UNIVERSITY OF CAPE TOWN

An Agent-Based Model of The Emergency Medical Services System in Nelson Mandela Bay Municipality

Minor Dissertation presented in partial fulfilment of the requirements for the degree of Master of Science specialising in Advanced Analytics and Operations Research, in the Department of Statistical Sciences.

Author:
Sky Cope

Student Number:
CPXSKY001

Supervisor: Assoc. Prof. Sheetal Silal

April 26, 2024

Plagiarism Declaration

I know the meaning of plagiarism and declare that all of the work in this dissertation, save for that which is properly acknowledged, is my own.

Signature: _____

Sky Cope.

CPXSKY001

Abstract

Inefficient EMS systems can lead to delays in accessing urgent medical care and increased mortality for critically ill or injured patients. In the Nelson Mandela Bay district of South Africa's Eastern Cape province, the public EMS system struggles to meet its own response time targets. In addition to long response times, staff and vehicles are not always allocated efficiently, as highly-skilled personnel and specialised vehicles are frequently used for responding to low priority or planned patient transport calls. This decreases the quality of medical care provided to the most critically ill patients.

The aim of this research is to improve patient outcomes in Nelson Mandela Bay's under-resourced public EMS system, which serves the majority of the local population, including those who are unable to afford private EMS. It therefore has the potential to improve access to EMS for the most underprivileged communities, and enhance healthcare equity in the region. To achieve this, the research provides decision-makers in the Eastern Cape Department of Health (ECDoH) with a set of evidence-based recommendations for reducing response times, and improving the efficiency of staff and vehicle allocations. These recommendations are sensitive to the resource-limited nature of the setting, and prioritises interventions that do not require additional staff or vehicles.

The EMS system was modelled using an agent-based simulation model, which enables multiple sources of variation in the system to be explicitly accounted for, and nuanced scenarios to be investigated. The model was built and validated using anonymised EMS call data, a smaller dataset of precise response times, and travel time estimates from Google Maps.

A key finding of this research is that the median response time of Priority 1 calls can be reduced to below the 30 minute target by implementing changes to dispatching, rerouting and prioritisation behaviour alone, and without increasing resources. These improvements come at the expense of substantial increases in median response times for lower priority calls, but these increases can be counteracted by moderately scaling up the number of staff employed. Improving the accuracy of dispatchers in triaging calls was identified as a particularly effective method of reducing response times, without considerable increases to response times for other call types.

A number of policy recommendations were formulated based on these results. These will be presented to management in the Eastern Cape Department of Health, aiming to guide policy interventions for Nelson Mandela Bay's EMS system.

Keywords: Emergency Medical Services, agent-based modelling, simulation modelling, operations research.

Acknowledgements

I am deeply grateful to my supervisor, Assoc. Prof Sheetal Silal, for guiding me through every step of this journey. Thank you for sharing your wealth of knowledge and passion for using modelling to positively affect the world, for being so generous with your time, and for pushing and encouraging me when I needed it. You are an exceptional supervisor and human.

I would also like to thank Nikhil Khanna, from the Clinton Health Access Initiative (CHAI), for encouraging me to do this research in the first place, and for giving me helpful suggestions and valuable feedback at several stages of the project. I am also grateful to CHAI and the Eastern Cape Department of Health (ECDoH) for supplying me with the main datasets used in this research.

A big thank you goes to Fadzai Munyanyi, from the Modelling and Simulation Hub, Africa (MASHA), for flying with me to Gqeberha, and helping me to capture thousands of hand-written EMS call slips over several days. Together we were able to capture far more data than I could have on my own.

I would like to thank Sibongiseni Ntengu from the ECDoH, for helping me to understand the most important aspects of Nelson Mandela Bay's EMS system, for answering my many questions, and for being an expert source for a number of parameter distributions.

I am grateful to my mother, Julia Martin, and my father Michael Cope, for the many helpful conversations we had about this work, for the constant support and encouragement, and for keeping me fed in these final few days before submitting. A special thank you goes to my mother, for meticulously proof-reading my work. My sister, Sophie Cope, also provided me with helpful and encouraging feedback at several points along the way, for which I am grateful. My cat, Mala, was also a source of infinite love and playfulness for most of this journey, and I will always be grateful for her.

Finally, I would like to thank Jared Norman from MASHA, for helping me connect to and use the fast computer called Toothless to run my results, and for developing the R Shiny application I used to generate parameter sets for the sensitivity analysis.

This research was funded in part by CHAI, and by ETDP SETA. I am grateful for this financial support. Any views expressed in this work are my own.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Emergency Medical Services (EMS) systems are mainly responsible for providing pre-hospital care to seriously ill or injured patients, and transporting them to tertiary facilities (Aboueljinane *et al.* 2013). EMS response time, measured by the time difference between a call being placed and an adequately-staffed vehicle arriving at the scene, is considered a key performance indicator for health systems. Response time has been shown to significantly affect patient outcomes, particularly for cardiac patients (Stein *et al.* 2015).

Nelson Mandela Bay Metropolitan Municipality is the setting of this research. This region has a population of 1.1 million people, according to the last census, and 1.2 million according to more recent mid-year population estimates (Stats SA 2011; Stats SA 2022). Between 2019 and 2021, the region received an average of over 998 EMS calls per week. These calls have three priority levels (Priority 1, 2 and 3), and two call types (Primary and Planned Patient Transport).

In the Eastern Cape, provincial targets for EMS response times are 30 minutes in urban areas and 60 minutes in rural areas for the most urgent Priority 1 calls (Kamnqa 2023). However, in the urban Nelson Mandela Bay region, these targets are rarely met in practice. In addition, the dispatching of staff and vehicles is not optimal. For example, Priority 1 calls are not always attended to by the most qualified EMS personnel, and Planned Patient Transport calls are at times responded to by ambulances instead of the more appropriate patient transport vehicles. This combination of issues adversely affects the quality of Emergency Medical Services delivered in the region, and collectively motivated this research.

1.2 Research Aims and Objectives

The research aims are what I hope to achieve through this work, and the research objectives are the steps taken to achieve these aims.

1.2.1 Aims

The Clinton Health Access Initiative (CHAI) and the Eastern Cape Department of Health (ECDoH) identified the need for evidence-based recommendations aimed at improving efficiency in Nelson Mandela Bay's EMS system. As part of an internship for the STA-CHAI Fellowship Programme, I was given a number of datasets collected by dispatchers in the region, with the hope that I would use the data to derive helpful recommendations using a modelling approach.

The main aim of this research is to identify ways to improve median response times for the most urgent calls in the Nelson Mandela Bay region. The most urgent calls are those that are classified as code Red or Blue by EMS personnel at the scene, and a good but imperfect proxy of call urgency is the priority assigned to each call by dispatchers. A second aim is to identify ways of increasing the probability that appropriate staff mixtures and vehicle types respond to calls. This is because it is not only essential that urgent calls are responded to in a timely

manner, but also that the staff and vehicles that arrive are capable of providing a high quality of care.

1.2.2 Objectives

The objectives of this research are summarised as follows:

- Collect and clean the data for EMS calls and current EMS staff and vehicle resources in the region. In addition, through consultation with experts in the ECDoH, gather information on the rules and behaviour followed by staff and dispatchers in responding to calls.
- Conduct an exploratory analysis of the data in order to form a clearer understanding of how the system works in practice.
- Build a conceptual model for the EMS system.
- Implement the model in code.
- Conduct verification, such that we can be reasonably sure the computer model behaves like the conceptual model.
- Conduct validation, such that we can be reasonably sure the conceptual model, represented by the computer model, behaves like the real-world system.
- Use the model to investigate a number of carefully-chosen scenarios, from which a set of recommendations can be derived.

1.3 Methods

This section outlines in more detail and contextualises some of the specific methodological decisions taken by this research.

To identify possible interventions for improving efficiency in Nelson Mandela Bay's EMS system, one approach is to conduct randomised experiments for a range of interventions, and assess the effects of each in terms of performance measures of interest. However, such an approach would be extremely costly and time-consuming, and therefore not feasible in practice.

Another approach – the one taken by this research – is to build a model of the system, use this model to conduct *in silico* experiments rapidly and cheaply, and use the results of these experiments to derive recommendations for improving the real-world system. Integer programming and simulation are the two main methods used in the literature for modelling EMS systems. I chose to build an agent-based simulation model due to the flexibility and interpretibility of this approach.

Call rates were estimated for each ward in the region. These rates vary depending on the time of day, day of the week and the type of call (which is either Priority 1, Priority 2 or Planned Patient Transport[1]). Incoming calls were simulated using these call rates. The model includes four types of agents, each with unique features and rules that govern their behaviour. These include vehicles (Ambulances and Patient Transport Vehicles), EMS personnel (Independent Practitioners and Supervised Practitioners), patients (differentiated by call type) and dispatchers, who coordinate the system's response to calls. Data from Google Maps were used for estimating inter-ward travel times. The empirical distribution of destination facilities for each

[1]These are abbreviated as P1, P2 and PPT, respectively.

incident location and call type was used by the model to select destination facilities. A number of more nuanced features of the system were also modelled, including the re-routing of vehicles to high priority incidents, and the call classification and prioritisation behaviour of dispatchers, among others.

This research is significant for its geographical and methodological contributions to the EMS modelling literature. To my knowledge, it is the first to apply a simulation model to any EMS system in the Eastern Cape, and the second in South Africa, after the work of Stein *et al.* (2015). In addition, because simulation models are rarely applied to EMS systems in Low and Middle Income Countries (LMICs), this research addresses a number of methodological challenges that are not widely seen in the literature, but have wide applicability.

1.4 Chapter Overview

This dissertation is organised into seven chapters. The following is a one sentence summary of each chapter's aims.

- **Chapter 2 – Literature Review**: contextualises the research by outlining the main themes and summarising some of the most influential studies in the EMS modelling literature.
- **Chapter 3 – Data**: describes the main datasets used to construct and validate the model, and explores these datasets with an exploratory data analysis.
- **Chapter 4 – Methods**: introduces the agents and algorithms used by the simulation model, and then describes the steps taken to verify and validate it.
- **Chapter 5 – Results**: presents the results of the verification, validation, scenario analysis and sensitivity analysis.
- **Chapter 6 – Discussion**: interprets the results, presents a set of recommendations, outlines the contributions of this research to the EMS modelling literature, highlights limitations and suggests directions for future research.
- **Chapter 7 – Conclusion**: summarises the main results and their significance, and presents final reflections.

Chapter 2

Literature Review

In this chapter, the two main EMS modelling paradigms are discussed, followed by a comparison of the two. Section 2.1 outlines the main integer programming approaches, section 2.2 outlines the main simulation approaches and section 2.3 briefly discusses the benefits and drawbacks associated with each class of approaches, and finally section 2.4 discusses the limited South African EMS simulation modelling literature. This chapter is by no means a systematic review of the EMS modelling literature – it aims only to characterise the main features of the field, and to highlight particularly significant studies.

2.1 Integer Programming Models

Spatial EMS demand is typically aggregated to discrete 'demand points' to simplify the problem (Li *et al.* 2011). A large number of deterministic models of EMS systems optimise for some measure of coverage, where a demand point is considered covered if it is less than a given distance or expected travel time from at least one EMS base (Li *et al.* 2011; Janosikova *et al.* 2021). The earliest examples of EMS coverage models are the Location Set Coverage Problem (LSCP) by Toregas *et al.* (1971) and the Maximum Coverage Location Problem (MCLP) by Church & ReVelle (1974). These models, when used for modelling EMS systems, aim to find optimal locations of EMS facilities, where optimality is measured using a coverage criterion (Li *et al.* 2011).

In the LSCP model, the objective function is the number of facilities in the system (Toregas *et al.* 1971). This is minimised subject to the constraint that there is full coverage in the region, where all demand points are reachable by at least one facility within a specified time or distance (Toregas *et al.* 1971). The purpose of the model is to find 'good' locations of facilities that are linked to demand points on a network (Toregas *et al.* 1971). It is therefore a model that can be applied generally to similar problems, and is not designed for EMS systems specifically (Toregas *et al.* 1971). One limitation of this model as applied to the EMS location problem is that in practice it is usually costly to obtain full coverage of a given region (Li *et al.* 2011). In some cases, a subset of the demand points are widely dispersed, and the solutions generated by the LSCP model have large numbers of facilities (Li *et al.* 2011).

In the MCLP model, instead of minimising the number of facilities required to obtain full coverage, the number of covered demand points is maximised, given a specified number of facilities (Church & ReVelle 1974). This is an improvement on LSCP, as decision-makers can specify the number of facilities, potentially resulting in solutions that are less costly to implement (Li *et al.* 2011). Additionally, Church & ReVelle (1974) weight each demand point by population or call volume, and maximise the weighted sum of demand points. The objective function is then given as simply:

$$\max \sum_{i \in \mathcal{I}} \varphi_i x_i,$$

where $\mathcal{I}$ is the set of demand points, φ_i is the demand at point i (measured by, for example the population at point i or the average number of calls received at point i per day), and x_i is a

binary variable which is 1 if demand point i is reachable by at least one facility within a given time or distance threshold, and 0 otherwise (Chen *et al.* 2021).

A limitation of both LSCP and MCLP approaches is that they do not account for the number of vehicles assigned to a facility, but only the locations of the facilities: when only one vehicle is assigned to a facility, the surrounding demand points could lose coverage when the vehicle responds to a call (Li *et al.* 2011). There is an implicit assumption that vehicles are always available to respond to calls, which can potentially result in the covered population being overestimated (Saydam *et al.* 1994).

A number of approaches have been taken to overcome this issue (Li *et al.* 2011). One approach is to maximise the population that is covered by at least two facilities, first seen in the Double Coverage model by Hogan & Revelle (1986). Gendreau *et al.* (1997) modified the Double Coverage model slightly by maximising the population covered at least twice within a small radius r_1, while adding the constraint that all demand points must be covered within a larger radius r_2.

Another class of approaches explicitly incorporate the probability that a facility is available to respond to a call within the specified time (Li *et al.* 2011; Berg & Essen 2019). For example, the Maximum Expected Coverage Location Problem (MEXCLP) model developed by Daskin (1983). In this model, the 'busy fraction' of facilities in the system is taken into account, and the expected coverage is maximised. The busy fraction is independent probability that each facility is busy and not able to dispatch a vehicle to a given call. The MEXCLP objective function is given as:

$$\max \sum_{i \in \mathcal{I}} \sum_{k=1}^{p} \varphi_i (1-\rho) \rho^{k-1} x_{ik},$$

where $\mathcal{I}$ is the set of demand points, p is the total number of facilities, φ_i is the demand at point i, ρ is the busy fraction, and x_{ik} is a binary variable which is 1 if at least k facilities can potentially cover demand point i, and 0 otherwise (Daskin 1983; Li *et al.* 2011).

While the MEXCLP model accounts for uncertainty in facility availability, Batta *et al.* (1989) identified some overly-simplifying assumptions. These include the assumption of equal busy probabilities at all facilities, regardless of location, and the assumption that facilities, or 'servers', operate independently. Batta *et al.* (1989) then relaxed these assumptions in the Adjusted Maximum Expected Coverage Location Problem (AMEXCLP) model. This uses the hypercube queueing model developed by Larson (1975) to account for dependence between servers and differing busy probabilities, while optimising for an adjusted estimate of expected coverage (Batta *et al.* 1989).

Many other extensions of the MEXCLP model have been made since its publication (Li *et al.* 2011). For example, McLay (2009) introduced MEXCLP2, which also uses a hypercube queueing model, but allows for multiple types of servers and multiple priority levels in the demand. While integer programming EMS models have typically been focused on optimising facility locations based on some measure of coverage, other decision variables and objectives have also been used. For example, Toro-Diaz *et al.* (2013) used an expected coverage model with an embedded hypercube queueing model to explore differing location and vehicle dispatching decisions simultaneously. Erdogan *et al.* (2010) used a modified MEXCLP model to explore both vehicle locations and staff shifts. The procedure used by Erdogan *et al.* (2010) has two steps, where the assignment of vehicles to existing facilities is first selected, followed by the assignment of staff to working schedules (Erdogan *et al.* 2010). Both vehicle locations and staff shifts are chosen in order to maximise expected coverage (Erdogan *et al.* 2010).

Alternative approaches to maximising coverage include studies that maximise expected survival (Erkut *et al.* 2007; Knight *et al.* 2012; Wajid & Nezamuddin 2022) and those that minimise staffing costs (Vile *et al.* 2016; Vermuyten *et al.* 2018; Horvat *et al.* 2020). Knight *et al.* (2012) introduced the Maximal Expected Survival Location Model for Heterogeneous Patients (MESLMHP), which allows for multiple patient types, and models their survival probability as a function of response time. Vile *et al.* (2016) developed a decision support system (DSS) for the Welsh Ambulance Service Trust, with the aim of improving staff scheduling. The DSS first provides an hourly demand forecast, over two priority classes (Priority 1 and lower priority calls), then determines the hourly staffing capacity required to meet demand, and finally returns a staffing schedule that minimises the number of staff required to meet demand. Vile *et al.* (2016) used Singular Spectrum Analysis to forecast hourly demand, and a queueing model was used to convert the demand into minimum staffing requirements for each hour. An integer programming model was then used to allocate staff to shifts, in order to meet the minimum staffing requirements at each time point while minimising total labour hours used (Vile *et al.* 2016).

2.2 Simulation Models

In the EMS modelling literature, a large number of studies choose simulation models over integer programming: a review by Aboueljinane *et al.* (2013) found that 83 simulation models of EMS systems were published between 1969 – 2013, and this number has likely grown substantially since then.

The first EMS simulation model was developed by Savas (1969) who, in the year of the Apollo 11 moon landing, described operations research as a "space age" method. Savas (1969) modelled the EMS system in New York City, by including a queue for when no vehicles are available, a stochastic spatial demand distribution from which calls are generated, and a fairly complex set of rules governing the behaviour of dispatchers and vehicles. This level of complexity and realism was arguably far ahead of the LSCP, MCLP and even MEXCLP integer programming models, all of which were published in later years (Toregas *et al.* 1971; Church & ReVelle 1974; Daskin 1983).

Since the work of Savas (1969), a number of different approaches have been used for simulating EMS systems, without following a clear trajectory of gradually improving on a core set of models over time (Aboueljinane *et al.* 2013). In this regard, it differs from the integer programming literature, where the progression of the field follows a clearer narrative. The simulation literature is therefore best discussed in terms of the different modelling decisions that have been taken, including how the calls, travel times and destination facilities have been simulated, and the types of scenarios explored (Aboueljinane *et al.* 2013).

When EMS demand is simulated, it is typically aggregated to the level of districts or wards in order to simplify the problem (Yang *et al.* 2019). Approaches differ in terms of whether different call types (for example, priority levels) or time-varying call rates are incorporated into the call generating process (Aboueljinane *et al.* 2013). For example, in the simulation model developed by Ingolfsson *et al.* (2003), call rates vary depending on the demand point, time of day, day of week, and call priority; however, in the model developed by Berlin & Liebman (1974), demand varies depending on the location only, and no call types are specified. As an exception to the majority of the literature that aggregates demand to discrete points, Yang *et al.* (2019) used a different approach, and instead simulated calls from a continuous spatio-temporal distribution of demand. Yang *et al.* (2019) had a dataset with precise GPS coordinates for each call, which enabled a Gaussian mixture model to be fit to the data and sampled from in the simulation

model (Yang *et al.* 2019).

Vehicle travel times have been incorporated in a number of ways (Aboueljinane *et al.* 2013). Simple approaches to travel time estimation have used Euclidean distance-based estimates, where the straight line distances between points are multiplied by a constant (Fujiwara *et al.* 1987; Silva & Pinto 2010). Lubicz & Mielczarek (1987) used a more sophisticated approach by creating a distance matrix using estimated shortest paths on the actual road network. Similarly, Boutilier & Chan (2020) modelled a detailed road network for Dhaka, Bangladesh, and gathered data using custom-made GPS devices for estimating travel times accurately, depending on the time of day and level of traffic (Boutilier & Chan 2020). The authors compared a deterministic mixed integer linear programming model with a simulation model (Boutilier & Chan 2020). They also used a regularised regression model to predict EMS demand in areas without sufficient data (Boutilier & Chan 2020).

Sampling of destination facilities is another way in which EMS simulation models differ. This refers to how the destination is chosen for transporting each patient. There are a number of different approaches for simulating this decision, a common one being selecting the facility that is nearest to the incident (Lee *et al.* 2012; van Buuren *et al.* 2012). In the model developed by Repede & Bernardo (1994), the destination is selected using the empirical distribution of destinations.

The simulation modelling literature explores a large number of scenarios. One kind of scenario involves increasing the resources available: for example, Savas (1969) investigated the effects of increasing the number of vehicles assigned to a particular hospital. Other approaches investigate more new nuanced scenarios, like changes to destination facility selection behaviour (Wears & Winton 1993) and changes to the rules governing how vehicles or staff are dispatched (Repede & Bernardo 1994).

