedure' and 'module' based structures), C# offers the user the benefits of concepts such as 'inheritance', 'polymorphism' and 'encapsulation' (Murach 2008, Wren 2007). Classes are created for objects, containing sequences of logic ('methods') and instructions on the actions befitting the individual objects.

- **Inheritance** – reduces development time and makes the coding more accurate by defining a new object in terms of an existing one. In this project, 'demand', 'station' and 'hospital' are all types of 'location' but with different attributes; therefore each of these three classes inherit (through polymorphism) all properties, methods and constructors from the parent class 'location'.

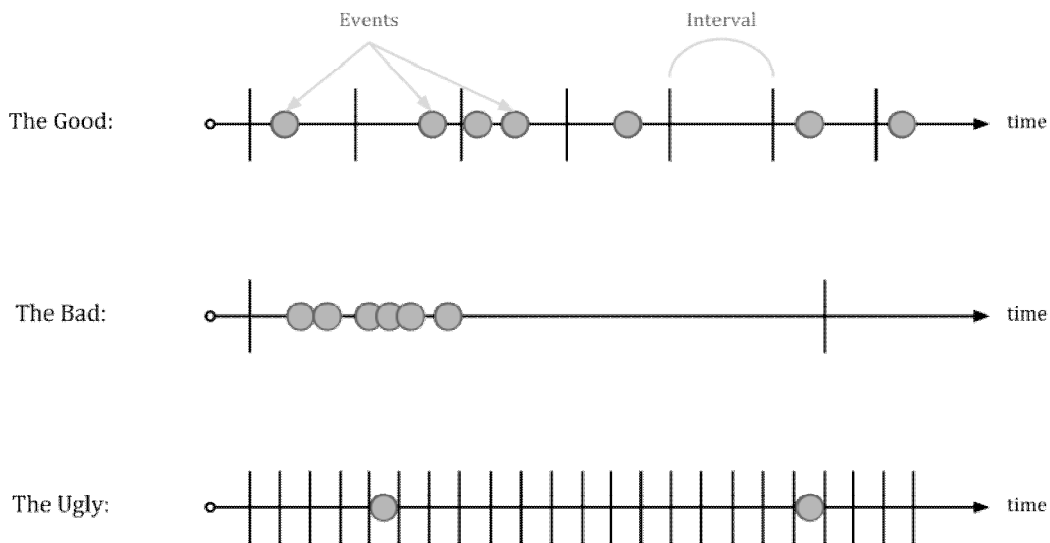
- **Polymorphism** – methods can accommodate more than one type of object but behave differently accordingly. Classes can provide inheritance to different objects due to this feature. It provides flexibility in usage and reduces the amount of code required to process objects during construction when their type may not be known until run-time.
- **Encapsulation** – enables better control of data within classes, where methods describe object actions. Methods can make changes to object instances but access to specific methods, modules and data associated with an instance of an object can only be viewed by other objects.
- **Reusability** – objects are created from a single class; the code is reusable anytime the same object type needs to be developed. Also, all methods contained in the class are accessible throughout the model where an object necessitates its access. Classes allow construction of new data types and can easily be adapted for implementation in other OOP languages or .NET applications.

Free-coding in Visual C# allows complete control and the ability to integrate the Travel Matrix Generator Tool into the simulation whilst capturing the exact level of detail required for the research. The tool interacts with the JavaScript source code behind the Google Maps web page and utilises the API as described in Chapter 5 and Appendix 5.1. Finally, the visual aspect that can be incorporated to a model built using Visual Studio, achieves the goal of creating an intuitive user-interface for WAST to implement in the future – navigated easily by non-technical users.

### 7.4.3 Time Handling

DES handles processes through time, where events indicate system actions. Various methods exist for dealing with time steps, including the three-phase, activity based and process based approaches (Pidd 2004). More commonly, the 'clock' of the simulation is advanced using the 'next event' technique, whereby the simulation skips forward in time (in unfixed increments) until it reaches a time state in which an event must be processed. The life cycle of the system is split into manageable parts and a time scan is performed to see if an event is due to occur; if so, the clock pauses, an event execution is triggered and a routine is called to deal with the state change associated with the event(s).

An alternative process for handling the progression of time in a simulation is 'time-slice' programming. It involves moving forward through time in equal steps or intervals, and reviewing requirements of the system at each step. The interval is denoted as having length  $\delta t$  and so the next review of the system occurs at time  $t + \delta t$  (Pidd 2004). It is important to correctly gauge the time slice when simulating so that the model gives a 'good' representation and that information and system conduct is not lost by too large an interval ('bad'), yet computer processing effort is not wasted unnecessarily by too small an interval ('ugly') (Figure 7.2).



**Figure 7.2** Examples of time-slice programming: The Good, the Bad and the Ugly

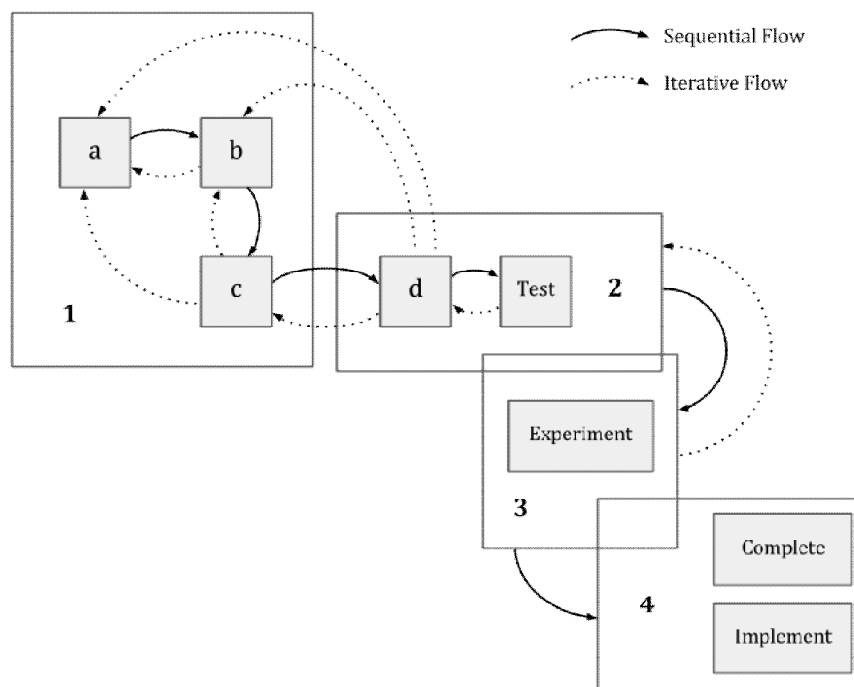
Next-event avoids 'slack' periods, speeds up processing time and performs more efficiently in periods of low demand than time-slicing. However, to accomplish next-event, more information must be provided for the simulation logic. Time-slicing is equally efficient in periods of high demand so long as the fixed time interval selected is suitable. In this simulation project, the time-slice approach is implemented, with a step interval of one minute (since the system generally witnesses a large number of events).

## 7.5 Simulation Design

### 7.5.1 Contribution to WAST

There is an abundance of reasons for the decision to simulate an EMS specifically. For WAST, the daily operational structure, for the most part, is designed and planned by hand. Vehicle locations (and allocations) are planned by the information analysts and decisions are based on the judgement of experienced individuals. Results are unlikely to be near optimal when devised in this manner. Simulation is used here to help plan static resource location strategies to give good starting allocations such that the expertise of the service planners can then be better exploited for any dynamic relocation requirements at an operational level.

The four main, iterative stages (Robinson 1994) of a project using simulation, conveyed by Figure 7.3, are: 1) Define the problem; 2) Build and test the model; 3) Experiment; 4) Complete and implement. The first two points are covered by the remainder of this chapter, the third point is explored in Chapter 8 and the fourth remains a hopeful achievement of this project.



**Figure 7.3** Stages of Simulation

Designing a simulation (1 and 2) consists of separate, sequential stages: a) definition; b) plan; c) conception; d) build. Design and build stages for this particular project are sizeable since a good amount of interaction with the tool is desired so WAST may later use it themselves. The model needs to convey policy and operational options to the client without intimidating users with unnecessary complexity. A generic design also permits extension to other regions and EMS systems.

### 7.5.2 Objectives

A commonly used framework (Robinson 1994) for project objectives is to consider, in context, three components: achievements, measurements and constraints (sections 7.5.7-7.5.9).

In brief, simulation is chosen as a methodology to model South East Wales' EMS operations for the following reasons, (based on the advantages of simulation discussed previously and the purpose of this particular modelling project):

- reduce risk to patients during strategy testing;
- gain understanding of the operations of the current system;
- reduce costs, lost ambulance hours at hospital and patient waiting time, whilst improving utilisation, patient experience and performance (saving money and lives);
- test impact of suggested system changes such as clinical outcome.

In determining the objectives, motivation is justified by the discussions of earlier chapters – negative reports of the service's operations show a desperate need to improve the customer experience and reduce response times in line with the current UK targets, whilst reducing yearly expenditure. Investigation of the turnaround aspect of WAST's operations is also ideally suited to simulation modelling. The Trust has little control over this service phase so cannot make real-world operational changes but can instead experiment through simulation to see if improvements of the interaction with A&E departments assist with EMS performance.

In addition to the main objectives, this study also considers vehicle utilisation across the region. This is obviously a by-product of vehicle positioning, but also of turnaround time.

The problem faced by WAST is redefined with respect to the possibilities of simulation, allowing the objectives for this part of the project to be stated as:

*Improving patient service and survival by minimising the response time and turnaround time experienced across the South East region by utilising optimal allocations and whilst maintaining existing resource levels.*

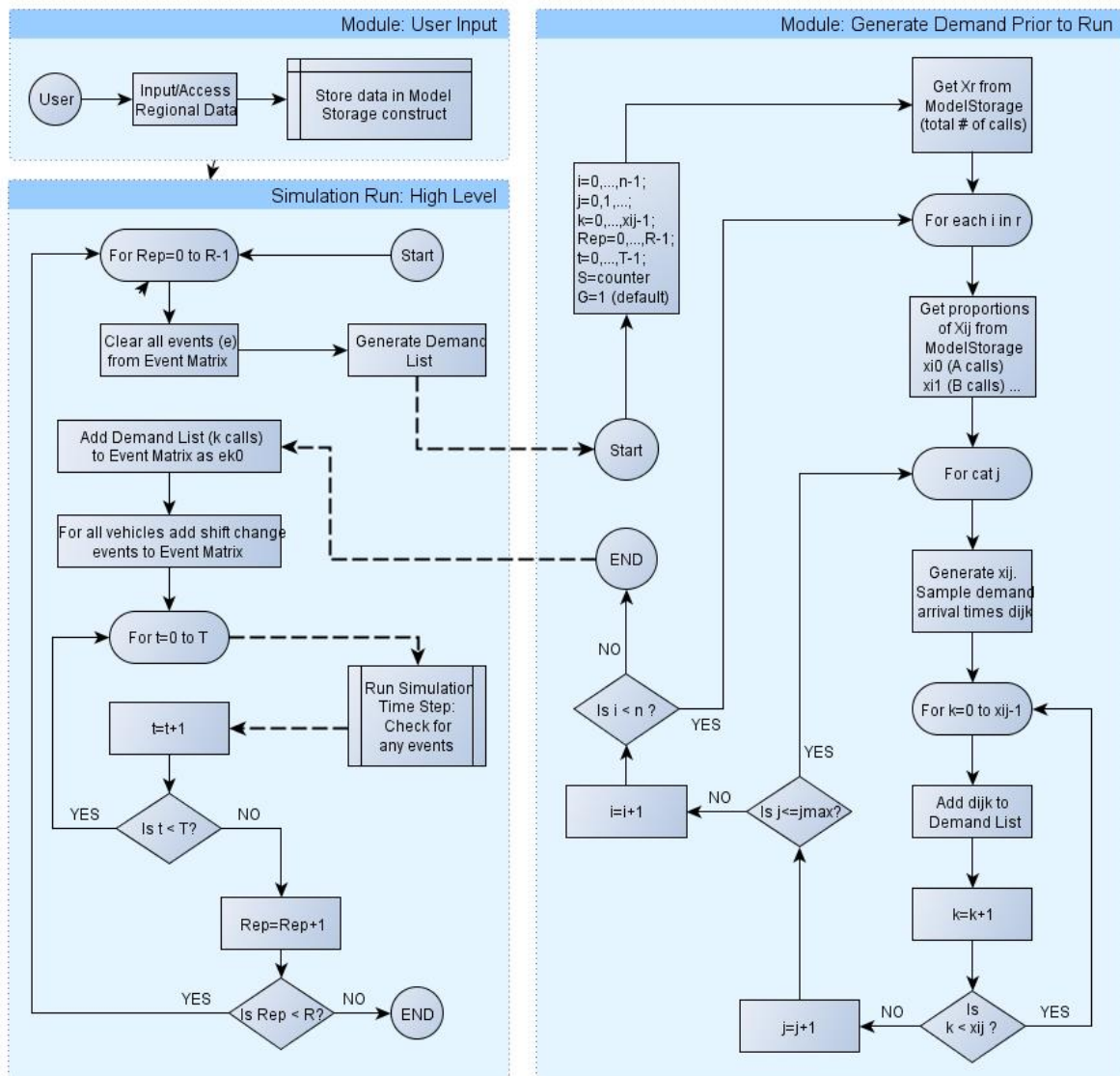
The general project objectives call for a high level of detail, especially following the static and deterministic efforts of the previously described location models (Chapter 6), to build on current information and explore system processes in depth.

### **7.5.3 Conceptual Modelling**

By conceptualising a model before building, the developer is able to consider both the 'scope' (breadth) and 'level' (detail) required of the simulation. For WAST, the model intends to cover the entire emergency system in a resource planning capacity whilst investigating closely specific phases of service (response and turnaround).

There exists a trade-off in the design of the EMS system simulation between the level of detail to include (since the problem is quite complex and much information is required for a good representation) and keeping the model as simple but accurate as possible to avoid unnecessary build time. The optimal scope and level for a specific model may be determined by judgement and "successive inclusion" or "successive exclusion" (Robinson 1994) of details along with the consideration of run time and time scale of the project.

The conceptual model, seen in Figures 7.4a and 7.4b, was created and verified through interaction with WAST control centre employees and the historical data received.



**Notation:**  
 $r$  = the total demand region  
 $X_r$  = number of calls per period in region  $r$   
 $n$  = number of postcode areas in  $r$   
 $i$  = postcode area;  $i = 0, \dots, n-1$   
 $j$  = category of call;  $j = 0, 1, \dots$  (e.g.  $j=0$  is category A;  $j=1$  is category B)  
 $x_{ij}$  = number of calls in area  $i$  for category  $j$   
 $k_{ij}$  = incident (or call);  $k_{ij} = 0, \dots, x_{ij}-1$  for all  $i, j$   
 $dk_{ij}$  = arrival time of call  $k_{ij}$  in area  $i$  for category  $j$   
 $p$  = event type, indicated priority of event;  $p = 0, \dots, 6$   
 $ek_p$  = event associated with call  $k$ , with priority  $p$   
 $Rep$  = replications;  $Rep = 0, \dots, R-1$  ( $R$  = total number of replications to run)  
 $t$  = simulation clock time;  $t = 0, \dots, T-1$  ( $T$  = length of run/number of time steps)  
 $S$  = number of stations in  $r$

**Figure 7.4a** Process flow design of the simulation model (high level)





An 'entity' is an individually identifiable object (in OOP terminology), element, resource or person in the simulation environment of interest, which changes and moves through time, obeying the control logic in place. In the case of the Welsh ambulance service, it can be thought that there are two types of entity of interest: the patient or incident (a 'temporary' entity) and the responding EMS unit (a 'permanent' entity). Each of these is followed through time in the system and changes are made according to the entities' location, status and history within the system. 'Attributes' are the characteristics of a particular entity that may be changed, and are summarised in Table 7.1. (Entities and C# attribute enumerators may be seen in the class structure - Appendix 7.1 and 7.2.)

**Table 7.1** Details of the simulation entities and their main attributes

Entity	Attribute	Class, Enumeration or Value of Attribute
Incident	Origin	Origin of call is Demand, inherits from Location class
	Type	Enumerates either A, B, C, AS2, AS3
	Time Stamps	From arrival time of call through to clear time of service vehicle (see Figure 4.1, Chapter 4 for all time stamps)
	Vehicles	Number of vehicles required (1 or 2)
	Survival	Probability of patient survival given response length
Vehicle	Base Station	Station inherits from Location class
	Type	Enumerates either EA or RRV
	Status	Enumerates Off-Shift, Busy, Free, Returning or Returning
	Utilisation	Total utilisation based on total busy and on-shift time
	Job List	List of all Incident entities served by the Vehicle entity

### 7.5.5 Assumptions

Including too much scope and detail into a simulation means increasing the amount of time required for development. By making some carefully considered simplifying assumptions, the integrity and accuracy of the model can be maintained whilst minimising design time and avoiding including unnecessary system aspects not critical to the problem objectives.

Assume the simulation represents a region with distinct, fixed and restricted (sub-region) boundaries. Other assumptions made during the building of this simulation include:

- infinite queue size – if all vehicles are busy, a patient will simply wait for the next available and suitable resource;
- as default, RRVs serve only category A (but the ability to change this exists and will be explored in Chapter 8);
- only EAs have the ability to transport patients to hospital facilities;
- nature of an incident influences the time spent on scene, as does the vehicle type responding;
- on-scene service length is not dependent upon decision to transport;
- travel times are temporally independent;
- pre-travel delay is dependent upon vehicle type and category (not station);
- turnaround times are independent of nature of emergency but dependent upon hospital;
- the modelled South East Wales region is self-contained – can also assume the principle of equal assistance rates as given in the discussion of Chapter 4, section 4.4.2.

### **7.5.6 Data Analysis**

The simulation model is built bearing in mind many of the results and discoveries of the Chapter 4 data analysis findings. All parameters and variable values of the simulation model are able to be amended according to a particular scenario, region or Trust being modelled. For the purposes of this study, the design and demonstration of the working model is based on the characteristics of the South East Wales regional data for 2009.

Stochastic elements are incorporated to reflect the unpredictable nature of an EMS. A deterministic model would not be ideal in this scenario since the distribution of arrivals and changing demand has an effect on emergency service. The model needs the ability to model variation and stochasticity around travel, service and turnaround phases of the system.

### 7.5.7 Input

Inputs and experimental 'factors' of this project are listed in Table 7.2 along with the possible range of values which they feasibly and realistically could take. The choice of data entry method is a crucial aspect to the input mechanism.

Data entry is granted through a combination of menu-driven options, manual user input, and loading of pre-generated text, comma-separated value (csv) or spreadsheet files (see Appendix 7.3). Values are stored and referred to in memory, where all values may be viewed and altered internally if required. Since the user-interface provides further details displayed for all data values and factor levels, it offers an advantage to making alterations to the external csv files. Any changes made within the tool are able to be saved as external csv files such that if a scenario is required to be run again exactly, it may be reloaded and all history of experimentation is accessible.

The data values and factor levels are not all pre-determined before the model building began; however, due to the iterative nature of simulation, the experimental factors and other such items may have been decided upon throughout the design process.

Fleet allocation to stations across a region per shift can be altered after loading the information into the model. Before running a trial, it is possible to modify the default or pre-generated shift allocation data manually or read in alternative shift information from other external files (Excel, xml or csv). The allocations are completely controllable and freely adjustable by the user. This feature is particularly useful since real-world vehicle allocations were unobtainable for the project; links can instead be made with the outputs of the location models presented in Chapter 6 to gain a suitable benchmark set-up. Allocation and capacity estimates can be explored within the simulation.

Some further variables, relating solely to conduct of a scenario, are also input to the simulation model, or can be specified within the tool by the user at run-time:

- run length;
- replications;
- warm-up length;
- choice of auto seed or specified seed;
- control seed (if auto seed selected).

**Table 7.2** Input data and variable types provided to simulation

Input Data (D) & Factors (F)	Details	Data Type	Default Values
Total Demand (D)	Average weekly demand	$\mathbb{Z} > 0$	$\sim N(3150,140)$
Speed Factor (D)	Percent to scale estimated travel time by vehicle and journey type	[0,100]	100
Travel Equation (D)	Regression model and parameters used to estimate travel time		Google Maps results
Shift (F)	Number of shift changes	$\mathbb{Z} > 0$	1
Locations (F)	List of addresses	Object	Demand, Station or Hospital type
Routes (D)	Dictionary of travel information between all pairs of locations		0 distance, 0 time
Category Demand (D)	Proportion of demand	[0,100]	
Response Targets (F)	Targets per category	$\mathbb{Z} > 0$	Current UK targets
Transport (F)	Proportion of category transported	$\mathbb{R} \geq 0$	100
On-scene (D)	Information of on-scene length by category and vehicle type	Distribution	15 minutes (fixed)
Pre-travel delay (D)	Information of pre-travel delay per category and vehicle type	Distribution	1 minute (fixed)
Hourly Demand (D)	Profile of demand per category	Empirical distribution	
Vehicles (F)	Number of vehicles assigned to each station location, per type	$\mathbb{Z} \geq 0$	0 EAs, 0 RRVs
Shift Vehicles (F)	Number of vehicles on shift at each shift change, per type	$\mathbb{Z} \geq 0$	
Turnaround (D)	Information for turnaround time at each hospital location	Distribution	20 minutes (fixed)

The dispatch algorithm (e.g. closest available), hospital facility choice algorithm (e.g. closest) and service policies (i.e. whether an RRV serves all categories or only some) can also be modified. These will be discussed further in Chapter 8 as they lend themselves to the experimentation stages of this simulation project; however they are mentioned here to illustrate that the possibility of such diversification is incorporated to the model design.

### 7.5.8 Google Maps API Input

In addition to the input data, travel information between all nodes of the network modelled is required for reference to compute vehicle journey times during the simulation. The Travel Matrix Generator Tool is utilised in conjunction with developed regression models for this purpose.

All locations to be represented in the simulation are passed to the Matrix Generator Tool (embedded within the simulation model interface) with a matrix of travel distances (and times) returned and stored ready for reference throughout the modelling process. This method speeds up simulation run-time, since the Google Maps API is accessed only once per location pair in advance of a trial, followed by a lookup process from a reference matrix as and when required during a run (i.e. Google Map requests do not have to be made for every occurring journey of the run, only once per possible journey).

During simulation, a choice can be made to use Google Map travel times directly, or a prediction method. Unfortunately, the travel times returned by Google Maps are not necessarily representative of journey times by emergency vehicles.

Distances over routes can be taken as constant for any vehicle type or journey purpose, so as previously explained, travel times can instead be calculated via regression models (Chapter 5) with Google Maps distance as the independent variable. Scaling parameters are applied to the models to represent variation in speed for the different EMS journey and vehicle types, all of which are provided as input information to the model.

Regional journey travel time  $Y$ , is modelled as

$$Y(t) \sim LN(\mu, \sigma) \cdot s - 1$$

where  $s$  is a scalar dependent on the journey and vehicle type,  $\sigma$  is the average standard deviation of all routes from the data (by category) and mean,  $\mu$ , is given by the developed regression models of Chapter 5, dependent upon vehicle type and incident type:

$$\mu = (a + bX^c) + \varepsilon \sim N(\mu_\varepsilon, \sigma_\varepsilon^2)$$

When a journey time needs to be calculated during a simulation run, a look-up in the Travel Matrix of the distance of the route is made and the value input to the corresponding regression model equation, returning an estimated travel time for the journey.

Utilising the models of Chapter 5 (Equations 5.8-5.14), with the correct parameter values for  $a$ ,  $b$  and  $c$ , for each of the incident types and vehicles types captured by the simulation model, the following equations (7.1-7.8) are input as default models for travel time in the South East region of Wales.

EAs per category:

$$A: Y(t) \sim LN([7.23 + 0.80X] + \varepsilon \sim N(1.84, 1.35), 6.12) \quad (7.1)$$

$$B: Y(t) \sim LN([7.47 + 0.82X] + \varepsilon \sim N(1.91, 1.28), 5.66) \quad (7.2)$$

$$C: Y(t) \sim LN([6.90 + 0.90X] + \varepsilon \sim N(1.92, 1.55), 5.54) \quad (7.3)$$

$$\text{Urgent: } Y(t) \sim LN([7.76 + 0.94X] + \varepsilon \sim N(2.33, 1.96), 6.57) \quad (7.4)$$

RRVs per category:

$$A: Y(t) \sim LN([-1.57 + 4.92X^{0.5}] + \varepsilon \sim N(1.71, 1.56), 6.83) \quad (7.5)$$

$$B: Y(t) \sim LN([-1.61 + 5.16X^{0.5}] + \varepsilon \sim N(1.88, 1.40), 5.90) \quad (7.6)$$

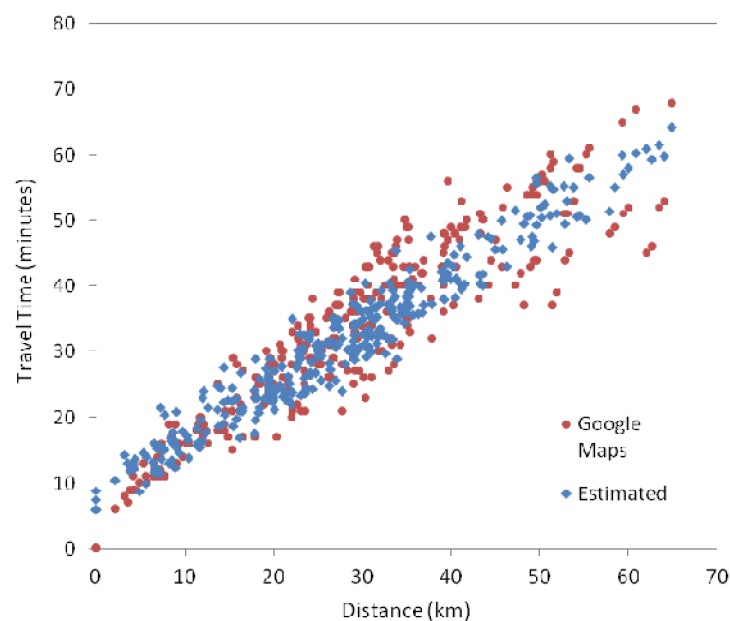
$$C: Y(t) \sim LN([-0.36 + 4.77X^{0.5}] + \varepsilon \sim N(1.90, 1.38), 6.11) \quad (7.7)$$

$$\text{Urgent: } Y(t) \sim LN([-0.25 + 5.68X^{0.5}] + \varepsilon \sim N(2.01, 1.66), 7.11) \quad (7.8)$$

Using a globally stored value, (seen later in Figure 7.8) the estimated travel time value is then scaled to represent the speed at which the vehicle is likely to travel given the type of journey and vehicle serving. For example, a vehicle on a response journey travels faster than a vehicle returning

to its base. From discussions with WAST employees, the consensus is that a vehicle transporting patients generally travel more slowly than when responding, even during a high priority service, since the vehicle aims to remain stable when the patient is on board no matter how urgently they need to get to hospital.

When a base is located at a demand node, travel time or distance given by Google Maps (stored in the Travel Journey Matrix), will have a value of zero. However, there is obviously still *some* travel time associated with a response journey from the base to the exact scene of the emergency, so the simulation accommodates this by taking minimum travel time to be one minute when using travel times directly. Goldberg and Paz (1991) set the distance of such instances to be 0.5 km before estimating the travel speed and response time.



**Figure 7.5** Google calculated travel time compared with predicted travel time for a category A transportation journey using only the mean component of Equation 7.1

A comparison of the travel times calculated by Google for a given journey, with those estimated using Equation 7.1 is given in Figure 7.5. This scatter plot represents the similarity between the developed predicted travel times of transportation journey for a category A patient by an EA (based on distance) and the unknown Google Maps value calculation method. During a transportation journey (represented by the plot), vehicles are unlikely to travel at high speeds so as not to cause



further trauma to the patient. The further need for prediction methods over exact Google Maps results arises when other vehicle types and journey types are considered.

### 7.5.9 Outputs

Outputs or responses identify the values useful to inform WAST. The method of reporting should be determined along with how to view the outputs (Robinson 1994). Responses refer to the way in which the measurements of the simulation, listed in Table 7.3, are portrayed. The output may also highlight ways in which the objectives have not been met and so identify potential room for improvement in the solutions.

**Table 7.3** Measurements of the simulation that highlight performance of the system

Measure	Details	Expected Impact On
Waiting Time	Time patient spends waiting from incident time to arrival of vehicle at scene	Survival
Response	Time from vehicle allocated to arrival at scene	Survival & Performance
Turnaround	Time taken to transfer the care of patient at the closest hospital facility	Utilisation & Response
Utilisation	Amount of time vehicle spends serving patients compared to total on-shift length	Response
Survival	Given a response time (from arrival to on scene) the corresponding survival probability	Performance

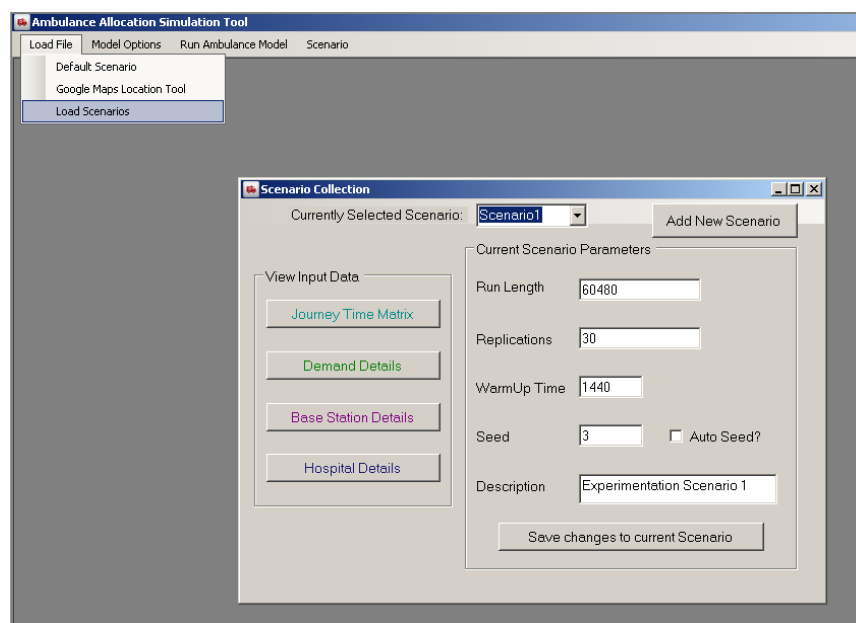
After running the simulation (for individual or batched scenarios), the tool generates summary output, all of which may be viewed internally, but additionally, all data generated by the model can be exported. The externally stored output enables further analysis to be conducted at a later stage or can be used for reference. The user has complete control over the location of the exported data; the style in which the data is saved allows the resulting file to be opened using programs such as Microsoft Excel or SAS for further exploration.