Another important way in which in EMS simulation models differ is with regard to the type of model used, which is typically either a Discrete Event Simulation (DES) model or an Agent Based Simulation (ABS) model (Aboueljinane *et al.* 2013). Both model classes have been used widely in the operations research literature for simulating complex systems and modelling human behaviour (Majid, Fakhreldin, *et al.* 2016). In DES, time passes in discrete intervals, and the state of the system changes over time according to a specified set of rules (Majid, Aickelin, *et al.* 2009). In ABS, time also typically progresses discretely, but an important difference is the inclusion of agents capable of interacting and communicating with one another (Majid, Fakhreldin, *et al.* 2016). While DES models have been shown to be capable of simulating complex human behaviour well (for example, in the work of Brailsford *et al.* (2006)) ABS models are considered to be more appropriate for modelling large-scale systems (Majid, Fakhreldin, *et al.* 2016). Aringhieri (2007) used an ABS model to simulate the behaviour of an EMS system in Milan, Italy. In this model, there are three types of agents: vehicles, the operation centre and the calls themselves (Aringhieri 2007). These agents interact and communicate in that the operation centre receives the calls and responds to them by making probabilistic decisions about which vehicles to dispatch (Aringhieri 2007). The vehicles in turn respond to the decisions made by the operation centre (Aringhieri 2007).

2.3 Benefits and Drawbacks of Each Approach

Integer programming models are valuable tools for optimising estimated measures of system performance such as expected coverage, given a limited range of input variables such as vehicle placements (Ridler *et al.* 2022). However, these approaches attempt to model complex EMS systems, which in reality involve multiple sources of stochasticity and interacting behaviour,

using an analytic approach with many simplifying assumptions (Yang *et al.* 2019). These assumptions limit the range of scenarios that can be considered for improving the system, and in some cases produce allocations that are inefficent in reality (Unluyurt & Tuncer 2016; Yang *et al.* 2019; Ridler *et al.* 2022).

Simulation models, by contrast, are capable of modelling the behaviour of a system to an arbitrary level of complexity (Klugl 2016). This potentially enables such models to closely approximate the underlying behaviour of the system, while accounting for the important sources of uncertainty and interactions (Yang *et al.* 2019). The flexibility of simulation models also potentially enables them to explore a wider range of nuanced scenarios for improving efficiency, as compared with integer programming approaches. In addition, Henderson & Mason (2004) note that simulation models are generally easier to explain to decision-makers than, for example, a MEXCLP model with an embedded hypercube queueing model. A drawback of the simulation approach is computational cost, since it typically takes longer to evaluate solutions generated by simulation models compared with those generated by integer programming ones (Ridler *et al.* 2022).

2.4 South African Literature

To identify EMS modelling studies conducted in South Africa, a search was performed using PubMed and Google Scholar, with the following search string: "South Africa" modelling emergency medical services ambulance. This revealed two relevant publications, by Stein *et al.* (2015) and Stassen *et al.* (2023). Stein *et al.* (2015) authors modelled the EMS system in Cape Town, South Africa, using a discrete event simulation model with multiple call priorities and multiple vehicle types, and data from the Emergency Control Centre's Computer Aided Dispatch (CAD) system (Stein *et al.* 2015). Stein *et al.* (2015) used four scenarios: two investigated the effects of different vehicle locations, and another two investigated changes to vehicle dispatching rules. Stassen *et al.* (2023) used a simulation model for Helicopter EMS (HEMS) services in Northern South Africa, and aimed to determine the distance beyond which HEMS would be faster than ground EMS. As the work of Stassen *et al.* (2023) is not focused on the optimisation of ground EMS specifically, it is less relevant for this research.

The fact that the work by Stein *et al.* (2015) is the first and only model of its kind conducted in South Africa is perhaps unsurprising, given the limited number of EMS simulation studies that are focused on LMICs in the international EMS simulation modelling literature. In a review of the literature published between 1969-2013 Aboueljinane *et al.* (2013) identified only one simulatation study that modelled an EMS system based in a LMIC, based on OECD classifications made at the time of each study's publication (OECD 2023). This study was the work of Fujiwara *et al.* (1987) which applied a simulation model to the EMS system in Bangkok, Thailand. Another more recent example of an EMS simulation model applied to an LMIC setting is the work of Boutilier & Chan (2020), discussed in section 2.2.

Chapter 3

Data

In this chapter, the three datasets used to construct and validate the model are first described in section 3.1. Then, the procedures used to collect and clean the data are outlined in section 3.2.2. Finally, the datasets are explored using various exploratory data analysis techniques in section 3.3.

3.1 Datasets

The three main datasets used in this research include call centre spreadsheets, a dataset I collected of call times and granular response times, and inter-ward travel times obtained from Google Maps. Two additional datasets, the staff and vehicle registers, were used to add staff qualifications and vehicle types to the dataset. The following subsections contain descriptions of each dataset.

Data collection practices were approved by the ethics boards of UCT's Science Facilty and CHAI. Approval letters are included in the Appendix.

3.1.1 Call Centre Spreadsheets

The Nelson Mandela Bay EMS system has electronically captured data for each call received for several years. I was provided with spreadsheets for each month from January 2019 to May 2021. A total of 154 267 unique calls are captured in these spreadsheets. All variables in this dataset are described in Table 3.1. It must be emphasised that this dataset is anonymised.

Variable name	**Description**
Date	Date on which the call was placed.
Call number	Number used to identify calls. Resets to 1 each day.
Code	Numeric incident classification.
Priority	Priority assigned to call.
Colour	Triage classification, according to EMS personnel at scene.
Time	Response time, discretised into 15-minute intervals.
F/T	Indicates whether patient was at the scene when EMS personnel arrived.
From	Suburb or clinic where the ambulance was requested.
To	Suburb or clinic where the patient was transported.
Reason	Longer typed reason for the call.
Calltaker	Name of calltaker.
Dispatcher	Name of dispatcher.
Ambulance	Identifying code of the ambulance.
Crews	Names of responding EMS personnel.
Data	Name of data capturer.
CRO	Name of centre manager.

Table 3.1: Variable names and descriptions found in the call centre dataset.

3.1.2 Collected Dataset

Together with Ms. Fadzai Munyanyi from MASHA, I collected a second dataset. The purpose was to assess congruency between the electronically captured records and the detailed paper call slips. In addition, because the time of day each call was placed is not included in the call centre dataset, I set out to collect these data, for estimating the intra-day temporal distribution of demand. Variables captured include date, call number, the time of day the call was placed, and time of arrival on the scene. We collected a total of 3 762 observations. Section 3.2.1 outlines the data collection process.

3.1.3 Inter-Ward Travel Times

As explained in section 4.2.5, inter-ward travel times were estimated using Google Maps. A set of 60 ward centroid coordinates for the Nelson Mandela Bay region (shown in Figure 3.4) was used as inputs to the Google Maps Distance Matrix API, and the estimated travel times between each coordinate were recorded in a 60×60 matrix (shown in Figure A.1 in the Appendix).

3.2 Collecting and Cleaning

3.2.1 Data Collection

The dataset outlined in section 3.1.2 was collected over several days in the Nelson Mandela Bay EMS archive, located in Dora Nginza Hospital. Hand-written paper call slips – one for each call – were transcribed and captured using Microsoft Excel.

Dates were sampled randomly for capturing from the call centre dataset, with dates from March 27th 2020 to April 30th 2020 excluded due to potentially distorting effects of the level 5 lockdown (Potgieter *et al.* 2021). The total number of calls captured by each date is shown in Figure 3.1:

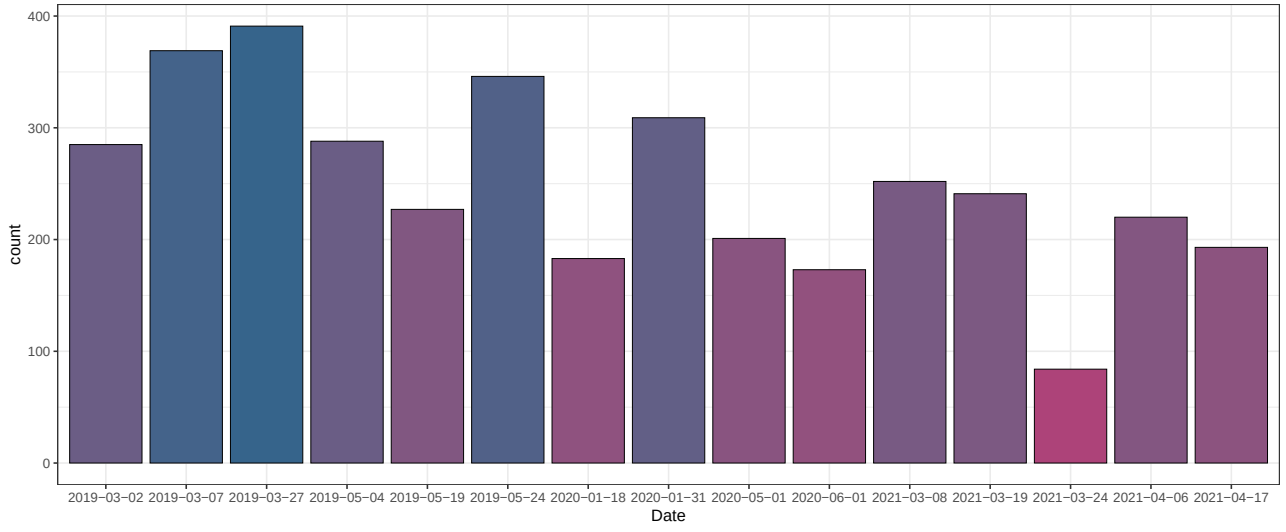


Figure 3.1: Total Number of Calls Captured, by Date.

There is clearly some variation in the number of calls captured, by call date [1]. However, we did capture all available call slips for each date shown in Figure 3.1. The dataset we captured was merged with the call centre dataset, using date and call number as matching variables.

[1] The same notably low number of calls on 2021-03-24 was also observed in the call centre dataset.

3.2.2 Data Cleaning and Integration

A number of data cleaning and integration steps were undertaken. The following is a list of these steps:

- For allowing into a single dataset, variable names were standardised between call centre spreadsheets.
- Rows with all of the following features were removed: invalid or missing date, invalid or missing time, and invalid or missing priority. Since this research does not consider choppers or ambulance transport to Cape Town, these calls were also removed.
- To have a single variable for Call Type, the Call Priority and PPT Status variables were collapsed into a new variable. Call Type is set to PPT for Planned Patient Transport calls, if not it is set to P1 for Priority 1 calls, and otherwise it is set to P2.
- In order to add staff types, staff qualifications were added by matching the Crew names on names in a staff register spreadsheet. Staff qualifications were collapsed into two categories: Independent Practitioners (IPs) and Supervised Practitioners (SPs). Personnel with Basic Life Support (BLS) qualifiations were considered SPs, while other personnel were considered IPs. These classifications are from the Health Professions Council of South Africa (HSPCA 2019).
- For including vehicle types, vehicle registrations were used to merge the vehicle asset register with the cell centre dataset. The vehicle type variable was simplified into two types. Patient Transport Vehicles (PTVs) were left unchanged, and other vehicle types (Ambulance, Response, and Rescue) were consolidated to the Ambulance type. Administrative (also called White Fleet), vehicles were excluded.
- For adding a column that classifies calls as Planned Patient Transport (PPT) or Primary calls, I classified calls as **PPT** if either of the the following conditions were met:

  1. The Reason column contains the keywords `renal`, `dialysis` or `home` **and** the Priority is not P1.
  2. The Code column is either `31` (Inter-Facility Transfer) or `34` (Discharges), **and** the Priority is not P1.

  If neither of these conditions were met, calls were classified as **Primary**.

Additionally, more precise locations were required in order to infer the spatial distribution of demand. Originally locations were presented in the dataset as manually-entered locations, which were not standardised. For collapsing multiple spellings of the same location (for example "GUSTAV LAMORE" and "GUSTAF LAMORE") I used the OpenRefine R package for fuzzy matching (Metaweb Technologies, Inc 2022).

Then, I used Google Maps to extract approximate coordinates for as many text locations as possible. I sorted the text locations in order of the frequency of appearances in the dataset, and worked through this list. Many clinics, hospitals, suburbs and townships are searchable on Google Maps. For example, Google Maps knows the perimeter of Motherwell Neighbourhood 1 (NU 1), a frequently-appearing location in the dataset. A screenshot showing this perimeter is presented in Figure 3.2. The coordinates for the Motherwell NU 1 region can then be obtained by finding the coordinates of the approximate centroid of this region.

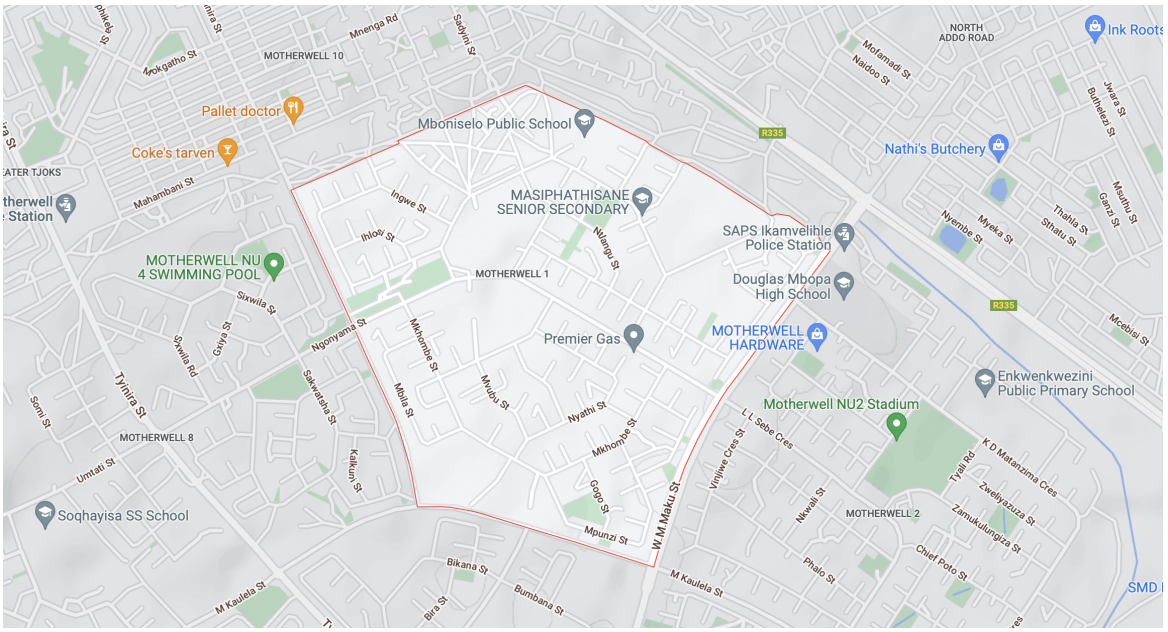


Figure 3.2: Google Maps interface, used for assigning GPS coordinates to location names.

Snapshots of raw data from the EMS call centre and vehicle register spreadsheets are shown in the Appendix (Figures A.7 and A.6 respectively). This process enabled 99 262 calls (64%) to have approximate coordinates for incident locations. Because locations were ranked from most to least frequent, the procedure became increasingly time-consuming as more calls were assigned approximate coordinates, due to low frequencies in the remaining locations. In addition, many of the remaining locations were imprecise, making it impossible to assign even approximate coordinates (for example, 498 calls listed 'HOME' as the pickup location). This final dataset included calls from January 1st 2019 to June 3rd 2021.

3.3 Exploratory Data Analysis

3.3.1 Demand Distribution

Demand for EMS services has both spatial and temporal components.

Spatial Demand

For spatial demand, the approximate coordinates obtained using Google Maps were used to plot the spatial distribution of EMS demand, in Figure 3.3:

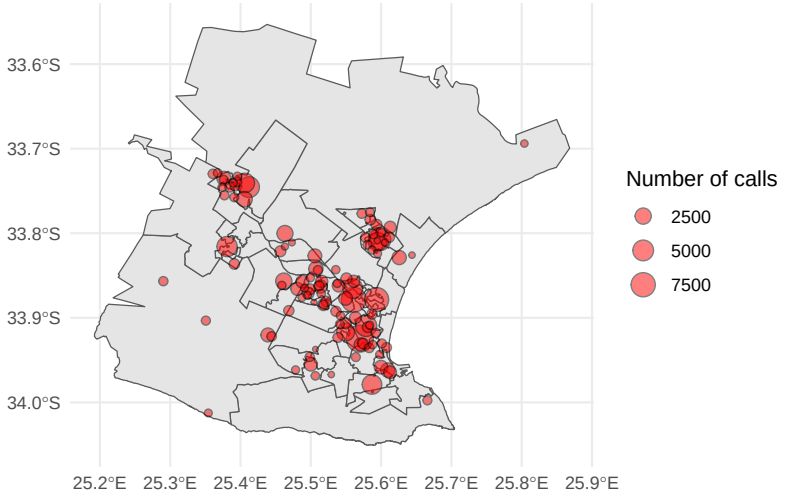


Figure 3.3: Overall spatial EMS demand distribution.

As expected, there are some clusters of demand. There are three main clusters: around the top left, in Uitenhage, around the top right in Motherwell and around the centre in Gqeberha.

In this research, instead of having a large number of spatial coordinates between which to estimate travel times, the spatial data shown in Figure 3.3 are aggregated to the level of wards. Each ward and ward ID is plotted in Figure 3.4:

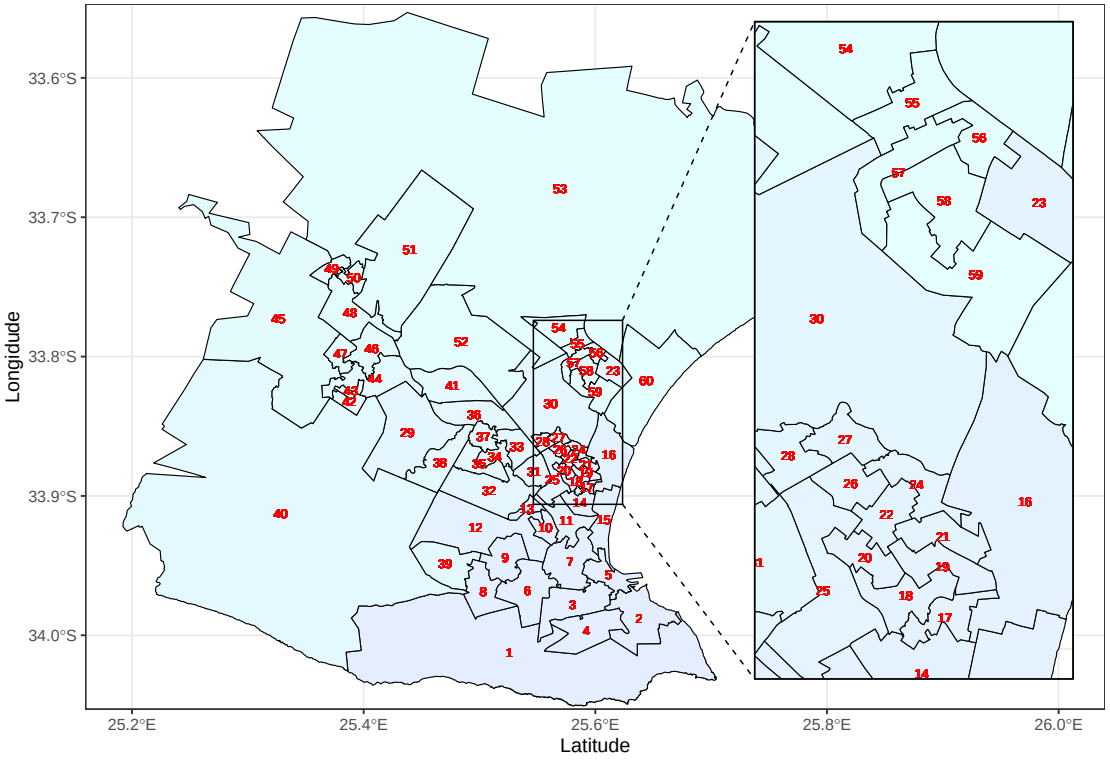


Figure 3.4: Locations of each ward in Nelson Mandela Bay, and ward IDs used in this research. Labels are plotted at the centroid of each ward. Ward shapefiles obtained from MDB (2020).

This allows one to plot the spatial demand distribution, aggregated to the level of wards, shown in Figure 3.5.

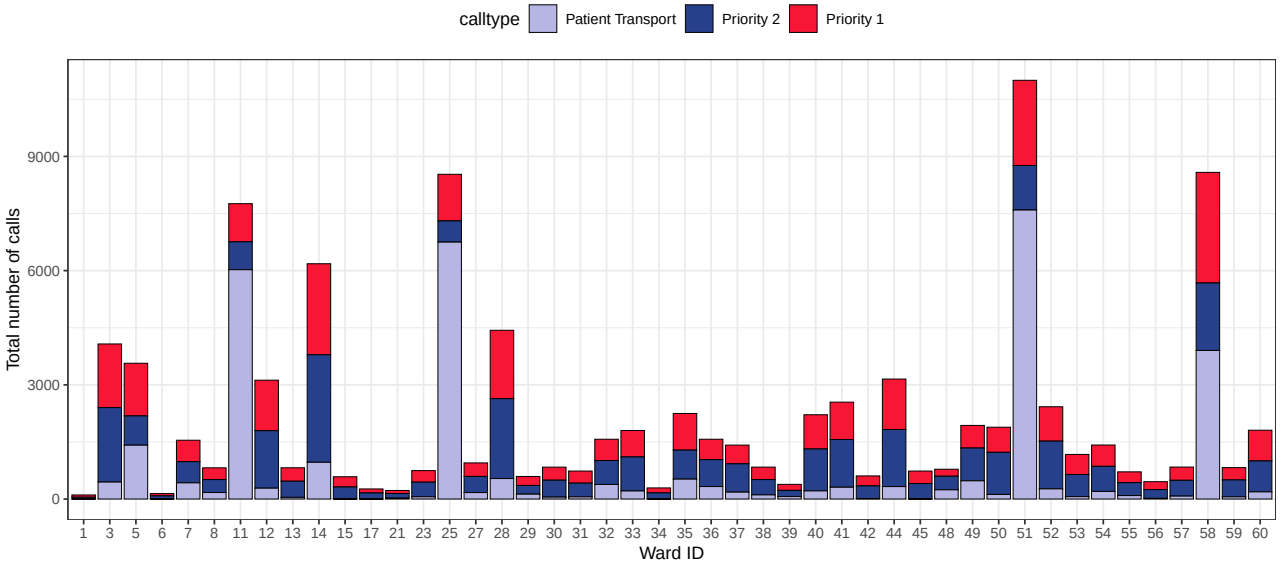


Figure 3.5: Distribution of Incident Ward IDs, by Call Type.

There are wards with notably high demand: Ward 51 (Uitenhage), Ward 58 (Motherwell), and then a number of other wards closer to central Gqeberha. There is also a striking variation in spatial EMS demand depending on the call type. For example, Ward 51 contains Uitenhage Provincial Hospital (UPH), and the majority of calls received in this ward are Patient Transport

calls. This may be the result of inter-facility transfers from UPH. It is therefore essential that spatial demand is differentiated by call type.

Temporal Demand

Temporal EMS demand refers to how call volume typically varies within days and between days, for each call type. Figure 3.6 shows the inter-day call volume distribution, for each call type.

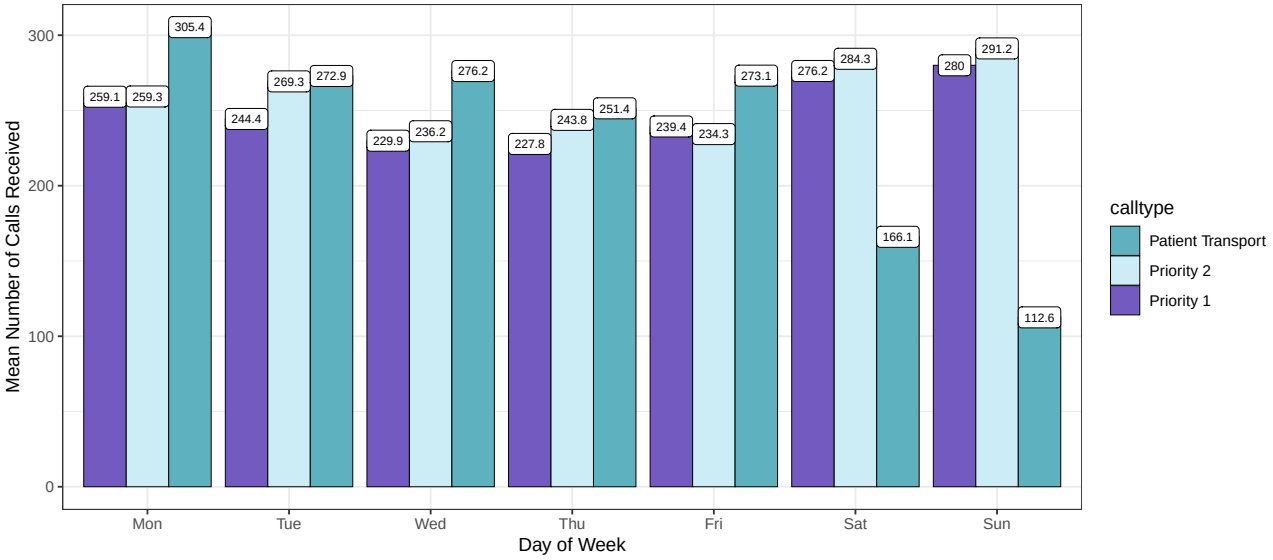


Figure 3.6: Inter-day EMS Demand Distribution, broken down by call type.

P1 and P2 call volume is highest on Saturdays and Sundays, while this is when Patient Transport call volume is at its lowest. Patient Transport call volume typically peaks on Mondays, when some clinics re-open after the weekend.

For intra-day call volume, the question is how call volume typically varies within a given day, for each call type. For this, the call times we collected were used, since the time of day each call was received was not recorded in the original call centre dataset. For each call type, the number of calls received at each hour of the day was counted, for each day. These counts were then smoothed using a cyclic cubic spline, fitted using the `mgcv` package in R. These profiles were scaled by the daily mean call rates for each call type, such that the integral of each profile equaled the totals in Figure 3.6 for each day and call type. All profiles were then scaled to provide the mean number of calls received per 10-minute window, per day and call type [2]. The final profiles are shown in Figure 3.7. An assumption of this approach is that the intra-day distributions do not differ in shape between days. In reality, call volumes may peak at slightly different times on weekends compared to weekdays, for example. However, we did not have enough data for each day of the week to be able to estimate the differing intra-day demand distributions with confidence.

[2]This is because, as explained in section 4.2, time passes in the agent-based model with discrete 10-minute intervals.

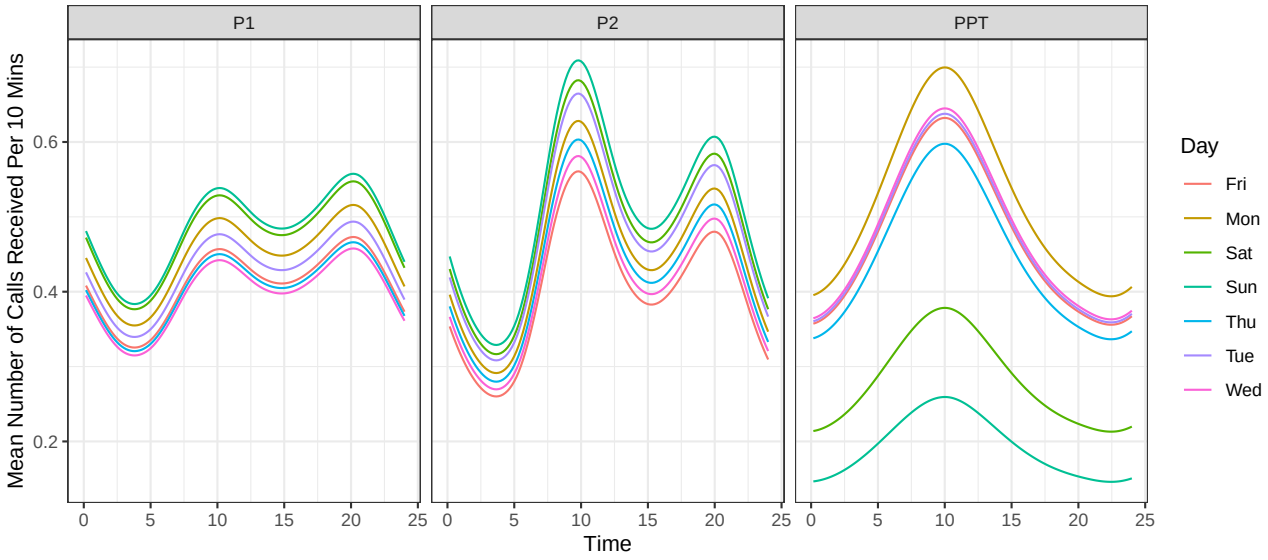


Figure 3.7: Intra-day call volume distribution, by call type and day of week. Profiles are generated using cyclic cubic splines.

The intra-day call distributions show that call volume tends to peak around 10h00 for all call types. For P1 and P2 calls, there is another peak around 20h00, but not for PPT calls.

3.3.2 Supply Distribution

The supply of services in an EMS system is composed of the vehicles and staff operating within it. Vehicles and staff are allocated to a limited number of bases (or sites). In Nelson Mandela Bay, vehicles are stationed at these bases but can theoretically be dispatched to a call anywhere else in the region. Once a vehicle has handed over a patient at a facility, it returns to the base. The locations of these bases and the numbers of staff and vehicles assigned to each base are therefore important aspects of the system.

Data were limited in this regard, as the call centre spreadsheet did not include data for the bases from which the staff and vehicles were dispatched. However, a Daily Operational Status spreadsheet (shown in Figure A.4 in the Appendix) was provided for March 10th 2022, which details the numbers of staff and vehicles assigned to each area on that day. The total numbers of staff and vehicles on duty in this Daily Operational Status (DOS) closely matched long term averages estimated from the call centre spreadsheets, implying that the DOS I was given is a typical allocation of staff and vehicles. The bases found in the DOS were PE, Central, West End, Motherwell and Uitenhage. PE and Central were re-classified as Gqeberha 1 and Gqeberha 2 respectively. These bases were then assigned to wards, and their coordinates are plotted in Figure 3.8.

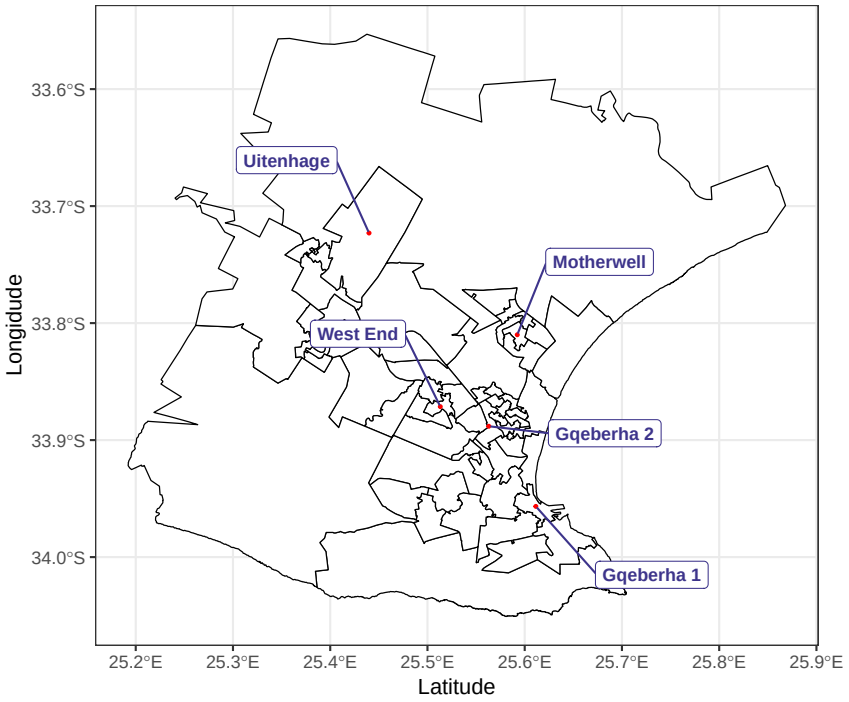


Figure 3.8: Approximate locations of EMS bases, used by the agent-based model.

The number of staff members and vehicles assigned to each of the above bases, as used in the base simulation model, are given in Table 3.2:

Resource	Base	Number
Ambulances	Gqeberha 2	4
	Gqeberha 1	1
	Uitenhage	3
	Motherwell	1
	West End	1
Patient Transport Vehicles	Gqeberha 2	4
	Gqeberha 1	0
	Uitenhage	2
	Motherwell	1
	West End	1
Supervised Practitioners	Gqeberha 2	3
	Gqeberha 1	1
	Uitenhage	2
	Motherwell	2
	West End	1
Independent Practitioners	Gqeberha 2	3
	Gqeberha 1	1
	Uitenhage	3
	Motherwell	2
	West End	1

Table 3.2: Number of staff and vehicles assigned to each base. Values used in the base scenario.

3.3.3 Vehicle Routes

After a vehicle picks up a patient, it hands over the patient to a facility. There is therefore a distribution of destination wards, which was computed in the same way as the distribution of incident wards. This distribution is shown in Figure A.5 in the Appendix.

The overall distribution of destination wards cannot be interpreted independently of the incident wards. For example, a patient requesting an ambulance in Uitenhage is more likely to be handed over to a facility in Uitenhage than one in Gqeberha. Destination ward distributions were therefore computed for each incident ward and call type, forming a series of three 60×60 matrices. The matrix for P1 calls is shown in Figure 3.9, and the other two matrices (Figures A.8 and A.9) are in the Appendix.

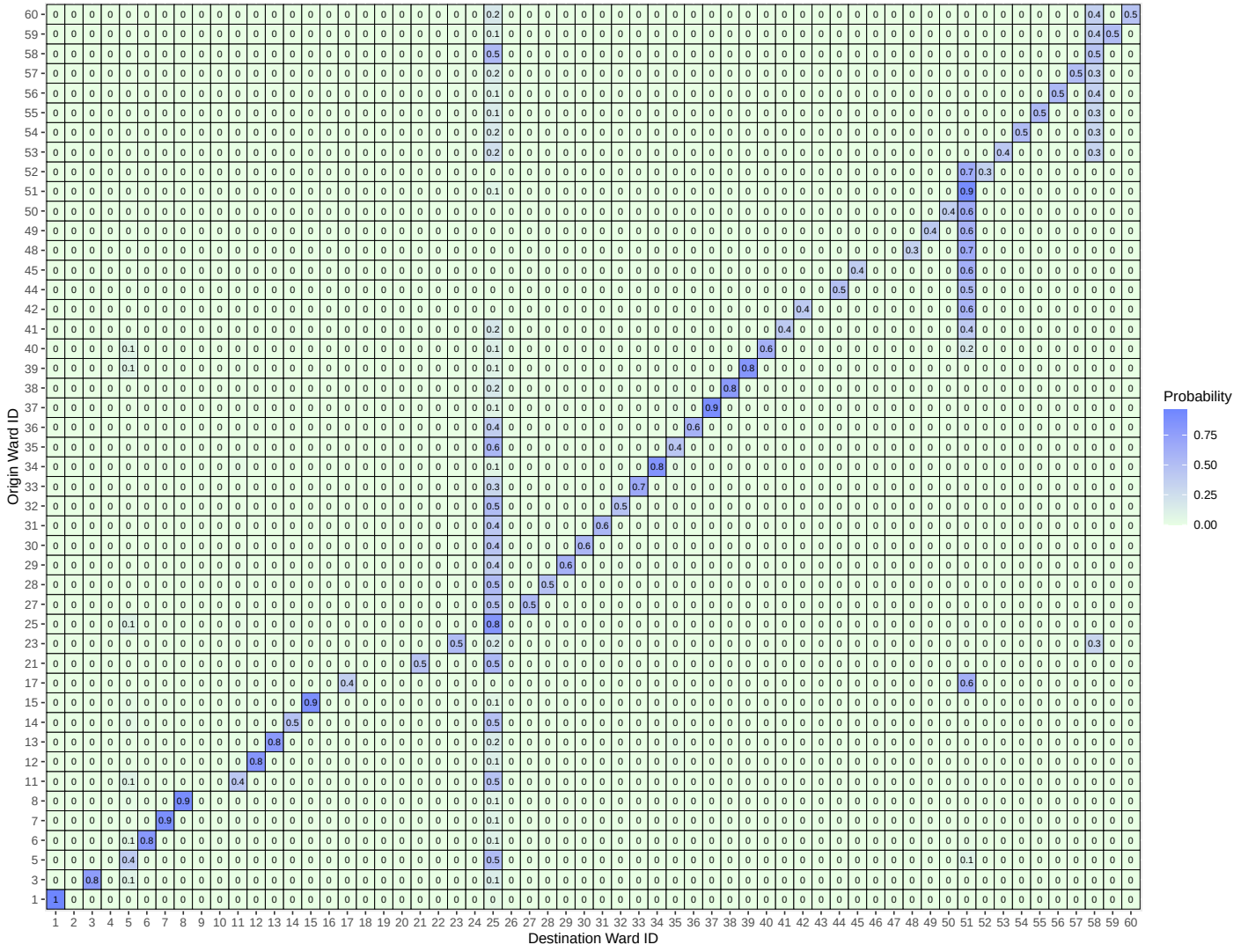


Figure 3.9: Priority 1 Destination Ward Probability Matrix. Each row provides the probability distribution over destination wards, conditional on a given origin ward.

Note: Rows for wards without any observed demand are omitted. In this figure, some rows do not sum to 1 due to rounding.

The non-zero probabilities along the diagonal indicate that many calls were handed over to facilities in the same ward. The vertical stripes at wards 25, 51 and 58 are likely the result of the three large provincial hospitals (Livingstone, Uitenhage and Dora Nginza provincial hospitals respectively) being located in these wards, and receiving patients from the entire region, especially for the more serious Priority 1 calls.

3.3.4 Response Time

This subsection compares the response times we captured to the response times in the call centre dataset, and draws some general conclusions about EMS response times in the NMB region.

In order to assess the reliability of the discrete response times in the call centre dataset, and to check for any data capturer bias, I plotted the discretised response times over time, for each data capturer (Figure A.2 in the Appendix). This suggested that there may be some data capturer bias in the data, as the recorded response times are strongly dependent on the data capturer.

However, the main issue with the supplied response time data is that they were discretised into the following categories (measured in minutes):

- $0 < RT \leq 15$.
- $15 < RT \leq 30$ or $15 < RT \leq 45$.
- $30 < RT \leq 60$ or $45 < RT \leq 60$.
- $60 < RT$.

This discretisation makes calculating precise summary statistics impossible. Because the response times were potentially unreliable and not continuous, it was necessary to capture new response times from EMS call slips.

The response times we captured suggested that the response times in the call centre dataset were likely deflated. In Figure 3.10, response times are compared between the call centre dataset and the new dataset we captured.

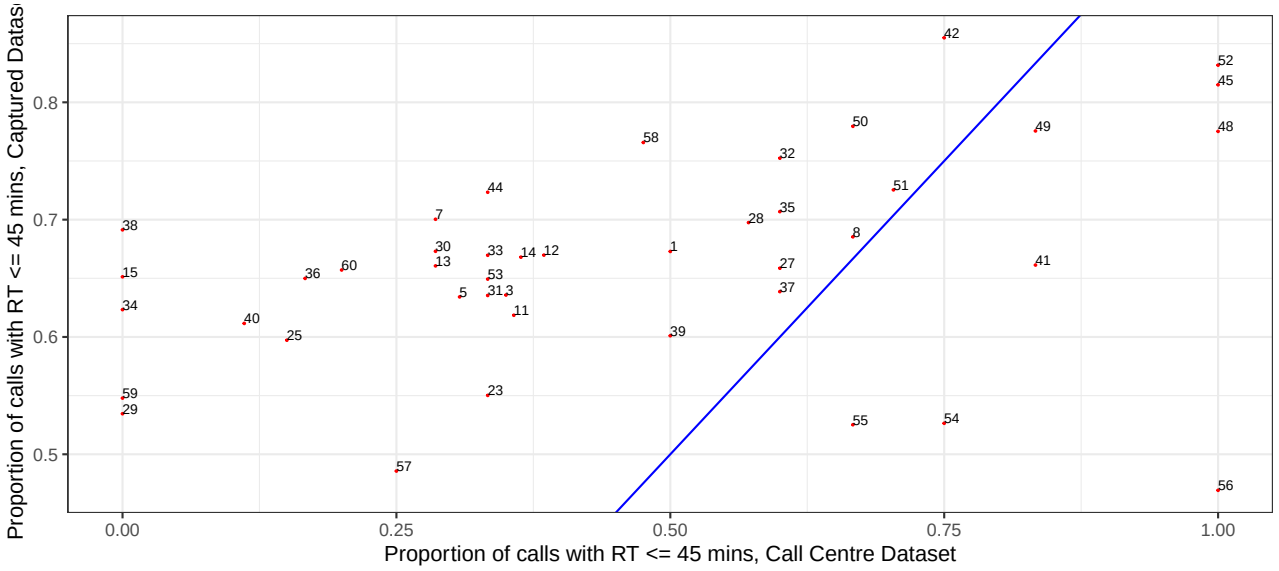


Figure 3.10: Comparison of overall response times between the call centre dataset and the dataset we captured. Scatterplot compares the proportion of calls with RT $\leq$ 45 minutes between the two datasets, for each ward. Numbers indicate ward IDs. Blue line is of slope 1, indicating equality between the two datasets.

Clearly, while there is a moderate positive correlation between the proportions observed in the two datasets ($r = 0.50, p < 0.001$), the relationship is somewhat noisy, and the majority of observations lie above the line of equality. This implies that there is a general pattern of

underestimating response times in the call centre dataset. For example, in ward 36, 65.0% of calls were recorded as having response times ≤ 45 minutes in old dataset, while this same figure was only 14.8% in the dataset we captured.

In Figure 3.11, the distribution of response time from the data we captured is discretised, and compared to the distribution of discrete response times in the call spreadsheet:

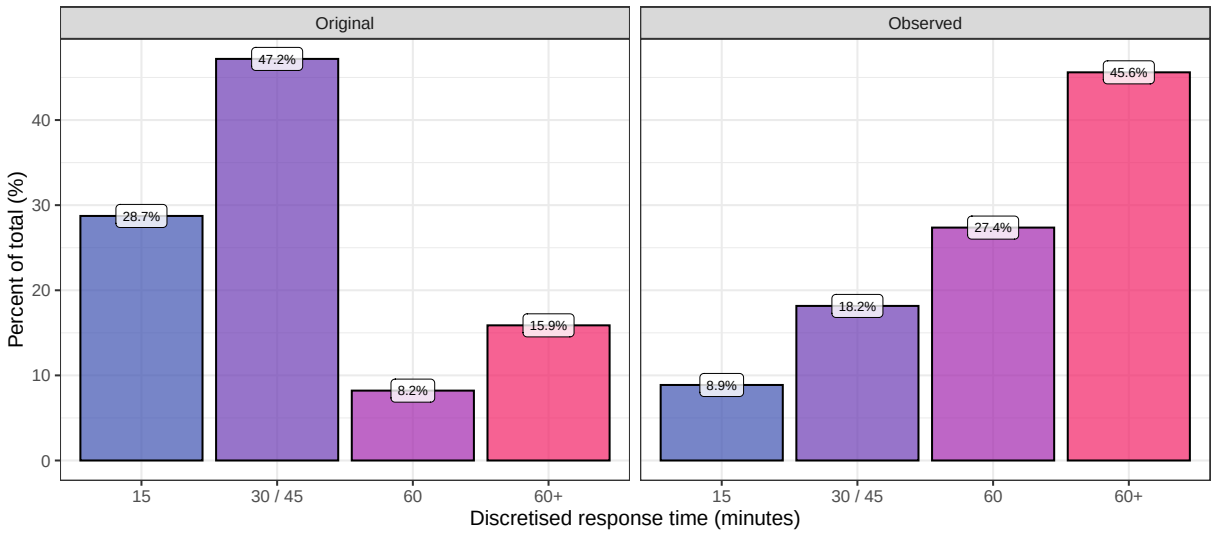


Figure 3.11: Distribution of response time in the call centre data (left) vs newly-captured distribution of response time (right).