Importantly, the recorded and exported data has an almost identical structure and format to WAST's data set in order to optimise comparisons. In-depth analysis may therefore be performed on the simulation results compared with historical data (analysis of Chapter 4) to investigate operations and contribute to the model validation processes. WAST's monthly response target reports would be comparable (if the data were accessible) to the modelled output since the tool produces similar content. One addition is that the simulation retains and stores information regarding individual vehicle utilisation and incidents.

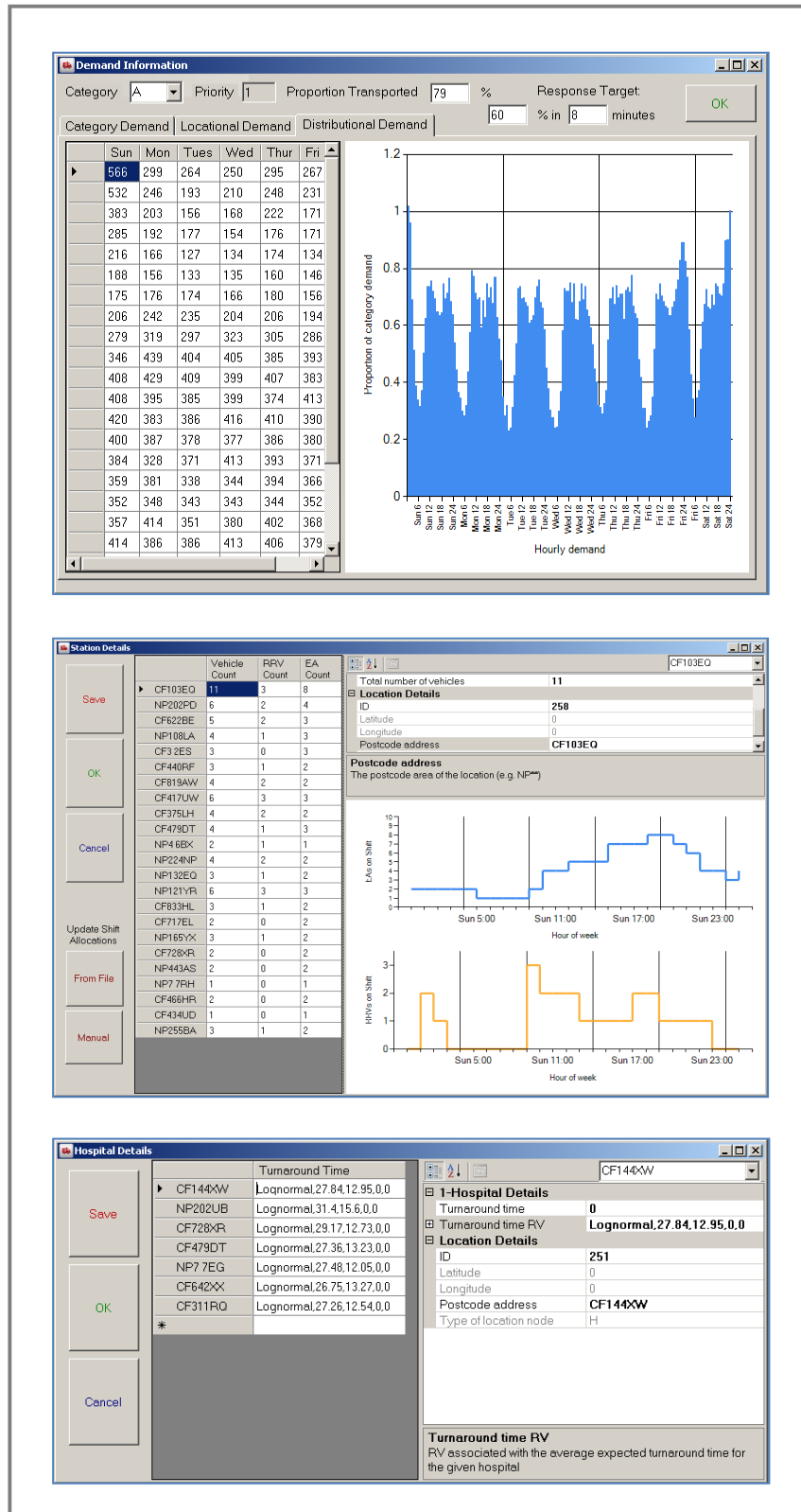
## 7.6 Program Processes

### 7.6.1 Process Introduction

All of the required input data and parameter values are stored as references and values in memory such that the model has immediate access to them. A main window displays menus directing the user to sub-forms of data selection and variable options. At this point the simulation itself is idle, waiting for the user to instruct on the activities to conduct, as in Figure 7.6. The subsequent sub-windows, Figure 7.7, accessed from the 'Scenario Collection' form, allow alterations of travel, demand, station and hospital parameter values, distributions and options via the user-interface.



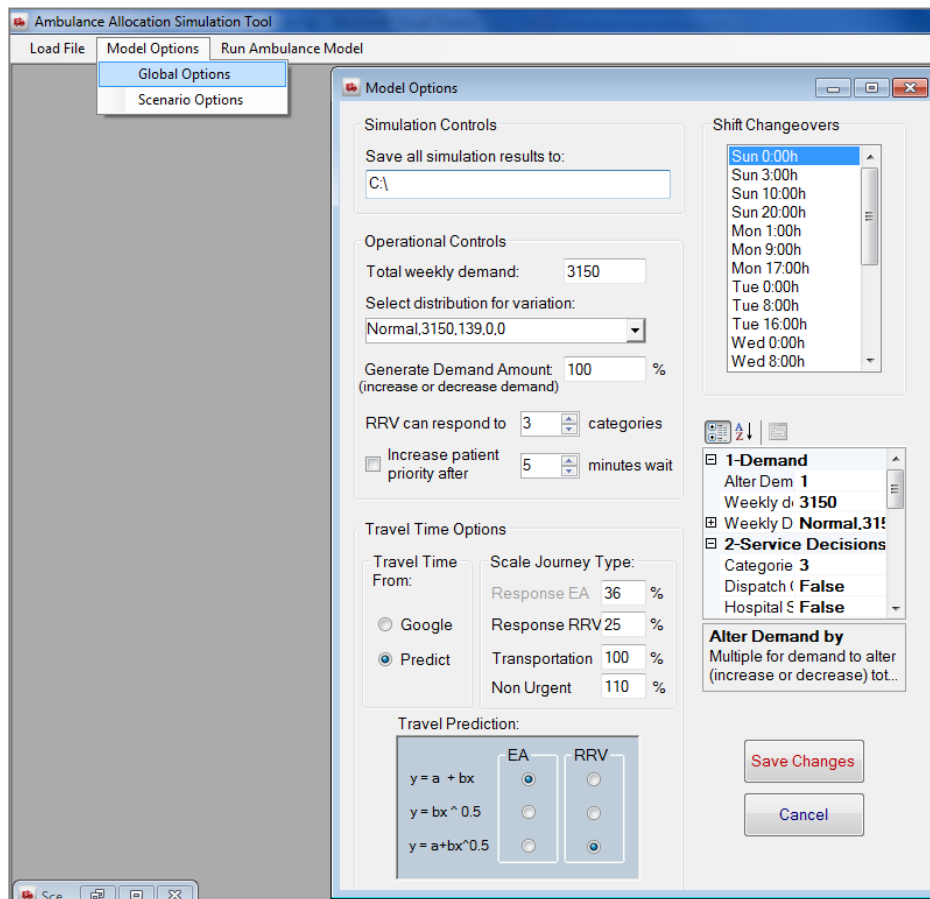
**Figure 7.6** User-interface of the simulation tool, in idle state, with scenario collection sub-window as displayed upon loading scenarios from the 'Load File' main menu



**Figure 7.7** Sub-windows ‘Demand Details’, ‘Base Station Details’ and ‘Hospital Details’ as accessed from the ‘Scenario Collection’ form of Figure 7.6 (Examples of the display given by the ‘Journey Time Matrix’ form were provided in Chapter 5, Figures 5.16 and 5.17)

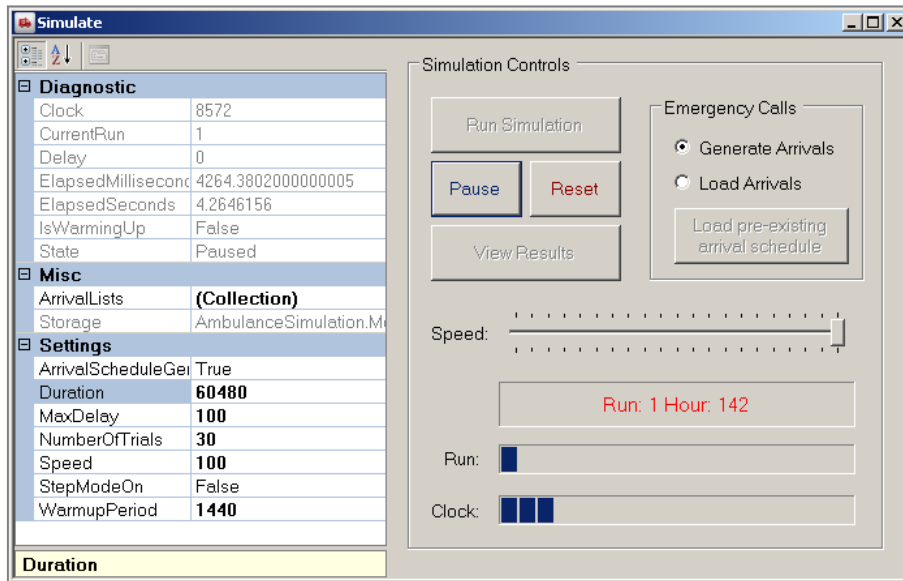
Additionally, the user may make further alterations to do with the model set-up and logic through the 'Model Options' menu and 'Global Options' form (Figure 7.8). Choices regarding shift times, overall demand amount to simulate, dispatch policies and whether to use Google or predicted travel values are facilitated.

Upon deciding to perform a simulation, the user initiates the start of this modelling process from the current window. By selecting the 'Run Ambulance Model' menu option, a final window is opened as in Figure 7.9, so that the user can observe the simulation's progress via the status bar, portraying the current weekday and hour being modelled and the proportion of overall runs so far completed during run-time. The speed of the simulation may be altered using the slide-bar mechanism, this facilitates debugging and judgement of process and scenario speed.



**Figure 7.8** 'Global Options' form displaying alternative choices from the default values (displayed) for model and system variable options

The program must perform some tasks before conducting any replications of the system, (such as the decision on whether to ‘generate arrivals’ or ‘load arrivals’ – assume the former option for the following explanations, the latter will be explained shortly, in section 7.6.3). The steps involved in a single run of the model were laid out in the process flow diagram of Figure 7.4b and will now individually be detailed, where pseudo-code for some of the subsequent section action can be referred to in Appendices 7.4a and 7.4b.

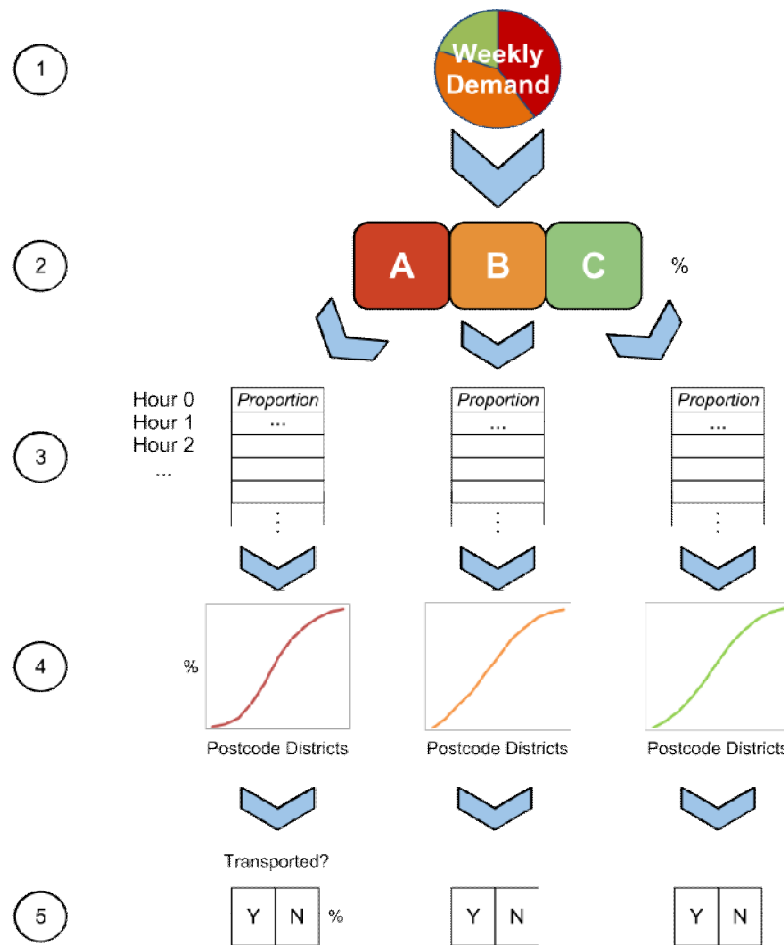


**Figure 7.9** User-interface during a trial of the simulation, showing settings and current state of the model

## 7.6.2 Generating Demand

The first action of the simulation is to generate the correct demand quantity ready to represent the arrivals of emergency calls to the region. The partially stratified (for call category) sampling process for demand is summarised in Figure 7.10, which depicts the method employed for sampling the characteristics or attributes of an emergency call (also known as an arrival).

The figure represents diagrammatically the steps of attribute sampling for generated calls for service at the beginning of each new run during a simulation, which are then stored as a call log, detailing all expected demand and their known attributes.



**Figure 7.10** Probability pathway of patient groups and call generation process used within simulation programming structure

Steps:

1. Calculate the number of calls expected to arrive throughout the run-length time period, dependent upon run length and weekly demand. Weekly demand is a global variable, with a given distribution from which to sample. Sample the total number of expected calls per week of the run-period; adjust the demand amount for any partially modelled weeks since run-lengths are not required to be exactly divisible by seven days.
2. The category of call is determined using known emergency type proportions from Chapter 4. Time of day dependency is not included for overall call category proportions, but arrival time is dependent upon category.

3. Arrival time of the call is sampled using input data detailing the average number of expected calls in any one hour of the week (hour of the day, by day of the week). Within the model, cumulative probability is calculated for arrival time. After random number generation, the hour of the week is sampled and assigned; the remainder of the random number denotes the minutes associated with the selected arrival hour.
4. Similarly, determine the origin of the call. Each postcode district has an associated expected number of calls. Proportions are cumulated so that a generated random number look-up method can represent an arrival from any one of the locations. The profile of demand by location differs by category, but not specifically by time of day or day of the week – any time dependency is captured by category not location in the simulation.
5. Finally, a decision of whether or not to transport the patient is made. It may be thought of as independent of location, time of day and all other factors, except for emergency type for the purposes of the simulation. Proportions are given per category and the decision is the result of a simple Bernoulli sample.

### **7.6.3 List Structures**

When a call log is generated, either using the method of the previous section or by accessing a pre-generated call log (containing an exact list of incidents and their attributes), all the incident objects are stored in a sequential-list in the order they were generated (due to the way in which the arrival time is sampled). During the simulation, it is preferable and more efficient to search a list in order of event occurrence (service order). An example of a generated call log is given in Figure 7.11 for a portion of the generated calls and a selection of their attributes. From the fourth column of the incident list the arrival time and day of the call is noticeably unordered temporally.

Category of Call	Event Stage of Call	Origin of Call	Arrival Day	Arrival Hour	Arrival Minute	Transported?	Time to Respond	On Scene Service Length	Transportation Length	Time to Turnaround	Overall Service Length
AS3	Arrival	NP23	Day 5, 14:36	134	36	True	0	0	0	0	0
A	Arrival	NP12	Day 6, 10:1	154	1	True	0	0	0	0	0
A	Arrival	NP11	Day 6, 21:52	165	52	False	0	0	0	0	0
AS3	Arrival	NP12	Day 3, 16:50	88	50	True	0	0	0	0	0
A	Arrival	NP11	Day 4, 3:57	99	57	True	0	0	0	0	0
B	Arrival	NP10	Day 6, 11:53	155	53	True	0	0	0	0	0
AS2	Arrival	NP10	Day 5, 12:25	132	25	True	0	0	0	0	0
C	Arrival	NP12	Day 0, 14:20	14	20	False	0	0	0	0	0
AS3	Arrival	NP10	Day 4, 10:10	106	10	True	0	0	0	0	0
A	Arrival	NP23	Day 4, 16:14	112	14	True	0	0	0	0	0
C	Arrival	NP11	Day 5, 18:11	138	11	True	0	0	0	0	0
A	Arrival	NP18	Day 3, 9:4	81	4	False	0	0	0	0	0
A	Arrival	NP11	Day 4, 4:18	100	18	False	0	0	0	0	0
AS3	Arrival	NP10	Day 0, 10:18	10	18	True	0	0	0	0	0
B	Arrival	NP10	Day 5, 2:25	122	25	True	0	0	0	0	0
B	Arrival	NP11	Day 0, 12:36	12	36	True	0	0	0	0	0
AS2	Arrival	NP23	Day 5, 21:44	141	44	True	0	0	0	0	0

**Figure 7.11** Example of a call log generated at the start of a simulation run

The next step of the simulation logic organises the call log by arrival time (day, hour and minute) and then by category of call, placing any calls with the exact same arrival time in order of service priority (highest to lowest). Each item in the new linked-list (which is a faster search structure than a sequential-list) is given a unique ID number to be able to track the incident object throughout the simulation and update the attributes of these entities as changes take place. The resulting list is copied into an 'event' list, referred to as the 'schedule', so that additional event type objects (not just incident arrivals) may later be included, utilising object timestamps.

#### 7.6.4 Event List

An event list (also a linked-list but with binary search, to speed-up processing), as communicated in the previous section, is a collection of all objects requiring action at a particular (discrete) time step. Within the simulation process, there are several event types that are launched during a run, requiring some form of control logic to take effect. Priority is given to certain events over others, so that in the ordered linked-list, events with the same time stamp will be dealt with in the natural or necessary processing order. This ensures resources are available accordingly and the processes with precedent in reality are carried out similarly in the model.



Events can be one of the following (in priority order):

1. vehicle goes on shift;
2. vehicle is clear;
3. vehicle returns to base;
4. vehicle arrives on-scene;
5. vehicle goes off-shift;
6. incident awaits an RRV;
7. incident awaits an EA;
8. arrival.

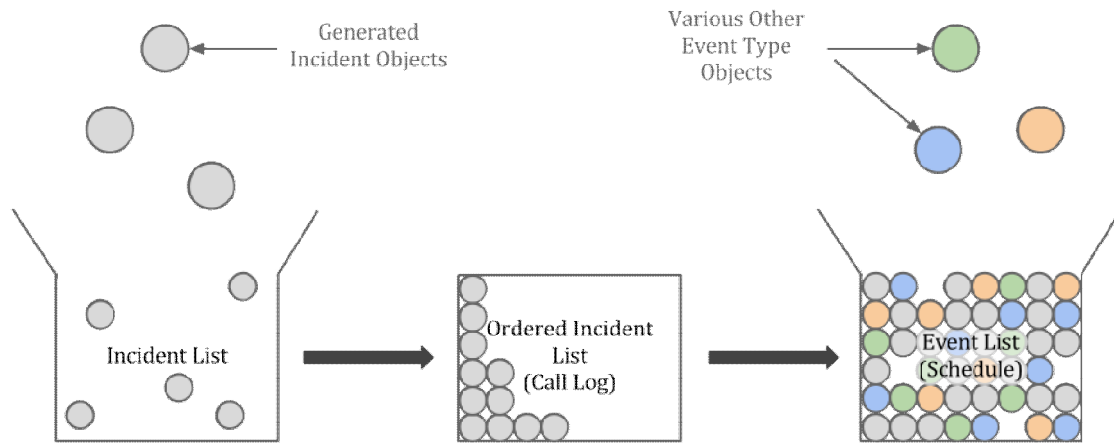
If events at a time step have the same priority, the order in which the event objects are dealt with is dependent upon the category. If the category is also identical, the length of time the patient has been waiting for service will indicate their sequence (this only has an influence on results if dealing with a waiting event). If by chance, every attribute compared is identical, the object with the earliest ID number is handled first.

Each individual event object references an incident object and vehicle object along with other attribute information such as the timestamp and event type (and priority).

In event approach DES, an event list is generally created at the beginning of a run and is populated initially with arrivals and subsequently with service events and vehicle shift changes throughout the run. If and when an event is dealt with or is no longer required, it is removed from the list.

Correctly scheduling events in the list or scanning the list during run-time can be quite demanding on computer processing time. There are some developed methods that combat this problem; however, due to the size of this particular problem, it is appropriate to schedule using a simple calendar queue data structure (Brown 1988).

The relationship between the various list structures housed in the model is demonstrated by the linking configuration of Figure 7.12.



**Figure 7.12** Relationship between the various list structures (call logs and schedule) for incident objects and the eight event objects in the model

### 7.6.5 Waiting Events

If no vehicle is available, an event is created of type 'incident awaiting vehicle'. If the incident is awaiting an EA response, the event created should be scheduled at the next time step that has an event already listed. When the clock reaches the planned time of this new event, a search for an available EA occurs; since no other system state or object changes occur until this next populated time step, there is no benefit in simply scheduling the waiting event at the next time step, at least one other event of some sort must occur first for the outcome of the search for available vehicles to differ. If however, the incident is of a type that requires service by an RRV and EA (e.g. category A) and an EA has already been dispatched but an RRV was not initially available, there is no advantage in scheduling a waiting event for an RRV at all if the EA is due to arrive on scene before the next available RRV. Since an RRV is usually sent as an initial responder with EA follow-up, a new event is not created in this situation. Both these waiting scenarios are possible in simulation since future events are known in advance (unlike in reality), through the use of an event list.

Additionally, when an EA arrives on scene, a check is made to see if an RRV is required but has not yet arrived on scene. If an RRV is scheduled, not at the scene but en route or dispatched, the RRV is then cancelled, since their use as an initial response is no longer necessary. This routine is based

on what is considered to be common practice in the control rooms of WAST after discussions with the analysts in the Trust. The cut-off for cancellation is one minute – when an RRV is a minute or less away from arrival on scene, the vehicle is not cancelled and the system continues as scheduled.

If a patient experiences a long wait, their priority level can be increased to account for the possible deterioration of their health and the heightened urgency with which the Trust would, in reality, attempt to respond to the incident. There are various options for incorporating this method in the simulation (no knowledge of the real conduct in such a situation is known); for example:

1. Increase priority by a maximum of one category level, if waiting time exceeds a pre-defined global maximum time length (for all but the highest priority patients).
2. Increase priority, up to the highest priority emergency type depending on length of wait. For every multiple of the global, pre-defined maximum waiting limit surpassed by the patient waiting time, increase their priority another level.
3. Increase priority, up to the highest priority emergency type if waiting time exceeds a pre-specified length of time, for all emergency priority levels as originally triaged.
4. Increase priority, up to the highest priority emergency type if waiting time exceeds a pre-specified length of time, for only specific categories (e.g. originally triaged category B patients have the potential for their condition to deteriorate over time, meaning their chance of survival decreases more than less critical emergency types).
5. Sample a maximum waiting limit from a distribution (e.g.  $\sim N(20,5)$ ). If the experienced waiting time is greater than this sampled value, increase the priority up to the highest priority emergency type (alternatively, increase only by a maximum of one level).

#### 7.6.6 Dispatch Method

When a vehicle is to be dispatched, expert EMS control room operators decide which vehicle is suitable; this is generally guided by a process or algorithm. In the simulation, this algorithm is less abstract and can be written down specifically; however, due to the nature of simulation models, the algorithm modelled is likely much more rigid than the actual operations in reality. This is one of

the downfalls of modelling, but still the logic coded can provide a good general approximation to the everyday decisions EMS operators face.

The dispatch method implemented in this project represents the decision of dispatching the closest, available and suitable EMS unit. The pseudo-code for the algorithm can be seen in Appendix 7.4c.

A vehicle is deemed:

- **available** if it is on-shift and either free, or returning to base after finishing servicing a previous call;
- **suitable** if it is of the required type to service the type of emergency given by the input rules of the model (e.g. RRVs only serve one category of call, as initial responders, unable to transport);
- and **closest** based on the travel distance between the vehicle's location and the demand node, based on the value stored in the Google Maps Journey Matrix.

Although in 2009, the official policy recommended an RRV as an initial responder to category A patients only (see earlier descriptions – Chapter 2 section 2.3.4 and Chapter 4 section 4.2.3 – on dispatching), vehicles are not always so strictly assigned to calls.

Considering only services where either an EA attends alone, or an RRV is dispatched as the initial responder with an EA follow-up, proportions are given in Table 7.4 of these service occasions per category. Category AS2 and AS3 are almost always served by a sole EA (or equivalent, e.g. HDU) and therefore are not included here. (All other vehicle combination possibilities and their proportions making up all 2009 services in South East Wales were given in Table 4.1.)

**Table 7.4** Service occurrences for single and double dispatch

Category:	A	B	C
1 EA Only	0.53	0.75	0.89
1 EA + 1 RRV	0.47	0.25	0.11

The data suggest RRVs are frequently dispatched to lower priority calls, and so, although policy suggests otherwise, the inclusion of these alternate dispatch rule proportions in the simulation is necessary.

During testing stages of the simulation build, it was found (through trial and error) that in fact, the proportions given by the data do not provide a simulated outcome matching the travel and response time distributions expected. The model's distribution profiles more closely match the historical data when RRVs are instead sent to category A, B and C calls (followed-up by an EA) on 60%, 30% and 15% of occasions respectively (as opposed to the data proportions of 0.47, 0.25 and 0.11). The reason these higher percentages improve the reliability of the model is that these numbers specify the number of times an RRV is attempted to be dispatched (assigned to a call); it does not illustrate the number of successful services by such a vehicle type as found from the data. The larger input proportions for dispatch policy account for times when an RRV is unavailable or an EAs arrives at the scene ahead of the RRV (which then might 'step down').

Dispatching a returning vehicle requires some further calculation. Since the definition of 'returning' in terms of the simulation means that the vehicle is en route back to its assigned base from some earlier call, its exact location is unknown. The starting location of the unit's journey (either a hospital or demand node) is known, and the length of time it has currently spent travelling back to base is also known. The vehicle's current location is taken to be whichever of the two locations (last service location and assigned base) is closest. From this, the travel time to the new incident location can be estimated.

A global setting in the simulation exists, so that interrupting vehicles on their return journey can be prohibited if desirable, preventing the need for this estimation of position and journey time, but increasing regional utilisation and service length, since vehicles will be considered 'busy' per call for longer.

### **7.6.7 Transportation Policy**

For the purposes of all experimentation in this study, patients requiring transportation are transferred to an ED at the nearest, open hospital facility to the scene of the incident. It is assumed that all hospital locations input to the simulation have the capacity and capability to deal with any patient type and will not turn patients, paramedics or EMS units away (except when specified during modelling of diversion tactic scenarios). With a simple addition to the simulation logic, it would be possible to include specialist units in the network and allow transportation of only certain patient categories to some facilities.

## 7.7 Sampling Methods

Simulation has the capacity to include levels of uncertainty in its processes, allowing for possible stochastic as well as deterministic modelling. The importance of incorporating this stochastic nature in healthcare is demonstrated with examples in Simul8 (Harper 2013, Simul8 2013).

There are multiple ways to model randomness, through data streams (obtained for example, from WAST's historical data – such as *loading* an incident schedule), or by user-defined distributions (such as *generating* demand from the stored time-dependent demand distribution) or through standard statistical distributions (as utilised for turnaround and service length).

The dangers of stochastic simulation are the issues of randomness, correlation and sampling errors. It is vital that variation between runs is unbiased. True randomness is unobtainable in computing, since a mathematically generated stream from which to read a number is required. For this reason, sampling uses pseudo-random number sequences – where an algorithm produces seeming random numbers. The numbers are predictable, but a good sequence will pass the randomness tests of uniformity and independence (correlation should be in-determinant).

By using a number stream with a specified starting point or 'seed' (as opposed to an automated one), simulation scenarios may be replicated with the same sequence of random numbers, enabling fair comparison between trials. This allows system changes to be accountable for identified differences in results rather than simply stochastic variation in sampled numbers. To introduce variability between trials with the same conditions, different seeds may be used. The particular pseudo-random number sampling method implemented in this project is the 'Mersenne Twister' (Matsumoto and Nishimura 1998).

One or more random numbers are generated and transformed into a value sampled from the required distribution, usually via top-hat sampling for discrete distributions (similar to look-up tables but of the CDF (Morgan 1984)) or analytical transformation for continuous distributions (such as inversion and rejection methods using the PDF). In this simulation a single pseudo-random number stream is used throughout a trial. Whenever a new random number is generated, the algorithm begins at the last finishing point of the stream to obtain the next value. This continuity maintains the integrity of the sampling and minimises the chance of dependence between generated numbers.

## 7.8 Model Validation

### 7.8.1 Introduction

Model validation ranges from testing assumptions about data and input, to testing that the structure responds similarly to reality. It is something that is not always possible but should always be attempted, in order to *"give credence to results [...] and instil confidence in extrapolations beyond the range of model experience"* (Fishman and Kiviat 1967). A model is simply either valid, or not valid (useful or not useful), but the validation process itself cannot provide such a conclusive answer. It is a version of proof by contradiction – by attempting to prove the model is wrong in some way, when results show it is similar to reality, or expectation, then confidence in the model grows. Testing continues until sufficient evidence and confidence exists in the model's accuracy, leading to the assumption of validity.

Graphical validation makes use of scatter plots to certify that the *"occurrence of random events is truly random"* (Robinson 1994). That is, one random event plotted against the previous random event should have no pattern or relationship, suggesting a good sampling technique is implemented.

Scope and level also need to be validated since precision and accuracy cannot be considered as the same thing. Too much detail does not result in more accuracy of the model compared with reality; the data may not be accurate enough to justify an increase in detail and build time may limit the detail included.

Verification and validation should be performed throughout the modelling process – as the design, build and testing of the model is iterative (Figure 7.3) so should the procedures for ensuring reliability and credibility.

### 7.8.2 Warm-up Period

A 'warm-up' is a tool for avoiding error. The usual necessity of a warm-up (or 'run in' or 'truncation') is due to the fact that the expected mean from a single run is not an unbiased estimator for the population mean if the simulation does not start in steady state. A second motive is that independence of the solution is required; effects on the results of a non-steady state system may

influence the outcomes, especially if the system starts 'empty and idle', where in reality it does not. A warm-up period reduces this initialisation bias.

In this simulation study, warm-up translates to allowing the system to become stable, with a realistic number of patients in the system and busy vehicles, and where resource utilisation is at a natural level. Only the steady-state output is desirable for result interpretation. Since WAST operates 24 hours a day, every day of the year, there is no point at which the system switches off – it is a non-terminating simulation. The initial transient (initialisation bias) must be dealt with, via a warm-up period of determinable length, so that investigation is only made on the stable system.

Rejecting the addition of a warm-up is acceptable in certain cases; alternative approaches include:

- truncation – the discarding or deletion of a portion of collected data;
- initialisation – starting the system off in a realistic state by providing typical queue lengths and system process values given by data or judgement.

There are many studies that suggest ways to improve simulation output analysis, such as investigating methods for determining these warm-up periods (Mahajan and Ingalls 2004), steady-state (Alexopoulos 2006) or truncation points (White et al. 2000). By 2009, there were already 44 different methods for determining steady-state points; the five main categories of warm-up period determination methods are (Robinson 2007):

1. **Graphical** – such as time series inspection and Welch's method (Welch 1983), where the output of a KPI on completion of the run provides an asymptotic graph from the point of steady-state;
2. **Heuristic** – for example, MSER-5 is a heuristic method that minimises the mean squared error of the batch means (size five) output data;
3. **Statistical**;
4. **Initialisation bias**;
5. **Hybrid methods**.

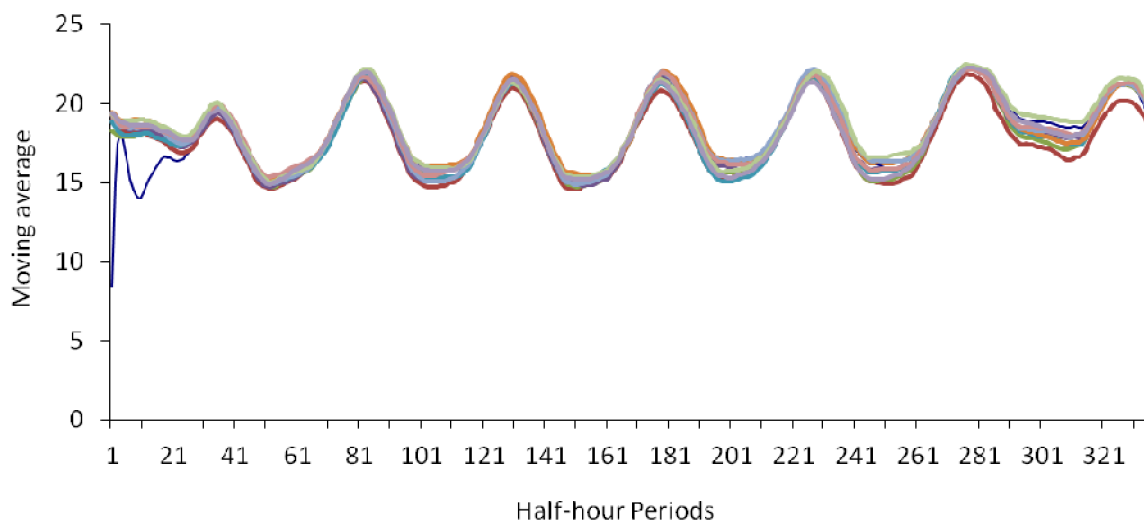
The simplest way of discovering a suitable warm-up period is by running the model until it reaches steady-state and studying the output. From this suitable point onwards during experimentation, responses can be collected and a range of accurate information obtained from the run. This time series inspection method, although simple and effective, is not necessarily a good approach when



demand is cyclic (or highly variable over time) – the point of steady-state is more difficult to spot in such systems, as in an EMS, where incident arrival rates are non-stationary. This is also an inefficient method when models take hours to run since the length of the warm-up period adds to the overall length of the run (although run-length is not a significant problem in this study). Overestimating warm-up, although increases stability in the output, is wasteful and information that could have been used for analysis is discarded, wasting computer power and run-time.

For this study, time-series inspection is assisted by determining batch means of the measured output statistics – average number of busy vehicles (by type) and average queue length for waiting patients, each per half hour time period. An experiment is conducted of the simulation, with a single long or ‘continued’ run (of ten weeks), replicated ten times. Different random number streams are used for each replication so that variation between runs is created.

Using Welch’s method, with batch window length of 15 periods (i.e. a moving average of 15 half-hour periods) and by splitting the corresponding results into weekly cycles, for this single run of the simulation, a trend can be seen despite the demand periodicity (Figure 7.13). A constant fleet is used (30 EAs and 5 RRVs), to avoid any unnecessary additional fluctuation from number of available resources on shift at different times of the week.



**Figure 7.13** Average number of busy EAs (from 10 runs) per week

Even though the system begins empty, it begins at midnight on a Sunday, where demand is low, so full and extensive analysis of the warm-up period may not be necessary since this is a natural lull in

the system. By midnight Monday, a full cycle of demand will have occurred and a second natural lull occurs. It is therefore expected that a warm-up period of maximum a day will suffice; this recommendation is supported by the non-steady behaviour in the first portion of the curve of Figure 7.13, where an obvious difference in trend occurs for the first 30 half-hour periods (15 hours) for only the first week of the run.

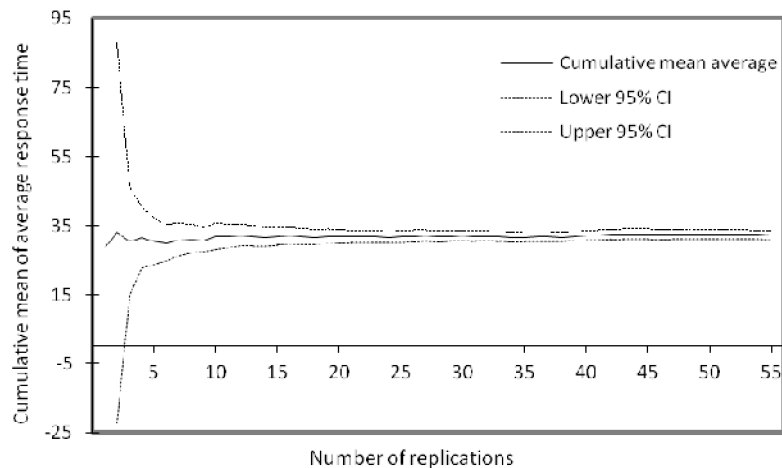
### 7.8.3 Run-Length and Replication Analysis

The decision for the length of run and number of replications to perform when experimenting is made simultaneously. A short run-length will usually require more replications to reduce the variation in the responses, and vice versa. In terms of reducing the deviation from the cumulative mean of the number of replications, it is considered better practice to have a longer run-length with fewer replications, than a short one (Mahajan and Ingalls 2004).

Run-length is a difficult simulation parameter to estimate. A rule of thumb for deciding on the length of run is to say it *"is sufficient when the most infrequent event has taken place on at least 10 to 20 occasions"* (Robinson 1994). Alternatively, if data are available, it is adequate enough to use a run-length similar to the sample data used or period covered by the historical data (Fishman and Kiviat 1967). For this simulation, the run length can be changed to be any length of time in minutes, plus a warm-up, for all desired number of replications; however, initial benchmark values are required before experimentation can begin.

Some test scenarios are performed to determine the two simulation parameters. The simulation responses used in the evaluation are the average response time and the average waiting times for both EA and RRV units. Again a constant fleet (30 EAs and 5 RRVs) is used in the scenario, with a single day warm-up period. The decision is then based on a result of less than 5% deviation of the 90% confidence interval from the mean.

Results indicate that experimentation would be adequate with a run length of six weeks and around 25-30 replications based on the stability and narrow range of the confidence intervals portrayed in Figure 7.14.



**Figure 7.14** Average response time for a trial with 55 replications, each with six weeks run-length

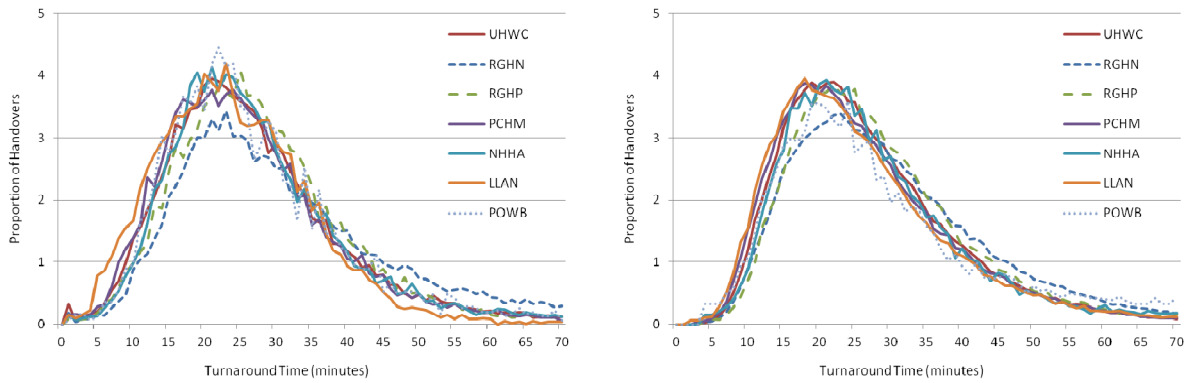
#### 7.8.4 Verification

Verification can be thought of as a micro-check of the simulation model and goes hand-in-hand with validation. Verification ranges from ensuring the pseudo-random number generator used does indeed generate independent pseudo-random variables, to verifying the behaviour of sub-structures of the model (Fishman and Kiviat 1967).

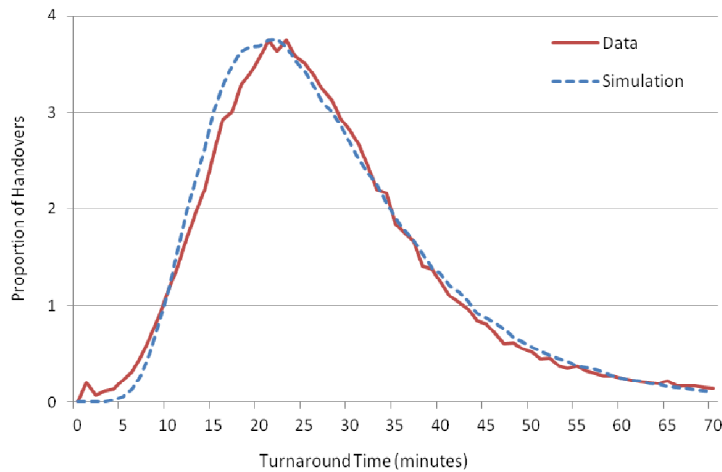
**Table 7.5** Verification approach and conclusion

Verification Tactic	Conclusion
Arrivals per category	Numbers as expected (output is not significantly different to data)
Total postcode district arrivals by hour	Distribution over time is not significantly different to data for 46 (out of 48) postcode districts
On scene length by category & vehicle type	No significant difference between simulated distributions and data (only category A analysed for RRVs), see Appendices 7.5a & 7.5b
Transportation time by category	No significant difference between simulated estimations and data distribution (using travel time models), see Appendix 7.5c
Turnaround time by hospital (& category)	Significant difference found (simulated output and data distributions are statistically different with lognormal sampling Figure 7.15, but for overall region are similar graphically, Figure 7.16 so are assumed suitably verified)

Deterministic checks are made to ensure that the numbers of patients entering the system are the same as the number added to the event list and the final list of the call log. Other similar authentication tests are made, listed in Table 7.5, to increase dependability and belief in the model design and set-up.



**Figure 7.15** Turnaround time distributions, per hospital, from data (left) compared with simulation results (right) with Lognormal sampling



**Figure 7.16** Turnaround time distribution for whole region, comparing data and simulated results with simulation sampling from a truncated Lognormal distribution per hospital

### 7.8.5 Validation

Much discussion surrounds the apparent lack of model validation in past EMS and other policy and operation-driven simulation literature, despite the wealth of validation techniques documented and methods provided generally (Finlay and Wilson 1987, Gass 1983, Goldberg et al. 1990, Green and Kolesar 1989, Kleijnen 1972).

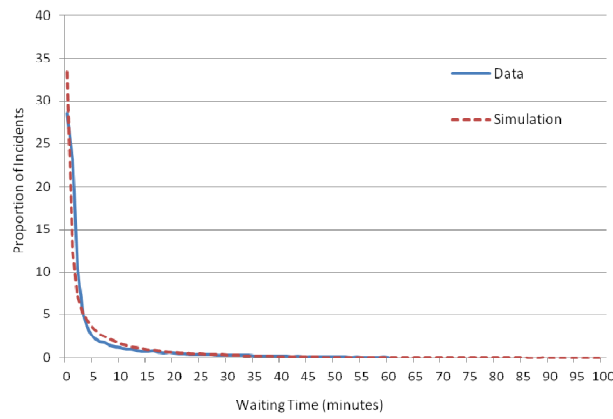
Validation, defined originally by Fishman and Kiviat (1967), is necessary since models typically make many assumptions to the design and structure of the system replicated. Three levels of validation are (Gass 1983, Robinson 1994):

1. **Face validity:** do decision makers and system experts agree the model has credibility?
2. **Replication or sensitivity validity:** do results change based on input variable and parameter changes;
3. **Prediction or hypothesis validity:** are the modelled system outcomes comparable with the real system?

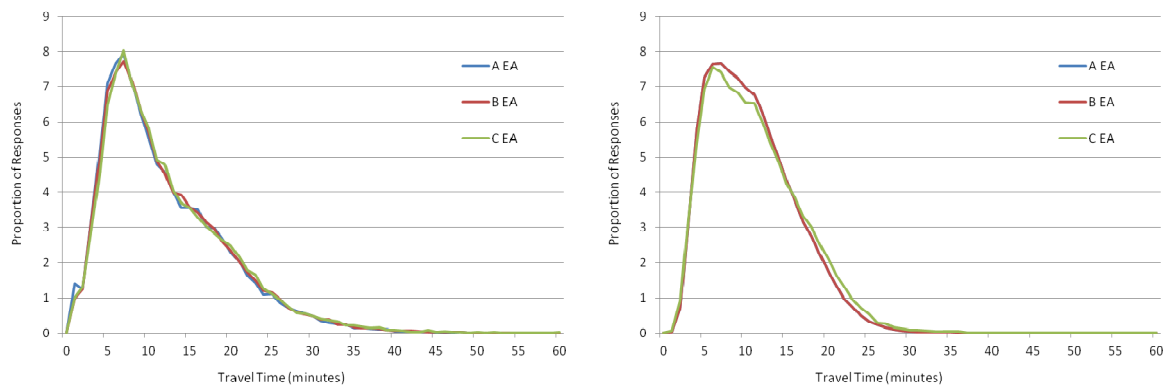
The verification and validation tests performed on this simulation model are detailed in Table 7.6, showing the conclusion of each individual investigation.

**Table 7.6** Validation approach and conclusion

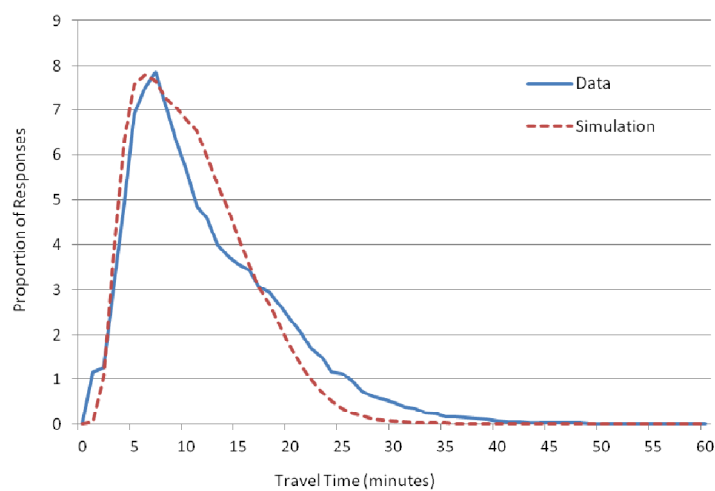
Validation Tactic	Conclusion
Waiting times of patients for response	Dependent upon fleet – benchmark fleet taken as average from WAST data. See Figure 7.17
Travel time by category & vehicle type	Significantly different simulated estimations compared with observed data distributions. Dependent upon scalar parameter used, fleet size and allocation. Figures 7.18 and 7.19 for best fit.
Response time by category & vehicle type	Significant difference between simulated response times and observed data distributions, Figure 7.20, but similar graphically, so assumed suitable.



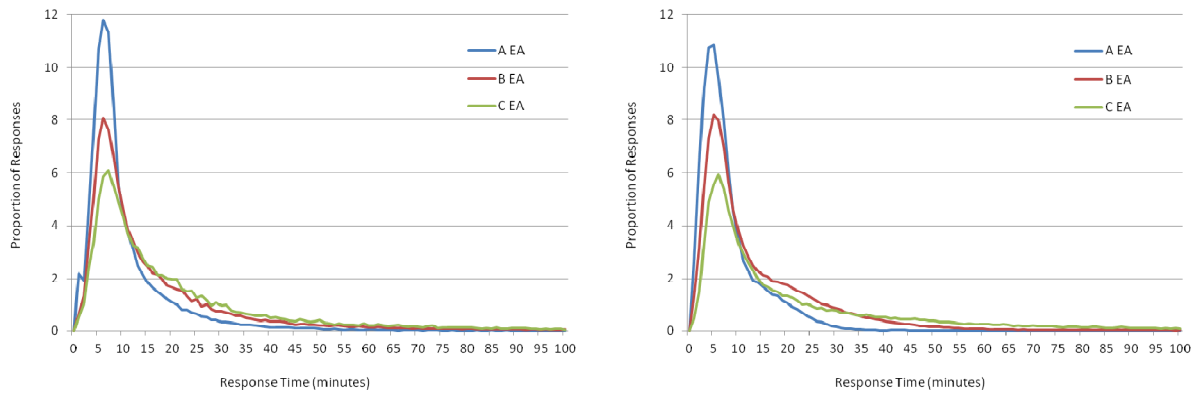
**Figure 7.17** Comparison of waiting time (including pre-travel delay) for combined category A, B and C for data distribution and simulation output



**Figure 7.18** Travel time to the scene from the data (left) and as results from the simulation (right)



**Figure 7.19** Comparison of travel time of EA vehicles to the scene of category A, B and C combined using average WAST fleet recommendation at 75% capacity and prediction method



**Figure 7.20** Comparison of response time from data distribution (left) to simulation output (right), using average WAST fleet recommendation at 75% capacity and travel times of Figure 7.18

## 7.9 Discussion and Extensions

Simulation can be used in many different forms. It is possible to use the approach to evaluate solutions of less precise modelling approaches, for example covering models, as is attempted in this project with regards to survival. In Chapter 8, this idea is utilised and the results from the survival covering models of Chapter 6 are implemented and compared to investigate the performance of the suggested allocations.

Simulation optimisation is the process of embedding the simulation in a search routine, whereby the best system solutions can be determined by testing different combinations of parameter values (Goldberg 2004). By extending the model's framework, this technique could be used to assist in the search routine for best fleet size, allocation and travel time estimation parameters.

With some adjustments to the code, it would also be possible to use the tool as a more real-time planning tool in the future, where decisions on relocation of vehicles could be tested (relatively quickly) to see the impact on service provided to the rest of the region, assisting control room operators in re-allocation decisions.

Since the tool was originally intended to be highly interactive during run-time – a Visual Interactive Model (VIM) – future endeavours would aim include such dynamic graphical aspects. 'Interactive' implies that a simulation run can be interrupted to allow the user to make changes to the decisions made by the model. The environment developed in this study, although not fully interactive, does

allow the user to watch the run in progress. If the ability existed to view the locations of vehicles or entities over time, it is likely buy-in of the tool and confidence in its replication of the real world would greatly improve; however, this level of detail significantly adds to design, build and run times. The current model, as it stands, takes steps in the direction of interactivity, offering more insight than many other models (particularly analytical types) but without the greater run-length.

There exists a belief that a model may never be fully validated (Holling 1978, Quade 1980), that strengths and weaknesses may be identified but exactness will not be proven. Nevertheless, validation is a crucial modelling aspect when simulating, and for the purposes of this thesis, the developed model is considered valid and verified based on the findings of sections 7.8.4 and 7.8.5. Despite some finding suggesting significant differences between the historical data and simulated outputs, based on graphical interpretation, the model seems suitably similar to reality, given the assumptions made and lack of knowledge of the real-world system for some procedures (dispatch algorithm and exact allocations).

One benefit of a simulation model is that the objectives of a study do not have to be specified before the model is run. Multiple investigations can be carried out simultaneously during a single experiment. The next chapter goes on to use the final model as described by this chapter in experimenting with various EMS scenarios of interest to the WAST. It also aims to validate the survival conclusions of Chapter 6, based on certain fleet allocations, showing the impact on the rest of the system, its performance and various other attributes.



## Chapter 8

# Simulation Results

### 8.1 Why Experiment?

Fundamentally, the strength of simulation lies in its ability to create a domain in which system managers can explore various scenarios of interest, investigating the impact of change on (multiple) specific aspects such as throughput, performance, efficiency and profit. For an EMS system simulation, many attributes can be changed; those that are of direct interest to WAST (and to academic research) are investigated in this chapter, including: service and dispatch policies, demand, transportation tactics, allocations and fleet size.

Large-scale assemblages of the public – festivals, concerts, sporting events – and disaster situations, undoubtedly place additional strain on any emergency service, so much so, that it is in WAST's interest to have strategies in place for such increased and likely localised emergency occurrences.

Additionally, WAST are moving towards a system that operates with a clinical outcome based performance measure, as opposed to the usual response-target driven system. Although these changes have already been implemented in some parts of England (London Ambulance Service © 2013), and are likely to develop across the rest of the UK in the next few years, it is important to understand fully the implications of such a change to policy. Demonstration is a powerful motivator, such that if improvements can be quantified, the strategy merits justification and other Trusts may also recognise the need for swift change in structure.

### 8.2 Model Set-up

#### 8.2.1 Introduction

Since operational vehicle allocations are difficult to decipher from historical data, it is not possible to compare simulation results with exact real-world WAST operations. Although the data provides the majority of simulation model input, without a known fleet capacity or its arrangement at any given time, the simulated results will not truly represent 2009 operations and so the performance is

not directly comparable. Instead, a benchmark scenario is established – where the operations match as closely as possible to what is known of the real South East Wales EMS system, and further modelling experiment results can be compared to this standard. The model setup described in the following sections aims to detail the input used for this benchmark scenario; factors that can later be altered for experimental scenarios will be indicated where existent.

### **8.2.2 Data**

The simulation is initially conducted with a single week's worth of data as a benchmark for further testing. When selecting this week of information from the historical data, desirable criteria include 'typical' system profiles, no major public events or holidays falling within the period, occurrence within typical school term time and avoidance of periods with the potential for extreme weather conditions. For this study, the 'typical' week chosen is therefore Sunday 10<sup>th</sup> to Saturday 16<sup>th</sup> May 2009 (as in Chapter 6), witnessing a total of 3041 unique incident records.

With the exception of the demand profile (and vehicle allocations – explained in section 8.3), all other distributions and variable parameters refer to the whole yearly averages, as analysed in Chapter 4. This enables typical variation, already investigated for different phases of service, to be incorporated into the simulation as opposed to week specific occurrences. Demand however, varies by hour, day and season; therefore, a snapshot of this variation is captured by using a single week's worth of data for experimentation. Alternatively, it is possible to model a full year's demand profile (or any time period) by inputting a schedule; although, this is deemed unnecessary since it is unlikely evaluation and planning for a whole year would be conducted in a single investigative instance.

### **8.2.3 Run Options**

Validated by the conclusions of Chapter 7, the chosen initial high-level model settings for use during experimentation are given in Table 8.1. This benchmark trial takes approximately 6 minutes in total to run, suggesting one replication every 12 seconds.

**Table 8.1** Benchmark model run-time options

Model Option	Decision Setting
Run-length	60480 minutes (6 weeks)
Warm-up period	1440 minutes (1 day)
Replications	30

#### 8.2.4 Parameters and Variable Values

In addition to the high-level settings, model options are available (that can be loaded automatically or changed manually by the user) that relate to chosen distributions, variable values and logic decisions for simulation.

The subsequent list, given in Table 8.2, constitutes what are known as the global parameter options and the values taken and distributions sampled from during the benchmark scenario experimentation process – later labelled as scenario 0.

Additional benchmark scenario values exist that do not refer to global model settings but to values specifically relating to incidents dependent upon emergency types (Table 8.3), the serving stations (also Table 8.3) and hospital transfers (Table 8.4). The variable values found in the following tables were originally defined as simulation model options in Chapter 7 (Table 7.2), but some of which are also elaborated on elsewhere in the thesis.

**Table 8.2** Global model options initialised for benchmark scenario

Global Option	Value Taken	Additional Details
Total Weekly Demand, $X$	$X \sim N(3150,140)$	Expected weekly demand for region (all categories)
Interruptible?	False	Are vehicles interruptible during a response journey to service a higher priority call?
EA Response Scale	0.36	Value to scale response journeys by (from a transportation journey estimate)
RRV Response Scale	0.25	Value to scale response journeys by (from a transportation journey estimate)
Non-urgent Scale	1.1	Value to scale non-urgent (return) journeys by (from a transportation journey estimate)
Travel Information	Predicted	Whether the model obtains a journey time value from Google Maps directly, or through prediction methods (models given in Chapter 7)
RRV Dispatch Policy (see also Section 7.6.6)	A: 60% B: 30% C: 15%	Are RRVs able to attend more than one category, and what proportion of each?
Maximum Wait?	False	Is there an implemented limit on the maximum length of time a patient waits before priority increases?
Maximum Wait Limit	n/a	If Maximum Wait is true, what is this waiting limit?
Transport Policy	Closest	Rule for transporting patients – which facility?
Shifts	21	Number of shifts per week (3 per day)

**Table 8.3** Data and input option values for benchmark scenario given emergency category

Variable/Parameter	Benchmark Option	A	B	C	AS2	AS3
Pre-travel delay EA	Exponential distribution sampling with mean:	0.668	0.660	0.675	0.661	0.756
Pre-travel delay RRV		0.508	0.660	0.675	•	•
Queueing Policy (section 7.6.5)	Priority (default)	1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>	4 <sup>th</sup>	5 <sup>th</sup>
Demand	Proportion of total demand	32%	35%	15%	15%	3%
On-scene EA	Distribution of on-scene length with	(20.7,12.8)	(19.4,14.3)	(20.4,14.5)	(18.5,12.4)	(15.1,14.8)
On-Scene RRV	$L \sim LN(\mu, \sigma)$ :	(31.3,20.7)	(20.4,15.9)	(19.4,14.4)	•	•
Transportation	Proportion of patients requiring transport	79%	72%	69%	91%	46%

**Table 8.4** Data and input option values for benchmark scenario by hospital

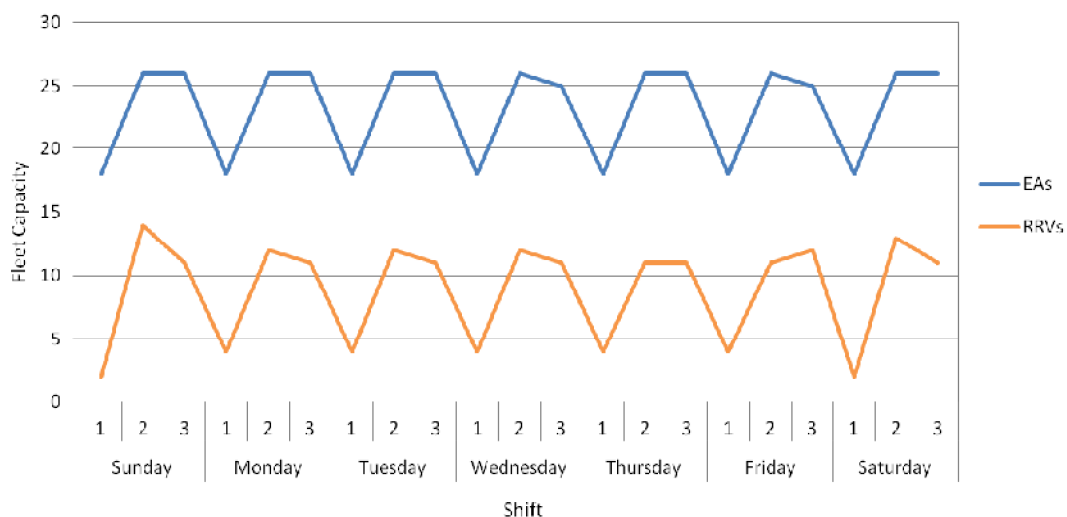
ID	Hospital	Truncated Lognormal Turnaround Distribution (to 70 minutes)
1	UHWC	$X \sim LN(27.8, 12.9)$
2	RGHN	$X \sim LN(31.4, 15.6)$
3	RGHP	$X \sim LN(29.2, 12.7)$
4	PCHM	$X \sim LN(27.4, 13.2)$
5	NHHA	$X \sim LN(27.5, 12.1)$
6	LLAN	$X \sim LN(26.8, 13.3)$
7	POWB	$X \sim LN(27.4, 12.5)$

### 8.3 Fleet Allocations

In the validation section (7.8.5), a constant fleet is commonly used in order to minimise the variation experienced from vehicles going on and off shift at numerous points in the week, so that the factor under scrutiny can be better analysed. To validate response and waiting time however, and during experimentation, the closest operations to the real-world system are desired.

Mentioned as a guideline in section 4.6 (Chapter 4), the WAST allocations should be scaled to approximately 70% of the fleet size to accommodate the inflated capacity estimates. Upon inspection of the created simulation model, the validation process in fact determined the best operational capacity to be a scaled version of WAST's available fleet with 67% EA and 50% RRV capacities. These values are supported in general by a comment made by WAST employees during discussions – that the Trust on the whole have around 50% total spare fleet capacity – assuming the data provided accounts for all vehicles owned (but not necessarily operational) by the Trust.

From the location models developed in Chapter 6, and the shift patterns detailed in Table 6.2, a benchmark allocation of vehicles to the stations in the region is obtained and used as fleet starting positions, input to the simulation model. The scaled EA and RRV fleets are reflected in Figure 8.1 for the vehicle types by shift of the week (as detailed initially in Chapter 6, section 6.8.5).