The distributions are not very similar. For the days on which we captured data, 26.7% of response times were classified correctly in the call centre spreadsheet. Overall, the original discretised response times appear to be biased towards shorter response, compared with the times we captured. For example, while 45.6% of response times are over 60 minutes in the data we captured, in the call spreadsheet only 15.9% of response times are over 60 minutes. While we did observe that response times were generally higher in the paper call slips than the electronic call centre dataset, it must be emphasised that this comparison is based on a limited subset of dates.

Using the data we captured, the overall average and median response times are 1h 13m and 52m respectively. This suggests that some extreme values are skewing the distribution of response times. Indeed, in examining the histograms of response times, broken down by priority, we see that the distributions are positively-skewed, where some response times are several hours. These histograms are shown in Figure 3.12.

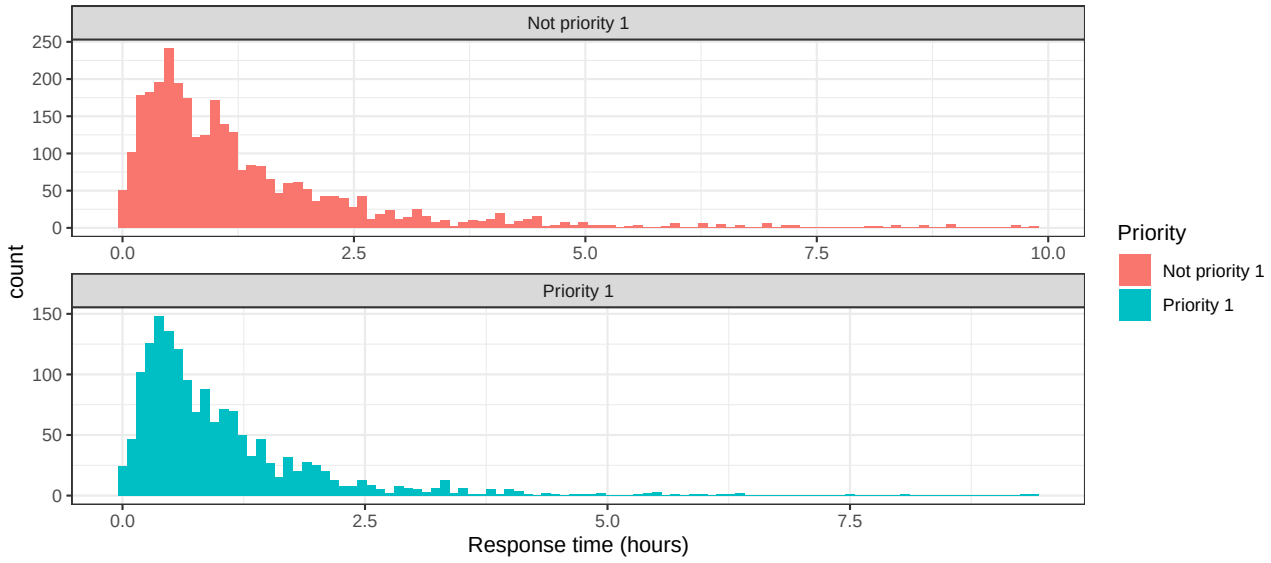


Figure 3.12: Histograms of response time (hours), by call priority.

There is no clear visual difference in the distributions of response times, on the basis of these histograms alone, but the mean response time for P1 calls is 61 minutes, while it is 1h 20m for non-P1 calls. Median response time is 43 minutes for P1 calls and 56 minutes for non-P1 calls.

Figure 3.13 shows the percentage of P1 calls with response times within 15-minute and 30 minute targets, for Gqeberha, Motherwell and Uitenhage.

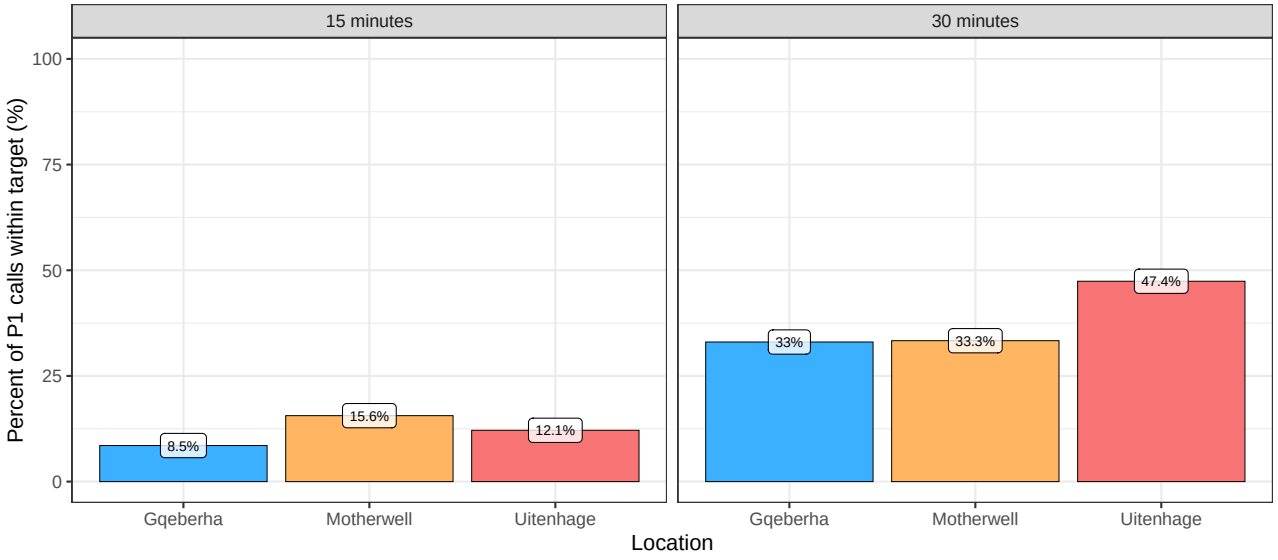


Figure 3.13: Percent of P1 calls within target RT, by location and target.

In the data we collected, 9.6% of P1 calls in Gqeberha were responded to within 15 minute, while the figure was 12.2% and 11.6% for Motherwell and Uitenhague respectively.

The target response time for P1 calls in urban areas is currently 30 minutes (Kamnqa 2023). However, a minority of P1 calls in the data we collected fall within this relaxed target. In Gqeberha, 33% of P1 calls were responded to in under 30 minutes, while this same figure was 33.3% in Motherwell and 47.4% in Uitenhage.

3.4 Prioritisation

For assessing how well prioritisation is being carried out, we can compare the call prioritisation and on-scene triage classification distributions. Triage classifications, according to the South African Triage Scale (SATS) are assigned on a colour scale, in order of urgency (Cheema & Twomey 2012). According to SATS, the colour scales are Green, Yellow, Orange, Red and Blue, where code Blue patients are deceased (Cheema & Twomey 2012). The ECDoH does not use Orange classifications.

Classification	Management
G (Green)	Refer to a designated area for **non-urgent** cases.
Y (Yellow)	Refer to a major facility for **urgent** management.
R (Red)	Refer to a resuscitation room for **emergency** management.
BOA (Blue On Arrival)	Refer to a doctor for certification.
BIA (Blue In Ambulance)	Refer to a doctor for certification.

Table 3.3: Management of Each Triage Classification. Management of each triage classification quoted verbatim from Cheema & Twomey (2012).

Figure 3.14 shows the univariate distributions of priority levels and triage classifications.

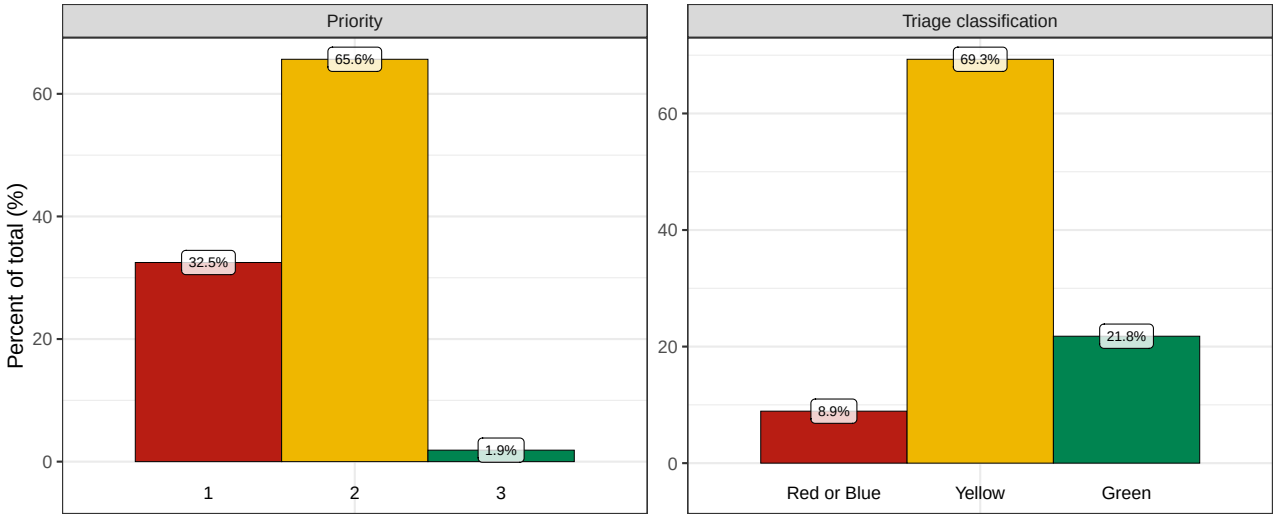


Figure 3.14: Priority and Triage Classification Distributions.

Approximately one third of calls are classified as Priority 1, but only 8.9% of calls are classified as code Red or Blue by EMS personnel on the scene. With optimal allocation, only code Red or Blue calls would be classified as P1.

Figure 3.15 examines the joint distribution of call type and triage classification:

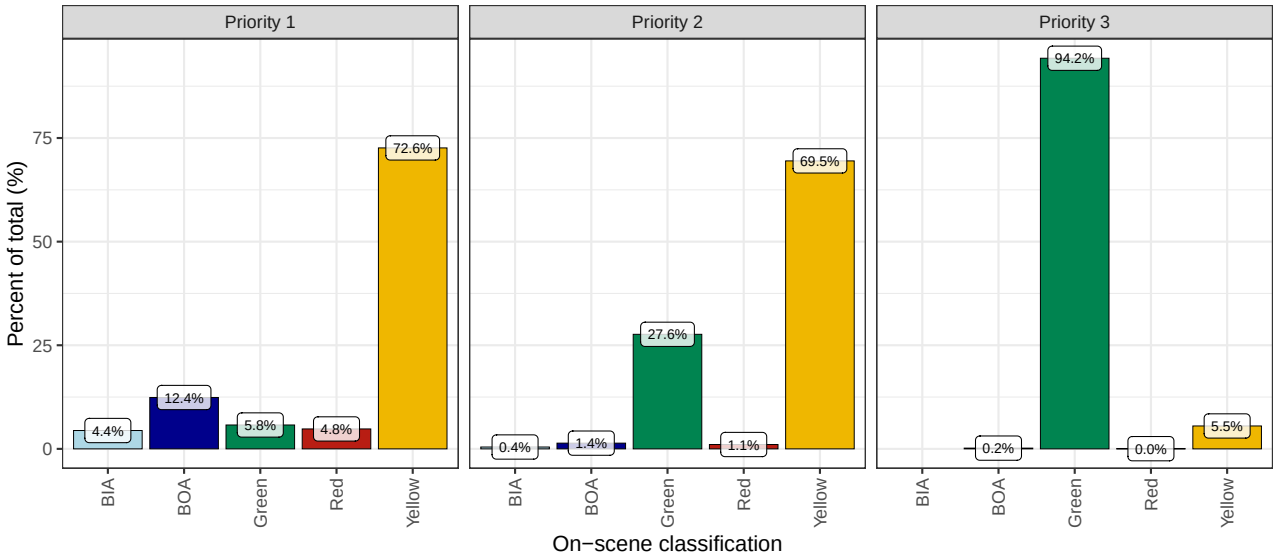


Figure 3.15: Triage Classification Distributions, by Priority.

72.6% of P1 calls were classified as yellow by EMS personnel on the scene, which implies that there is a general tendency of dispatchers or data capturers to overestimate the priority of calls.

3.4.1 Staff and Vehicles

This final section examines the current utilisation of vehicles and staff.

The distributions of staff qualifications for each call type is shown in Figure 3.16:

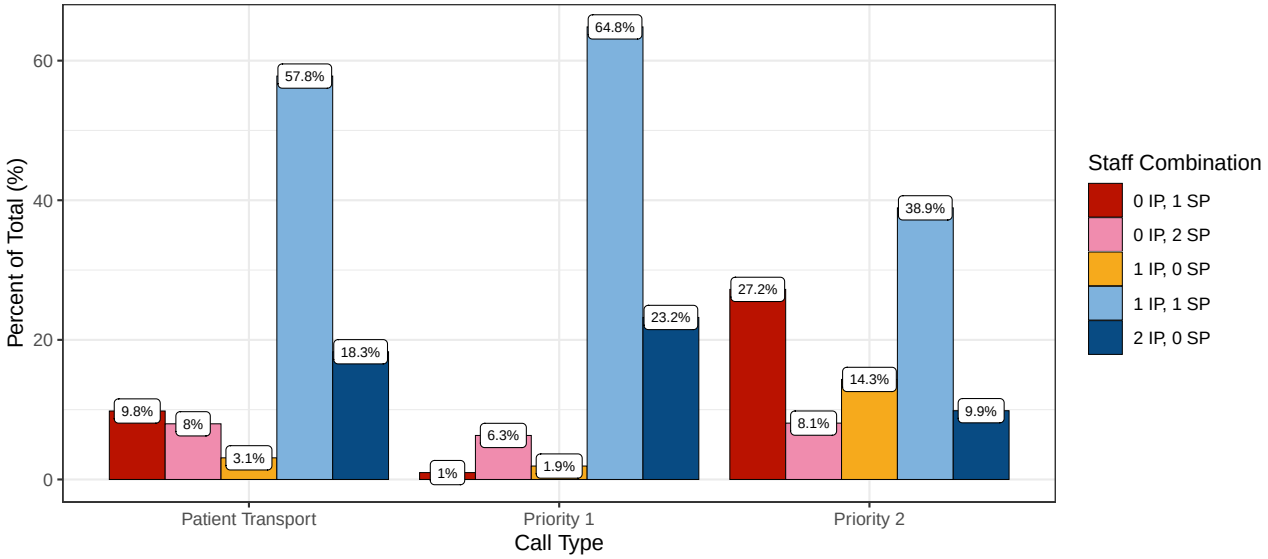


Figure 3.16: Staff Mixture Qualification Distribution, by Call Type. IP refers to Independent Practitioner, and SP refers to Supervised Practitioner.

EMS regulations require that Supervised Practitioners (SPs) must be supervised by Independent Practitioners (IPs) (HSPCA 2019). However, 7.3% of P1 calls were responded to by SPs who were not acting under the supervision of IPs. The majority of calls were responded to by one SP and one IP, and 18.3% of Patient Transport calls were responded to by two IPs.

Figure 3.17 shows how the vehicle type distribution varies depending on the call type.

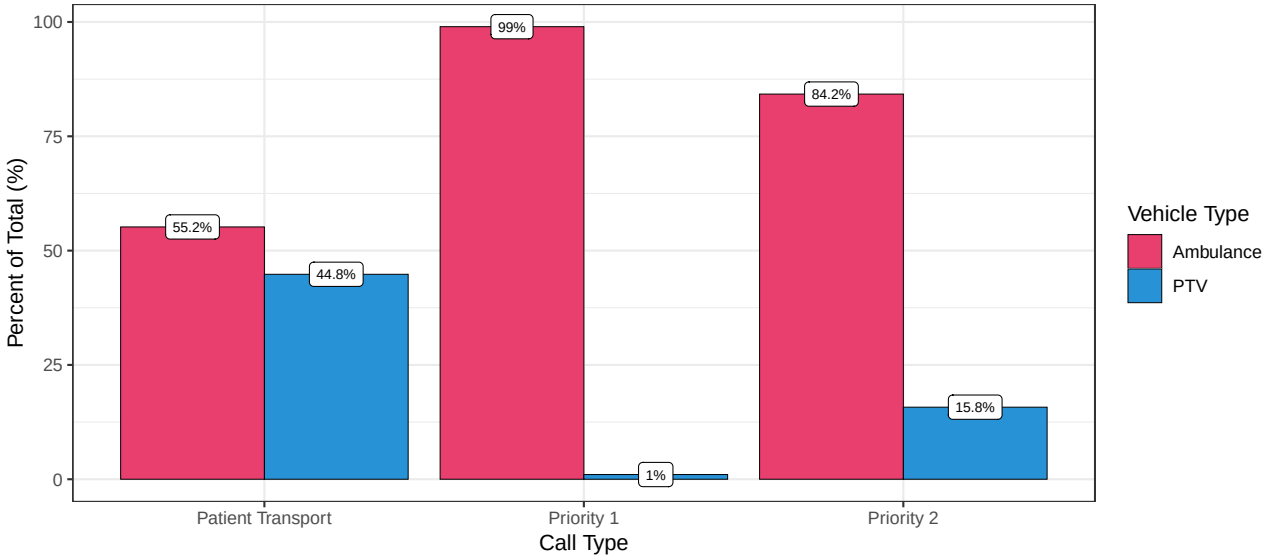


Figure 3.17: Vehicle category distribution, by Call Type.

The clear majority of P1 calls are responded to by ambulances. Approximately 1% of P1 calls are responded to by patient transport vehicles (PTVs). This could be the result of data capturing errors.

In general, ambulances are best used to respond to urgent medical emergencies rather than less urgent PPT calls. However, Figure 3.17 shows that ambulances respond to 55.2% of PPT calls, a figure that is higher than expected[3]. Some PPT calls involve transporting patients from one facility to another, which are referred to as inter-facility transfers. These calls are more likely to use ambulances compared with other types of PPT calls: 79.2% of inter-facility transfer calls are responded to by ambulances. For the remaining PPT calls, 89.2% of calls are responded to by PTVs.

[3]In this research, no Priority 1 calls can be classified as Planned Patient Transport (PPT). See the classification criteria for PPT calls in sub-section 3.2.2.

Chapter 4

Methods

4.1 Introduction

Simulation models are well-suited to modelling complex human systems and capable of explicitly incorporating multiple sources of variability. I therefore used a simulation model to model Nelson Mandela Bay's EMS system. In addition, because Agent-Based Models (ABMs) are particularly well-suited to modelling large-scale systems with complex interactions between individual agents, I used this type of simulation model. The model I developed has four types of agents: the patients, the vehicles, the EMS personnel, and the operation centre which coordinates the system's response to calls [1]. Time passes in discrete 10-minute intervals. On a high level, calls are placed at each time point by patients, and the operation centre responds to these calls by determining probabilistically the staff and vehicles to dispatch. If there are not enough staff or vehicles, the operation centre places calls in a First-Come, First-Served (FCFS) queue. The full explanation of the model, including all agent types and the rules governing their behaviour, is contained in this chapter.

First, the conceptual model of the EMS system is introduced in section 4.2. This explanation begins with a most basic ABM and builds on it sequentially, justifying the inclusion of each additional feature. The following sections contain explanations and justifications for the steps taken to conduct the validation, verification, scenario and sensitivity analyses.

The model was coded in Python 3.11, with heavy reliance on the Numpy and Pandas libraries (Van Rossum & Drake 2009; Harris *et al.* 2020; pandas development team 2020). The sensitivity and scenario analyses were run in parallel on 50 CPU cores. The code used to run the model and perform sensitivity and scenario analyses is provided in a public GitHub repository[2].

[1] These agent types are similar to those seen in the ABM developed by Aringhieri (2007).

[2] URL: `https://github.com/skycope/ems_thesis/tree/main`.

4.2 Conceptual Model

4.2.1 A Most Basic Model

On the most basic level, the purpose of an EMS is to provide urgent care to people experiencing medical emergencies, and if needed to transport them to appropriate tertiary medical facilities (Al-Shaqsi 2010). Any model of an EMS then needs at least to include patients, vehicles and facilities. The one vehicle, one facility, one patient model is illustrated in Figure 4.1:

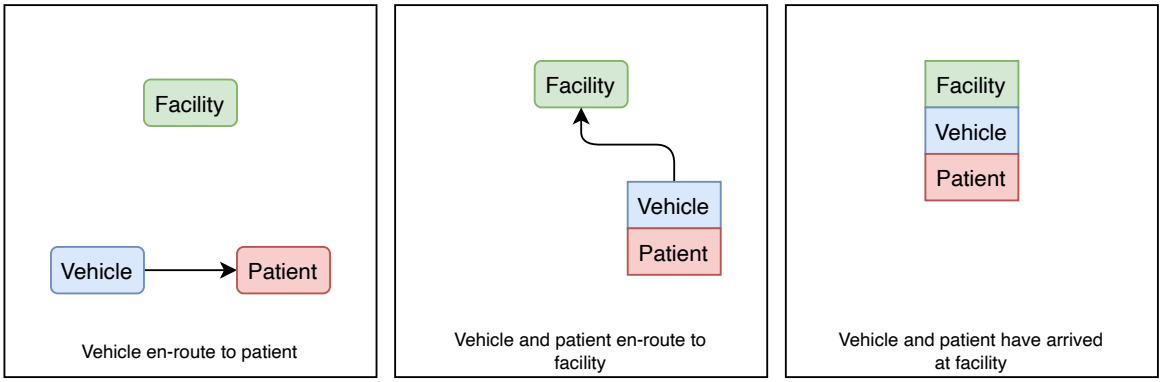


Figure 4.1: Three time steps of a one vehicle, one facility, one patient model.

Even in this most basic case one needs to accommodate travel times, from the vehicle to the patient, and then from the scene to the facility. Also, because agent-based models are discrete time simulations, time passes in discrete ticks. I used ticks of 10-minute intervals. By using an agent-based model, travel times can be sampled from some appropriate distribution rather than being fixed at a single value. Travel time sampling is explored in section 4.2.5. However, it is apparent that this model is a radical oversimplification of reality: we need a model that can also accommodate multiple patients, multiple facilities and multiple vehicles. For this, a number of changes need to be made, including the addition of queueing.

4.2.2 Queueing

When there are multiple patients, at times a patient will call and there will not be enough idle vehicles available to respond. For these cases, it is necessary to introduce a queueing system. Each patient is assigned a unique ID. If there are insufficient vehicles available to respond, the patient ID, location, and time of placing call, are appended to a queue list. This list is sorted by time spent in queue, such that patients are responded to on a first-come-first-served (FCFS) basis.

Figure 4.2 shows an example of a multiple vehicle, multiple facility and multiple patient model with queueing.

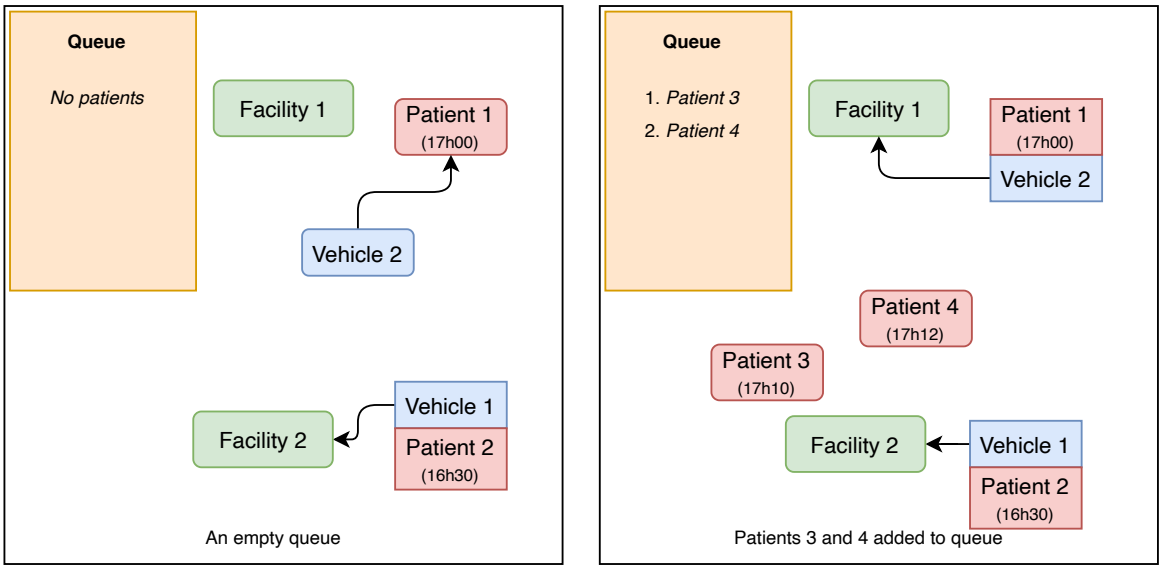


Figure 4.2: Two time steps of a basic queueing model with multiple vehicles, multiple facilities and multiple patients. Time of calling EMS shown in brackets.

In the left panel, vehicle 2 is en-route to patient 2, and vehicle 1 is transporting patient 2 to facility 2. There are no additional patients so the queue is empty. In the right panel, some time has passed, and patients 3 and 4 have called. Unfortunately the two vehicles are both unavailable, so the additional patients are placed in a queue. Patient 3 called before patient 4, so is placed first in the queue. When there is more than one vehicle available to respond to a call, the vehicle with the shortest expected travel time is dispatched to the call.

Some aspects of this model are explained in subsequent sections: the spatio-temporal arrival process for the calls (explained in section 4.2.8), and how vehicles determine which facility to drive towards, after picking up the patient (explained in section 4.2.9).

While the model described in this section does allow for multiple facilities, patients and vehicles, it assumes they are all of the same type. In reality there are multiple call types: Priority 1 (P1), Priority 2 (P2), Priority 3 (P3) and Planned Patient Transport (PPT) calls. There are also Ambulances (AMBs) and Patient Transport Vehicles (PTVs), and there are also multiple types of facilities, from provincial hospitals to renal clinics.

4.2.3 Call Types and Prioritisation

For incorporating multiple call types, type can simply be added as an attribute to the patient agent. The queue behaviour can then be altered in order to implement the prioritisation regime P1 > P2 > PPT, where P1 calls are highest priority, followed by P2, followed by PPT calls. This is achieved by sorting the queue first by call type, and then by time spent in queue. This is illustrated in Figure 4.7.

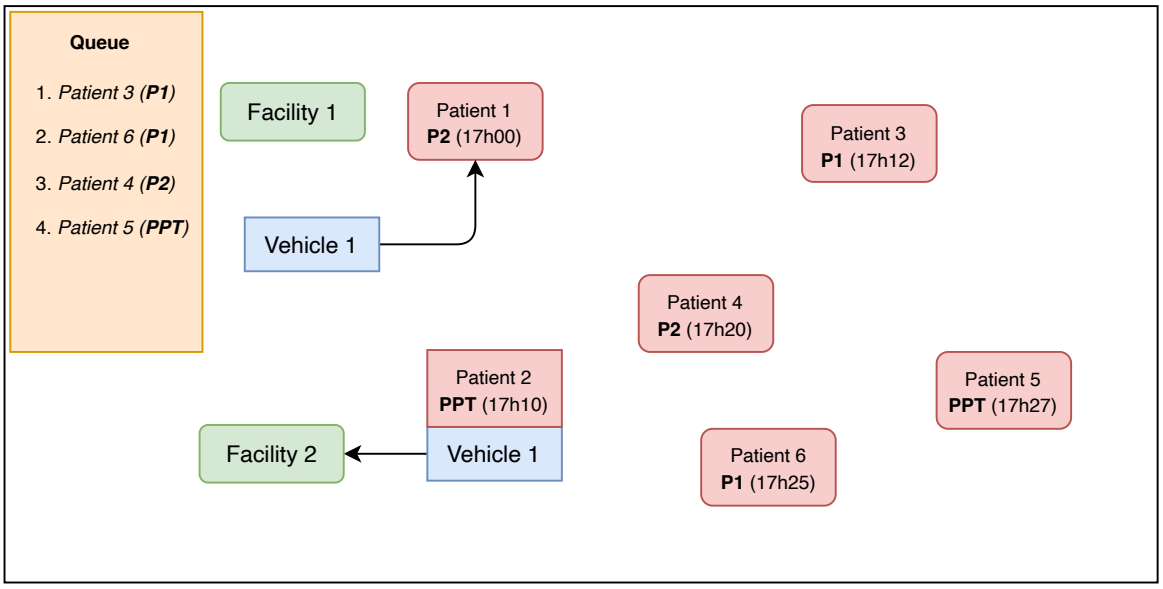


Figure 4.3: A model with multiple patient types. Time of calling EMS shown in brackets. Call type shown in bold.

This shows the queue sorting prioritisation in effect, where patient 6's call is prioritised over patient 4's because it is a P1 call, even though it was received after the P2 call. This form of prioritisation is effective, but it fails to account for two key aspects of prioritisation. Firstly, such rules are not followed 100% of the time in any real-world system, so the model should be able to incorporate varying probabilities of the prioritisation rules being followed. This is addressed in section 4.2.12. Secondly, the EMS system in NMB uses re-routing as another means of prioritising P1 calls. For example, in Figure 4.7, Vehicle 1 is en-route to patient 1, which is a P2 call, but patient 3 is a P1 call: this vehicle should be able to re-route to patient 3, and return patient 1's call to the queue. This is addressed in section 4.2.7. Another key omission is staff members, or EMS personnel, and multiple vehicle types.

4.2.4 Staff and Vehicle Types

The supply side of the EMS system in NMB is usually either bottlenecked by staff or vehicles: more commonly there are enough staff members to respond to a call but not enough vehicles, while at other times there are enough vehicles but not enough staff. It is therefore important that the model incorporates both staff and vehicle agents. The model uses two staff types: Independent Practitioners (IPs) and Supervised Practitioners (SPs). SPs include Basic Life Support (BLS) personnel, while IPs are more qualified, and include Advanced Life Support (ALS) and Intermediate Life Support (ILS) practitioners. The model uses two vehicle types: Ambulances (AMBs) and Patient Transport Vehicles (PTVs).

In terms of the agent-based model, the addition of staff involves adding staff agents, each with a type attribute. The addition of vehicle types involves adding a type attribute to the vehicle agent. The staff and vehicle types can then be used to implement rules regarding which call types a given vehicle or staff member can respond to. For example, we can add the rules that PTVs cannot responded to P1 calls, or that a P1 call needs to be responded to by at least one IP.

The dispatching algorithm, with added vehicle and staff types, is illustrated in Figure 4.4.

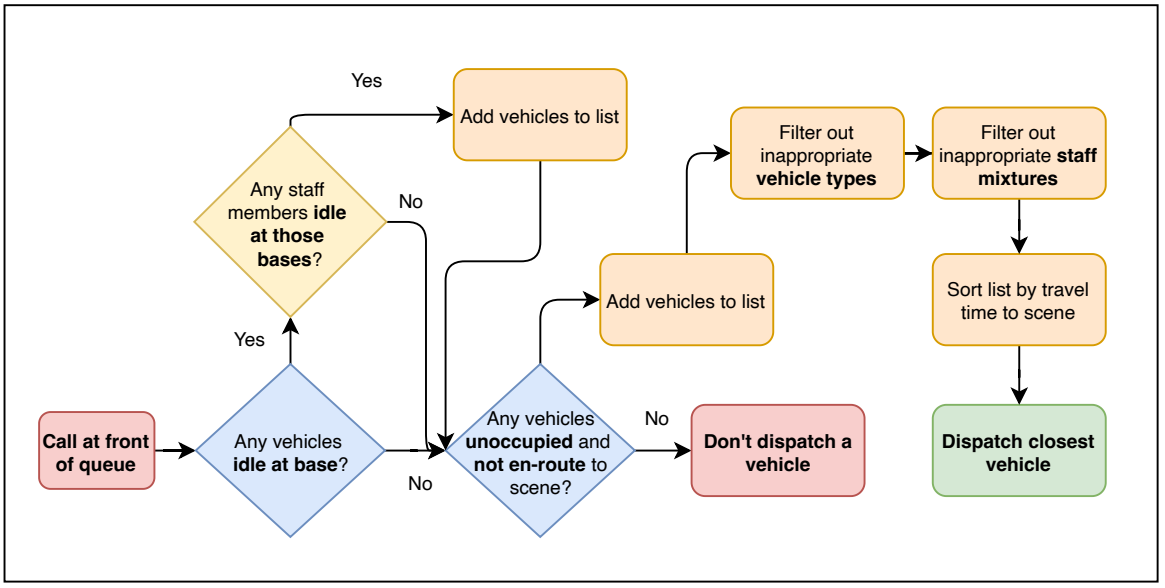


Figure 4.4: Flowchart showing the dispatching algorithm.

In the model illustrated in Figure 4.4, the vehicle and staff filters are applied 100% of the time, implying the rules are always followed. This is relaxed with the addition of staff and vehicle rule probabilities (explained in section 4.2.12).

4.2.5 Travel Times

I used 60 wards for the Nelson Mandela Bay region, calculated the coordinates of each ward's centroid, and then used the Google Maps Distance Matrix API to get estimated travel times between each ward centroid. Road quality was identified as an important factor affecting travel times in the Nelson Mandela Bay region. Google Maps was chosen partly because it incorporates road quality in its estimates. Each incident, EMS base and facility is then aggregated to the level of wards, such that any incident occurring within a given ward is assigned the same coordinates. This simplification allows for the estimation of travel times, and also allows one

to determine, based on expected travel times, which vehicle is nearest to any incident. How the travel time matrix is used to determine which vehicle to dispatch is outlined in Figure 4.5

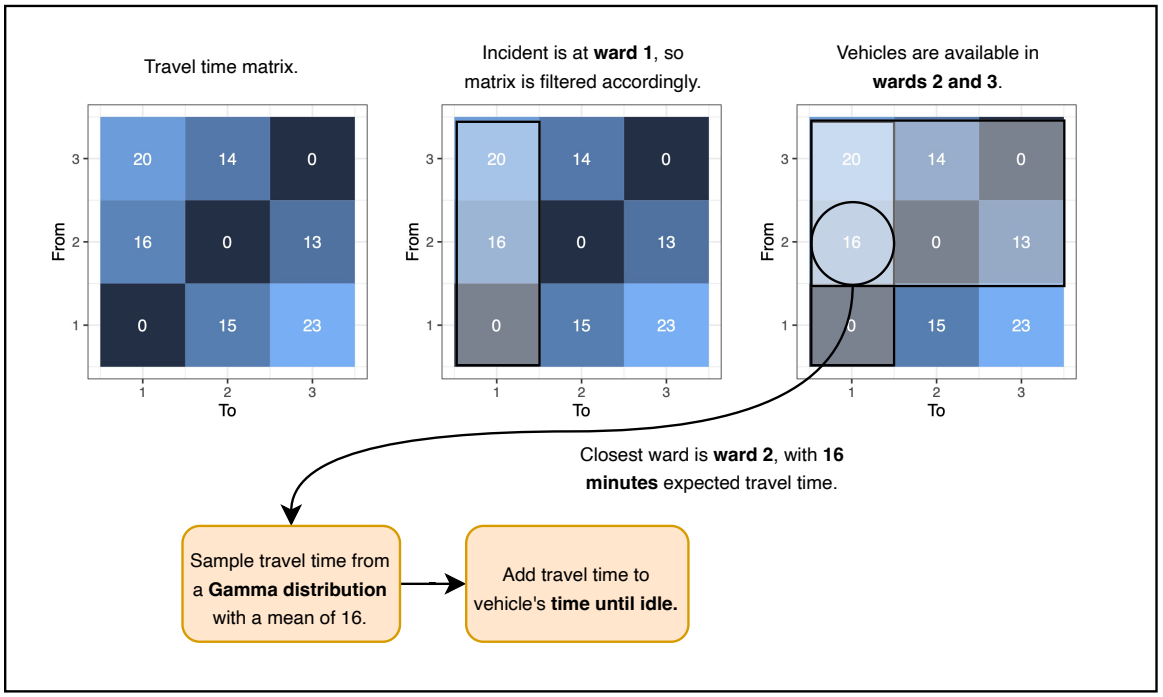


Figure 4.5: Demonstration of how travel times are sampled, with hypothetical data. 'From' and 'To' axes are hypothetical ward indices for origin and destination wards, respectively.

Once the relevant expected travel time is taken from the travel time matrix, to obtain some small variability around this mean, travel time is sampled from a Gamma distribution with a standard deviation of 0.01 ticks.

4.2.6 Scene and Facility Wait Times

Travel time is one component of the time during which a vehicle is unavailable, but there is also the time a vehicle spends at the scene attending to the patient, and the time spent at the facility completing the patient handover.

T_{engaged} is the total time a vehicle is considered unavailable to respond to other calls from the moment it arrives at the scene to the moment it completes the handover of the patient to the facility. It does not include the travel time to the scene. T_{engaged} is given as

$$T_{\text{engaged}} = S + D + F,$$

where S is the time spent on scene, D is the travel time from scene to facility and F is the time spent at the facility. S is sampled from a uniform distribution with specified parameters, which vary depending on the call type:

$$S \sim \begin{cases} U(a_1, b_1) & \text{for P1 calls} \\ U(a_2, b_2) & \text{for P2 calls} \\ U(a_p, b_p) & \text{for PPT calls} \end{cases}$$

F, the time a vehicle spends at the facility, is also sampled from uniform distribution, which is the same for all call types. D, the travel time from scene to destination, is calculated once the

destination ward has been selected. These distributions were, in practice, chosen in consultation with an expert source, Ms. Sibongiseni Ntengu from the ECDoH. A uniform distribution was chosen because the most likely times within the specified ranges were not obtained from the expert. The sampling of travel times is explained in section 4.2.5.

4.2.7 Re-routing

Re-routing is done in the EMS system in NMB, where vehicles en-route to non-P1 calls can be diverted to P1 calls. In the dispatching algorithm, re-routing is achieved by first checking if a call is a P1 call, and then including in the available vehicles list vehicles that are en-route to non-P1 calls (but not vehicles that are already occupied with a patient). Then, the usual dispatching algorithm applies: the nearest available vehicle is dispatched to the call, which could potentially be a vehicle that was originally en-route to a lower priority call. If a vehicle was re-routed, the patient to which the vehicle was travelling is returned to the queue, with queue position determined by wait time and call type.

These additions to the dispatching algorithm are illustrated in 4.6:

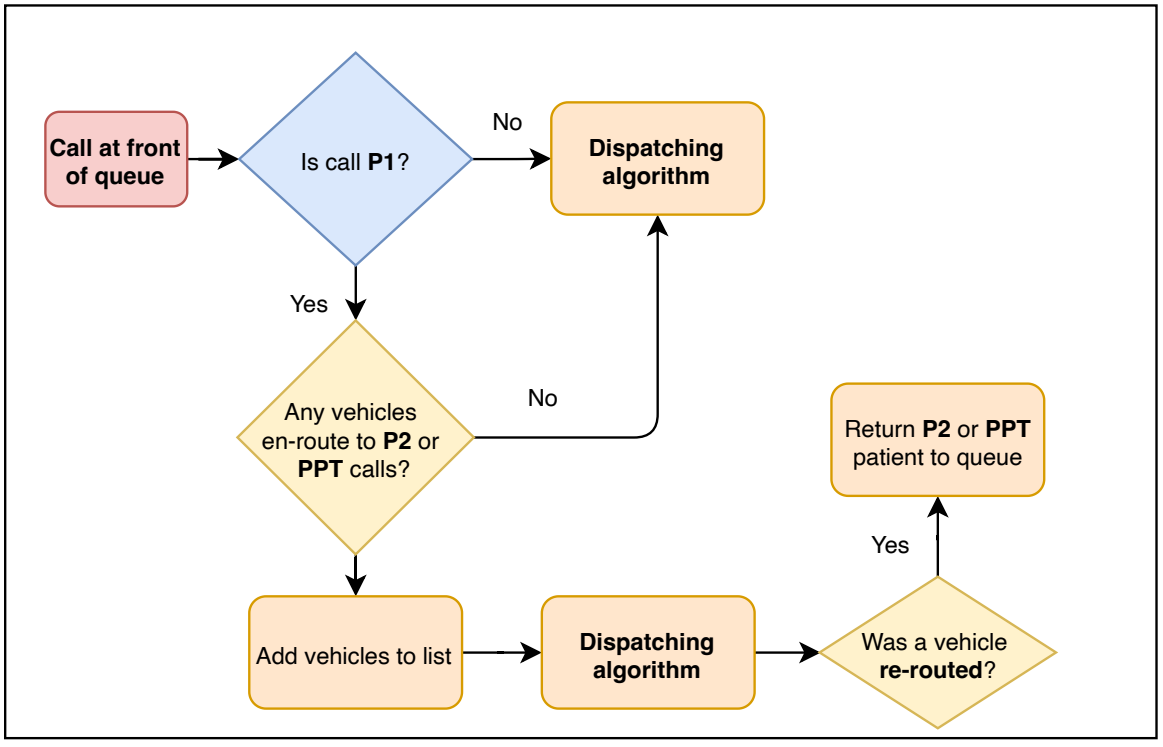


Figure 4.6: Flowchart showing the addition to the dispatching algorithm introduced by the re-routing algorithm.

The dispatching algorithm aims to dispatch the nearest available vehicle to each patient, with nearness determined by expected travel time. Vehicles are either idle, en-route to a patient, at the scene, transporting the patient, or at a facility handing over the patient. When re-routing is not a part of the dispatching algorithm, vehicles can only be dispatched if they are idle. This makes calculating travel times simple: vehicles are idle when they are either stationed at a facility or an EMS base. Each base and facility is located in a known ward, and each ward centroid has known coordinates, which are used to estimate expected travel times in the procedure outlined in section 4.2.5.