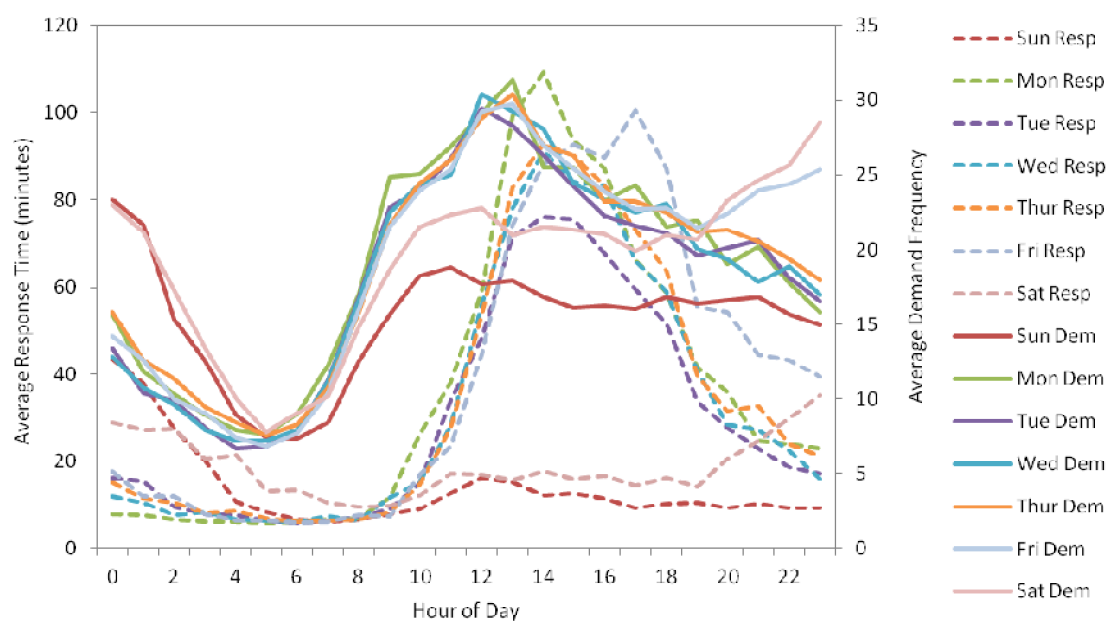
**Figure 8.1** Benchmark allocation of operational vehicles based on average available WAST fleet per shift, with 67% EA and 50% RRV scalar

## 8.4 Simulation Scenarios

### 8.4.1 Benchmark Scenario Results

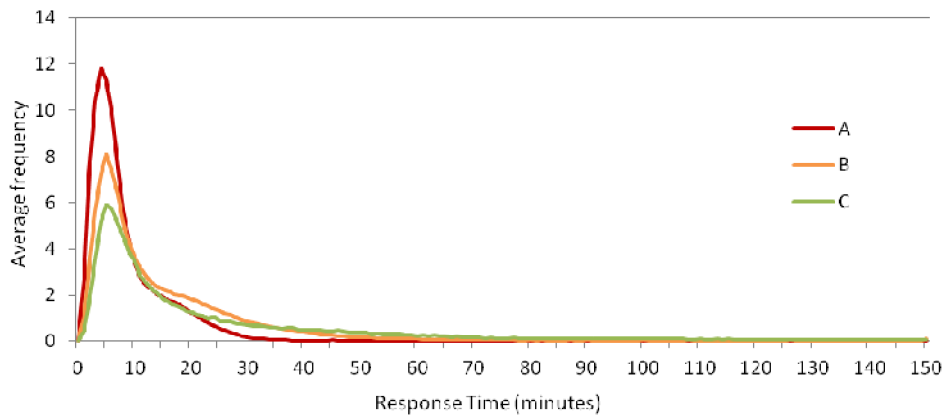
Some results are now shown for the benchmark investigation, scenario 0, with a further, more concise summary of its results depicted later in Table 8.6. For all other experimental scenarios, the same summary analyses are performed following run completion. A developed spreadsheet tool (created in Microsoft Excel and formulated around a VBA program) expedites the process of post-simulation analysis to enable effortless replication for the end user.

Initial curiosity surrounds the output average response time in the region relative to the level of demand across each day. Overall average response time reaches extreme levels shortly after a peak in demand (following the demand profile input from Chapter 4, Figure 4.10), fitting with the larger waiting times experienced by patients requiring emergency assistance when the system is already busy. Interestingly, the average weekend response times are relatively low compared with demand, showing saturation in the weekday demand but a manageable level during the weekend day shift.



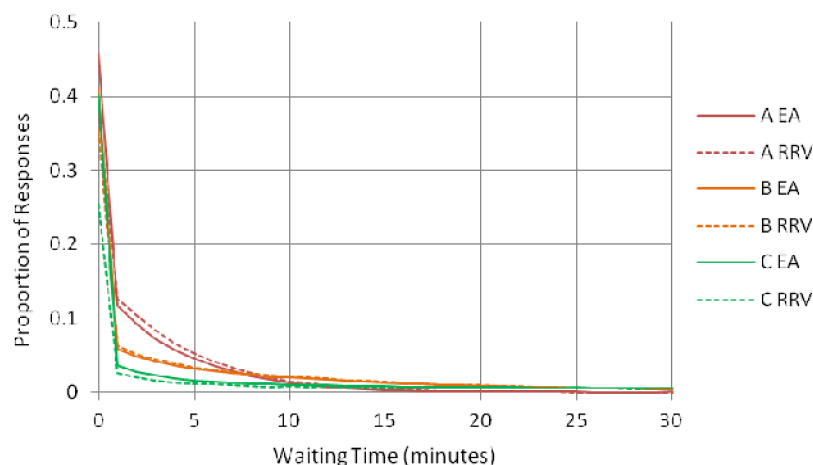
**Figure 8.2** Average demand and response for benchmark scenario trial

The overall category A, B and C response distributions are displayed in Figure 8.3, highlighting heavy and problematic service phase tails of the simulated system. In reality, such outliers are contributors to the high average response time peaks seen between 2pm and 4pm on weekdays in the graph of Figure 8.2.



**Figure 8.3** Average response time frequency, per category, for benchmark trial

As explained previously, response time is made up of two main components – waiting time and travel time. Since travel time for the benchmark scenario is estimated via Google Maps distance and regression models, and validated to be similar to the distributions found in the data, the waiting time phase is of more interest during experimentation – as this aspect is affecting response distribution characteristics.

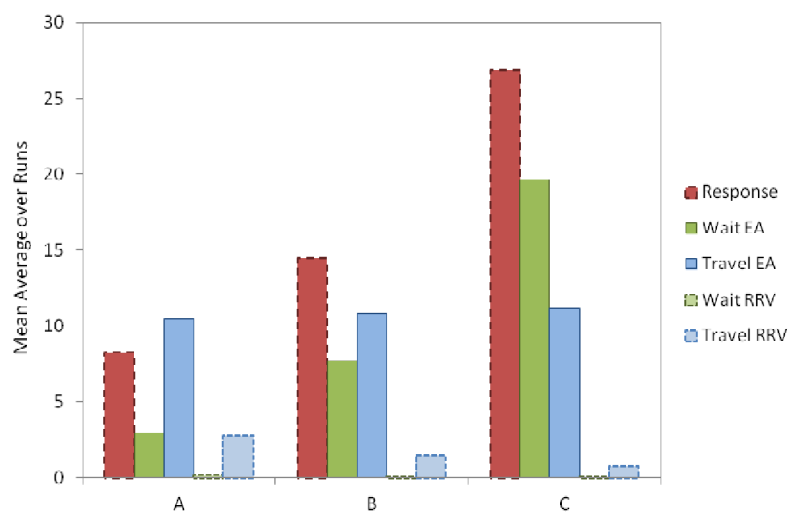


**Figure 8.4** Average waiting time, per category, for benchmark trial



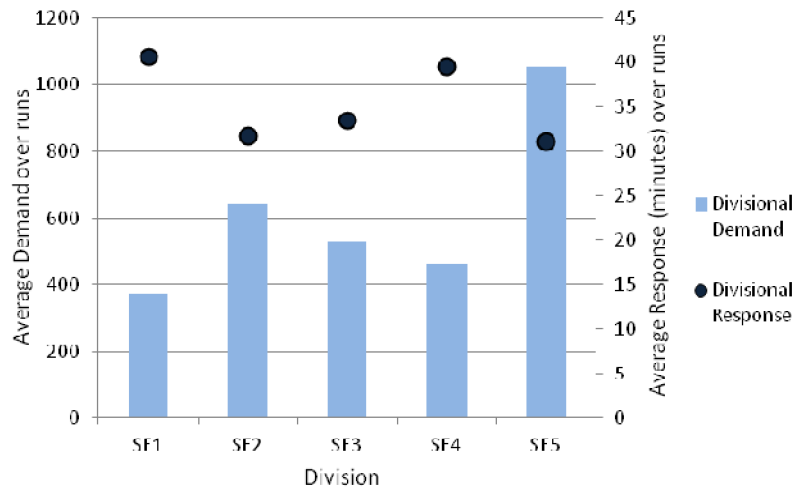
Figure 8.4 shows the distribution of waiting times experienced by patients of type A, B and C, when served by each vehicle type. The overall shape of this distribution was validated in Chapter 7, but proves here to represent fairly small waiting times for all individual emergency incident types.

Gaining further insight to the contribution of waiting time to the overall average response time, in comparison to the actual travel time of the initial responder comes from Figure 8.5. The average response time and wait time for an EA dispatch both increase somewhat linearly as incident priority decreases, whilst average travel time for an EA remains constant (as expected). However, wait time and travel time for an RRV decrease with a decrease in priority; this is attributed to the prevalence in which such vehicle types are required.



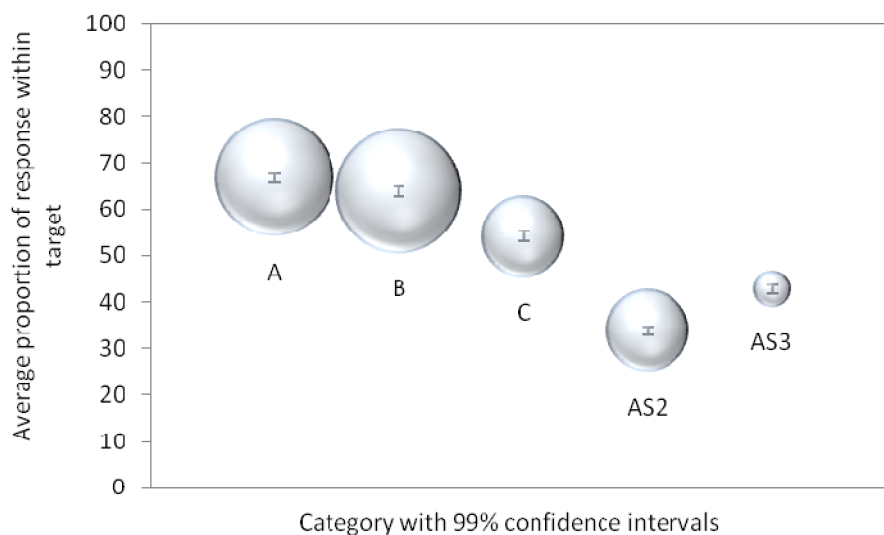
**Figure 8.5** Average response phase summary statistics, per category, for benchmark trial

Similarly to Figure 8.2, average response time of a trial compared with demand can again be portrayed – Figure 8.6 – but now with respect to the division of incident location. The key feature of this graph is the difference in ratios of response. That is, for SE2 and SE5, the difference in average response time is fairly small, whereas demand is around two thirds higher in SE5 (Cardiff); however SE3 and SE4 have a relatively small difference in average demand, but average response time differs by more than five minutes. Both these comparisons show how response time is not only a factor of demography but also of travel and geography.



**Figure 8.6** Average over runs of divisional demand and response time, for benchmark trial

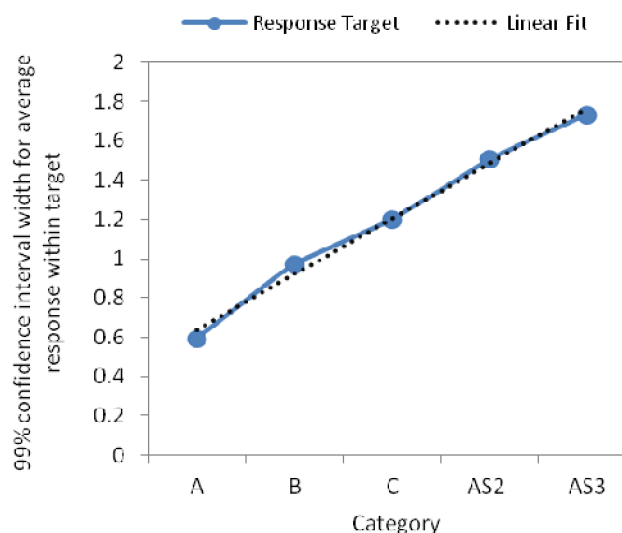
Of more interest to an EMS than the average response time, is in fact the average proportion of responses met within the individual category target times. For this reason, Figure 8.7 collates information in relation to expected demand volume per category and the confidence in proportion of responses meeting the target for the benchmark scenario trial. From the 99% confidence intervals, it is possible to witness the small spread of results for all categories over simulation runs. The size of the bubble in the graph represents the proportion of regional demand occurring throughout the system attributed to each category.



**Figure 8.7** Proportion of demand and average proportion of within-target responses, per category, for benchmark trial

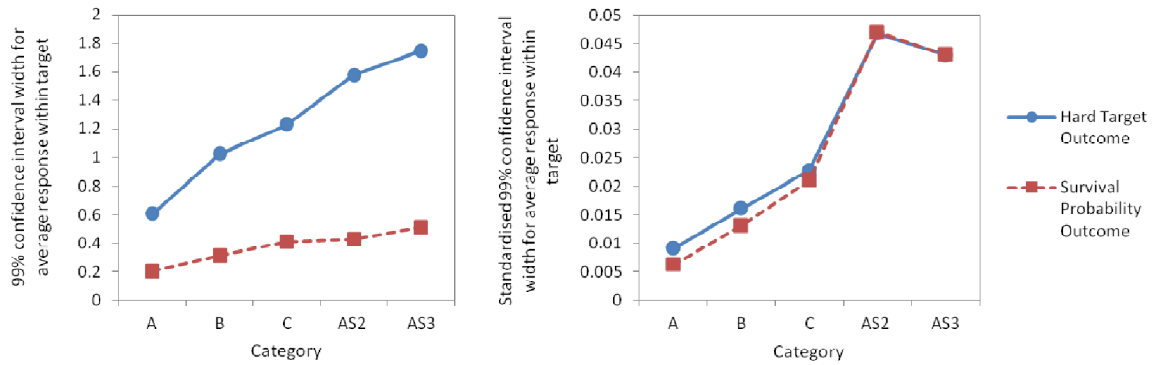
Little difference lies in the average response times for categories A and B, with a more noticeable reduction when incident priority reaches classification C. The reason for the lower priority AS3 calls being responded to within the target time more successfully than AS2 calls (which have the same target) can be put down to the fact that calls for service of AS3 incidents are infrequent and much less common than AS2 calls. Although outliers may be larger during peak hours, often requests for service originate from hospitals, meaning responses can credibly be very short. All five categories demonstrate a small confidence interval width at the 0.01 significance level.

Although the confidence interval widths are small across emergency types, it is interesting to see that in fact the intervals increase almost linearly with respect to priority (Figure 8.8). That is, the spread of response data in the system is linearly dependent upon the queueing structure implemented for servicing emergency calls.



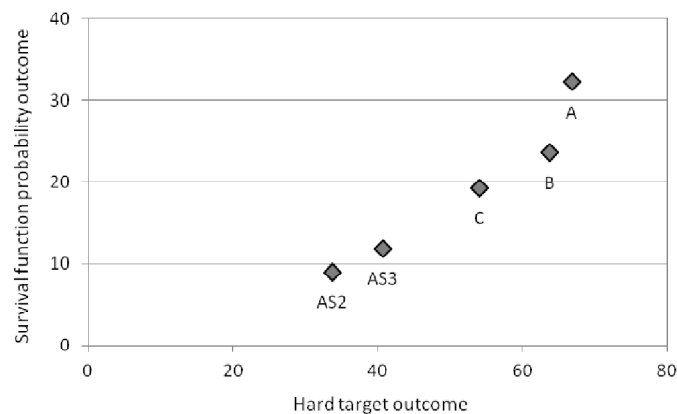
**Figure 8.8** The 99% confidence interval half-widths (two-sided 0.01 significance level) for emergency categories for average proportion of responses met within corresponding target time

Similar confidence interval width line graphs may be produced for both average hard target patient outcome (equivalent to response targets) and average survival probability outcome with the same trends as seen in the Figure 8.7. A mathematical idiosyncrasy appears when standardised – the half-widths produce an almost identical trend (confidence) over emergency types for the two different performance approaches (Figure 8.9), so that even though the intervals increase as priority decreases (due to larger variation in waiting and therefore response times) they are always small.



**Figure 8.9** The 99% confidence interval half-widths compared with standardised half-widths of average patient outcome over runs for total hard target survival and total survival probability

The scatter plot of Figure 8.10 shows that there is a non-monotonic, non-linear trend between survival and hard outcomes by priority. Patient outcome for both types increases as the critical nature of the incident increases for A, B and C incidents due to the order in which patients are served and responses given, but the sequence is lost for AS2 and AS3 incidents. This is likely due to the very small number of AS3 calls witnessed often originates from a hospital facility, meaning any recently cleared vehicles at the same hospital will be able to respond almost immediately. Although the relationship may seem linear between the two approaches if priority is unordered, Figure 8.9 reassures that the spread of outcomes for each approach per category is still similar – the priority queueing structure has the same form of impact on responses whichever performance measure is implemented, despite differing in the ratio and number of positive outcomes.



**Figure 8.10** Scatter plot of average hard target outcome versus average survival function probability outcome for response time, for each emergency type of benchmark scenario trial

It is important to realise that lower priority patients are in fact at less risk of experiencing a worsening medical outcome (given a longer response period), their conditions are triaged as non-life threatening, and therefore the risk to survival is relative only to their waiting time, not their health.

#### **8.4.2 Experimental Scenarios: What if?**

The intention of a benchmark is so that other experimental scenarios have a baseline comparison from which improvements, or lack thereof, in system performance and system user outcome can be made. Table 8.5 details all the experiments performed on the South East Wales EMS system simulation set-up. Results and comparisons follow in sections 8.4.3 - 8.4.10.

Results of the experiments are comparable, as the same underlying set-up is used for all scenarios and all measurements are calculated consistently; where differences do exist, they are as such described.

The scenarios are identified by letter and number, grouping similar system set-ups and scenarios that act upon the same phase or operational policy of the system. Scenarios 1 a, b and c involve dispatch and response policy; scenarios 2 a, b and c affect demand; scenarios 3, 4, 5a and 5b examine the service phases of turnaround and transportation, i.e. affecting the service of patients who require hospital admission; the penultimate grouping – scenario identification 6 and 7 – utilise allocations obtained from the location models of Chapter 6 and compare the patient outcomes (and responses) with the current benchmark system. Finally, scenario 8 delves deeper into the capacity issues faced by WAST but will be explained in more detail later in section 8.4.10.

**Table 8.5** Scenario descriptions for all simulation experiments performed

Scenario	Scenario Name	Approach
0	Benchmark	Standard model with parameters and options as sections 8.2.4 RRVs attend 60% category A, 30% category B, 15% category C
1a	RRV to A	RRVs attend only category A calls, 60%
1b	RRV to all A	RRVs attend all category A calls only (as is described policy)
1c	Fixed Travel	Deterministic travel times; RRVs attend all category A calls only
2a	Increased demand	10% increase of entire regional demand
2b	Altered demand	20% demand increase for categories A, B and C, Saturday noon-midnight; 10% location specific increase for CF10
2c	Catastrophe	Increase CF10 Monday 1pm demand with an extra 200 (50%) calls
3	Diversion	Emergency admission refusal at UHWC (e.g. A&E department closes); diversion to next closest hospital for transfer of care
4	Transportation	10% reduction in transportation - category A from 79% to 71%, B from 72% to 65% and C from 69% to 62%
5a	Turnaround	Reduction of all turnaround distributions to 20 minute average at all hospital facilities (corrected standard deviation)
5b	Ideal Turnaround	Truncation of all sampled turnarounds to 20 minutes or less (resample if longer) using original distributions
6	Hard Allocation	MESLMHPHF allocation (optimised with hard targets) with fleet capacity equivalent to WAST capacity as used in benchmark ( <i>6a is comparable to benchmark scenario, 6b to scenario 1b and 6c to 1c</i> )
7	Survival Allocation	MESLMHPHF allocation (optimised with survival function) with fleet capacity equivalent to WAST capacity as used in benchmark ( <i>7a is comparable to benchmark scenario, 7b to scenario 1b and 7c to 1c</i> )
8	Capacity	Increases to fleet capacity for all shifts are made and allocations from MESLMHPHF are used to position the fleet

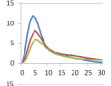
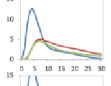
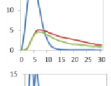
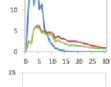
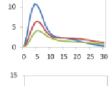
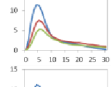
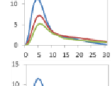
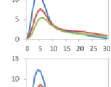
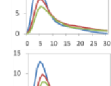
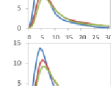
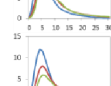
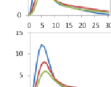
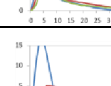
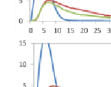
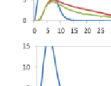
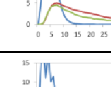
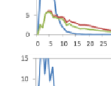
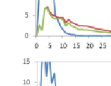
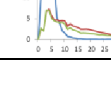
Handover of care at the hospitals is a forefront concern for the Trust. Since many ambulance hours are lost during this phase of service it is an area for investigation via simulation. Another main focus of the output results lies in response time performance as this is the common measure for EMS systems and comparable across the country; however, since some ambulance services, in particular WAST, are moving to clinical outcome based measures, it is also of interest to see if survival between different scenarios alters. Later, the comparison of survival and response performance measures will be made more explicitly, but for now, it is enough to realise that, for example, 60% target-met responses equates to 60% survival when using hard target measures; it is therefore simple and intuitive to directly compare response target performance with survival outcome.

### **8.4.3 Results Summary**

The main descriptive performance results obtained from the simulation for each of the executed experiments (compared to the benchmark) are given in Table 8.6, informing of the percentage increase or decrease in certain system aspects given operational and strategic changes.

Interestingly, all scenarios run are able to at least meet the 60% category A response time target of 8 minutes; however, the 95% targets for B and C calls are rarely met. It would appear that comparing to the data, the benchmark scenarios reach similar proportions of all category calls within the targets, confirming the simulation validation of Chapter 7.

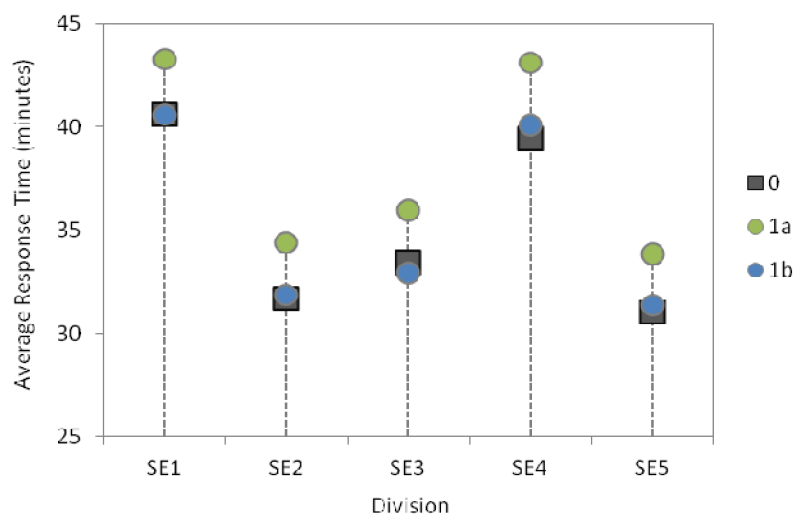
**Table 8.6** Experimentation results for scenarios 1-7 (percentage difference comparisons)

Scenario	Response Time Results (category)			Response Graph	Expected Survivors (survival function)			Turn-around in target	EA Utilisation	RRV Utilisation
	A	B	C		A	B	C			
<b>0</b>	<b>66.8%</b>	<b>63.8%</b>	<b>54.1%</b>		<b>32.3%</b>	<b>23.7%</b>	<b>19.3%</b>	<b>32.4%</b>	<b>86.3%</b>	<b>41.6%</b>
1a	1.3%	-15.5%	-8.5%		0.5%	-6.9%	-3.7%	-0.1%	0.0%	-14.1%
1b	24.5%	-14.3%	-7.3%		7.4%	-6.6%	-3.4%	0.1%	-0.4%	3.8%
1c	26.3%	-4.0%	3.5%		9.9%	-1.6%	2.0%	0.1%	-2.8%	3.6%
2a	-5.1%	-11.1%	-14.4%		-1.8%	-3.5%	-4.7%	0.1%	6.0%	5.2%
2b	-0.9%	-1.8%	-1.9%		-0.4%	-0.6%	-0.5%	0.0%	0.0%	0.4%
2c	-1.5%	-2.4%	-2.7%		-0.6%	-0.8%	-0.8%	0.1%	0.8%	0.7%
3	-1.2%	-3.0%	-3.5%		-0.5%	-1.0%	-1.2%	0.4%	1.5%	30.3%
4	2.1%	5.0%	6.1%		0.7%	1.5%	1.9%	-0.1%	-2.2%	-0.3%
5a	4.9%	12.0%	14.8%		1.7%	3.7%	4.7%	30.7%	-5.0%	-0.6%
5b	8.3%	19.4%	23.9%		2.7%	5.9%	7.6%	67.7%	-8.3%	-1.6%
6a	0.3%	-0.1%	-0.1%		0.0%	0.0%	0.0%	0.0%	-0.8%	0.4%
7a	0.7%	0.4%	0.6%		0.2%	0.2%	0.2%	0.0%	-0.8%	-0.1%
<b>1b</b>	<b>91.3%</b>	<b>49.5%</b>	<b>46.8%</b>		<b>39.7%</b>	<b>17.1%</b>	<b>16.0%</b>	<b>32.4%</b>	<b>86.3%</b>	<b>41.6%</b>
6b	0.2%	-0.8%	-0.8%		0.0%	-0.3%	-0.3%	-0.1%	-0.9%	4.3%
7b	0.3%	-0.6%	-0.6%		0.1%	-0.1%	-0.2%	0.0%	-0.8%	4.0%
<b>1c</b>	<b>93.1%</b>	<b>59.8%</b>	<b>57.6%</b>		<b>42.2%</b>	<b>22.1%</b>	<b>21.3%</b>	<b>32.5%</b>	<b>83.5%</b>	<b>45.2%</b>
6c	0.6%	0.9%	1.0%		0.1%	0.3%	0.3%	0.0%	-1.2%	0.0%
7c	0.8%	0.8%	0.9%		0.2%	0.5%	0.5%	0.0%	-1.3%	-0.4%



#### 8.4.4 Dispatch Policy Results

The key factor of interest in any simulation experiment investigating dispatch and service policy is the effect on response performance within the region. Restricting emergencies for which a vehicle type may serve, or increasing prevalence of dispatch of a sub-fleet, may impact the number of incidents met within target, and so subsequently the number of ‘survivors’. Figure 8.11 compares average response to demand within each division (defined in Chapter 4) from scenarios 1a and 1b (RRVs attend a proportion of category A calls) with the benchmark (where RRVs are dispatched to a set proportion of A, B and C calls). Similar graphs can be generated for survival outcomes.



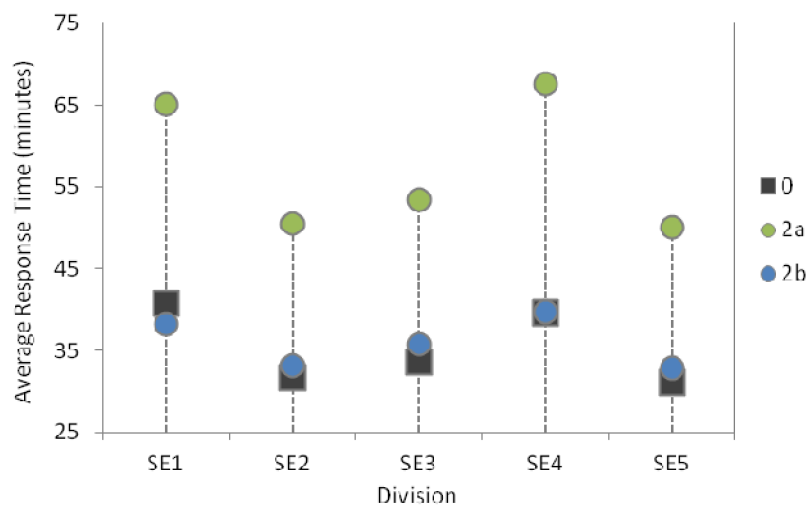
**Figure 8.11** Average overall response times per division, comparing scenarios 0, 1a and 1b

By sending RRVs to category A calls only (at the benchmark rate – scenario 1a) there appears to be little improvement in response across divisions; however with a category A only, automatic double-dispatch policy (RRVs always dispatched to A emergencies – scenario 1b), the response performance improves across all divisions. Reviewing Table 8.6, EA utilisation decreases by 0.4% with this policy, leading to an approximate three minute improvement in all divisional responses. Category A survival also increases by 7.4%, equating to an extra 75 ‘survivors’ in a typical week! Even the relatively small improvement of 0.5% category A survivors in scenario 1a compared to the benchmark refers to approximately five more survivors a week.

Scenario 1c differs somewhat from 1a and 1b, in that travel is assumed to be deterministic. The experiment’s purpose is to see the affect of allocation on response without the interference of travel variation. This comparison is explored later in conjunction with scenarios 6 and 7.

### 8.4.5 Demand Scenario Results

It is obvious that as demand increases on the service that performance may suffer. Although this conclusion is fairly intuitive, it may be that a small increase in demand leads to a large discrepancy in response when the system is already near its capacity. The system is beyond steady-state, which in this case is detrimental to patients entering the system. From Figure 8.12 and Table 8.6, it can be seen that comparing scenarios 2a (10% demand adjustment), 2b (location and time specific demand adjustment) and later 2c (catastrophe) with the benchmark, an increase in demand does indeed reduce response performance (similarly for patient survival). An increase of ten percent in demand across the region leads to a reduction of target-met responses between 5 and 15 percent. Divisional service is also affected, where response to the more populated and higher demand areas (SE5) has more consequence than some of the more rural regions (SE1 & SE4) in scenario 2b.



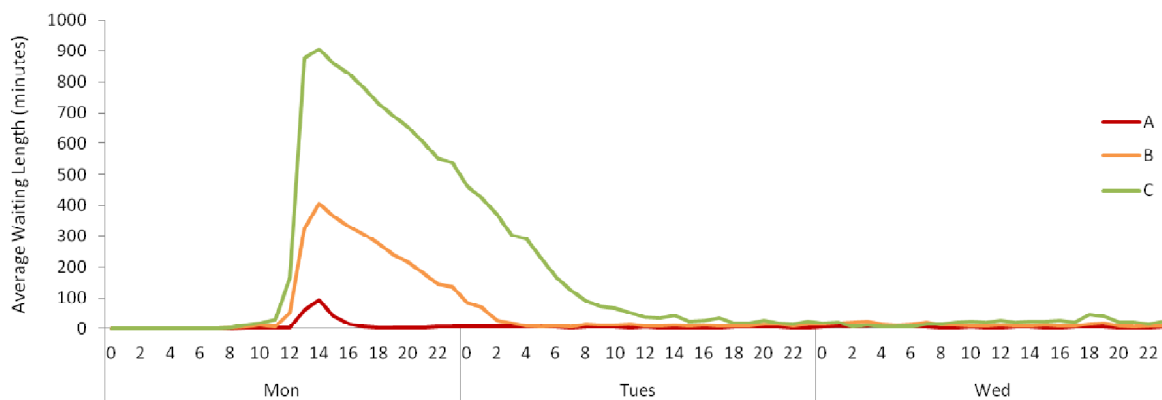
**Figure 8.12** Average overall response times per division, comparing scenarios 0, 2a and 2b

Large scale events, such as sporting fixtures, may increase widespread demand whilst simultaneously bringing a higher concentration of localised demand between specific hours. A great decrease in response performance is found when a major event occurs within the region, even if localised, compared with a smaller magnitude of increased widespread demand. This shows that the system can accommodate an increase in general demand more easily than it can spontaneously deal with a large localised incident.

### 8.4.6 Catastrophe Scenario Results

When a catastrophic event occurs in the region, as in the case of scenario 2c included in the analysis, the response performance deteriorates even further than the examples of small scale demand-increase of the previous section.

A simulation trial is run to demonstrate the impact of a dramatic increase in emergencies in one particular area (CF10, Cardiff centre) during a small period of time near the beginning of the run length (Monday lunchtime). This allows the consequences to be monitored through waiting time and response time graphs, as in Figure 8.13, and the knock-on effect within the region over time.

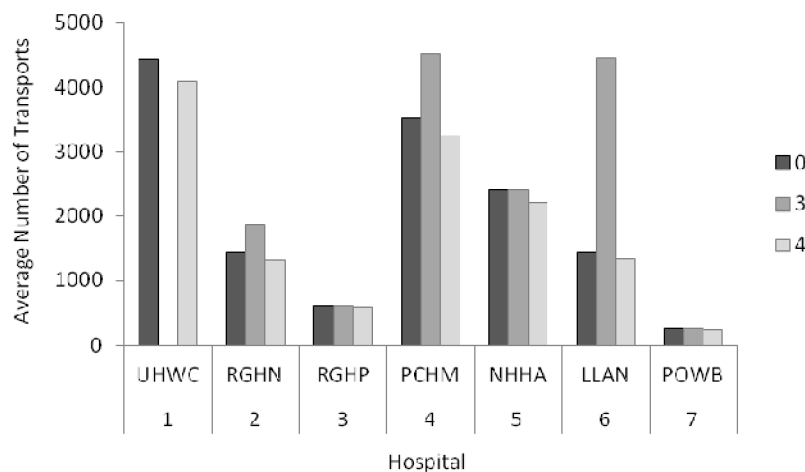


**Figure 8.13** Average waiting lengths per hour of patients for each emergency type following the occurrence of a catastrophic event, scenario 2c trial

As can be seen, it takes a considerable amount of time for the vehicle utilisation to return to normal, allowing service to resume as usual within the region, despite only an hour of increased demand. The peak waiting time for category C calls is around 900 minutes, meaning on average during the aftermath of the incident, these patients can expect to wait around 15 hours for service. Obviously, in reality, extra services would be utilised, and additional crews and vehicles deployed to account for the disaster, yet the simulation demonstrates how long the system takes to recover from such an incident.

### 8.4.7 Transportation Policy Results

Despite the closure of the University Hospital of Wales Cardiff (UHWC) A&E department in scenario 3 (such that all potential transports are diverted to the next closest hospital) leading to an increase in target-met turnarounds (Table 8.6), the overall response times per category are slightly worse than the benchmark due to the risk associate with diversion (Patel et al. 2006, Redelmeier et al. 1994). On the surface it could be thought that if turnaround times were lower in general, response times may also improve due to the positive knock-on effect on utilisation and vehicle availability; however, UHWC, accepts the largest proportion of emergency patients in the region and is centrally located for high demand areas, so, by sending patients to alternative facilities (mainly to the other Cardiff hospital and a neighbouring one in Merthyr, Figure 8.14), vehicles spend longer in the transportation phase of service, leading to overall increases in utilisation in both sub-fleets and waiting times of subsequent patients.

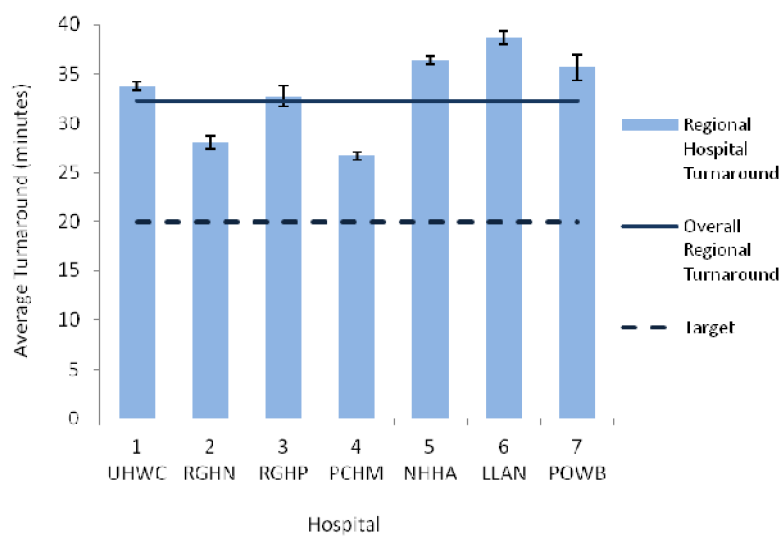


**Figure 8.14** Average transportation numbers to each hospital, comparing scenario 0, 3 and 4

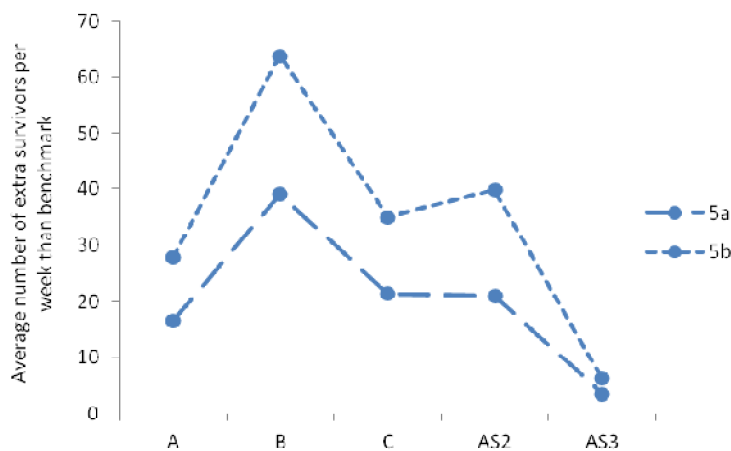
By reducing the proportion of patients actually transported, response performance can be seen to improve (scenario 4); however this policy would be better conducted with respect also to the dispatch policies. WAST and other UK Trusts are currently attempting to transform their response models by reducing conveyance along with the number of automatic double-dispatches (National Audit Office 2011), so that more patients are treated in the community by their initial responders, minimising the number of unnecessary care transfers. This approach would require consideration of training provided and specialist crews and so would be best investigated if further fleet information could be obtained from WAST.

### 8.4.8 Turnaround Time Results

Summary results regarding turnaround for scenarios 5a and 5b are shown in Table 8.6 but can be seen more coherently in Figure 8.15 and 8.16. Little variation surrounds the distribution at each hospital for handover, given by the input values; however, slight fluctuation between hospitals across the region exists following the data analysis investigations of Chapter 4 (Figure 4.23). If these discrepancies could be standardised (scenario 5a), or if turnarounds could in fact be reduced to a maximum of the 20 minute target (scenario 5b), then Figure 8.16 shows the dramatic impact of better vehicle utilisation on patient outcome compared with the benchmark scenario.



**Figure 8.15** Average turnaround time at regional hospitals with 99% confidence intervals

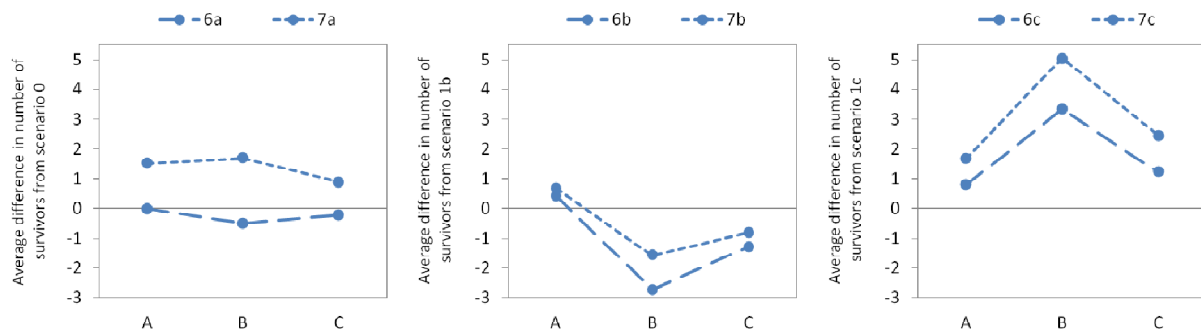


**Figure 8.16** Average extra number of survivors than benchmark scenario 0 per category, comparing scenarios 5a and 5b

### 8.4.9 Location Model Allocation Comparisons

For all scenarios 6, 7 and 8, allocations are taken from the Maximal Expected Survival Location Model with Heterogeneous Patient and Heterogeneous Fleet (MESLMHPHF) of Chapter 6, for hard target optimisation and separately for survival probability maximisation. The fleet capacity constraint is based on the average WAST fleet, per shift, scaling EAs to 67% and RRVs to 50% of the given size. Summary results may be seen in Table 8.6, but more detailed analysis of the survival outcomes per scenario, compared with the equivalent standard models, are now given.

Although it initially appears there is minimal improvement in using the location model allocations over the benchmark allocation, the crux of the matter is that allocating vehicles across a region with a survival maximising approach does indeed produce improved performance results over a hard target maximising approach. The comparisons of scenarios 6 and 7 with each other, rather than against the benchmark (for which it appears on the surface of this particular EMS system setup to perform worse – Table 8.6), validate the conclusions drawn in Chapter 6, where the location model designs themselves were tested.



**Figure 8.17** Comparison of average expected extra survival per category for scenarios 6 and 7 with equivalent benchmarks (equivalent dispatch and service policies)

Figure 8.17 shows that the scenarios 6 and 7 are almost always better than their equivalent WAST average allocation systems operating with the same policies. Even where little improvement exists (or results are worse for lower priority calls) on the equivalent benchmark (be it 0, 1b or 1c), scenario 7 (survival maximising) is always better, substantially, than scenario 6 (response target optimisation).

Although the location models optimise vehicle allocations, when this fleet set-up is put into a non-deterministic model, the stochastic nature means it may not always perform to its full potential when high variation in system phases exist. The third graph in Figure 8.17 – the comparison of 6c and 7c with benchmark 1c – demonstrates that when travel in the simulation model is fixed (no variation, deterministic travel times) – the allocations obtained by the location models are indeed superior to the one estimated from WAST’s data (supporting Chapter 6).

#### 8.4.10 Capacity Results

Scenario 8 aims to make comparisons between various increased fleet sizes and the benchmark simulation model setup. Capacity is increased in line with a given percentage for one or both vehicle type sub-fleets in turn. Three capacity combinations are explored, seen in Table 8.7; results with regards to response and survival are given in Table 8.8 alongside the original results of the benchmark comparison of scenario 0 shown earlier (Table 8.6).

**Table 8.7** Resulting scenario version from increased sub-fleet capacity combinations

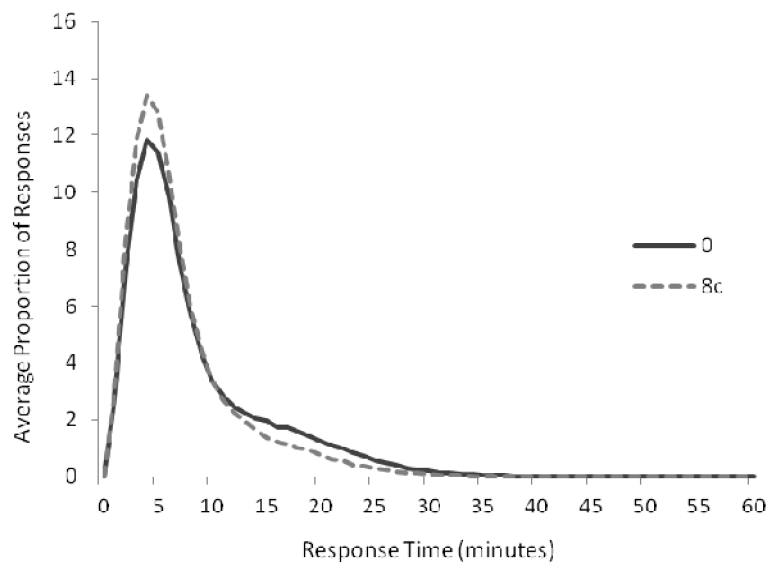
		EA	
		0%	10%
RRV	% Increase		
	0%	0	8a
	10%	8b	8c

As can be seen from the results, and as is expected, as fleet size increases so does the ability of the service to improve performance, whether it be with respect to response or clinical outcome based objectives. Scenario 8c is compared to the benchmark in Figure 8.18, showing the higher peak of the average response time distribution when the number of operational vehicles (per type) is increased by only 10%. This percentage corresponds to a maximum of only three additional vehicles per shift at peak times during the week, and on some occasions, a rounded integer increase in vehicles of 10% means that where the original number of vehicles on shift were less than 5, there will be no increase in the new scenario (i.e. during lower demand periods).

**Table 8.8** Experimentation results for scenario 8 (percentage difference comparisons)

Scenario	Response Time Results (category)			Expected Survivors (survival function)			EA Utilisation	RRV Utilisation
	A	B	C	A	B	C		
<b>0</b>	<b>66.8%</b>	<b>63.8%</b>	<b>54.1%</b>	<b>32.3%</b>	<b>23.7%</b>	<b>19.3%</b>	<b>86.3%</b>	<b>41.6%</b>
8a	6.7%	14.7%	17.8%	2.2%	4.6%	5.7%	-7.2%	-1.0%
8b	1.2%	0.8%	1.1%	0.5%	0.4%	0.4%	-1.2%	-3.6%
8c	7.4%	15.3%	18.6%	2.5%	4.8%	6.0%	-7.6%	-4.6%

A small increase in operational vehicles can have a large impact on overall expected number of survivors. By simply adding approximately two or three more EAs per shift (scenario 8a), a category A survival percentage increase of 2.2% is witnessed, which equates to more than twenty additional survivors per week.

**Figure 8.18** Average response time distributions for scenario 0 and 8c trials

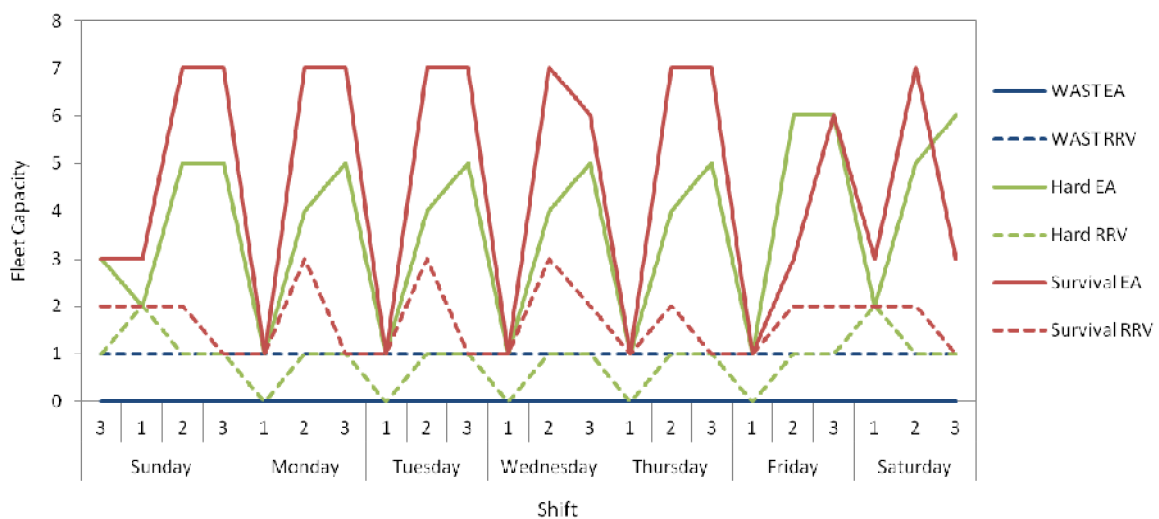


## 8.5 Conclusions

### 8.5.1 Allocation Insight

Some interesting points to note from the allocations used in the experimentation are that, firstly, the average WAST allocation utilised in the benchmark, positions only RRVs at station 1 (SABW – Blackweir in Cardiff) and never EAs; however, for both the survival and hard target allocation model approaches this is the best located and so most frequented station for both vehicle types, housing the largest proportion of vehicles for the shifts throughout a typical week (Figure 8.19). It would be possible to include station-specific capacity constraints within the location models in future work to enable such restrictions or even the preferences of WAST to be taken into account.

In the survival allocation scenarios (scenarios 7), only one station, station 23 (HQ) is not ever occupied by EAs; yet for the hard target approaches, all stations are used at some point during a week period for EA situation.



**Figure 8.19** Comparison of sub-fleet capacity at station 1 (SABW) for average WAST fleet and location model allocations per shift

One unanticipated discovery is that the hard target allocation scenario, where RRVs attend A, B and C calls (scenario 6a), does not produce markedly better results than the benchmark, which since scenario 0 attempts to reflect current WAST operations, suggests WAST are able to achieve better results without any mathematical optimisation assistance. However, the location models optimise

based on average service time and deterministic travel, not taking into account full service. It was shown in Chapter 6, that if the current WAST allocation was input to the same model, in fact the location models are indeed able to find better allocations. In the simulation, the optimised allocations may not enhance performance compared with the average WAST setup due to the impact of other service phases. For example, if the new allocations mean vehicles are positioned further from hospitals, the time they spend en route to and from facilities may increase, worsening utilisation, decreasing availability and subsequently affecting response. As mentioned, the location model allocations also exploit deterministic travel times, whereas the simulation typically runs based on predicted travel time, scaled by vehicle type. It is this variation that may therefore limit the impact of the allocation on patient survival (given a hard target approach); hence the inclusion of a contrasting set of location model allocations with a fixed travel benchmark (scenario 1c), demonstrating improved performance over the benchmark, as expected.

Particular interest lies not in the discrimination between policies, but within them – that is, the comparison between hard and survival approaches as opposed to location model allocations versus a true WAST allocation. Since the scaled average benchmark allocation is essentially an educated guess, the real daily operations of WAST are not known exactly. Although scenario 0 is not expected to be a fully accurate representation, its uncertainty may prove misleading if further experimental results are to be directly connected to reality; therefore, scenario 0 is used in place of true operations as a standard from which to measure impact of system changes.

A sensitivity analysis could be conducted, altering some of the allocations to see the impact on performance. A simple exploration in this vein showed results do differ, although only slightly, but further investigation should be sought after as validation before applying the model to other EMS systems. The slight difference in results is of course expected, since location is known to affect response, yet it might also indicate a lack of robustness in the model. For the purposes of this study, the models are deemed robust enough to take reliable solutions for the region investigated, even given slight sensitivity to input changes. This is combated through experimentation (where all other parameter values are kept constant), validation and multiple runs.

### **8.5.2 Diversion**

During the investigation of diversion (scenario 3), patients who would have been transported to UHWC are instead mainly sent to Llandough (both located in Cardiff), which in reality, would not always be a possible alternative tactic. University Hospital Llandough is much smaller and operates mainly outpatient clinics, being classified as a 'major acute', not a 'major accident and emergency' facility. Extensively increasing its emergency admissions in this manner would not necessarily be feasible. Further experimentation within the simulation could investigate transportation policy modifications; however, to do so, additional information regarding capacities of hospitals and their functionality would be essential.

Further work utilising the developed simulation tool could extend to the inclusion of interruptible lower priority calls (the structure for such operational control logic is already in place in the simulation program). At a recent conference (NISCHR February 2013), a Welsh paramedic commented that crews assigned to lower priority calls (category C and below) are at risk of being stood down in order to attend any higher priority calls. Currently, the simulation operates where only return journeys are interruptible but vehicles on response journeys may also be diverted in reality.

### **8.5.3 Payoff**

Culminating the investigations of this chapter leads to many recommendations and insights regarding the procedures of WAST for standard operations, forming the basis of discussion in the following chapter. Before going forwards, one valuable fact to adopt from the experimentation process is that, as hoped, it is not necessary to increase fleet capacity in furtherance of performance. A similar level of improvement can be made with simple policy changes, from hard response time targets, to instead locating from optimised patient survival probability. Not much of a difference lies between the proportion of survivors per category for scenario 8b (increasing the number of operational RRVs per shift by 10%) than for scenario 7a (survival approach allocation, with RRVs attending A, B and C calls); although there is a small discrepancy, the latter scenario comes at no additional cost to WAST, improves both response performance and expected clinical outcome whilst also reducing utilisation from the current setup.

## Chapter 9

# Conclusion

## 9.1 Discussion

### 9.1.1 Introduction

An array of evidence, throughout the literature sections and the data analysis of Chapter 4, suggests WAST are indeed struggling to meet current Government set (national) targets, despite improving on their own efficacy over a number of years. The data and publications prove that there is scope to progress, particularly in response and turnaround service phases.

The question posed by this research was whether the system's performance could be enhanced, complying with the demographic, geographic, resource and monetary constraints faced by the Trust, by simply suggesting ways WAST could better allocate vehicles across regions. In short, the answer to this is, yes. However, further enquiry emerges from the idea that response time is perhaps an inadequate way of evaluating such a public service (South Wales Argus 2013, Wankhade 2011) and that in fact turnaround time may be a more highly contributing factor to poor performance than immediately recognised (BBC News 2011a, Goldhill 2013, Knight and Harper 2012).

Operational Research facilitates the investigation process of such matters, not only in healthcare and EMS systems, but also in various public and private sector situations. This was made evident in Chapter 3 through the models, research and successful application of techniques to location problems. Common methodologies of Location Analysis, Queueing Theory and Simulation are employed harmoniously throughout this study. The following discussions aim to highlight the main successes of the techniques exploited and key findings of the research, culminating in an offering of recommendations to WAST or to any EMS hoping to gain insight of their service.

### 9.1.2 Objectives

At the outset, this investigation aspired to suggest better resource allocations across the South East of Wales in order to enhance utilisation and response performance whilst demonstrating the potential benefit gained from applied OR techniques. It intended to explore the proposed change to clinical outcome based performance measures across the UK and by looking at the system as a whole, see the impact of such policy alterations on current operations. An auxiliary goal of the investigation was to develop generic tools that could be utilised by EMS managers and analysts for future planning purposes within their own Trusts, so that region specific solutions can be discovered for nationwide EMS problems.

Referring back to the introduction of Chapter 1, the outcomes of the study address the presented objectives (section 1.2) and successfully provide system insight not only mathematically and academically but also to service managers, planners and users. In summary, the original objectives and resultant outcomes are:

- Investigate if improvements to WAST's performance can be made with regards to response and turnaround phases, whilst maintaining current capacity;
  - Improvements are shown with better allocation of the operational fleet;
  - Additionally, a reduction in the turnaround service phase improves utilisation, availability and response;
- Investigate current policy impact on patient survival;
  - Through the location theory models survival given a hard target performance measure approach is calculated;
  - Using the simulation, the expected number of survivors, given the benchmark system setup can be evaluated;
- Suggest ways in which to improve survival probability;
  - Better allocation also proves beneficial to patient outcome;
- Support WAST's move to clinical outcome based measures;
  - Advantages over hard target measures are found via the allocation models, and supported by the simulation outcomes;

- Develop generic tools that may be utilised by EMS managers for future planning purposes, in areas dealing with demand, fleet allocation and capacity;
  - Travel time matrix generator utilising the Google Maps API;
  - Two allocation models and the provision of four novel modelling approaches;
  - User-friendly simulation tool with full graphical interface.

Further insights that materialise throughout the course of this study are discussed in the following sections, relating the method of discovery with its findings.

### 9.1.3 Modelling Conclusions

#### *Travel Time Matrix Generator*

In order to model any network formulated around a road structure, travel times or distances between all points on the network must be known; if unknown, they must be estimated. Chapter 5 presents a tool utilising the Google Maps API providing the user two travel time calculation options for the subsequent EMS modelling approaches, either:

1. exploit the Google Maps calculated travel times directly;
2. or, utilise the Google Maps measured distances to better estimate travel time in conjunction with developed regression models for all journeys conducted by a heterogeneous EMS fleet.

The resulting matrices, of route distances and travel times, are used as input to both the location and simulation modelling endeavours. Such an extensive estimation process is conducted since the classic Euclidean and linear calculation methods explained in the literature are poor predictors of response time of emergency service vehicles, particularly in rural areas. Furthermore, the tool's generic structure allows it to be invoked easily and comparably by other EMS Trust regions.

#### *Location Models*

After providing a typical data structure to the newly developed mathematical programming location models (of Chapter 6), allocation optimisation of a fleet (whether it be homogeneous or heterogeneous) to maximise the expected patient survival probability is conducted for a particular

region – in this case, South East Wales. Two approaches (although extendable further) are integrated in each of the four model designs; firstly, the models are run whereby the system operates with hard targets measuring successful service of patients (as is current practice) and secondly where survival is calculated as a probability given a certain response time.

Drawing conclusions from the experiments within and between the four location models shows that a survival-maximising approach produces better overall results than hard target operational conduct for life-threatened patients. Those lower in risk witness a worse survival probability, but this does not necessarily lead to unfavourable outcomes in reality, just to the experience of slightly delayed responses in trade-off for more critical-patient lives being saved.

### ***Simulation***

Thirdly, since the location models are only able to explore the influence allocation has on response phases of the system, simulation is used in order to investigate consequences overall, and which other service phases impact performance. The model is also used to support the conclusions of the location models – that a survival-maximising approach is superior for patient outcome to a hard-target based approach when optimising operational fleet allocation. This verdict was suggested in the literature and supports prior findings. A survival founded allocation in the simulation produces consistently better results than a hard allocation, for all categories, not just high priority patients (as is the bias present in the location model comparisons). This more favourable outcome of a survival-maximising approach comes from the improved allocation impact on sub-fleet utilisation and response. Due to the deterministic nature of the location models, despite the attempts of the iterative versions, the full system stochasticity can only be captured through simulation.

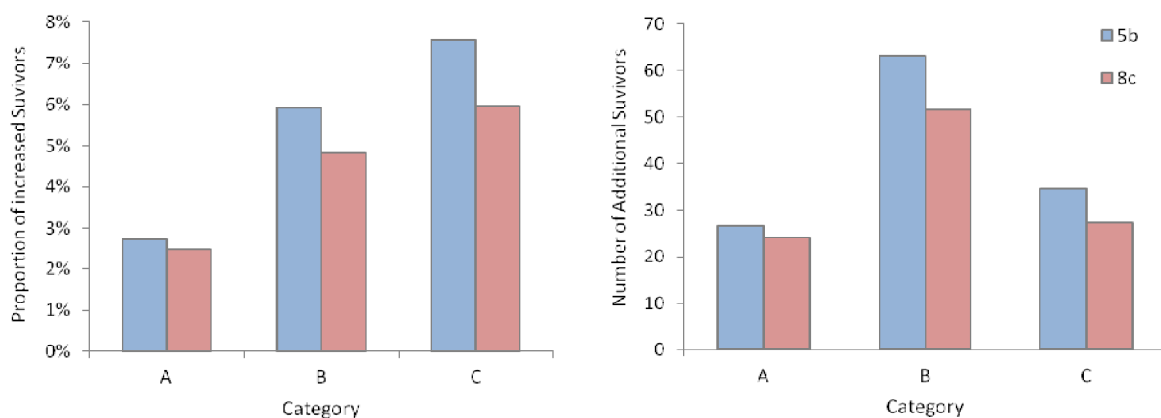
In addition to evaluating the conclusions of Chapter 6, the simulation experiments enable insight to the problematic phases of service. Experiencing an increase in regional demand (simulation scenarios 2a and 2b) is on the whole detrimental to the population as would maybe be speculated. The consequence of reducing turnaround time however (scenarios 5a and 5b), which is found from the data to be on average far above the recommended length, is extensively positive. Freeing up vehicles at hospital quicker, allowing them to respond to new calls earlier and reducing vehicle utilisation all lead to a substantial increase in survivors. Similarly, by treating more patients in the community, reducing proportion of conveyance per category (scenario 4), performance is improved for both response and survival measures.

### 9.1.4 Investigative Conclusions and Considerations

It is indeed possible, as initially hoped, to improve response, survival, patient service and performance within an EMS system. The tools provided by OR and applied by OR analysts can assist in the development process for individual EMS Trusts, utilising insight and designs gained from this and similar research.

Performance has been shown to improve with sophisticated fleet positioning and that optimising allocation with respect to patient outcome produces better EMS results than for response target led optimisation. The scale of these responses does vary with fleet size, but a good level of efficiency can be maintained without increasing capacity or number of operational vehicles per shift. Such a discovery contradicts the original notion voiced by WAST employees in section 2.3.7 that the ambulance service lacks resources.

The objective of this study has not been to recommend a solution of when or where to increase staff or operational crews, as this demands an increase in monetary investment; instead, the suggestions hold at their heart the understanding that performance can improve if the resources are just utilised more effectively, as hidden slack often exists in system capacity. It is demonstrated (Figure 9.1 and in Chapter 8) that indeed, an immediate solution would be to increase the number of operational vehicles by two or three crews per shift (simulation scenario 8c), but with just a reduction in turnaround time (scenario 5b), the performance can be improved to an even greater degree.



**Figure 9.1** Comparison of additional survivors (proportion and numerical) per category for a typical week, resulting from scenario 5a and 8c compared with benchmark scenario 0



If turnaround time could be improved, the overall performance of the system would benefit. Perhaps one of the most potentially influential findings of this study is the fact that turnaround time reduction in fact has more of an impact on system user survival than allocation. A small decrease in turnaround time at hospitals, relieving EMS resource blockages at this point in the system, reduces utilisation, increasing vehicle availability and so subsequently impacts response time. By simply improving turnaround time (although, pragmatically, this will not be an easy task), all other system phases benefit due to less strain on the resources, leading to better overall system conduct and so consequently a higher degree of favourable patient outcome.

As system managers know, removing strain from one aspect often leads to an increase in pressure elsewhere. In this case, a quick fix to relieve WAST's turnaround problem would shift the problem to Accident and Emergency (A&E) departments across the country. The EMS turnaround issues actually originate from the congestion problems within hospitals and the inability to transfer quickly the care of the patient from paramedics to A&E doctors.