However, when re-routing is possible, vehicles can sometimes be dispatched when they are en-route to calls. It then becomes necessary to know the locations of vehicles at each point along their trip. For this purpose, I assumed that trip trajectories were straight lines between the origin and destination ward centroids. Current locations were estimated at each time step using the percentage of the trip completed, as well as the origin and destination coordinates. For example, if a trip from ward A to ward B is 50% completed, the current location of the vehicle driving between these points is estimated to be at the midpoint of the straight line between points A and B.

For using the inter-ward travel time matrix, these current coordinates then need to be converted to current wards. This is simply done by assigning each vehicle's current ward as the ward with the centroid that has the shortest Euclidean distance between the vehicle's current coordinates. This process is illustrated in Figure 4.7.

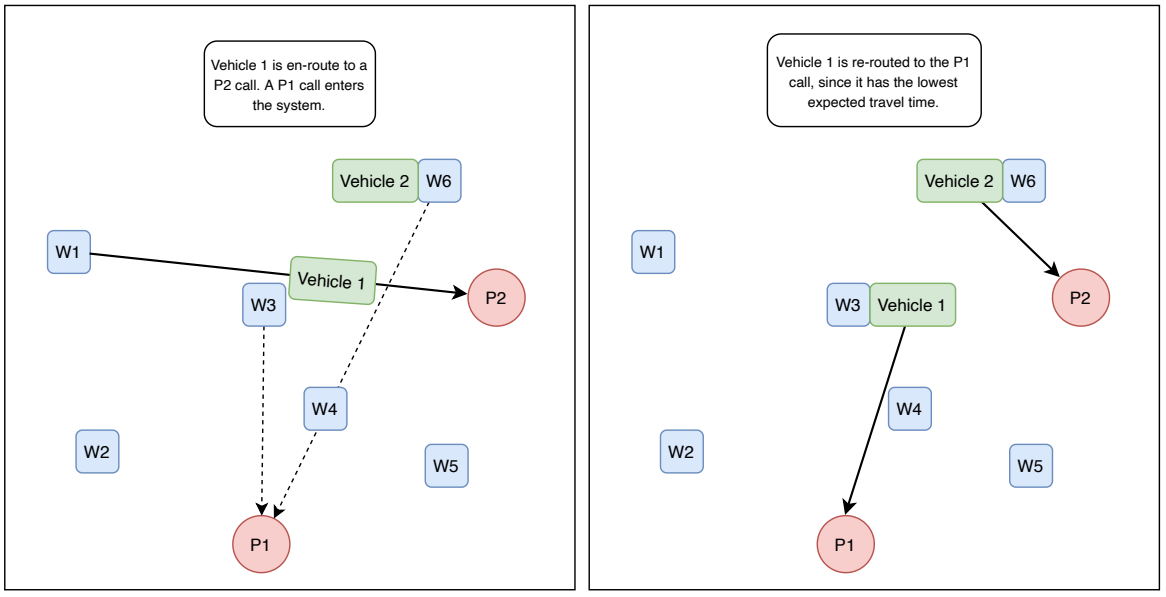


Figure 4.7: An illustration of the re-routing algorithm in practice. P1 and P2 refer to wards with outstanding Priority 1 and Priority 2 calls. W1 to W6 refer to ward centroids.

In the left panel, Vehicle 1 is en-route to P2, a Priority 2 call. It is driving from Ward 1, and its route trajectory is a straight line from W1 to P2. Vehicle 2 is idle at a base in Ward 6. A Priority 1 call has just entered the system. In this case, it is possible to re-route Vehicle 1 since it is only en-route to a P2 call. The system dispatches the vehicle with the shortest expected travel time, so it calculates the expected travel time for both vehicles 1 and 2. For Vehicle 1, the nearest ward centroid, judged by Euclidean distance, is W3. The expected travel time for Vehicle 1 is then estimated as the expected travel time from ward 3 to the P1 patient. The expected travel time for Vehicle 2 is greater, so Vehicle 1 is re-routed, and the P2 call is returned to the queue. It is then responded to by the nearest idle vehicle, which is Vehicle 2.

4.2.8 Demand Distribution

To simulate demand such that calls are sampled from a representative spatio-temporal demand distribution, the available call volume data must first be used to estimate call rates for each location, each call type, and each time point in the model. These rates can then be used to draw counts from an appropriate distribution. The model uses ticks of 10-minute intervals. It is necessary to estimate mean call rates from the available data for each 10-minute interval. The

model has also been chosen to run for two weeks, or 2016 10-minute iterations. The process used to calculate these rates is illustrated in Figure 4.8:

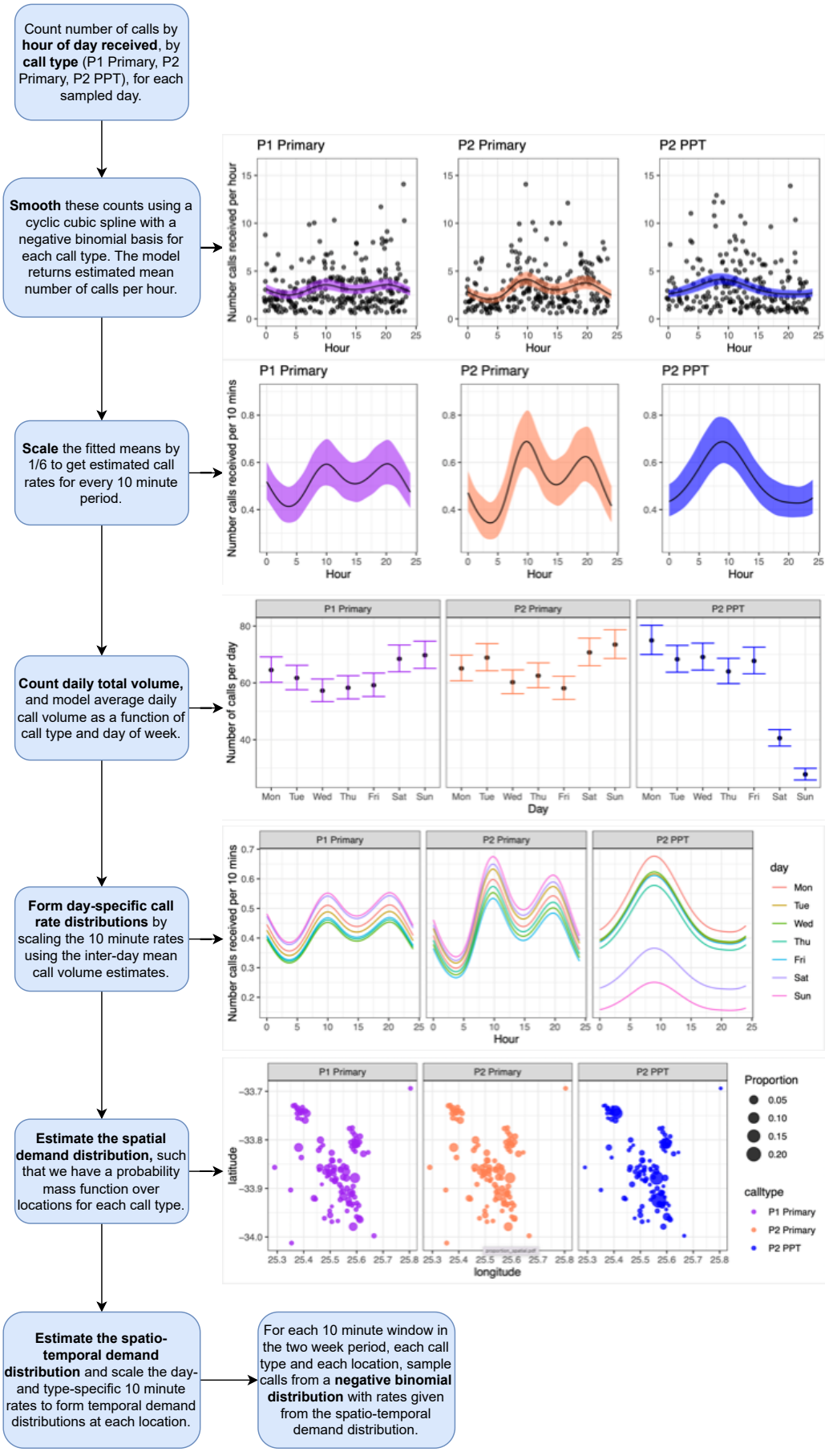


Figure 4.8: How spatio-temporal demand distribution is estimated.

4.2.9 Drop-off Distribution

Once a vehicle has arrived at a scene, a facility is selected by EMS personnel. This is the facility to which the patient will be transported. To choose drop-off locations, the model samples once from the following conditional probability mass function, for each call:

$$P\left(W = w|c, l\right),$$

where w is a given drop-off ward, c is the call-type (P1, P2 or PPT) and l is the call location. There are expected differences in $P\left(W = w\right)$ for varying levels of c: for example P1 calls will likely be dropped off primarily at large provincial hospitals which can handle more severe cases, and some renal PPT calls are dropped off at smaller dialysis clinics. It also clear that this distribution will likely vary with changes in l, because vehicles typically choose to drop off patients at facilities that are near the incident.

Because there are incident and facility ward locations, as well as call types, available in the data, this probability distribution can be estimated by using the empirical frequency distributions from the data, conditional on a given incident ward and call type. For example, we can count in the data a certain number of P1 calls where patients were transported from ward 1 to ward 2. Using the empirical distribution of destination ward counts, holding ward 1 and P1 constant, we can sample a destination ward. This process is outlined in Figure 4.9.

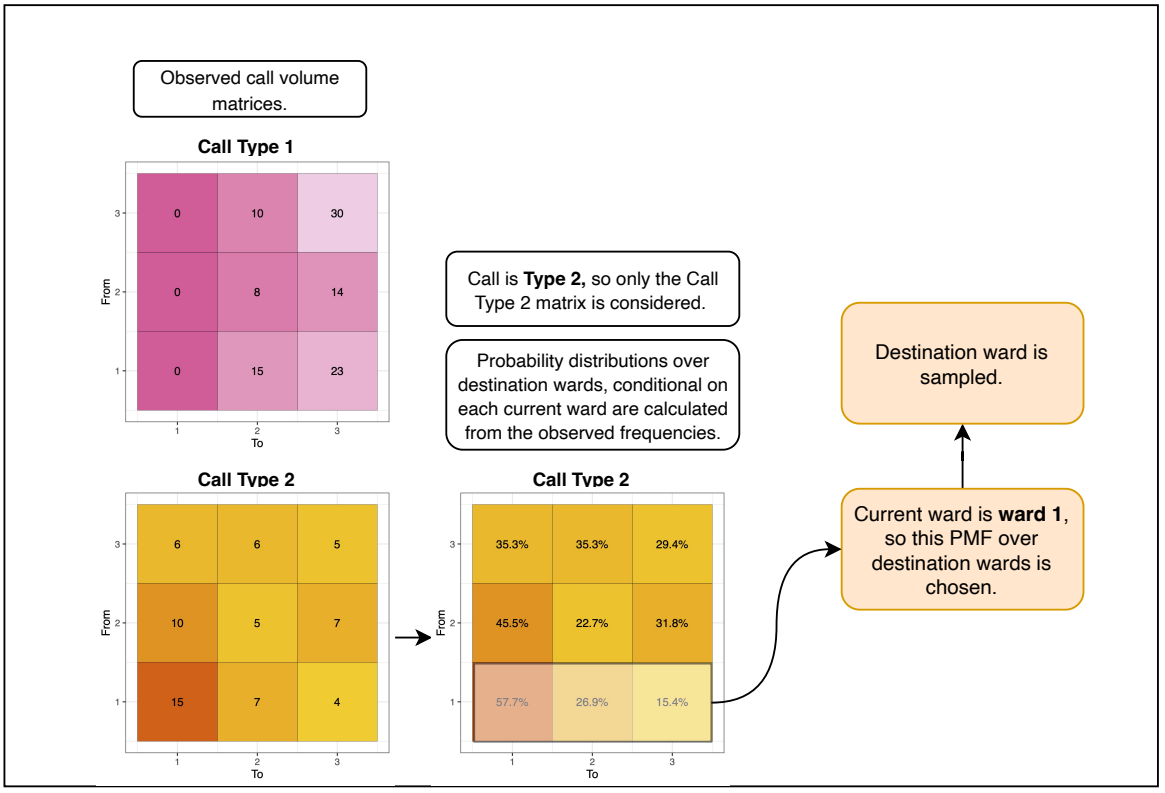


Figure 4.9: Demonstration showing how $P\left(W = w|c, l\right)$ is estimated and sampled from, using hypothetical data.

4.2.10 Failure to Convey Rates

In some cases, vehicles arrive at the scene and the patient is either not present or refuses to be conveyed. This is a function of response time: the greater the response time, the less likely the patient will be there when EMS personnel arrive. I used a logistic regression model to estimate the Failure To Convey (F/T) probability for each call type as a function of response time. The model is specified as:

$$FT_i \sim \text{Bernoulli}(p_i)$$
$$\text{logit}(p_i) = \beta_0 + \beta_1 \text{ResponseTime}_i$$

where FT_i is 1 if the i-th call is classified as Failure To Convey (F/T) and 0 otherwise, ResponseTime_i is the response time in minutes of the i-th call, and p_i is the probability that the i-th call is classified as F/T.

In the model, as soon as EMS personnel arrive at the scene, these profiles are used in the model to determine whether a call is F/T, given the call type and patient response time. This process is outlined in Figure 4.10, for hypothetical data:

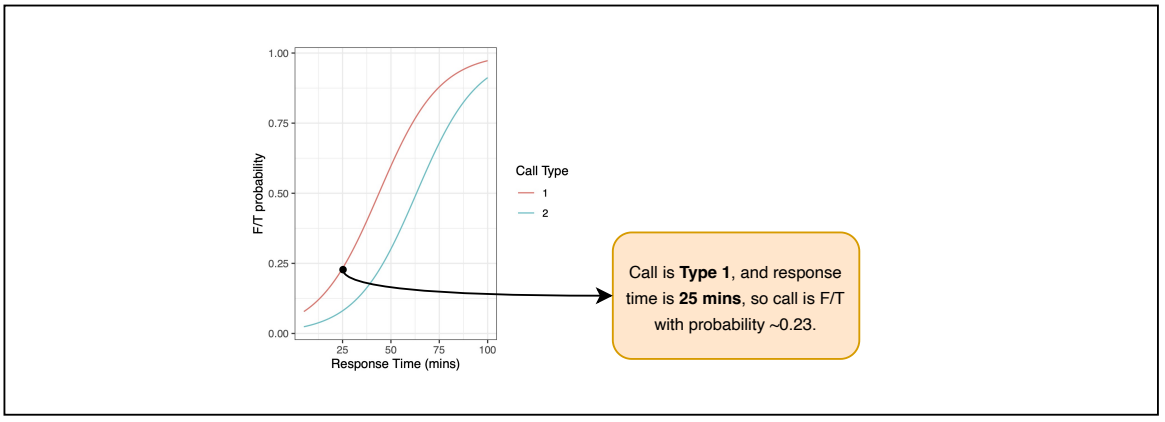


Figure 4.10: Demonstration showing F/T probability is calculated, using hypothetical data.

4.2.11 Triage Classification Rates

There are two main classifications of call severity: call type, as determined by the dispatcher, and triage classification, as determined by the EMS personnel at the scene. As shown in section 3.3, call type and triage classification are correlated but not identical. The triage classification probability mass function is estimated from the data, such that there are three PMFs from which to sample, depending on the call type. For example, $P(T = t|C = \text{P1})$, the PMF for triage classification conditional on the call being classified as P1 by the dispatcher, is given as

$$P(T = t|C = \text{P1}) = \begin{cases} p_{r1} & \text{for } T = \text{Red.} \\ p_{y1} & \text{for } T = \text{Yellow.} \\ p_{g1} & \text{for } T = \text{Green.} \end{cases}$$

where the probabilities p_{r1}, p_{y1} and p_{g1} are estimated using relative frequencies from the data. Equivalent probabilities are estimated for P2 and PPT calls, and for each call the triage classification is sampled from the appropriate PMF.

4.2.12 Affecting System Behaviour with Probabilities

In reality, many of the rules coded into the model (for example, re-routing) are not followed with probability 1. It is then necessary to add probabilities with which these rules are followed, at each time point. These probability parameters can also be tweaked in scenarios, to subtly affect the behaviour of the system. The aspects of the system where probabilities were introduced include vehicle behaviour, staff behaviour, call prioritisation and dispatcher classifications.

Vehicle Behaviour

P1 calls generally require ambulances (AMBs) and PPT calls generally require patient transport vehicles (PTVs) to respond to them, but these rules are not followed at all times in the data. As seen in sub-section 3.4.1, almost half of PPT calls were responded to by AMBs. Therefore, instead of specifying rules such as "P1 calls at all times can only be responded by an AMB", these rules were enforced, for each call type, with a given probability at each time step. Three parameters were introduced – a *Require AMB Probability* for each call type. For example, if the P1 Require AMB Probability were 0.9, then with probability 0.9 only AMBs could be dispatched to P1 calls at a given time step, and with probability 0.1 either a PTV or AMB (whichever is nearest) could be dispatched.

Staff Behaviour

Similarly, P1 calls ideally should be responded to by at least one Independent Practitioner (IP). However, this rule is not rigidly enforced in practice, and three additional probability parameters were introduced – *Require IP Probability* for each call type.

Rarely, a single staff member responds to a call. In order to allow single staff members to respond to calls occasionally, three more probability parameters were introduced – *Single Staff Probability* for each call type. For example, if the P1 Single Staff Probability were 0.1, then with probability 0.1 it would be possible for a vehicle containing only one staff member to respond to P1 calls at a given time step. Otherwise, only vehicles containing two staff members are eligible to respond to calls.

Prioritisation

The double sorting approach to prioritisation, outlined in 4.2.3, where calls in the queue are sorted first by call type and second by time spent in the queue, is effective. P1 calls are responded to first, then P2 calls, then PPT calls. However, this process is not always followed in reality, and it is necessary to introduce additional probabilities. The *Prioritise P1 Probability* specifies the probability with which, in a given time step, the P1 calls are brought to the front of the queue and responded to before the other calls. The *Prioritise P2 Probability* is the same as above, but for P2 calls. These are ordered in the model code such that the P1 calls are sorted first, and then the P2 calls. If both parameters are 1, the prioritisation behaviour collapses into the double sorting approach.

Dispatcher Classification

For investigating the effects of improved dispatcher classifications of call types, the *Dispatcher Correction Probability* was introduced. This parameter specifies the probability with which, in a given time step, the call types are corrected to match the triage classifications. If call types are corrected, then the following logic is executed: if the triage classification is Red and the call type is not patient transport, the call type is set to P1; otherwise if the triage classification is not Red and the call type is not patient transport, the call type is set to P2.

4.3 Verification and Validation

Model verification and validation are used in simulation-based modelling to test two key questions: whether the computer model reflects the conceptual model and whether the conceptual model, as implemented in the computer model, reflects reality (Thacker *et al.* 2004). Model verification is done first to test whether the computer model is an accurate representation of the conceptual model (Xiang *et al.* 2005). Once there is reasonable confidence that the computer model is coded correctly, model validation is done to test whether the model accurately represents the real-world system being modelled (Thacker *et al.* 2004).

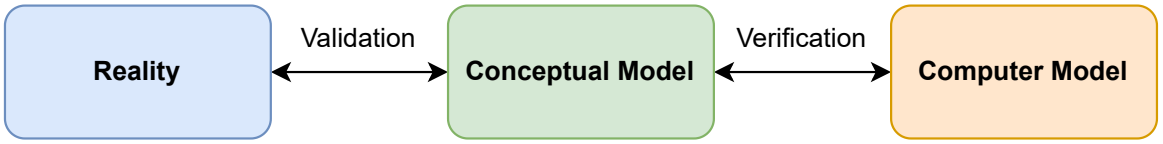


Figure 4.11: Illustration of Verification and Validation, adapted from Thacker *et al.* (2004)

In verification, the general procedure involves comparing simulation outputs to those expected by the conceptual model, and using these comparisons to identify bugs in the model code. As I built the model, a big challenge early on was getting it to execute without runtime errors. Once the model did run successfully, it was difficult to get it to produce results that remotely made sense, according to the conceptual model. Some example issues, among many include:

- Hypersensitivities to input parameters resulting in unstable results.
- Some patients being left in the queue for days without being responded to, while other patients were responded to timeously.
- Basic rules in the dispatching algorithm not being followed (for example, sometimes distant vehicles were dispatched to patients when closer vehicles were available).
- The same staff member being assigned twice to a vehicle, as both the driver and passenger.

As each issue was fixed, new ones were at times introduced. Some issues necessitated changes to the conceptual model. For example, the prioritisation probabilities were introduced because the system was overly efficient at prioritising P1 calls, resulting in unrealistically low response times. However, I eventually got the model to behave as expected by the conceptual model. The extent of agreement between the computer model's behaviour and the behaviour expected by the conceptual model is captured in Table 5.2.

For validation, model behaviour was compared with the data using side-by-side bar plots and histograms, in section 5.3.

4.4 Scenario Analysis

Changes were made to the baseline parameter values to create a series of scenarios, each one aimed at achieving one or more objectives. Bias-Correacted Bootstrapping with 150 000 resamples was used to obtain 95% confidence intervals for median response times, and percentages of staff and vehicles. These results form the basis of recommendations. The parameter combinations used to achieve each scenario are shown in table 5.3, and the results of each scenario are discussed in detail in section 5.4.