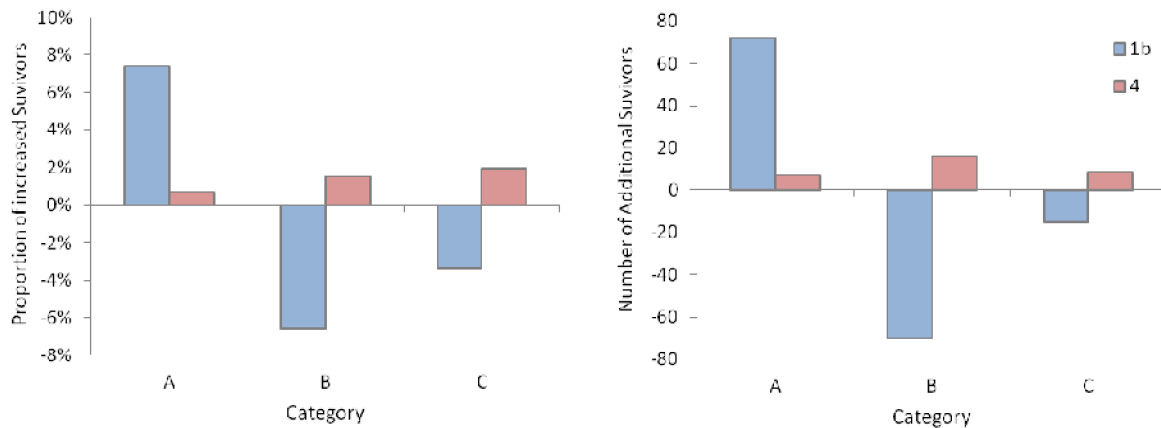
*"Barriers to swift handovers come in the form of capacity issues, patterns of accessing services and bed management across the whole of the NHS system" – Audit Committee (2009).*

Such blockages at hospitals increase the likelihood of EMS diversions (Fatovich et al. 2005), which have already been shown to reduce survival probability of critical patients, generating a vicious cycle of detriment.

Community treatment policies for certain emergency conditions (as explored in scenario 4, Chapter 8), alleviate some of the lost ambulance hours during transfer at hospital, ultimately saving more lives than default transportation rules. This simultaneously improves A&E congestion problems, where other non-critical patients now experience smaller waiting times at A&E and patients reside temporarily in corridors due to a shortage of emergency beds.

Previous studies have shown that cardiac outcomes were better in a targeted response system where paramedics only service critical incidents, compared with a uniform system where every incident receives a paramedic staffed ambulance (Persse et al. 2003); this is also supported by the findings of scenarios 1a and 1b of this experimentation process. Additionally, it is argued (Thakore et al. 2002) that by sending paramedic 'lights and sirens' responses to only critical patients the risk of EMS vehicles and crews becoming involved in road accidents themselves is reduced.

Conveyance minimisation policies (simulation scenario 4) combined with the reduction of double-dispatches (scenario 1b – RRVs only sent to category A patients) would lead to a significant increase in the number of category A patients using the simulation results, Figure 9.2; yet, of course, the lower priority patients do appear worse when applying the same survival function to all categories.



**Figure 9.2** Comparison of additional survivors (proportion and numerical) per category for a typical week, resulting from scenario 1b and 4 compared with benchmark scenario 0

If information could be obtained so that the possibility of modelling fully integrated systems with OR was attainable, rather than individual components, swift transition of care between systems could be greatly improved. However, overall, it is likely the turnaround time issues faced by WAST and the congestion problems occurring at many of the country's hospitals can be reduced, if not solved, without the need for, nor implementation of, sophisticated analysis or mathematical modelling tools. Stephen Thornton's insightful personal view published in the British Medical Journal (2007) highlights the issues surrounding London's provisions for stroke victims, stating that interaction between hospitals and ambulance services is necessary when making arrangements for treatment in the community or for urgent conveyance. It is also recommended better local clinical pathways are needed – communication between ambulance services and GPs – to increase the uptake of thrombolysis in the community, helping patients chances of recovery (Bloe et al. 2009) .

Better communication between separate Trusts and NHS departments, and the formation of alliances between Trusts should be encouraged and assisted by common targets instead of system-selfish motives. If communication and cooperation between systems could improve, if patients

were the foundation of evaluation measures, then each system may be able to perform to its best with little change.

OR offers the means to finding the best operational, strategic and tactical solutions, but final absolute achievements can only be obtained through client buy-in and the understanding of the power of OR by practitioners. Without convincing industry managers that even the simplest models can be effective, there will be slim chance of implementation of high-level, cutting edge technological advances from the research community. Here lies the difficulty in firstly bridging the gap with practitioners (Brailsford et al. 2013) and gaining acceptance by healthcare managers. This problem is accentuated by constant changing roles of healthcare professionals. The support of a 'champion' from within the service is essential if successful academic research is to be implemented in practice. Communication between Operational Researchers and service managers is just as vital as communication between services.

This research builds on the idea of making the developed models as generic as possible, so that the tools are not problem specific. Gaining support from the original and project based users, may lead to the implementation of the research, results or tools. If this is successful, the generic characteristics of the models expedite the simple application to other EMS systems. Even use of the location models and simulation to other location and priority queueing type problems, in fields such as transportation and logistics, would require minimal additional work to make them suitable.

## **9.2 Model Limitations**

### ***Travel Time Matrix Generator***

Since Google Maps restricts the number of requests made by a user to around 25,000 per day (at a certain rate), its use is limited practically by the level of granularity and size of the region explored.

The Google Maps API algorithm for calculating journey time between two locations is unknown, and there exist inaccuracies in the WAST journey data, so it is not possible to completely validate the travel time values obtained and utilised in the modelling process. Additionally, the speed assumed by Google Maps in the calculation process applies only to regular vehicles. The application of the developed regression models and determined scalar values to other EMS regions is therefore somewhat limited.

Some adjustments (via trial and error) were necessary in the simulation model in order to scale estimated travel times to EMS vehicle travel, which must also differentiate between EA and RRV travel as the nature of these sub-fleets implies rather different travel patterns over the network. However, on comparing predicted journey times with Google Maps calculated times (Chapter 7, Figure 7.5) and validating simulated response journey travel times against the original data (Chapter 7, Figure 7.19) the estimation methods are deemed adequate enough for the purposes of this study.

### ***Location Analysis***

Given that the multi-objective optimisation is dependent upon the weights used in the objective function, the results are sensitive to changes in these values. Since the decision for the value of these preference weights is entirely subjective – the choice is down to the user – if these models are to be applied to multiple EMS Trusts in the UK, it would be best to make use of a standard weighted preference so that comparisons can be made for the implemented service strategy in the optimisation process. This however, leads to an ethical dilemma as to the importance of patients within each service group and is not a straightforward quantifiable decision.

An alternative approach is that of ‘non-composite’ multi-objective optimisation, which does not make use of preference weights, but still requires the user to choose the best overall Pareto-optimal solution. Again, if more than one Trust were to use the models, problems arise in the subjective decision of the ‘best’ solution; however, at least in this case the best option is chosen with all knowledge of outcomes by the analyst, whereas with a weighted objective function, the scale of category preferences are input without all outcome information, unless extensive sensitivity analysis is conducted. The discussions of section 6.8.7 (Chapter 6) allude to the difficult decision in patient-class importance and the assigned preference values used in the investigations of this study.

To gain full compliance tables for a network, the models all have large run times (anywhere from a couple of hours to several days, dependent upon the selected model and problem size). Instead, to be more useful practically, the non-iterative versions can be run in around 15 minutes for a given fleet size – although this is still a substantial time period and so shows how these tools are best used for planning purposes not dynamic decision making.

A second limitation in implementing these tools within an EMS Trust is the use of a software package general purpose GA. For the user to run the location models, the Palisade Evolver package

is required. However, to get around this problem, it might be possible to code a purpose built GA, or implement an alternative and more easily accessible heuristic.

Finally, the deterministic nature of the location models means their potential insight is limited and so should ideally be used in conjunction with further modelling approaches, such as the developed simulation.

### ***Simulation***

The main drawback to the simulation model is the amount of data and input required in the tool. Although much of this information can be loaded from a pre-generated default file, for the model to be representative of the investigative region, many parameters would ideally need to be adjusted.

Both the location models and simulation ignore the requirement of a vehicle being staffed by a suitable crew when assigning an EMS unit to an emergency incident. In reality, the decision for dispatch is not based upon availability of a vehicle alone, nor solely in addition to the correct vehicle type, but also that the chosen vehicle is also manned by the suited crew type. For practical implementation of such modelling tools in an EMS trust, some indication of operational vehicles, categorised not only by vehicle type but also by crew type should be considered. This would necessitate the use of rostering tools alongside the suite of tools provided by this research.

## **9.3 Model Extensions**

### ***Location Analysis***

Unfortunately, the heterogeneous fleet models are limited to just two vehicle types. An immediate extension to the MESLMHPHF model of Chapter 6 would be to include more than two sub-fleets. The formulation of this model would be similar, but further mathematical exploration of the interactions between sub-fleets and information regarding the dispatch and service policies they follow would be required for instances where combinations of heterogeneous vehicles are used to serve different incident types.

The primary difficulty in this extension is that EMS Trusts may have contrasting operational policies. Many of the service strategies are designed and approved by the individual Trust managers. For the

purposes of this study, the location model design was maintained at a more generic and simple level in order to allow possible widespread usage.

A variety of survival functions for the different emergency categories could be included in the models in order to more realistically predict population survival rates for fleet capacities and allocation. More specific functions per category would better reflect the actual outcomes of the patients of each group, meaning evaluation of survival within the location and simulation models would not necessarily portray such large scale detriment to the lower priority classes as seen in Figures 9.1 and 9.2. The choice of coefficients in the survival function however, should remain constant if comparisons across systems are to be made (Eisenberg et al. 1991).

Average service time per vehicle type was implemented in the models. Given the deterministic setup of the model, segregation of service components and the inclusion of category dependent service times in the modelling process would give results matching more closely with reality. During data analysis stages, travel time was found not to be dependent upon category, but some discrepancy exists between service times and so the location model could be made more accurate with such information. Differentiation between category services is rarely witnessed in the literature and a general service time approximation still gives suitable solutions to the models used in this capacity.

### ***Simulation***

The very nature of simulation means that some simplifying assumptions of the real-world system are made during modelling design stages. In most cases these are suitably representative; however, in some instances, the inclusion of further information of the real system would be beneficial to the validity of experimental conclusions and scope of insight.

Aspects of an EMS system that have not been included in the simulation study, but lend themselves to such experimentation structures, include:

- dynamic events – updates to schedule throughout the run to represent dynamic demand;
- dynamic redeployment – to provide consistently equitable service;
- reallocation – interrupt service to reassign vehicles to higher priority calls;
- special practitioner service.

Search algorithms within the simulation program could be altered to provide different experimental dispatch and service strategies as desired by the Trust. For example, the current method only searches free or returning vehicles for dispatch to a newly arriving call; however, in calendar-queue DES, many future vehicle events are already scheduled. If an anticipated clearing event of a vehicle at a nearby hospital is due to follow soon after logging the new incident, the response time (including the waiting period) may be shorter for this clearing vehicle than for response by any currently available vehicles. In reality, it is probable that control-room operators' judgement and experience does influence such dispatch decisions.

Instead of searching through *all* stations in the region for the best available responder option every time a vehicle is required, a selective search could also heighten efficiency. For example, if only  $x \in \mathbb{Z}$  closest stations, or those within a specific time standard were searched for free and interruptible fleet members, simulation run-time would be reduced; however, paramedics are known (NISCHR February 2013) to attend emergencies anywhere in Wales if they are the only or best available vehicle – the current simulation operational procedure abides by this policy for a modelled region.

An important consequence of the search processes is choosing whether to interrupt vehicles, i.e. interrupting vehicles returning to base and assigning instead to new calls, or making low priority response journeys interruptible. The algorithm would need to ensure the same vehicle, crew and incident are not interrupted recurrently.

Alternative dispatching rules may also be investigated. Bandara et al.(2012) discovered that patient survival probabilities can be increased by not automatically sending the closest vehicle but basing the choice upon call priority.

Systems Dynamics (SD), as mentioned in Chapter 7, section 7.3, looks quantitatively at the interaction of systems. A larger region could easily be investigated in one model due to the high-level data requirements and a non micro-simulation approach. It would deal easily with the non-symmetric flows of service across regions and trust boundaries. Such a manner of modelling would enable investigation of the contribution of EMS to hospital admissions (Lane et al. 2000) and impact on other NHS systems – a more integrated system approach (Brailsford et al. 2004). There have even been efforts to implement hybrid methods in healthcare, combining SD and DES (amongst

other) techniques, to enlighten as to the interaction between separate NHS systems (Brailsford et al. 2010).

## 9.4 Implementation

There are three main means of OR implementation (Robinson et al. 2009):

- implement findings of a study;
- implement a model;
- implement as learning.

It seems from the small number of academic papers that show successful implementation of work in industry or public sectors, or even support from such organisations at the outset of any study, that the hardest of these three types is implementing models. Learning is commonly shared with academic communities and findings are regularly presented back to organisations where collaboration exists at the outset, which may even lead to future process changes for a particular scenario modelled. However, getting developed models to be regularly used within the workplace by controllers and analysts, after the conclusion of the study, is often much more difficult.

At the time of writing, discussions with WAST with regards to implementation of the resultant tools of this study were underway. The mention of trialling the tools in resource planning departments in the country and the engagement of the WAST informatics team throughout the research process has been encouraging. An arranged forthcoming meeting promises hopeful collaboration in the future between the Trust and OR academics. Hope lies that these tools will be able to offer foundations for regular experimentation of new suggested policies and ideas for WAST.

## 9.5 Final Reflections

The original posed question of this study asks, can mathematical modelling and OR techniques enable an EMS system to operate at a higher level, improving service to patients and increasing the probability of a positive clinical outcome? In this applied research study, developed models and tools are demonstrated to work successfully in this endeavour for the South East Wales EMS Trust.



Frequently, it is discovered through OR investigations that the examined system could cope well with new proposals or alterations in its current state, and that better understanding would alleviate the need for restructure, redesign or further investment.

Given the discussed conclusions, it would be all too easy for blame to be placed with A&E departments for EMS underachievement; instead the findings should be in a positive manner by both parties – improving turnaround time would relieve blocking outside A&E, which is not only a problem for WAST in terms of lost ambulance hours, but the congestion issue also gives individual hospitals bad publicity and puts pressure on staff dealing with the daily chaos. Ultimately, this study shows that better EMS fleet utilisation and so response performance (by whichever measure) can be achieved through reduction of automatic double-dispatches, better allocation, increased community treatment prevalence and careful consideration of need for conveyance.

Connected systems should not be operated entirely independently (Chen and Decker 2005), especially where the service user – the most important element in the system – and information flow from one to another. Follow-up of discussed policy changes should lead EMS Trusts into partnership with A&E departments and other NHS Trusts – often, a new way of thinking of an existing system, rather than a new idea in a new system, can enhance performance and benefit the service user. At the forefront of healthcare decision processes and OR modelling projects should be the ideas of *continuity of care*, *cooperative objectives* and *consolidated insight*.

## Appendix

### Appendix 4.1 WAST 2009 emergency service data set: field headers and details

Data set Header	Description	Range of Variable Values
Incident Date	Date (dd/mm/yyyy) operators log call	01/01/2009 - 31/12/2009
Incident Time	Time (hh:mm) operators log call	00:00:00 - 23:59:00
Unique ID	Code given to an incident for all response vehicles	174,665 unique entries
Vehicle Order	Order in which multiple vehicles are dispatched	From 1 to 11
Vehicle Type	Type of vehicle dispatched	17 types observed
Vehicle Station	Station at which vehicle is allocated (not positioned)	170 unique observed
Postcode District	Origin of the emergency as provided by the caller	50 unique observed
Nature	Initial medical details given on the emergency type	Various
MPDS Priority	Priority based on colour codes used by WAST	Red/Amber/Green
Time Allocated	Vehicle instructed to attend an incident	Time (hh:mm)
Time Mobile	Crew and vehicle begins the response journey	Time (hh:mm)
Time at Scene	Arrival of vehicle at the scene of the emergency	Time (hh:mm)
Time Left Scene	Vehicle leaves the scene with or without patient	Time (hh:mm)
Hospital Attended	Assigned hospital based on proximity to incident	170 unique observed
Time at Hospital	Vehicle and patient arrive at the assigned hospital	Time (hh:mm)
Time Clear	Vehicle becomes free to attend subsequent incidents	Time (hh:mm)
Incident Type	Type of emergency: emergency, urgent or routine	AS1/AS2/AS3
PCT Code	Primary Care Trust in which the incident originates	25 codes observed
Stood Down	Does vehicle stop service before completion	Yes/No
Reason Stopped	Reason why vehicle or incident process was stopped	Text

**Appendix 4.2** Hospital preference proportion of postcode districts for commonly used facilities

Division	PC Area	UHWC	RGHN	RGHP	PCHM	NHHA	LLAN	POWB	Proportion of Division	Total Proportion
SE1	NP11		93.88			5.99			17.74	
	NP12		90.51		5.76	3.73			15.79	
	NP13		1.31			98.65			11.00	
	NP22				49.71	49.82			13.25	
	NP23					99.59			19.62	
	NP25		2.03			97.97			6.37	
	NP7		8.93			91.07			13.50	
	NP8		0.70			99.30			2.73	92.82
SE2	NP10		99.60						6.48	
	NP15		5.96			94.04			1.99	
	NP16		95.77			4.23			5.46	
	NP18		99.27			0.73			4.39	
	NP19		99.57						17.34	
	NP20		98.71			1.29			24.09	
	NP26		99.19			0.81			6.34	
	NP4		50.71			49.29			16.03	
	NP44		97.40			2.60			17.88	92.96
SE3	CF44				99.74				20.00	
	CF45				100				9.35	
	CF46				99.54				5.17	
	CF47	14.86			84.95				12.46	
	CF48				99.93				11.83	
	CF81		2.71		96.90				8.09	
	CF82	2.13	6.83		91.04				8.23	
	CF83	78.34	21.52						22.27	
	NP24				99.54				2.58	87.45
SE4	CF31							100	2.81	
	CF32							100	1.89	
	CF35							99.67	5.21	
	CF37			100					20.37	
	CF38			100					6.62	
	CF39			99.85					14.47	
	CF40			100					13.01	
	CF41			100					6.31	
	CF42			99.95					7.80	
	CF43			100					6.79	
	CF61							99.84	2.63	
	CF71			39.83				60.17	2.56	
	CF72			96.83				3.17	9.55	94.18
SE5	CF10	97.85					2.15		10.82	
	CF11	76.18					23.82		9.01	
	CF14	89.17					10.83		14.58	
	CF15	93.61					6.39		2.21	
	CF23	96.93					3.07		7.29	
	CF24	95.52					4.48		11.86	
	CF3	96.12					3.88		8.57	
	CF5	72.09					27.91		14.31	
	CF62	59.12					40.88		8.22	
	CF63	66.63					33.37		5.13	
	CF64	67.70					32.30		8.00	93.63

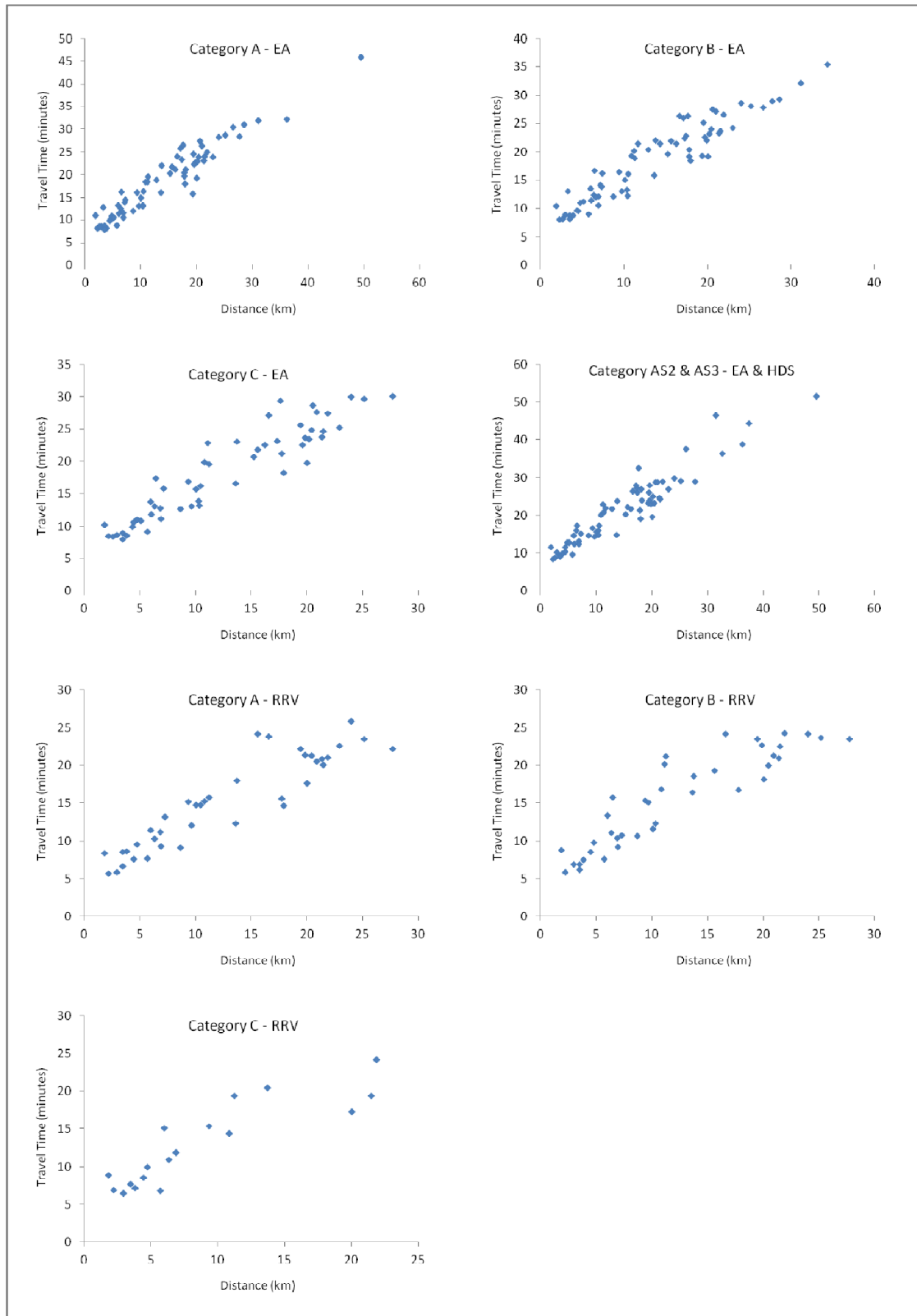
---

**Appendix 5.1** Travel Matrix Generator Tool Pseudo Code

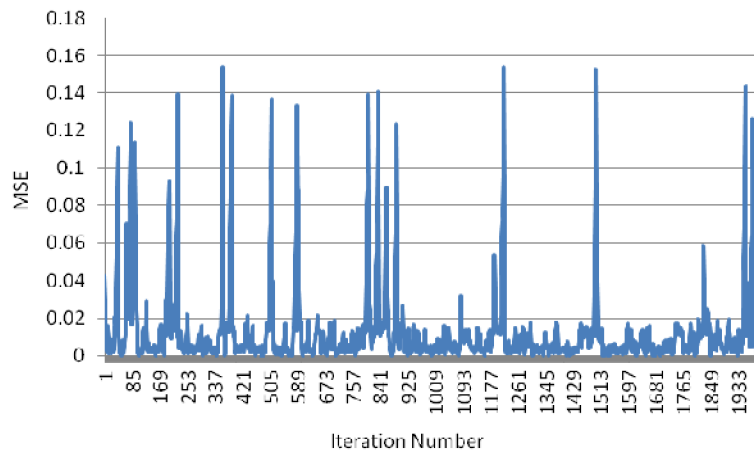
```
> Access text files
> Create empty location list
> For each location entry in text file
>     Add to location list
> Loop

> For index i = 1 to location list length
>     origin = address of location list item i
>     For index j = 1 to location list length
>         destination = location address of location list item j
>         Key = location pair (origin, destination)
>         Send Key to Google Maps API
>         Google Maps returns Route with:
>             Route Key = Key
>             Route time = fastest journey time between origin and destination
>             Route distance = equivalent journey distance between origin and destination
>         Add Route to Matrix for display purposes
>     Next j
> Next i
```

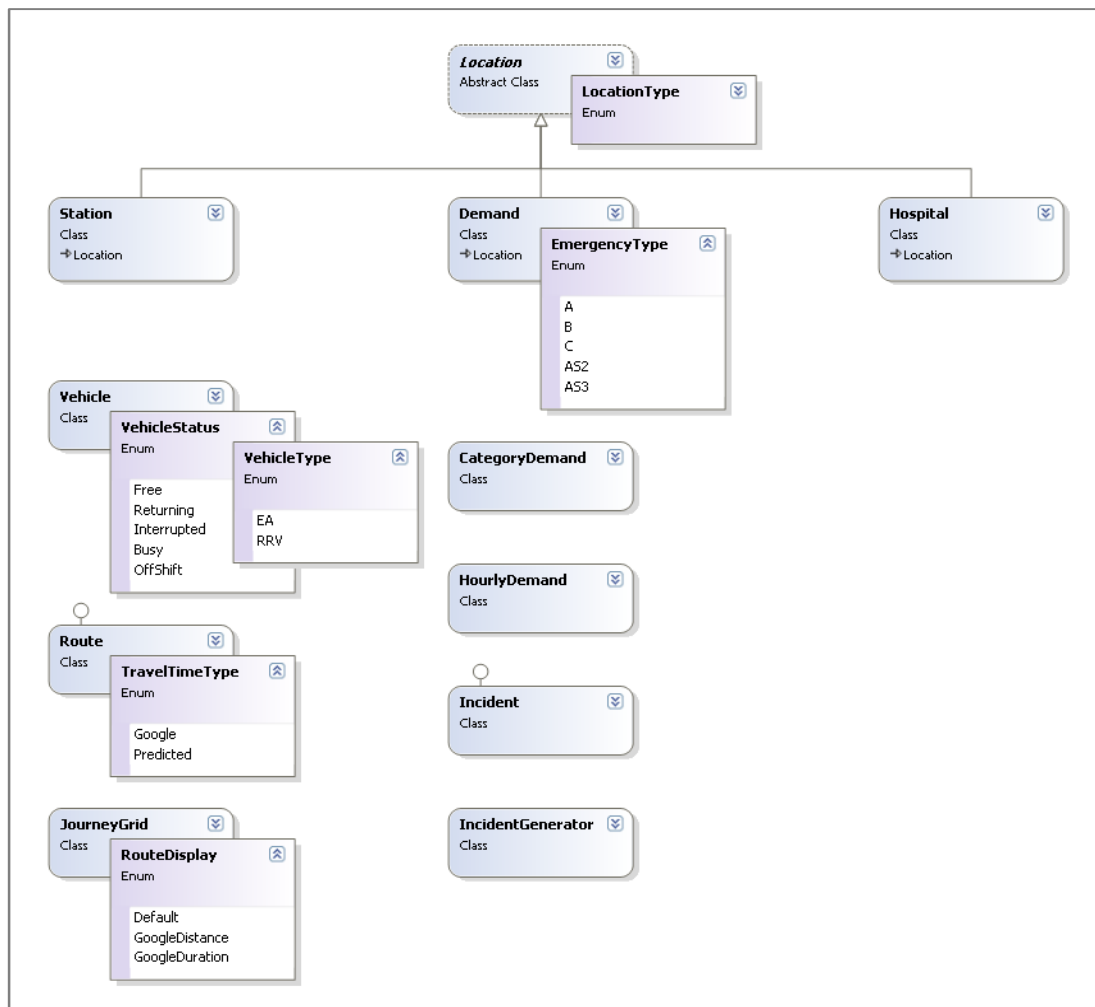
## Appendix 5.2 Scatter plots of Google Maps distance against average travel time data for demand to hospital routes per category and vehicle type



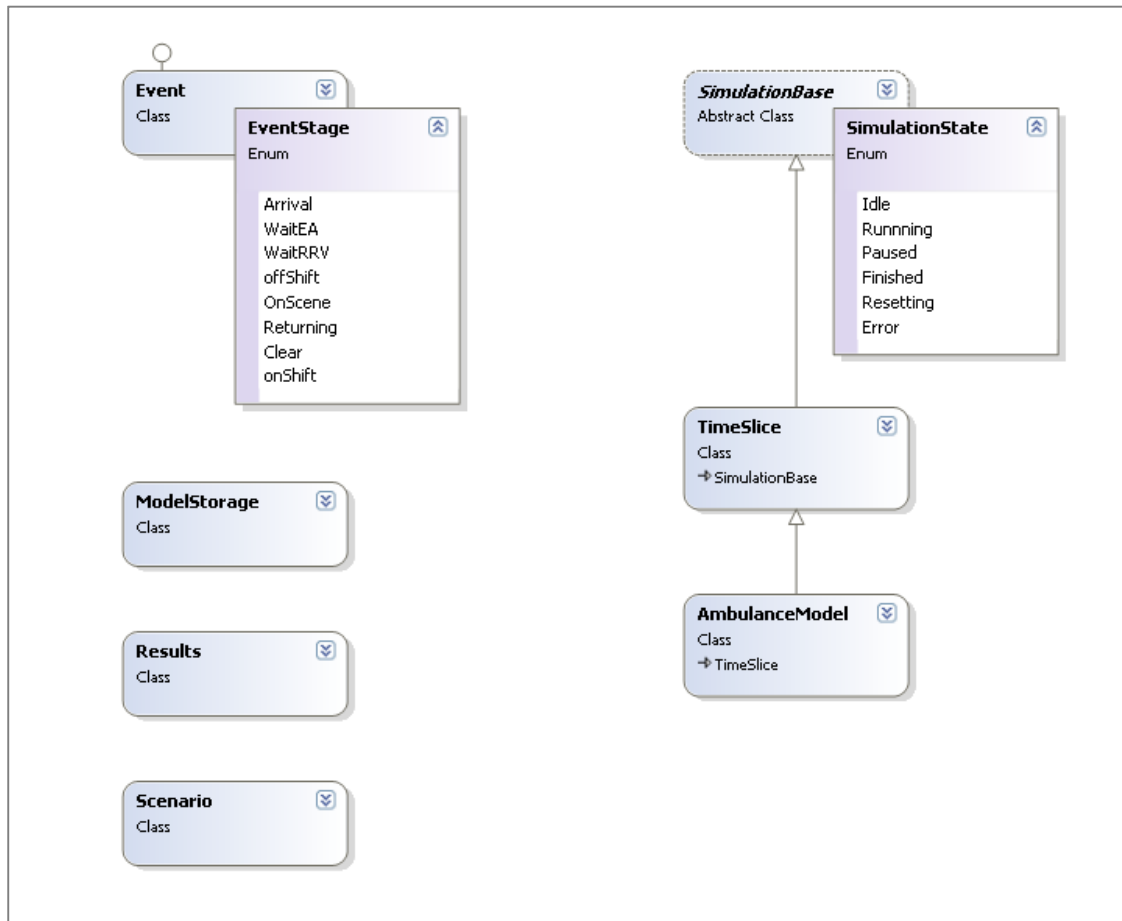
## Appendix 6.1 MSE of input utilisation and resulting utilisation after an iteration of MESLMHP-I for an experimental subset of the South East dataset



## Appendix 7.1 Class diagram for entities and all associated classes (blue) and enumerators (purple) in the simulation model (inheritance is implied by arrows)



**Appendix 7.2** Class diagram for non-object associated classes (blue) and enumerators (purple) in the simulation model (inheritance is implied by arrows)



**Appendix 7.3** Example of text template file require as input to simulation model with load input option

```

PostcodeDemandByCategory.txt
1 CF10|0|1|2333|2560|791|216|5,
2 CF11|0|1|1634|1853|862|539|14,
3 CF14|0|1|2376|2759|1186|2813|1191,
4 CF157PE|0|1|521|543|210|161|4,
5 CF23|0|1|1301|1440|634|568|23,
6 CF24|0|1|2032|2673|1075|691|12,
7 CF3|0|1|1642|1763|655|725|13,
8 CF31|0|1|244|262|107|70|1,
9 CF32 9|0|1|148|187|67|60|0,
10 CF355NP|0|1|453|484|158|176|8,
  
```

**Appendix 7.4a** Call generation pseudo-code

```

> Calculate appropriate amount of demand using run length:
>   number of weeks = round up (run length / 10080)
>   For 1 to number of weeks
>     randomly sample total demand
>     For 1 To total demand
>
>         generate incidents:
>             create new incident
>             sample priority
>             sample arrival hour:
>                 random value = (random number between 0 and 1)*100
>                 hour of call = top-hat sample using random value
>                 minute of call = (decimal part of random number)*60
>             sample location
>             sample on scene length
>             sample pre-travel delay
>             sample transportation decision
>             If transportation = yes
>                 choose hospital:
>                     Optional: closest facility based on minimum distance
>                 sample turnaround time at hospital
>                 calculate transport travel time
>             End If
>             Add new incident to incident list
>
>         Next
>     Next
>   Sort incident list by arrival time and incident priority
>   For Each incident in incident list
>       If arrival time > run length
>           remove incident from list
>       End If
>   Next

```

**Appendix 7.4b** Simulation logic pseudo-code

```

> Set clock = time zero;
> Do Until clock > run-length
>
>   If clock = time scheduled for first object in event-list
>       Do While event time = clock
>           If event type = arrival
>               Do arrival tasks:
>                   If incident type requires 1 vehicle

```



```

> Search for the closest available (or returning) EA
> If an available vehicle exists
>     arrival time at scene = call arrival time + waiting time
>         + pre-travel delay time + travel length
>     Create new event with event type = arrival at scene
>     Add new event to event list
> Else
>     Create new event with event type = incident awaiting EA
>     new event time = time of next event in event list
>     Add event to event list
>     Optional: increase incident priority If wait > limit
> End If
> Else If incident type requires 2 vehicles
>     Search for the closest available (or returning) RRV and EA
>     If available vehicles exist
>         arrival time at scene = call arrival time + waiting time
>             + pre-travel delay time + travel length
>         Create new events with event types = arrival at scene
>         Add new events to event list
>     Else
>         Create new event(s) with type(s) = incident awaiting EA/RRV
>         new event time = time of next event in event list
>         Add event(s) to event list
>         Optional: increase incident priority if wait > limit
>     End If
> End If
> End Do

> Else If event type = arrival at scene
>     Do service tasks:
>         If vehicle type = EA
>             cancel RRV if due but not already on scene
>             leave scene time = arrival time at scene + EA scene length
>             If transportation required
>                 arrival time at hospital = leave scene time + transport length
>                 leave hospital time = arrival time at hospital + turnaround length
>                 finish time = leave hospital time
>             End If
>         Else If vehicle type = RRV
>             leave scene time = arrival time at scene + RRV scene length
>             finish time = leave scene time
>         End If
>     response time = minimum arrival time at scene – call arrival time
>     Create new event with event type = returning to base
>     new event time = finish time
>     Add new event to event list
> End Do

```

```

> Else If event type = returning to base
>     Look-up travel distance of route from current location back to station
>     return travel time = route distance converted to time
>     vehicle status = returning
>     create new event with type = vehicle clear
>     new event time = finish time + return travel time
>     Add new event to event list

> Else If event type = vehicle clear
>     vehicle status = free
>     vehicle location = station

> Else If event type = incident awaiting RRV
>     Do waiting tasks:
>         Search for the closest available (or returning) vehicle
>         If an available vehicle exists
>             Determine time vehicle will arrive at scene
>             Create new event with event type = arrival at scene
>             Add new event to event list
>         Else
>             Create new event with event type = incident awaiting vehicle
type
>             new event time = time of next event in event list
>             Add event(s) to event list
>             Optional: increase incident priority if wait > limit
>         End If
>     End Do

> Else If event type = incident awaiting EA
>     Do waiting tasks (as above)

> Else If event type = vehicle starts shift
>     Do on-shift tasks:
>         vehicle status = free
>         vehicle shift start-time = clock

> Else If event type = vehicle ends shift
>     Do off-shift tasks:
>         If vehicle status = free or returning
>             vehicle status = off-shift
>             total vehicle on-shift length = total + (clock - vehicle shift start-time)
>         Else If vehicle status = busy or interrupted
>             Create event with type = vehicle ends shift
>             new event time = next time step with an event in event list
>             Add event to event list
>         End If
>     End Do

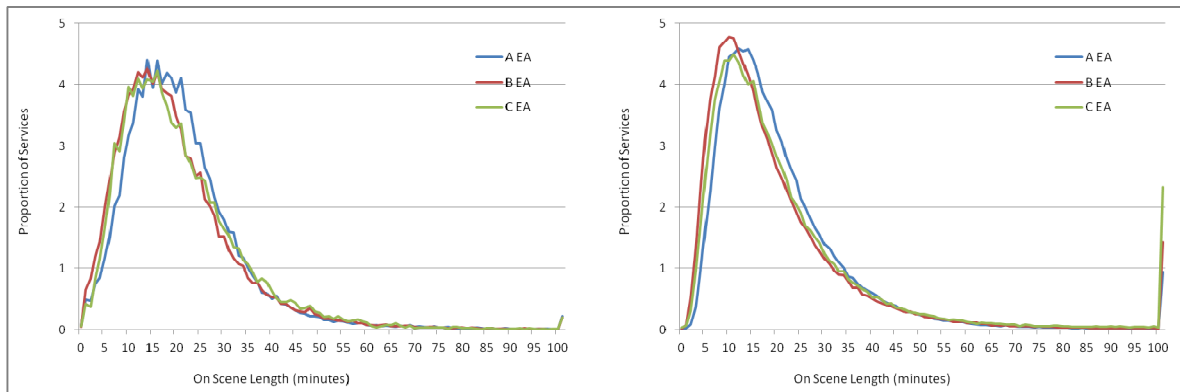
```

- 
- > Go To next event in event-list
  - > End
  
  - > End Do
  - > End If
  
  - > Increment clock one step
  - > Loop Do

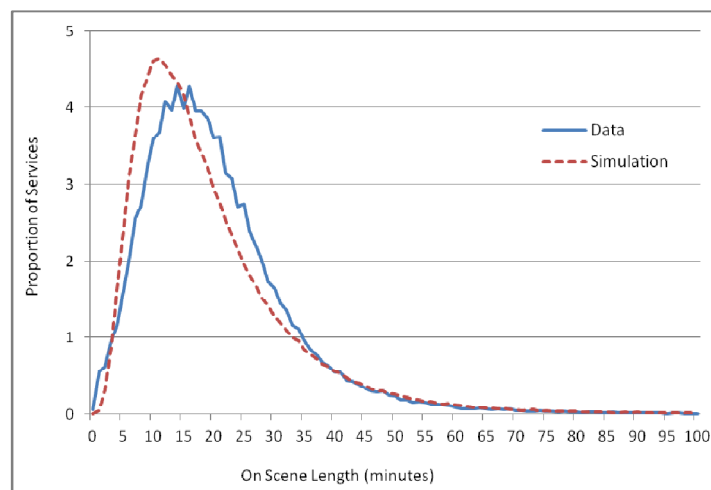
#### **Appendix 7.4c** Available vehicle search pseudo-code

- > For an incident
- > Best route distance = null
  
- > For each station in station list
  
- > If suitable vehicle is available
- > Current route = route from station to incident location
- > If current route distance < Best route distance
- > Best route = current route
- > Best vehicle = vehicle
- > Else if vehicle is returning
- > Current route = route from estimated vehicle location to incident location
- > If current route distance < Best route distance
- > Best route = current route
- > Best vehicle = vehicle
- > End
  
- > Next station
  
- > Best vehicle status = Busy
- > Travel time = best route distance converted to time

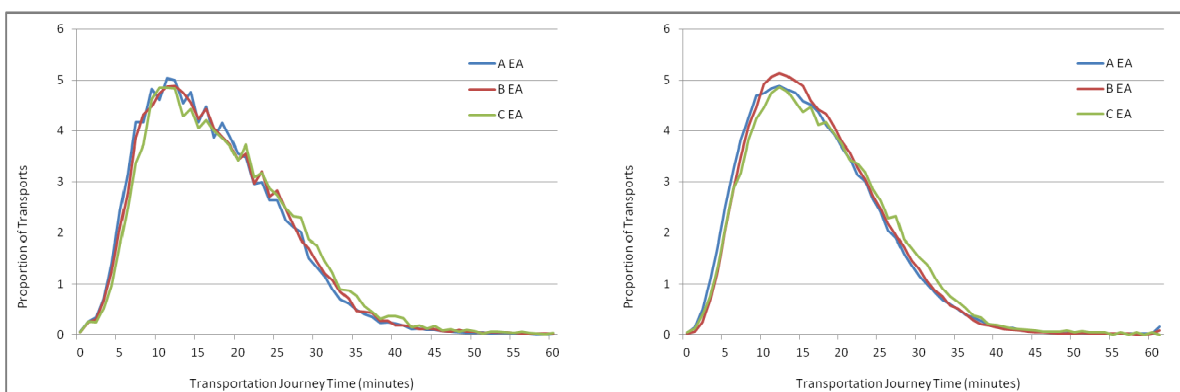
**Appendix 7.5a** On-scene service distributions for data (left) compared with simulation results (right)



**Appendix 7.5b** On-scene service distribution, for category A, B and C combined for EA vehicles, for data compared with simulation results



**Appendix 7.5c** Transportation journey time distributions for data (left) compared with simulation results (right)



## Glossary

<b>A&amp;E</b>	Accident and Emergency	<b>GA</b>	Genetic Algorithm
<b>ABS</b>	Agent Based Simulation	<b>GIS</b>	Geographical Information Systems
<b>ALS</b>	Advanced Life Support	<b>GP</b>	General Practitioner
<b>AMPDS</b>	Advanced Medical Priority Dispatch System	<b>GPS</b>	Global Positioning System
<b>API</b>	Application Programming Interface	<b>HDU</b>	High Dependency Unit
<b>AVLS</b>	Automatic Vehicle Location System	<b>KPI</b>	Key Performance Indicator
<b>BLS</b>	Basic Life Support	<b>LHB</b>	Local Health Board
<b>CAD</b>	Computer Aided Dispatcher	<b>MSLP</b>	Maximum Survival Location Problem
<b>CPR</b>	Cardiopulmonary Resuscitation	<b>NHS</b>	National Health Service
<b>DES</b>	Discrete Event Simulation	<b>OOP</b>	Object Oriented Programming
<b>EA</b>	Emergency Ambulance	<b>OR</b>	Operational Research
<b>ED</b>	Emergency Department	<b>PCS</b>	Patient Care Service
<b>EMD</b>	Emergency Medical Dispatcher	<b>RRV</b>	Rapid Response Vehicle
<b>EMS</b>	Emergency Medical Service	<b>SD</b>	System Dynamics
<b>EMT</b>	Emergency Medical Technician	<b>SP</b>	Special Practitioner
<b>FIFO</b>	First-in, first-out	<b>WAST</b>	Welsh Ambulance Service NHS Trust

## References

- Air Ambulance Service (2012) *The History of the Air Ambulance*. [www.airambulanceservice.com/history.html](http://www.airambulanceservice.com/history.html) Accessed: 13<sup>th</sup> September 2012
- Alanis R, Ingolfsson A and Kolfal B (2013) A Markov Chain model for an EMS system with repositioning. *Production and Operations Management* 22 (1) 216-231.
- Alexopoulos C (2006) A comprehensive review of methods for simulation output analysis. In Perrone, L F, Wieland, F P, Liu, J, Lawson, B G, Nicol, D M and Fujimoto, R M (Eds.), *Winter Simulation Conference*. Monterey, California.
- Alsalloum O I and Rand G K (2006) Extensions to emergency vehicle location models. *Computers and Operations Research* 33 2725-2743.
- Altay N and Green W G (2006) OR/MS research in disaster operations management. *European Journal of Operational Research* 175 457-493.
- Alves D W, Allegra J R, Cochrane D G and Cable G (2003) Effect of lunar cycle on temporal variation in cardiopulmonary arrest in seven emergency departments during 11 years. *European Journal of Emergency Medicine* 10 (3) 225-228.
- Andersson T and Varbrand P (2007) Decision support tools for ambulance dispatch and relocation. *Journal of the Operational Research Society* 58 (2) 195-201.
- Atkins B (2003) *More Than a Game*. Manchester University Press.
- National Assembly for Wales (2009) *Ambulance Services in Wales Inquiry*. National Assembly for Wales.
- Azizan M H, Lim C S, Hatta L W M and Gan L C (2012) Application of OpenStreetMap data in ambulance location problem. *Computational Intelligence, Communication Systems and Networks*. IEEE Computer Society.
- Ballou R H, Rahardja H and Noriaki S (2002) Selected country circuitry factors for road travel distance estimation. *Transportation Research Part A* 36 843-848.
- Bandar D, Mayorga M E and McLay L A (2012) Optimal dispatching strategies for emergency vehicles to increase patient survivability. *International Journal of Operational Research* 15 (2) 195-214.
- Barkley K (1978) *The Ambulance: The story of emergency transportation of sick and wounded through the centuries*. New York: Exposition Press.
- BBC News (2006a) Ambulance chief quits over merger. 16<sup>th</sup> March 2006.
- BBC News (2006b) New chief for ambulance service. Wales. 7<sup>th</sup> August 2006.
- BBC News (2009a) A&E transfer records criticised. Wales Politics. 23<sup>rd</sup> April 2009.
- BBC News (2009b) Ambulance flaws 'costing lives'. 7<sup>th</sup> July 2009.
- BBC News (2010) Welsh ambulance service chief to stand down. 10<sup>th</sup> March 2010.
- BBC News (2011a) Patient handovers at hospitals 'delaying ambulances'. Wales Politics. 17<sup>th</sup> January 2011.
- BBC News (2011b) Weather blamed for ambulance response times in December. Wales. 25<sup>th</sup> January 2011.
- BBC News (2011c) Welsh ambulance service winter response concern. Wales. 6<sup>th</sup> January 2011.
- BBC News (2011d) Winter cost ambulance service extra £400,000. England. 3<sup>rd</sup> February 2011.
- BBC News (2012a) Do government cuts mean some now 'dial 999 and wait'? Health. 3<sup>rd</sup> September 2012.

- BBC News (2012b) North East Ambulance Service faces major reorganisation. England. 2<sup>nd</sup> February 2012.
- BBC News Health (2009) Stroke campaign boosts awareness. 9<sup>th</sup> November 2009.
- Benveniste R (1985) Solving the combined zoning and location problem for several emergency units. *Journal of the Operational Research Society* 36 (5) 433-450.
- Berman O, Drezner Z and Krass D (2010) Generalized coverage: New developments in covering location models. *Computers and Operations Research* 37 1675-1687.
- Berman O, Krass D and Drezner Z (2003) The gradual covering decay location problem on a network. *European Journal of Operational Research* 151 474-480.
- Berman O, Larson R C and Parkan C (1987) The stochastic queue p-median problem. *Transportation Science* 21 207.
- Bevan G and Hamblin R (2009) Hitting and missing targets by ambulance services for emergency calls: Effects of different systems of performance measurement within the UK. *Journal of the Royal Statistical Society Series A* 172 (1) 161-190.
- Blackett P M S (1962) *Studies of War: Nuclear and Conventional*. First American Edition ed. NY Hill and Wang.
- Bloe C, Mair C, Call A, Fuller A, Menzies S and Leslie S J (2009) Identification of barriers to the implementation of evidence-based practice for pre-hospital thrombolysis. *Rural and Remote Health* 9.
- Brailsford S C, Bolt T, Bucci G, Chausalet T, Connell N A D, Harper P R, Klein J H, Pitt M and Taylor M A (2013) Overcoming the barriers: A qualitative study of simulation adoption in the NHS. *Journal of the Operational Research Society* 64 157-168.
- Brailsford S C, Desai M S and Viana J (2010) Towards the Holy Grail: Combining system dynamics and discrete-event simulation in healthcare. In Johansson, B, Jain, S, Montoya-Torres, J, Hagan, J and Yucesan, E (Eds.), *Winter Simulation Conference*. Baltimore, Maryland: INFORMS.
- Brailsford S C and Harper P R (2007) Editorial. *Journal of the Operational Research Society* 58 141-144.
- Brailsford S C, Lattimer V A, Tarnaras P and Turnbull J C (2004) Emergency and on-demand health care: Modelling a large complex system. *Journal of the Operational Research Society* 55 (1) 34-42.
- Breen N, Woods J, Bury G, Murphy A W and Brazier H (2000) A national census of ambulance response times to emergency calls in Ireland. *Journal of Accident & Emergency Medicine* 17 392-395.
- British Heart Foundation (2012) Hands-only CPR. BHF.org.uk:
- British Pathe (c1915) Dog Sleds Transport Wounded.
- Brotcorne L, Laporte G and Semet F (2003) Ambulance location and relocation models. *European Journal of Operational Research* 147 (3) 451-463.
- Brown R (1988) Calendar Queues: A fast  $O(1)$  priority queue implementation for the simulation event set problem. *Communications of the ACM: Algorithms and Data Structures* 31 (10) 1220-1227.
- Budge S, Ingolfsson A and Erkut E (2009) Approximating vehicle dispatch probabilities for emergency service systems with location-specific service times and multiple units per location. *Operations Research* 57 (1) 251-255.
- Budge S, Ingolfsson A and Zerom D (2010) Empirical analysis of ambulance travel times: The case of Calgary Emergency Medical Services. *Management Science* 56 (4) 716-723.
- Cadigan R T and Bugarin C E (1989) Predicting demand for emergency ambulance service. *Annals of Emergency Medicine* 18 (6) 618-621.
- Campbell J F (1992) Selecting routes to minimize urban travel time. *Transportation Research Part B* 26 (4) 261-274.

- Cardiac Scient (n.d.) *AED Usage: Statistics*. [www.cardiacscience.com](http://www.cardiacscience.com).
- Carson Y M and Batta R (1990) Locating an ambulance on the Amherst campus of the state university of New York at Buffalo. *Interfaces* 20 43-49.
- Carter G M, Chaiken J M and Ignall E (1972) Response areas for two emergency units. *Operations Research* 20 (3) 571-594.
- Chaiken J M and Larson R C (1972) Methods for allocating urban emergency units: A survey. *Management Science* 19 110-130.
- ChainofSurvival.com (2012) *Timing is Everything*. [http://www.chainofsurvival.com/cos/COSOverview\\_detail.asp](http://www.chainofsurvival.com/cos/COSOverview_detail.asp) Accessed: 1<sup>st</sup> November 2012
- Chan T C Y, Li H, Lebovic G, Tang S K, Chan J Y T, Cheng H C K, Morrison L J and Brooks S C (2013) Identifying locations for public access defibrillators using mathematical optimization. *Circulation*.
- Chang J C and Schoenberg F P (2009) A statistical analysis of Santa Barbara ambulance response in 2006: Performance under load. *Western Journal of Emergency Medicine* 10 (1) 42-47.
- Chanta S, Mayorga M E, Kurz M E and McLay L A (2011) The minimum  $p$ -evny location problem: A new model for equitable distribution of emergency resources. *IIE Transactions on Healthcare Systems Engineering* 1 (2) 101-115.
- Chelst K and Jarvis J P (1979) Estimating the probability distribution of travel times for urban emergency service systems. *Operations Research* 27 (1) 199-204.
- Chen W and Decker K S (2005) The analysis of coordination in an information system application - Emergency Medical Services. *Annals of Information Systems* 3508 36-51.
- Chien D-K, Tsai S-H and Chang W-H (2011) Comparing outcomes of out-of-hospital cardiac arrest between prehospital basic life support and advanced life support. *International Journal of Gerontology* 5 (2) 131-132.
- Church R and ReVelle C (1974) The Maximal Covering Location Problem. *Papers of the Regional Science Association* 32 (1) 101-118.
- Clave S A (2007) *The Global Theme Park Industry*. CABI.
- Clawson J, Olola C, Heward A, Patterson B and Scott G (2008) The medical priority dispatch system's ability to predict cardiac arrest outcomes and high acuity pre-hospital alerts in chest pain patients presenting to 999. *Resuscitation* 78 298-306.
- Coats T J and Davies G (2002) Prehospital care for road traffic casualties. *British Medical Journal* 324 1135-1138.
- Cook T M and Russell R A (1980) Estimating urban travel times: A comparative study. *Transportation Research Part A* 14 173-175.
- Cooper L (1963) Location-allocation problems. *Operations Research* 11 (2) 331-343.
- CreativeTeaching (2011) *Creative Teaching Site*. [www.creativeteachingsite.com](http://www.creativeteachingsite.com) Accessed: 17<sup>th</sup> January 2013
- Cretin S and Willemain T R (1979) A model of prehospital death from ventricular fibrillation following myocardial infarction. *Health Services Research* 14 (3) 221-234.
- Dahl O J (2002) *The Roots of Object Orientation: The Simula Language*. Springer.
- Dale J, Higgins J, Williams S, Foster T, Snooks H, Crouch R, Hartley-Sharpe C, Glucksman E, Hooper R and George S (2003) Computer assisted assessment and advice for "non-serious" 999 ambulance service callers: the potential impact on ambulance despatch. *Emergency Medicine Journal* 20 (2) 178-183.
- Daskin M S (2008) What you should know about location modelling. *Naval Research Logistics* 55 (4) 283-294.



- Daskin M S and Haghani A (1984) Multiple vehicle routing and dispatching to an emergency scene. *Environment and Planning A* 16 (10) 1349-1359.
- Daskin M S and Murray A T (2012) Modeling Public Sector Facility Location Problems. *Socio-Economic Planning Sciences: Modeling Public Sector Facility Location Problems* 46 (2) 111.
- De Maio V J, Stiell I G, Well G A and Spaite D W (2003) Optimal defibrillation response intervals for maximum out-of-hospital cardiac arrest survival rates. *Annals of Emergency Medicine* 42 (2) 242-250.
- Dean S F (2008) Why the closest ambulance cannot be dispatched in an urban emergency medical services system. *Prehospital Disaster Med.* 23 (2) 161-165.
- Deb K (2005) Multi-Objective Optimization. In: Burke, E K and Kendall, G (eds.) *Search Methodologies*. Springer.
- Department of Health (2005) Taking Healthcare to the Patient: Transforming NHS Ambulance Services. (2009) *Toolkit for improving urgent and emergency care pathways by understanding increases in 999 demand*.
- Donaldson L J, Rutter P D, Ellis B M, Greaves F E C, Mytton O T, Pebody R G and Yardley I E (2009) Mortality from pandemic A/N1H1 2009 influenza in England: Public health surveillance study. *British Medical Journal* 339.
- Eaton D J, Sanchez U H M L, Lantigua R R and Morgan J (1986) Determining ambulance deployment in Santo Domingo, Dominican Republic. *Journal of the Operational Research Society* 37 (2) 113-126.
- Eisenberg M S, Bergner L and Hallstrom A (1979) Cardiac resuscitation in the community: Importance of rapid provision and implications for program planning. *The Journal of the American Medical Association* 241 (18) 1905-1907.
- Eisenberg M S, Cummins R O, Damon S, Larsen M P and Hearne T R (1990) Survival rates from out-of-hospital cardiac arrest: Recommendations for uniform definitions and data to report. *Annals of Emergency Medicine* 19 (11) 1249-1259.
- Eisenberg M S, Cummins R O and Larsen M P (1991) Numerators, denominators and survival rates: Reporting survival from out-of-hospital cardiac arrest. *The American Journal of Emergency Medicine* 9 (6) 544-546.
- Erkut E, Ingolfsson A and Budge S (2008a) Maximum availability/reliability models for selecting ambulance station and vehicle locations: A critique. *Unpublished*.
- Erkut E, Ingolfsson A and Erdoğan G (2008b) Ambulance location for maximum survival. *Naval Research Logistics* 55 (1) 42-58.
- Fatovich D M, Nagree Y and Sprivilis P (2005) Access block causes emergency department overcrowding and ambulance diversion in Perth, Western Australia. *Emergency Medical Journal* 22 351-354.
- Felder S and Brinkmann H (2002) Spatial allocation of Emergency Medical Services: Minimising the death rate or providing equal access? *Regional Science and Urban Economics* 32 27-45.
- Field J M, Hazinski M F, Sayre M R, Chameides L, Schexnayder S M, Hemphill R, Samson R A, Kattwinkel J, Berg R A, Bhanji F, Cave D M, Jauch E C, Kudenchuk P J, Neumar R W, Peberdy M A, Perlman J M, Sinz E, Travers A H, Berg M D, Billi J E, Eigel B, Hickey R W, Kleinman M E, Link M S, Morrison L J, O'Connor R E, Shuster M, Callaway C W, Cucchiara B, Ferguson J D, Rea T D and Vanden Hoek T L (2010) Part 1: Executive Summary American Heart Association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care science: . *Circulation* 122 (suppl. 3) S640-S656.
- Finlay P N and Wilson J M (1987) The paucity of model validation in Operational Research projects. *The Journal of the Operational Research Society* 38 (4) 303-308.
- Fitch J (2005) Response times: Myths, measurement and management. *Journal of Emergency Medical Services* 30 (9) 46-56.
- Fitzsimmons J A (1973) A methodology for emergency ambulance deployment. *Management Science* 19 (6) 627-636.
- Forrester J W (1961) *Industrial Dynamics*. Productivity Press.

- French E and Casali G L (2008) Ethics in Emergency Medical Services - Who cares? An exploratory analysis from Australia. *Electronic Journal of Business Ethics and Organization Studies* 13 (2) 44-53.
- Fujiwara O, Makjamroen T and Gupta K K (1987) Ambulance deployment analysis: A case study of Bangkok. *European Journal of Operational Research* 31 9-18.
- Gallivan S and Utley M (2004) Devil's advocacy and patient choice. In Dlouhy, M (Ed.), *29th Meeting of the EURO Working Group Operational Research Applied to Health Services*. Prague.
- Gass S I (1983) Decision-aiding models: Validation, assessment and related issues for policy analysis. *Operations Research* 31 (4) 603-631.
- Geroliminis N, Karlaftis M and Skabardonis A (2006) Generalised Hypercube Queueing Model for locating emergency response vehicles in urban transportation networks. *85th Annual Meeting Transportation Research Board*. Washington D.C.:
- Geroliminis N, Karlaftis M G and Skabardonis A (2009) A spatial queueing model for the emergency vehicle districting and location problem. *Transportation Research Part B* (43) 798-811.
- Goldberg J (2004) Operations Research models for the deployment of emergency service vehicles. *EMS Management Journal* 1 (1) 20-39.
- Goldberg J, Dietrich R, Chen J M, Mitwasi M, Valenzuela T and Criss E (1990) A simulation model for evaluating a set of emergency vehicle base locations: Development, validation and usage. *Socio-Economic Planning Sciences* 24 (2) 125-141.
- Goldberg J and Paz L (1991) Locating emergency vehicle bases when service time depends on call location. *Transportation Science* 25 (4) 264-280.
- Goldberg J and Szidarovszky F (1991) Methods for solving nonlinear equations used in evaluating emergency vehicle busy probabilities. *Operations Research* 39 (6) 903-916.
- Goldhill O (2013) Slow ambulance transfer times could threaten lives, AMs warn. *WalesOnline*. Wales News.
- Gonzalez R P, Cummings G R, Mulekar M S, Harlan S M and Rodning C B (2009) Improving rural emergency medical service response time with Global Positioning System navigation. *The Journal of Trauma* 67 (5) 899-902.
- Google (©2013) *Map Data*. [online] maps.google.co.uk.
- Green L and Kolesar P (1989) Testing the validity of a queueing model of police patrol. *Management Science* 35 (2) 127-148.
- Hakimi S L (1964) Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations Research* 12 450-459.
- Hakimi S L (1965) Optimum distribution of switching centers in a communication network and some related graph theoretic problems. *Operations Research* 13 462-475.
- Harewood S I (2002) Emergency ambulance deployment in Barbados: A multi-objective approach. *Journal of the Operational Research Society* 53 (2) 185-192.
- Harper P R (2013) The Importance of Variability: A health care systems example. YouTube.
- Hassan A and Ferrell B (c2009) Forklift routing in warehouses using dual-commands and stackable pallets. Unpublished: Material Handling Industry of America (MHIA).
- Health Maps Wales (2004) Rural Urban Classification. NHS Wales Informatics Service (NWIS).
- Pre-Hospital Emergency Care Council (2010) *Demand analysis and tactical deployment of ambulance services in the National Ambulance Service South East Region*. Pre-Hospital Emergency Care Council.
- Healthcare Simulation South Carolina (2013) *A brief history of simulation*. Accessed: 17<sup>th</sup> January 2013

- Henderson S G and Mason A J (1999) Estimating ambulance requirements in Auckland, New Zealand. In Farrington, P A, Nembhard, H B, Sturrock, D T and Evans, G W (Eds.), *Winter Simulation Conference*. Phoenix, Arizona.
- Henderson S G and Mason A J (2000) BartSim: A tool for analysing and improving ambulance performance in Auckland, New Zealand.
- Henderson S G and Mason A J (2004) Ambulance Service Planning: Simulation and Data Visualisation. In: Brandeau, M L, Sainfort, F and Pierskalla, W P (eds.) *Operations Research and Health Care: A Handbook of Methods and Applications*. Boston: Kluwer Academic.
- Herlitz J, Engdahl J, Svensson L, Young M, Angquist K-A and Holmberg S (2004) Can we define patients with no chance of survival after out-of-hospital cardiac arrest? *Heart* 90 1114-1118.
- Hill A V and Benton W C (1992) Modelling intra-city time-dependent travel speeds for vehicle scheduling problems. *Journal of the Operational Research Society* 43 (4) 343-351.
- Hill III E D, Hill J L and Jacobs L M (1984) Planning for emergency ambulance service systems. *The Journal of Emergency Medicine* 1 331-338.
- Hill R R and McIntyre G A (2001) Applications of discrete event simulation modeling to military problems. In Peters, B A, Smith, J S, Medeiros, D J and Rohrer, M W (Eds.), *Winter Simulation Conference* (pp. 780-788). Virginia.
- Hillsman E L (1984) The p-median structure as a Unified Linear Model for location-allocation analysis. *Environment and Planning A* 16 (3) 305-318.
- Hogg J (1968) The siting of fire stations. *Operational Research Quarterly* 19 275-287.
- Holling C S (1978) *Adaptive Environmental Assessment and Management*. Chichester: Wiley.
- Holmberg M, Holmberg S, Herlitz J and Gardelov B (1998) Survival after cardiac arrest outside hospital in Sweden. *Resuscitation* 36 29-36.
- Hong N C and Ghani N A (2006) A model for predicting average ambulance service travel times in Penang Island. In *Proceedings of the 2nd IMT-GT Regional Conference on Mathematics, Statistics and Applications* (Vol. 4?). Universiti Sains Malaysia, Penang, Malaysia.
- Horn M E T (2000) Efficient modelling of travel in networks with time-varying link speeds. *Networks* 36 (2) 80-90.
- Health Protection Agency, HPA (2009) *Pandemic (H1N1) 2009 in England: An overview of initial epidemiological findings and implications for the second wave*.
- Health Protection Agency, HPA (2009) *HPA Board Meeting: Wales Report*. HPA, Health Protection Agency.
- Hughes O (2009) Crews waste 30,000 hours at A&E. *Daily Post*. Wales. 27<sup>th</sup> March 2009. p. 8.
- Hulshof P J H, Kortbeek N, Boucherie R J, Hans E W and Bakker P J M (2012) Taxonomic classification of planning decisions in health care: A structured review of the state of the art in OR/MS. *Health Systems* 1 129-175.
- Ingolfsson A (2013) EMS Planning and Management. In: Zaric, G S (ed.) *Operations Research and Health Care Policy*. Springer.
- Ingolfsson A, Budge S and Erkut E (2008) Optimal ambulance location with random delays and travel times. *Health Care Management Science* 11 (3) 262-274.
- Ingolfsson A, Erkut E and Budge S (2003) Simulation of single start station for Edmonton EMS. *Journal of the Operational Research Society* 54 736-746.
- Iwami T, Kawamura T, Hiraide A, Berg R A, Hayashi Y, Nishiuchi T, Kajino K, Yonemoto N, Yukioka H, Sugimoto H, Kakuchi H, Sase K, Yokoyama H and Nonogi H (2007) Effectiveness of bystander-initiated cardiac-only resuscitation for patients with out-of-hospital cardiac arrest. *Circulation* 116 2900-2907.