4.5 Sensitivity Analysis

Sensitivity analyses investigate the relationship between the parameters and model outcomes – a necessary step, because agent-based models do not have closed-form expressions linking the parameters to the outputs (Borgonovo *et al.* 2022). In particular, sensitivity analyses investigate which parameters most strongly affect the outcome, and how parameters affect the outcome both individually and jointly through interactions (Chen *et al.* 2021; Borgonovo *et al.* 2022). This process helps one to better understand the inner workings of the model, and to test the robustness of the model outcomes, and any conclusions reached, to changes in the parameters (Borgonovo *et al.* 2022).

For performing a sensitivity analysis, it is necessary to have a set of parameter values, run the model for each parameter set, and then investigate the model outputs (Collins *et al.* 2013). The generation of a parameter set is simple for the case of a single parameter: one can just sample values from some appropriate distribution to form the parameter set (Collins *et al.* 2013). However, parameter set generation becomes more difficult as the number of parameters increases: for example, if one performs a grid search over possible values, the required number of parameter combinations increases exponentially as more parameters are added (Collins *et al.* 2013). McKay *et al.* (1979) introduced an approach called Latin Hypercube Sampling (LHS) for parameter set generation. In LHS, a distribution for each parameter is specified (Pereira & Broed 2006). Parameters are then sampled from equally-likely intervals, based on the specified distributions (Pereira & Broed 2006). This approach achieves a high degree of coverage over the parameter space while requiring far fewer parameter combinations compared with earlier approaches such as Random Sampling (McKay *et al.* 1979). For this research, 3390 parameter sets were sampled using LHS. The distributions from which the samples were generated using LHS are shown in section 5.5.

Random forests are well-suited as metamodels for sensitivity analyses of complex agent-based models with many parameters (Granato & Li-Jessen 2020). I used the Ranger package in R to fit the random forest (Wright & Ziegler 2017). The predictor variables were the model parameters, with values obtained from the LHS procedure. The response variable was median P1 response time for each run, obtained by running the model for each parameter set. Hyperparameters were estimated by using the `train` function from the Caret library in R (Kuhn & Max 2008). The hyperparameter values are shown in section 5.5.

Chapter 5

Results

5.1 Introduction

This chapter has four sections. Section 5.2 presents the results of model verification, describes all parameters used in the model and presents the parameter values used in the base scenario. Section 5.3 details the results of model validation. Section 5.4 outlines in detail the main findings of this research, obtained by running the model using differing parameter combinations, in order to investigate the impact of carefully-chosen scenarios on the outcomes of interest. It details the parameter combinations used for each scenario, and presents the impact of each scenario on response times, staff usage and vehicle usage. Finally, section 5.5 contains the results of the sensitivity analysis, which examines how each parameter affects response time.

5.2 Model Verification

In simulation modelling, verification is typically done first to test whether the computer model is an accurate representation of the conceptual model (Xiang *et al.* 2005). Once there is reasonable confidence that the computer model is coded correctly, model validation is done to test whether the model accurately represents the real-world system being modelled (Thacker *et al.* 2004).

Before discussing the expected effects of key parameters in Table 5.2, it is necessary to introduce the model parameters. In Table 5.1, each parameter is described, and each parameter value used in the base model is shown.

Group	Parameter	Description	Base Value
Number of AMBs	Gqeberha 2	Number of AMBs in Gqeberha 2.	4
	Gqeberha 1	Number of AMBs in Gqeberha 1.	1
	Uitenhage	Number of AMBs in Uitenhage.	3
	Motherwell	Number of AMBs in Motherwell.	1
	West End	Number of AMBs in West End.	1
Number of PTVs	Gqeberha 2	Number of PTVs in Gqeberha 2.	4
	Gqeberha 1	Number of PTVs in Gqeberha 1.	0
	Uitenhage	Number of PTVs in Uitenhage.	2
	Motherwell	Number of PTVs in Motherwell.	1
	West End	Number of PTVs in West End.	0
Number of SPs	Gqeberha 2	Number of SPs in Gqeberha 2.	3
	Gqeberha 1	Number of SPs in Gqeberha 1.	1
	Uitenhage	Number of SPs in Uitenhage.	2
	Motherwell	Number of SPs in Motherwell.	2
	West End	Number of SPs in West End.	1
Number of IPs	Gqeberha 2	Number of IPs in Gqeberha 2.	3
	Gqeberha 1	Number of IPs in Gqeberha 1.	1
	Uitenhage	Number of IPs in Uitenhage.	3
	Motherwell	Number of IPs in Motherwell.	2
	West End	Number of IPs in West End.	1
AMB Probability	P1	Prob a given P1 call requires an AMB.	0.98
	P2	Prob a given P2 call requires an AMB.	0.60
	PPT	Prob a given PPT call requires an AMB.	0.30
IP Probability	P1	Prob a given P1 call requires ≥ 1 IP.	0.70
	P2	Prob a given P2 call requires ≥ 1 IP.	0.00
	PPT	Prob a given PPT call requires ≥ 1 IP.	0.10
SP Probability	P1	Prob a given P1 call requires ≥ 1 SP.	0.00
	P2	Prob a given P2 call requires ≥ 1 SP.	0.50
	PPT	Prob a given PPT call requires ≥ 1 SP.	0.50
Mean Wait Times	P1 Scene	Mean time on scene for P1 calls.	20 mins
	P2 Scene	Mean time on scene for P2 calls.	25 mins
	PPT Scene	Mean time on scene for PPT calls.	20 mins
	Facility	Mean time spent at facility for all calls.	25 mins
Time Buffers	Scene	Half range of scene wait times.	10 mins
	Facility	Half range of facility wait times.	10 mins
Single Staff Prob	P1	Prob a given P1 call requires ≤ 2 staff.	0.05
	P2	Prob a given P2 call requires ≤ 2 staff.	0.30
	PPT	Prob a given PPT call requires ≤ 2 staff.	0.30
Prioritising Prob	P1	Prob P1 calls are responded to first.	0.30
	P2	Prob P2 calls are responded to first.	0.15
	Correction Prob	Prob that dispatcher classification is corrected, to match triage classification, on a given iteration.	0.00
	Reroute Prob	Prob that re-routing is allowed on a given iteration.	0.20

Table 5.1: Descriptions of each parameter, and the value of each parameter in the Base Scenario

As described in more detail in section 4.1, the mean wait times and time buffer Descriptions of each parameter, and the value of each were chosen in consultation with an expert source, Ms. Sibongiseni Ntengu from the ECDoH. The numbers of AMBs, PTVs, SPs and IPs assigned to each location in the base scenario were set using the daily operational status spreadsheet (see Figure A.4). The daily totals for each staff and vehicle type in this spreadsheet were cross-checked against historical averages from 2019-2021 obtained from the call centre spreadsheets. The remaining parameters were chosen to ensure that the response times and staff and vehicle behaviour observed in the data were well approximated in the base scenario.

Parameters were chosen and varied in the base model to test whether the conceptual model

outlined in section 4.2 had been coded correctly. Parameters were varied, and I recorded whether the model behaved as expected in response to each set of parameter changes. Table 5.2 shows the model verification results, and all changes to model parameters resulted in expected behaviour.

Parameter to vary	**Expected effect(s) on model outcomes**	
Number of AMBs	Increasing parameter decreases RT for all call types, when there is a staff surplus.	✓
Number of PTVs	Increasing parameter decreases RT for all call types, when there is a staff surplus.	✓
Number of IPs	Increasing parameter decreases RT for all call types, when there is a vehicle surplus.	✓
Number of SPs	Increasing parameter decreases RT for all call types, when there is a vehicle surplus.	✓
AMB probability	Increasing parameter increases RT for all call types. Increasing parameter increases proportion of calls responded to by AMBs.	✓ ✓
IP probability	Increasing parameter increases RT for all call types. Increasing parameter increases proportion of calls responded to by IPs.	✓ ✓
Mean wait times	Increasing each wait time parameter increases RT for all call types.	✓
Single staff prob	Increasing parameter decreases RT for all call types. Increasing parameter increases proportion of calls responded to by a single staff member.	✓ ✓
Prioritisation prob	Increasing parameter decreases RT for prioritised call types. Increasing parameter increases RT for other call types.	✓ ✓
Correction prob	Increasing parameter decreases RT for code Red call types. Increasing parameter decreases the number of calls classified as P1.	✓ ✓
Reroute prob	Increasing parameter decreases RT for code P1 calls. Increasing parameter increases RT for other call types.	✓ ✓

Table 5.2: Model Verification Results. Green checkmarks indicate whether the expected behaviour was observed.

This shows that the computer model is likely a good representation of the conceptual model. Validation is the next step.

5.3 Model Validation

This section contains the validation results, where the simulated outputs are compared with the available observed data for assessing whether the conceptual model is a good match of reality. The aim of validation is not to demonstrate that that model replicates the data perfectly. Rather, it is to show that the model replicates the data sufficiently well that it can be used to inform decisions (Pullum & Cui 2012).

Twenty simulation runs were used to compute these results, an equivalent of 40 weeks worth of simulations. The results were relatively stable with small numbers of simulations, so twenty simulations was chosen due to the low computational burden, while also serving as a representative sample of the model's behaviour. This validated the model with respect to response time distributions, staff and vehicle distributions, call type distributions, failure to convey (F/T) distributions and drop-off distributions, as well as intra- and inter-day call volume distributions.

The model's simulated response times should be similar to those observed in reality. Figure 5.1 compares the observed and simulated response time distributions.

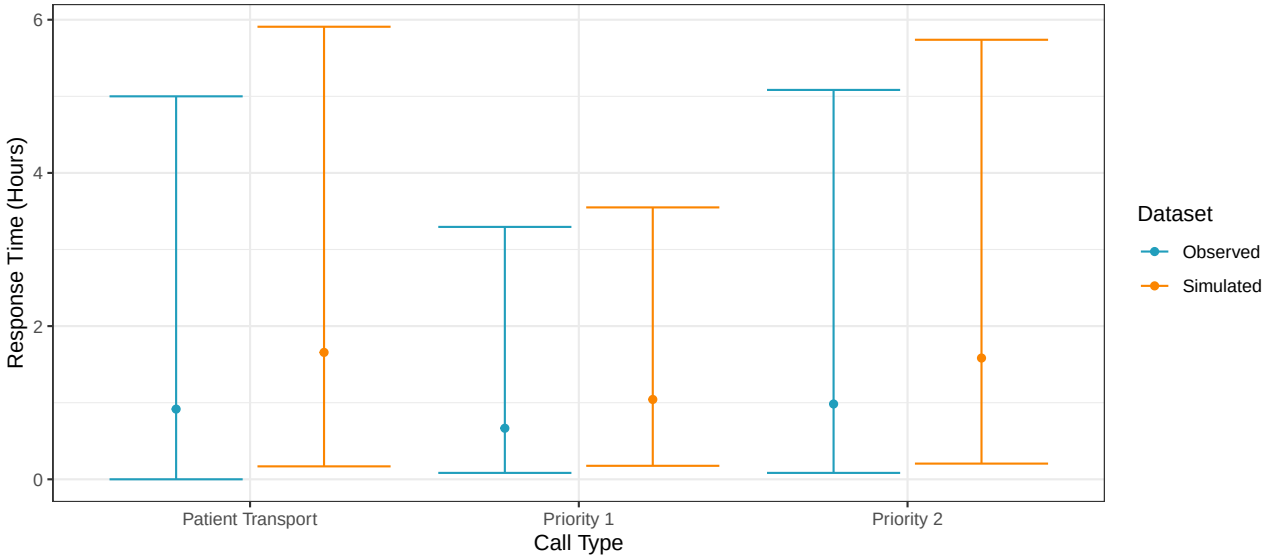


Figure 5.1: Comparison of Observed and Simulated Response Times. Points are medians and error bars are 95% quantiles for each variable.

There are some minor differences between observed and simulated response times, and a general tendency for simulated response times to be slightly greater than those observed in the data, as measured by the differing medians. However, the distributional spread, measured by the 95% quantiles, is very similar between the two datasets. Response times in the base model may be slightly overestimated on average, but the distributions of response times appear to be sufficiently similar between the two datasets.

The staff mixture, vehicle type, call type and failure to convey (F/T) distributions are compared in Figure 5.2.

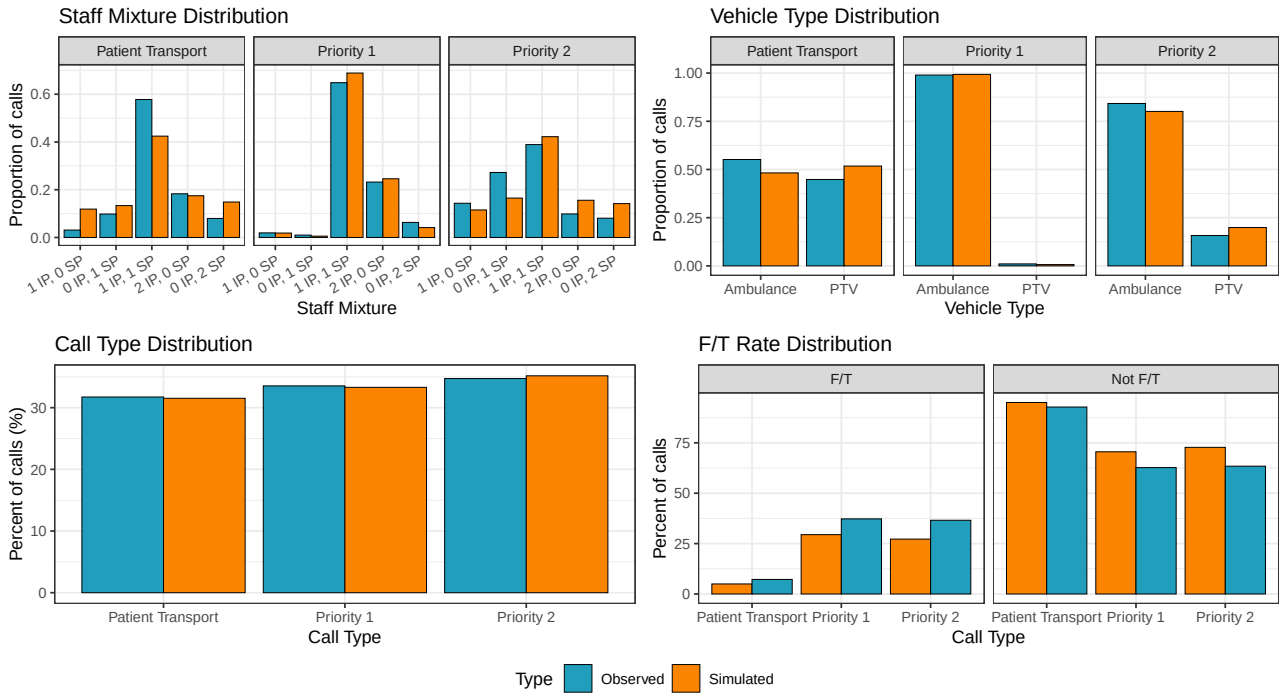


Figure 5.2: Staff Mixture, Vehicle Type, Call Type and Failure to Convey (F/T) Distributions, for simulated and observed datasets.

While there are some small differences in the proportions of certain staff mixtures between the observed and simulated distributions, there are no major differences. This indicates that the model with baseline parameter values produces a similar staff mixture distribution for each call type compared with the observed data.

For vehicle types, the two distributions are very similar, indicating that the model with baseline parameter values produces a good representation of the vehicle type distribution for each call type, compared with the observed data.

The distributions are almost identical for call types. Despite some small differences in the two distributions, with slightly more observed calls being F/T than in the observed data, the two distributions are similar overall.

When patients are handed off to a facility, the destination ward is recorded. The distributions of ward IDs are compared for the simulated and observed data in Figure 5.3, as well as the inter-day call volume distribution.

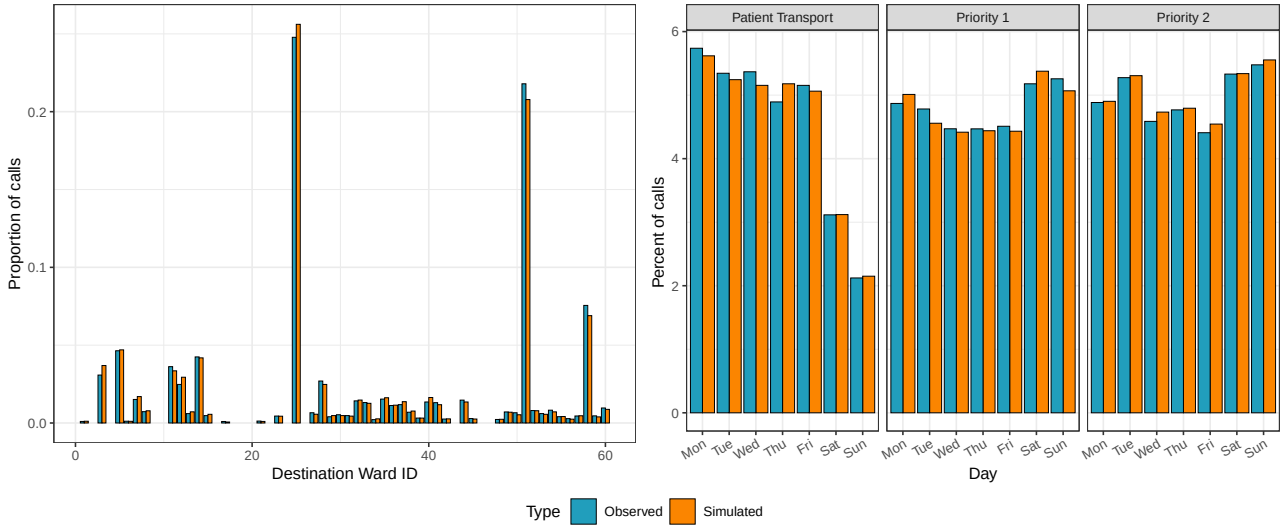


Figure 5.3: Incident Ward ID and Intra-Day Call Volume Distributions, for simulated and observed datasets.

In both cases, despite some small deviations, the two distributions are very similar.

Finally, the intra-day demand distributions are compared in Figure 5.4. The intra-day demand is presented in the following way: observed counts in each 10 minute period are binned, and these counts are smoothed using a Generalised Additive Model (GAM) with a negative binomial link function. 95% confidence intervals using GAM are plotted for the observed data, and the simulated 10-minute averages are overlaid as points, for each call type:

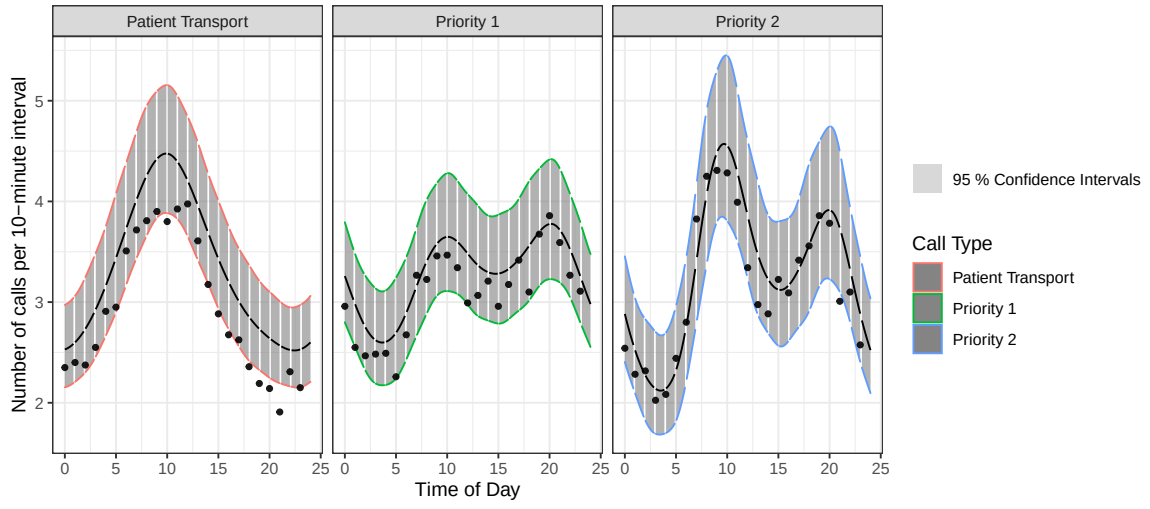


Figure 5.4: Comparison of intra-day demand between simulated and observed data, by call type.

Some of the simulated call volumes for patient transport calls lie below the 95% confidence interval for observed call volumes, resulting in slightly lower overall simulated call volumes for patient transport calls than expected. This could lead to a minor underestimation of the overall call demand in the system. However, all of the simulated means lie within the 95% confidence intervals for P1 and P2, indicating that the intra-day demand distributions produced by the base model are similar to the observed distributions. Given that response times are well calibrated, and that this research more focused on P1 and P2 response times, the effects of the minor underestimation of PPT demand on the overall results are expected to be minimal.

The model can be considered successfully validated, as it has passed a rigorous process of comparing its simulated outputs with observed outputs. Given that the model has also been successfully verified, we can be confident that the conceptual model has been coded correctly, and that this conceptual model is a reasonable approximation of the real-world system. We can now use the model to investigate ways to improve this system.

5.4 Scenario Analysis

Each scenario was designed, in discussion with CHAI, with three goals in mind: to lower response times for the most urgent calls, to improve adherence to staff regulations, and to allocate vehicles more appropriately. Some scenarios target individual goals, while others target multiple goals simultaneously.

In this section, each scenario is briefly described, while the exact parameter combinations used are detailed in Table 5.3. Here, the impacts of each scenario on response times and, where relevant, staff and vehicle allocations, are examined. To compute confidence intervals around each estimate, bias-corrected bootstrapped confidence intervals are used. Detailed tabular versions of each figure can be found in the Appendix, and where relevant these are referenced in the main text.

Scenarios were each run 150 times, or 300 weeks, totalling 302 400 iterations per scenario. In order to have minimal uncertainty about the model's behaviour in the base scenario I performed a total of 903 runs for this scenario. 150 runs were performed for other scenarios because the bootstrapping procedure, which was used to construct 95% confidence intervals for response times and other metrics, yielded sufficiently narrow confidence intervals with this number of runs. The objective was to perform enough runs of each scenario such that the bootstrapped confidence intervals were narrow enough to be able to make meaningful comparisons between each scenario and the base scenario. With 150 runs, the confidence intervals for each scenario rarely overlapped those of the base scenario. Additional runs were not performed due to computational cost. Median and average response times for all call types were recorded for each run.

Scenario	Change(s) Relative to Base Scenario
1. Improving vehicle allocation	P1 AMB probability increased to 1. PPT AMB probability decreased to 0.
2. Improving staff allocation	P1 IP probability increased to 1. PPT IP probability decreased to 0. P1 single staff probability decreased to 0. P2 single staff probability decreased to 0.2.
3. Improving prioritisation	Prioritise P1 probability increased to 0.8. Prioritise P2 probability increased to 0.8.
4. Improving classification	Correction probability increased from 0 to specified value, shown in figure.
5. Improving prioritisation & classification	Correction probability increased from 0 to specified value, shown in figure. P1 prioritisation probability increased to 0.5. Reroute probability increased to 0.3.
6. Increasing AMBs	Number of AMBs assigned to specified location are increased by 1.
7. Increasing PTVs	Number of AMBs assigned to specified location are increased by 1.
8. Increasing IPs	Number of IPs assigned to specified location are increased by 1.
9. Increasing SPs	Number of SPs assigned to specified location are increased by 1.
10. Adding 1 IP and 1 SP	Number of SPs and IPs assigned to specified location are both increased by 1.
11. Low Cost 1	Prioritise P1 probability increased to 0.8. Prioritise P2 probability increased to 0.8. Priority correction probability increased to 0.3. Reroute probability increased to 0.3.
12. Low Cost 2	Same changes as Low Cost 1. P1 AMB probability increased to 1. PPT AMB probability decreased to 0. P1 IP probability increased to 1. PPT IP probability decreased to 0.
13. Medium Cost 1	Same changes as Low Cost 1. One IP added to Uitenhage.
14. Medium Cost 2	Same changes as Low Cost 2. One IP added to Uitenhage.
15. High Cost 1	Same changes as Low Cost 1. One IP and one SP added to Uitenhage.
16. High Cost 2	Same changes as Low Cost 2. One IP and one SP added to Uitenhage.

Table 5.3: Changes in each parameter for each scenario, relative to the Base Scenario. IPs refer to Independent Practitioners, SPs refer to Supervised Practitioners. P1 refers to Priority 1 calls, P2 to Priority 2 and PPT to Planned Patient Transport. PTVs refer to Patient Transport Vehicles, and AMBs to Ambulances.

5.4.1 Optimising existing resources

A number of scenarios aim to optimise the system's existing resources, without hiring additional staff members or acquiring additional vehicles. Two such scenarios target the goals of improved vehicle staff allocations. An additional two scenarios target response times, by improving the prioritisation of P1 calls and increasing dispatcher call classification accuracy.

Improving dispatching of staff and vehicles
The first scenario aims to improve the joint distribution of vehicle types and call types, by increasing the proportion of P1 calls responded to by AMBs, and increasing the proportion of PPT calls responded to by PTVs.

The second scenario aims to improve the joint distribution of staff types and call types, by increasing the proportion of P1 calls responded to by at least one IP, decreasing the proportion of PPT calls responded by IPs, and reducing the proportion of non-PPT calls responded to by a single staff member. The impact of each scenario on median response times is examined in Figure 5.5.

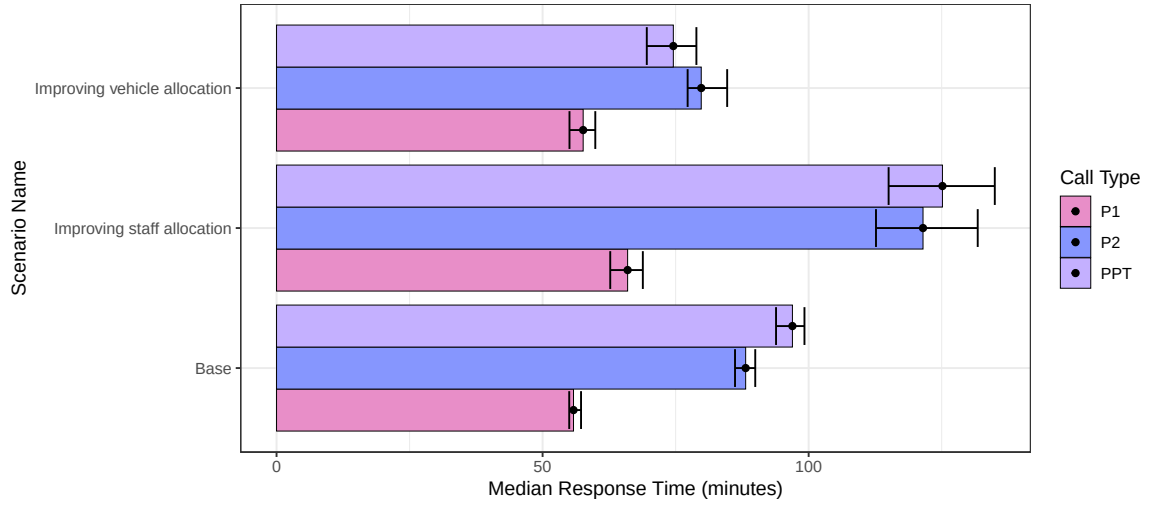


Figure 5.5: Median response time, by scenario and call type: scenarios aiming to improve allocations of staff and vehicles. Error bars show 95% bootstrapped BCa confidence intervals.

Median response times in the base scenario are 55.81 minutes (95% CI: 55.00, 57.23) for P1 calls, 88.16 minutes (95% CI: 86.16, 89.97) for P2 calls and 96.95 minutes (95% CI: 93.87, 99.21) for PPT calls. For the tabular version of this figure, and for all bootstrapped confidence intervals see Table B.2.

In the *Improving vehicle allocation* scenario, there are no large differences in P1 response times relative to the base scenario. However, median response times for P2 and PPT calls are lower in this scenario relative to their values in the base scenario. Response times in the *Improving vehicle allocation* scenario are 79.82 minutes (95% CI: 77.26, 84.69) for P2 calls and 74.57 minutes (95% CI: 69.58, 78.91) for PPT calls. While these differences are not large, they are noteworthy due to non-overlapping 95% confidence intervals.

By contrast, in the *Improving staff allocation* scenario, response times are considerably higher than their values at baseline. There are particularly notable increases in median response times for P2 and PPT call types in this scenario. Response times in the *Improving staff allocation* scenario are 121.49 minutes (95% CI: 112.66, 131.77) for P2 calls and 125.14 minutes (95% CI: 115.03, 134.98) for PPT calls. This increase in response time is due to the additional constraints

placed on the system by more rigidly enforcing staff and vehicle rules. For example, if at least one IP is required to respond to a P1 call, then sometimes an idle and staffed vehicle will be unable to be dispatched because it does not have the correct staff mixture. This results in increased response times, but in return more appropriate staff mixtures and vehicles respond to calls.

Figure 5.6 examines the differences in the joint distributions of vehicles and call types, for the base scenario and the *Improving vehicle allocation* scenario.

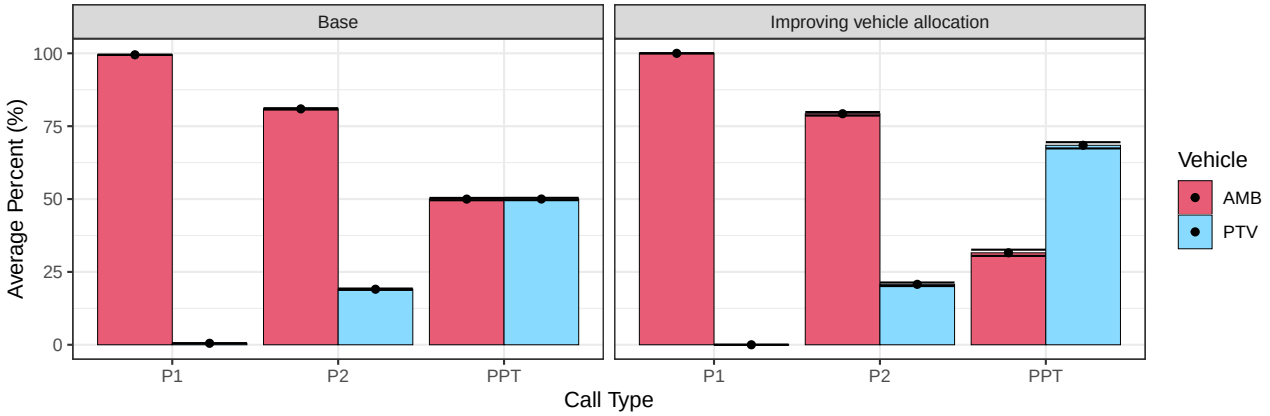


Figure 5.6: Joint distributions of vehicle types and call types, for two scenarios. Error bars show 95% bootstrapped BCa confidence intervals.

In the base scenario, an average of 50.0% of PPT calls are responded to by ambulances (95% CI: 49.6%, 50.4%). In the *Improving vehicle allocation* scenario, only 31.6% (95% CI: 30.5%, 32.6%) of PPT calls are responded to by ambulances. All other percentages are comparable, but it is notable that P1 calls are never responded to by PTVs in the *Improving vehicle allocation* scenario, a result of increasing the P1 AMB probability to 1. Refer to Table B.7 for the tabular version of this figure.

Figure 5.7 examines the differences in the joint distributions of staff and call types, for the base scenario and the *Improving staff allocation* scenario.

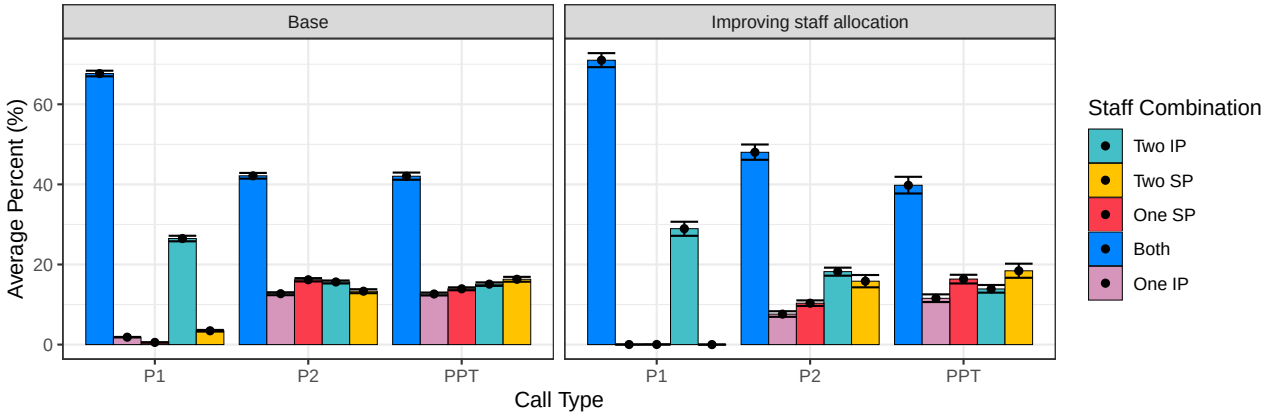


Figure 5.7: Joint distributions of staff mixtures and call types, for two scenarios. Error bars show 95% bootstrapped BCa confidence intervals.

Calls can either be responded to by two independent practitioners (Two IP), two supervised practitioners (Two SP), one supervised practitioner only (One SP), one independent practitioner

only (One IP) or one independent practitioner and one supervised practitioner (Both).

The following features of the *Improving staff allocation* scenario can be observed: a reduction in the percentages of One SP and Two SP for P1 and P2 calls relative to the base scenario, and a reduction in the percentages of Two IP and One IP for PPT calls relative to the base scenario.

Most notably, the percentage of P1 calls responded to by One SP is 0% in the *Improving staff allocation* scenario. The percentage of P1 calls responded to by Two SPs is also 0%, implying that P1 calls are always responded to by at least one IP. These results indicate that the parameter changes for this scenario have been successful at improving staff allocations, at the expense of increased response times.

Improving prioritisation

An additional scenario investigates the effects on response times of more rigidly enforcing the prioritisation rules, and obtaining decreases in P1 and P2 response times at the expense of PPT response times. The impact of this scenario on median response times is examined in Figure 5.8:

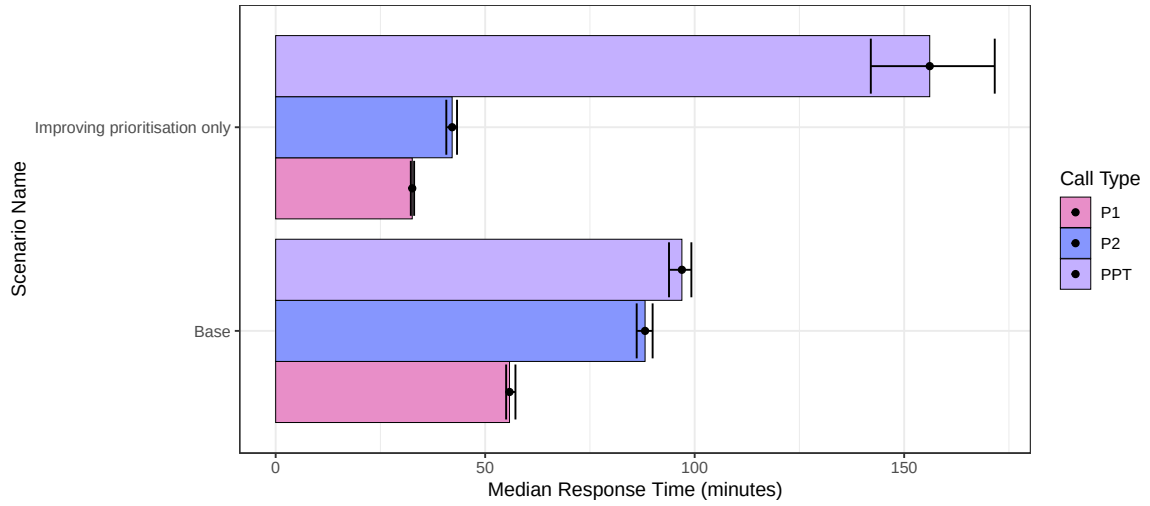


Figure 5.8: Median response time, by scenario and call type, for a scenario aiming to improve prioritisation of P1 and P2 calls. Error bars show 95% bootstrapped BCa confidence intervals.

This scenario results in considerable decreases in response times for P1 and P2 calls relative to the base scenario, but is mirrored by an equally dramatic increase in response times for PPT calls. In this scenario, median response times are 32.60 minutes (95% CI: 32.26, 33.05) for P1 calls, 42.12 minutes (95% CI: 40.73, 43.28) for P2 calls and 156.10 minutes (95% CI: 142.04, 171.62) for PPT calls.

Improving dispatcher call classification

A final set of three sub-scenarios aimed at reducing response times for the most urgent calls involve improving the dispatcher call type classification accuracy. With a given probability (for these scenarios probabilities of 0.2, 0.5 and 0.8 are used), the classifications assigned to calls by dispatchers are corrected to match the triage classifications assigned by EMS personnel at the scene. In the base scenario, the priority correction probability is 0, and in these three sub-scenarios, the effects of increasing it are explored.

For understanding better how these changes affects the behaviour of the modelled system, Figure 5.9 shows how the distribution of call types varies between the four scenarios presented above.

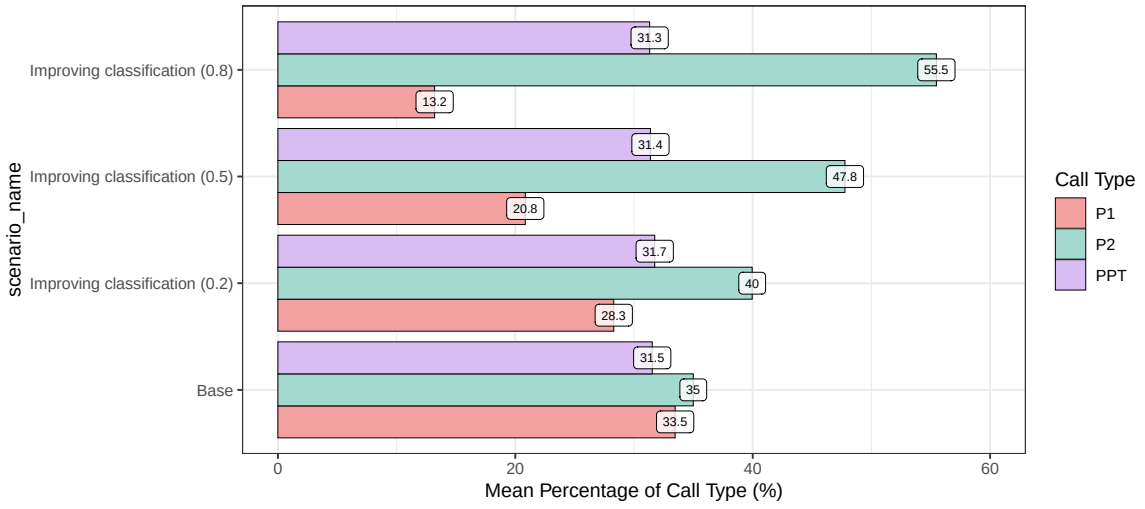


Figure 5.9: Distribution of call types, for four scenarios.

As the correction probability increases, the net effect is that more calls are classified as P2 and fewer are classified as P1. In the base scenario, an average of ∼33.5% of calls are classified as P1, but this number drops to only ∼13.2% in the *Improving classification (0.8)* scenario. This is because as the correction probability increases, the dispatcher call classification distribution better matches the triage classification distribution, and the percentage of P1 calls converges to the percentage of code red calls. Only ∼8.9% of calls are classified as code red (see Figure 3.14). When there are fewer P1 calls to respond to, the system is able to prioritise them more effectively. Prioritising P1 calls also has less of an effect on the response times of P2 and PPT calls when there are fewer P1 calls.

As the percentage of calls classified as P1 declines, the effect of strong P1 prioritisation measures on P2 and PPT response times also declines. However, the effect of the prioritisation measures on P1 response times is unchanged. The impact of this scenario, for a range of probabilities,

on median response times is examined in Figure 5.10.

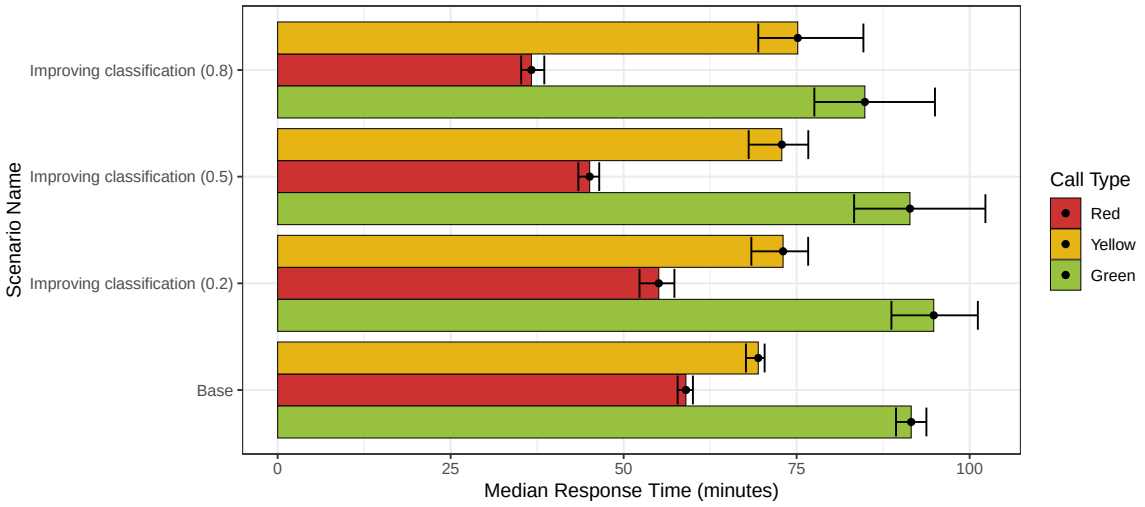


Figure 5.10: Median response time, by scenario and triage classification. Error bars show 95% bootstrapped BCa confidence intervals.

As the classification correction probability increases, response times for code Red calls decline relative to their values in the base scenario, as they become more likely to be classified as P1 calls and accordingly prioritised. Median response times for code Red calls are 55.08 minutes (95% CI: 52.42, 57.35) for the *Improving classification (0.2)* scenario, 45.10 minutes (95% CI: 43.45, 46.44) for the *Improving classification (0.5)* scenario and and 36.68 minutes (95% CI: 35.19, 38.53) for for the *Improving classification (0.8)* scenario.

Code Yellow response times increase slightly as the correction probability increases: in the *Improving classification (0.8)* scenario, median response times for code Yellow calls are 75.16 minutes (95% CI: 69.46, 84.07), compared to 69.43 minutes (95% CI: 67.83, 70.41) in the base scenario. This