- Jenkins D (2012) *Modelling the Welsh Ambulance Service Trust for a Local Health Board*. Cardiff University.
- Jewkes E (2011) Optimal Ambulance Location Tiered Response Time Standards. *Operational Research Group Seminar Series*. Cardiff University School of Mathematics.;
- Jones A (2011) Slow ambulance turnarounds cost NHS more than £10m. *BBC News Wales*. 26<sup>th</sup> August 2011.
- Keeney R L (1972) A method for districting among facilities. *Operations Research* 20 (3) 613-618.
- Kindler E (2007) Simulation and 40 years of object-oriented programming. *World Academy of Science, Engineering and Technology* 9.
- Kleijnen J P C (1972) The statistical design and analysis of digital simulation: A survey. *Management Informatics* 1 (2) 57-66.
- Kleinbaum D G, Kupper L L, Nizam A and Muller K E (2008) *Applied Regression Analysis and other Multivariable Methods*. 4th International Student Edition ed. Brooks/Cole.
- Klugman C M (2007) Why EMS needs its own ethics. Cygnus Business Media. p. 4
- Knapp B J, Kerns B L, Riley I and Powers J (2009) EMS-initiated refusal of transport: The current state of affairs. *The Journal of Emergency Medicine* 36 (2) 157-161.
- Knight V A and Harper P R (2012) Modelling emergency medical services with phase-type distributions. *Health Systems* 1 58-68.
- Knight V A, Harper P R and Smith L (2012a) Ambulance Allocation for Maximal Survival with Heterogeneous Outcome Measures. *Omega* 40 (6) 918-926.
- Knight V A, Williams J E and Reynolds I (2012b) Modelling patient choice in healthcare systems: Development and application of a discrete event simulation with agent-based decision making. *Journal of Simulation* 6 92-102.
- Knowles E, Mason S and Colwell B (2011) An initiative to provide emergency healthcare for older people in the community: The impact on carers. *Emergency Medical Journal* 28 316-319.
- Kolesar P and Blum E H (1973) Square root laws for fire engine response distances. *Management Science* 19 (12) 1368-1378.
- Kolesar P, Walker W and Hausner J (1975) Determining the relation between fire engine travel times and travel distances in New York City. *Operations Research* 23 (4) 614-627.
- Krarup J and Pruzan P M (1990) Ingredients of Locational Analysis. In: Mirchandani, P B and Francis, R L (eds.) *Discrete Location Theory*. John Wiley & Sons Inc.
- Lane D C, Monefeldt C and Rosenhead J V (2000) Looking in the wrong place for healthcare improvements: A Systems Dynamics study of an Accident and Emergency Department. *The Journal of the Operational Research Society* 51 (5) 518-531.
- Larsen M P, Eisenberg M S, Cummins R O and Hallstrom A P (1993) Predicting survival from out-of-hospital cardiac arrest: A graphic model. *Annals of emergency Medicine* 22 (11) 1652-1658.
- Larson R C (1974) A Hypercube Queuing Model for facility location and redistricting in urban emergency services. *Computers & Operations Research* 1 (1) 67-95.
- Larson R C and Stevenson K A (1972) On insensitivities in urban redistricting and facility location. *Operations Research* 20 (3) 595-612.
- Lenoir T and Lowood H (2005) Theaters of War: The Military-Entertainment Complex. In: Schramm, H, Schwarte, L and Lazardzig, J (eds.) *Collection, Laboratory, Theater: Scenes of Knowledge in the 17th Century*. Berlin.

- Lerner E B, Rea T D, Bobrow B J, Acker III J E, Berg R A, Brooks S C, Cone D C, Gay M, Gent L M, Mears G, Nadkarni V M, O'Connor R E, Potts J, Sayre M R, Swor R A and Travers A H (2012) Emergency medical service dispatch cardiopulmonary resuscitation prearrival instructions to improve survival from out-of-hospital cardiac arrest: A scientific statement from the American Heart Association. *Circulation* 125 648-655.
- Health Commission Wales & WAST (2010) *Efficiency Review of The Welsh Ambulance Services NHS Trust*. Chair Associates, L N and Tilly, B: Health Commission Wales & WAST.
- Lin Y H, Batta R, Rogerson P A, Blatt A and Flanigan M (2012) Location of temporary depots to facilitate relief operations after an earthquake. *Socio-Economic Planning Sciences: Modelling Public Sector Facility Location Problems* 46 (2) 112-123.
- London Ambulance Service (2010) Capital's cardiac arrest survival rate goes up more than six times in a decade. London Ambulance Service NHS Trust.
- London Ambulance Service (© 2013) *Clinical Quality Indicators*. [www.londonambulance.nhs.uk/about\\_us/how\\_we\\_are\\_doing/clinical\\_quality\\_indicators.aspx](http://www.londonambulance.nhs.uk/about_us/how_we_are_doing/clinical_quality_indicators.aspx) Accessed: 10<sup>th</sup> April 2013
- Love and Morris (1972) Modelling inter-city road distances by mathematical functions. *Operational Research Quarterly* 23 (1) 61-71.
- Love R F and Morris J G (1979) Mathematical models of road travel distances. *Management Science* 25 (2) 130-139.
- Lowthian J A, Cameron P A, Stoelwinder J U, Curtis A, Currell A, Cooke M W and McNeil J J (2011) Increasing utilisation of emergency ambulances. *Australian Health Review* 35 (1) 63-69.
- Mahajan P S and Ingalls R G (2004) Evaluation of methods used to detect warm-up period in steady state simulation. In Ingalls, R G, Rossetti, M D, Smith, J S and Peters, B A (Eds.), *Winter Simulation Conference*. Washington D.C.
- WAST, (2010) *Service, Workforce and Financial Strategic Framework: Response care bundle*. WAST, .
- Majzoubi F, Bai L and Heragu S S (2012) An optimization approach for dispatching and relocating EMS vehicles. *IIE Transactions on Healthcare Systems Engineering* 2 (3) 211-223.
- Maranzana F E (1964) On the location of supply points to minimize transport costs. *Operational Research Quarterly* 15 (3) 261-270.
- Marianov V and ReVelle C (1994) The Queuing Probabilistic Location Set Covering Problem and some extensions. *Socio-Economic Planning Sciences* 28 (3) 167-178.
- Marianov V and ReVelle C (1995) Siting emergency services. *Facility Location: A survey of applications and methods* 1 199-223.
- Marianov V and ReVelle C (1996) The Queueing Maximal Availability Location Problem: A model for the siting of emergency vehicles. *European Journal of Operational Research* 93 (1) 110-120.
- Mason A J (2013) Simulation and Real-Time Optimised Relocation for Improving Ambulance Operations. *Handbook of Healthcare Operations Management*. Vol. 184.
- Matsumoto M and Nishimura T (1998) Mersenne Twister: A 623-dimensionally equidistributed Uniform pseudo-random number generator. *ACM: Transactions on Modeling and Computer Simulation* 8 (1) 3-30.
- Mayer J D (1979) Paramedic response time and survival from cardiac arrest. *Social Science and Medicine* 13D 267-271.
- McLay L A, Boone E L and Brooks J P (2011) Analyzing the volume and nature of emergency medical calls during severe weather events using regression methodologies. *Socio-Economic Planning Sciences* In Press.
- McLay L A and Mayorga M E (2010) Evaluating Emergency Medical Service performance measures. *Health Care Management Science* 13 (2) 124-136.
- Mickevicius G and Valakevicius E (2006) Modelling of non-Markovian queueing systems. *Technological and Economic Development of Economy* XII (4) 295-300.

- Microsoft Corporation (©2010) Microsoft Visual Studio.
- Miyagawa M (2009) Rectilinear distance in rotated regular point patterns. *Forma* 24 111-116.
- Monmouthshire Beacon (2008) Wife's ambulance station plea. 8<sup>th</sup> October 2008. p. 4.
- Monroe C B (1980) A simulation model for planning emergency response systems. *Social Sciences and Medicine* 14 (1) 71-77.
- Moretti M A, Cesar L A M, Nusbacher A, Kern K B, Timerman S and Ramires J A F (2007) Advanced cardiac life support training improves long-term survival from in-hospital cardiac arrest. *Resuscitation* 72 458-465.
- Morgan B J T (1984) *Elements of Simulation*. Chapman and Hall Ltd.
- Murach J (2008) *C# 2008*. Mike Murach & Associates.
- NAEMT News (2010) *The Star of Life*. <http://www.emsworld.com/article/10319356/the-star-of-life> Accessed: 23<sup>rd</sup> November 2012
- Naess A C and Steen P A (2004) Long term survival and costs per life year gained after out-of-hospital cardiac arrest. *Resuscitation* 60 57-64.
- Nance R E and Sargent R G (2002) Perspectives on the evolution of simulation. *Operations Research* 50 (1) 161-172.
- National Army Museum (2012) *War Horse: fact and fiction*. [www.nam.ac.uk/microsites/war-horse/explore/roles/pulling/people](http://www.nam.ac.uk/microsites/war-horse/explore/roles/pulling/people) Accessed: 13<sup>th</sup> September 2012
- National Audit Office (2011) Transforming NHS ambulance services. In: Health, D o ed.
- Neebe A W (1988) A procedure for locating emergency service facilities for all possible response distances. *Journal of the Operational Research Society* 39 (8) 743-748.
- Newman H (2012) 'Chain of Survival' saves Barry cardiac victim. Big Medicine.
- NHS (2009) Stroke: Act F.A.S.T. awareness campaign. *Department of Health*. Department of Health.
- NHS Choices (2011a) *NHS ambulance services*. [www.nhs.uk/nhsengland/aboutnhservices/emergencyandurgentcareservices/pages/ambulanceservices.aspx](http://www.nhs.uk/nhsengland/aboutnhservices/emergencyandurgentcareservices/pages/ambulanceservices.aspx) Accessed: 11<sup>th</sup> September 2012
- NHS Choices (2011b) *NHS History*. <http://www.nhs.uk/NHSEngland/thenhs/nhshistory/Pages/NHShistory1948.aspx> Accessed: 23<sup>rd</sup> November 2012
- NHS Choices (2011c) *The NHS Structure*. [www.nhs.uk/NHSEngland/thenhs/about/Pages/nhsstructure.aspx](http://www.nhs.uk/NHSEngland/thenhs/about/Pages/nhsstructure.aspx) Accessed: 10<sup>th</sup> September 2012
- NHS Wales (2009) *NHS in Wales: Why we are changing the structure*.
- NHS Wales (2012a) *Our Services*. [www.wales.nhs.uk/ourservices](http://www.wales.nhs.uk/ourservices) Accessed: 2<sup>nd</sup> February 2012
- NHS Wales (2012b) *Structure*. <http://www.wales.nhs.uk/nhwalesaboutus/structure> Accessed: 23<sup>rd</sup> November 2012
- NHTSA EMS (2012) *The Star of Life*. <http://www.ems.gov/star.htm> Accessed: 23<sup>rd</sup> November 2012
- Nicholl J, Hughes S, Dixon S, Turner J and Yates D (1998) The costs and benefits of paramedic skills in pre-hospital trauma care. NHS R&D Health Technology Assessment Programme.
- Nicholson H J (2001) *The Knights Hospitaller*. Boydell and Brewer.
- NISCHR (February 2013) Making and Impact: what's new in emergency prehospital care research. *Personal communication*.

- O'Keeffe C, Nicholl J, Turner J and Goodacre S (2010) Role of ambulance response times in the survival of patients with out-of-hospital cardiac arrest. *Emergency Medical Journal*.
- Office for National Statistics (2012) Population ageing in the United Kingdom, its constituent countries and the European Union.
- Ong M E H, Chiam T F, Ng F S P, Sultana P, Lim S H, Leong B S-H, Ong V Y K, Tan E C C, Tham L P, Yap S and Anantharaman V (2010) Reducing ambulance response times using geospatial-time analysis of ambulance deployment. *Academic Emergency Medicine* 17 951-957.
- Operations Research Center (1975) *Optimization in stochastic systems with distinguishable servers*. Operations Research Center.
- OPALS (2004) *Ontario Prehospital Advanced Life Support Study*. [http://www.ohri.ca/emerg/research\\_archive/opals/#Introduction](http://www.ohri.ca/emerg/research_archive/opals/#Introduction) Accessed:
- Optima (2013) *Optima Solutions*. [www.theoptimacorporation.com/optima-solutions](http://www.theoptimacorporation.com/optima-solutions) Accessed:
- Ortiz C J M (1998) The Revolutionary Flying Ambulance of Napoleon's Surgeon. *U.S. Army Medical Department Journal* 4 17-25.
- Owen S H and Daskin M S (1998) Strategic facility location: A review. *European Journal of Operational Research* 111 423-447.
- Page R L (2000) Brief history of flight simulation. In *SimTecT*. Sydney: SimTexT Organising and Technical Committee.
- Palisade Corporation (2010) *Guide to Using Evolver: The genetic algorithm solver for Microsoft Excel*. New York.
- Palisade Corporation (2011) *DecisionTools Suite*.
- Pan A S, Tian J and Wang Y (2012) A quick and efficient algorithm to the emergency supplies distribution centers location problem. *Third Global Congress on Intelligent Systems*. Beijing, China: CPS.
- Panorama (2012) Dial 999... and Wait? BBC One: BBC.
- Patel P B, Derlet R W, Vinson D R, Williams M and Wills J (2006) Ambulance diversion reduction: The Sacramento solution. *American Journal of Emergency Medicine* 24 206-213.
- Peleg K and Pliskin J S (2004) A geographic information system simulation model of EMS: Reducing ambulance response time. *American Journal of Emergency Medicine* 22 (3) 164-170.
- Penguin Reference (2003) *English Dictionary*. In: Allen, R ed. *The Penguin Dictionary*. Penguin.
- Persse D E, Key C B, Bradley R N, Miller C C and Dhingra A (2003) Cardiac arrest survival as a function of ambulance deployment strategy in a large urban Emergency Medical Services system. *Resuscitation* 59 97-104.
- Petrovic S (2010) *Heuristics and Approximation Algorithms: Evolutionary Computation*. NATCOR. University of Nottingham.
- Petwave (2012) *Siberian Husky - History and Health*. [www.petwave.com/Dogs/Dog-Breed-Center/Working-Group/Siberian-Husky/Overview.aspx](http://www.petwave.com/Dogs/Dog-Breed-Center/Working-Group/Siberian-Husky/Overview.aspx) Accessed: 17<sup>th</sup> September 2012
- Pidd M (2004) *Computer Simulation In Management Science*. 5<sup>th</sup> ed. England: John Wiley & Sons.
- Pons P T, Haukoos J S, Bludworth W, Cribley T, Pons K A and Markovchick V J (2005) Paramedic response time: Does it affect patient survival? *Academic Emergency Medicine* 12 (7) 594-600.
- Pons P T and Markovchick V J (2002) Eight minutes or less: Does the ambulance response time guideline impact trauma patient outcome? *The Journal of Emergency Medicine* 23 (1) 43-48.
- Portz K, Newell R and Archibong U (2012) Rising ambulance life-threatening call demand in high and low socioeconomic areas. *Journal of Psychological Issues in Organizational Culture* 3 (3) 5-19.

- Price L (2006) Treating the clock and not the patient: Ambulance response times and risk. *Quality and Safety in Healthcare* 15 127-130.
- Public Health Wales (2010) Flu Pandemic 2009: Surveillance data for Wales. In: Division, H P ed. NHS Wales.
- Qi C, Shu H and Xu A (2006) Formal properties of cognitive distance in geographical space. In *16th International Conference on Artificial Reality and Telexistence*. Hangzhou, China.
- Quade E D (1980) Pitfalls in Formulation and Modelling. In: Majone, G and Quade, E S (eds.) *Pitfalls of Analysis*. Chichester: Wiley & Sons.
- Rajajee V and Saver J (2005) Prehospital care of the acute stroke patient. *Techniques in Vascular and Interventional Radiology* 8 (2) 74-80.
- Rea T D, Eisenberg M S, Culley L L and Becker L (2001) Dispatcher-assisted cardiopulmonary resuscitation and survival in cardiac arrest. *Circulation* 104 2513-2516.
- Redelmeier D A, Blair P J and Collins W E (1994) No place to unload: A preliminary analysis of the prevalence, risk factors, and consequences of ambulance diversion. *Annals of Emergency Medicine* 23 (1) 43-47.
- Rennard J-P (2007) *Handbook of Research on Nature Inspired Computing for Economics and Management*. Idea Group Reference.
- Repede J F and Bernardo J J (1994) Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky. *European Journal of Operational Research* 75 567-581.
- Restrepo M, Henderson S G and Topaloglu H (2009) Erlang loss models for the static deployment of ambulances. *Health Care Management Science* 12 (1) 67-79.
- ReVelle C and Hogan K (1988) A reliability-constrained siting model with local estimates of busy fractions. *Environmental Planning 8: Plan. Des.* 15 143-152.
- ReVelle C and Hogan K (1989) The Maximum Availability Location Problem. *Transportation Science* 23 192-200.
- ReVelle C S and Eiselt H A (2005) Location Analysis: A synthesis and survey. *European Journal of Operational Research* 165 (1) 1-19.
- Robinson S (1994) *Successful Simulation: A Practical Approach to Simulation Projects*. England: McGraw-Hill Book Company.
- Robinson S (1999) Three sources of simulation inaccuracy (and how to overcome them). In Farrington, P A, Nembhard, H B, Sturrock, D T and Evans, G W (Eds.), *Winter Simulation Conference*. Phoenix, Arizona.
- Robinson S (2007) A statistical process control approach to selecting a warm-up period for a discrete-event simulation. *European Journal of Operational Research* 176 332-346.
- Robinson S, Davies R, Pidd M and Cheng R (2009) NATCOR Simulation Course. University of Warwick.
- Rosen K R (2008) The history of medical simulation. *Journal of Critical Care* 23 157-166.
- SAS (2011) Enterprise Guide. SAS Institute Inc.
- Sasaki S, Comber A J, Suzuki H and Brunson C (2010) Using genetic algorithms to optimise current and future health planning: The example of ambulance locations. *International Journal of Health Geographics* 9 (4).
- Sastry K, Goldberg D and Kendall G (2005) Genetic Algorithms. In: Burke, E K and Kendall, G (eds.) *Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques*. New York: Springer.
- Savas E S (1969) Simulation and cost-effectiveness analysis of New York's Emergency Ambulance Service. *Management Science* 15 (12) 608-627.



- Saydam C, Rajagopalan H K, Sharer E and Lawrimore-Belanger K (2013) The dynamic redeployment coverage location model. *Health Systems*.
- Sayre M R, Berg R A, Cave D M, Page R L, Potts J and White R D (2008) Hands-only (compression-only) cardiopulmonary resuscitation: A call to action for bystander response to adults who experience out-of-hospital sudden cardiac arrest: A science advisory for the public from the American Heart Association Emergency Cardiovascular Care Committee. *Circulation* 117 2162-2167.
- Schelling T C (1971) Dynamic models of segregation. *Journal of Mathematical Sociology* 1 143-186.
- Schmid V and Doerner K F (2010) Ambulance location and relocation problems with time-dependent travel times. *European Journal of Operational Research* 207 (3) 1293-1303.
- SciVerse (© 2013) *ScienceDirect.com*[Online] Elsevier.
- Setzler H, Saydam C and Park S (2009) EMS call volume predictions: A comparative study. *Computers & Operations Research* 36 1843-1851.
- Shiah D and Chen S (2007) Ambulance Allocation Capacity Model. *9th International Conference on e-Health Networking, Application and Services, 2007*.
- Silva F and Serra D (2008) Locating emergency services with different priorities: The Priority Queueing Covering Location Problem. *Journal of the Operational Research Society* 59 (9) 1229-1238.
- Simul8 (2013) 4 steps to reducing variation in healthcare and improving patient outcomes. *IHI Conference*. Simul8 Corporation.
- Singer M and Donoso P (2008) Assessing an ambulance service with queuing theory. *Computers & Operations Research* 35 (8) 2549-2560.
- Sisiopiku V P and Roupail N M (1994) Toward the use of detector output for arterial link travel time estimation: A literature review. *Transportation Research Record*.
- Skandalakis P N, Lainas P, Zoras O, Skandalakis J E and Mirilas P (2006) "To afford the wounded speedy assistance": Dominique Jean Larrey and Napoleon. *World Journal of Surgery* 30 (8) 1392-1399.
- Smith H K, Laporte G and Harper P R (2008) Locational Analysis: Highlights of growth to maturity. *The Journal of the Operational Research Society* 60 (1).
- Smith L, Harper P R, Knight V A, Vieira I and Williams J E (2011) Google Maps Travel Matrix Generator: A user guide to a travel matrix generator tool with Google Maps Application Programming Interface (API). *Operational Research, Healthcare Modelling*. Cardiff University, School of Mathematics.
- Smith R (2010) The long history of gaming in military training. *Simulation & Gaming* 41 (1) 6-19.
- South Wales Argus (2013) Ambulance response times are 'useless way' to measure success, says union. 22<sup>nd</sup> March 2013.
- Stiell I G, Wells G A, Field B, Spaite D W, Nesbitt L P, De Maio V J, Nichol G, Cousineau D, Blackburn J, Munkley D, Luinstra-Toohy L, Campeau T, Dagnone E and Lyver M (2004) Advanced cardiac life support in out-of-hospital cardiac arrest. *The New England Journal of Medicine* 351 647-656.
- Stomp W, Fidler V, ten Duis H J and Nijsten M W (2009) Relation of the weather and the lunar cycle with the incidence of trauma in the Groningen region over a 36-year period. *Journal of Trauma* 67 (5) 1103-1108.
- Studnek J R, Garvey L, Blackwell T, Vandeventer S and Ward S R (2010) Association between prehospital time intervals and ST-Elevation Myocardial Infarction system performance. *Circulation* 122 (15) 1464-1469.
- Summers G J (2004) Today's business simulation industry. *Simulation and Gaming* 35 208-241.

- Swoveland C, Uyeno D, Vertinsky I and Vickson R (1973) Ambulance Location: A probabilistic enumeration approach. *Management Science* 20 (4 Part II) 686-698.
- Thakore S, McGugan E A and Morrison W (2002) Emergency ambulance dispatch: Is there a case for triage? *Journal of the Royal Society of Medicine* 95 126-129.
- The Order of St. John (2012) *The Order of St. John working to improve the health and wellbeing of people across the world*. [www.orderofstjohn.org](http://www.orderofstjohn.org) Accessed: 13<sup>th</sup> September 2012
- The Rand Corporation (1967) *Digital computer simulation: Statistical considerations*. California: The Rand Corporation.
- Thornton S (2007) *Health for London: Are we now on the right route?* British Medical Journal 335. 14th July. p. 98.
- Toregas C, Swain R, ReVelle C and Bergman L (1971) The location of emergency service facilities. *Operations Research* 19 1363-1373.
- Trowbridge M J, Gurka M J and O'Connor R E (2009) Urban sprawl and delayed ambulance arrival in the U.S. *American Journal of Preventive Medicine* 37 (5) 428-432.
- Medical Care Research Unit, University of Sheffield (2002) *The performance of Staffordshire ambulance service: A review*. Medical Care Research Unit, University of Sheffield.
- Medical Care Research Unit (2006) *The costs and benefits of changing ambulance service response time performance standards*. University of Sheffield: Medical Care Research Unit.
- Valakevicius E (2007) The application of phase type distributions for modelling queueing systems. *Journal of Systemics, Cybernetics and Informatics* 5 (6) 28-32.
- Valenzuela T D, Kern K B, Clark L L, Berg R A, Berg M D, Berg D D, Hilwig R W, Otto C W, Newburn D and Ewy G A (2005) Interruptions of chest compressions during emergency medical systems resuscitation. *Circulation* 112 1259-1265.
- Valenzuela T D, Roe D J, Cretin S, Spaite D W and Larsen M P (1997) Estimating effectiveness of cardiac arrest interventions: A logistic regression survival model. *Circulation* 96 3308-3313.
- Valenzuela T D, Roe D J, Nichol G, Clark L L, Spaite D W and Hardman R G (2000) Outcomes of rapid defibrillation by security officers after cardiac arrest in casinos. *The New England Journal of Medicine* 343 (17) 1206-1209.
- Vandeventer S, Studnek J R, Garrett J S, Ward S R, Staley K and Blackwell T (2011) The association between ambulance hospital turnaround times and patient acuity, destination hospital and time of day. *Prehospital Emergency Care* 15 (3) 366-370.
- Vile J L, Gillard J W, Harper P R and Knight V A (2012) Predicting ambulance demand using singular spectrum analysis. *Journal of the Operational Research Society* 63 1556-1565.
- Volz R A (1971) Optimum ambulance location in semi-rural areas. *Transportation Science* 5 (2) 193-203.
- Voronoi G (1907) Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *Journal für die Reine und Angewandte Mathematik* 133 97-178.
- Wachowski A and Wachowski L (1999) *The Matrix*.
- Wales Air Ambulance (2009) [www.walesairambulance.com](http://www.walesairambulance.com) Accessed: 27<sup>th</sup> September 2012
- Wales Rural Observatory (2012) Powys. Cardiff School of Planning and Geography.
- Wankhade P (2011) Performance measurement and the UK emergency ambulance service: Unintended consequences of the ambulance response time targets. *International Journal of Public Sector Management* 24 (5) 384-402.
- Wales Audit Office (2006) *Ambulance Services in Wales*. Chair General, A: Wales Audit Office.
- WAST (2008) Field Name and Descriptions. Headquarters St. Asaph:

- WAST (2012a) *999 calls and emergency ambulances*. [www.ambulance.wales.nhs.uk/Default.aspx?pagelD=55&lan=en](http://www.ambulance.wales.nhs.uk/Default.aspx?pagelD=55&lan=en) Accessed: 20<sup>th</sup> May 2012
- WAST (2012b) *About Us*. [www.ambulance.wales.nhs.uk/Default.aspx?pagelD=8&lan=en](http://www.ambulance.wales.nhs.uk/Default.aspx?pagelD=8&lan=en) Accessed: 20<sup>th</sup> May 2012
- WAST (2012c) *Treating People Fairly*. [www.ambulance.wales.nhs.uk/Default.aspx?pagelD=30&lan=en](http://www.ambulance.wales.nhs.uk/Default.aspx?pagelD=30&lan=en) Accessed: 27<sup>th</sup> September 2012
- WAST (2012d) *Working for Us*. [www.was-tr.wales.nhs.uk/Default.aspx?pagelD=5&lan=en](http://www.was-tr.wales.nhs.uk/Default.aspx?pagelD=5&lan=en) Accessed: 4<sup>th</sup> May 2012
- WAST (March 2011) National Ambulance Performance Standards. *Personal communication*.
- Watkins J and Price I (2010) WAST South East Resource Department Meeting. *Personal communication*.
- Welch P D (1983) The Statistical Analysis of Simulation Results. In: Lavenberg, S (ed.) *The Computer Performance Modeling Handbook*. New York: Academic Press.
- Welsh Government (2013) Ambulance Services - Number of emergency ambulance calls and responses, by UA, LHB, ambulance region and Wales. In: Unit, H S a A ed. StatsWales.
- Welsh Government HSA (2012) Ambulance Services: Responses to emergency calls by LA, ambulance region and Wales.
- Welsh Government KAS (2009) Population by place of residence, area size and population density; mid-2000 onwards.
- White J and Case K (1974) On covering problems and the Central Facility Location Problem. *Geographical Analysis* 6 281-293.
- White K R J, Cobb M J and Spratt S C (2000) A comparison of five steady-state truncation heuristics for simulation. In Joines, J A, Barton, R R, Kang, K and Fishwick, P A (Eds.), *Winter Simulation Conference*. Florida.
- Wong H T and Lai P C (2012) Weather inference and daily demand for emergency ambulance services. *Emergency Medical Journal* 29 60-64.
- University of Cambridge, Computer Laboratory (2007) *Relationships for object-oriented programming languages*. Cambridge: University of Cambridge, Computer Laboratory.
- Wright P D, Liberatore M J and Nydick R L (2006) A survey of operations research models and applications in homeland security. *Interfaces* 36 (6) 514-529.
- Wrigley H, George S, Smith H, Snooks H, Glasper A and Thomas E (2002) Trends in demand for emergency ambulance services in Wiltshire over nine years: Observational study. *British Medical Journal* 324 646-647.
- Wu C H and Hwang K P (2009) Using a Discrete-Event Simulation to balance ambulance availability and demand in static deployment systems. *Academic Emergency Medicine* 16 (12) 1359-1366.
- Zayas-Caban G, Lewis M E, Olson M and Schmitz S (2013) Emergency Medical Service allocation in response to large-scale events. *IIE Transactions on Healthcare Systems Engineering* 3 (1) 57-68.
- Zuckerman D W and Horn R E (1970) *The Guide to Simulation Games for Education and Training*. Cambridge, Massachusetts: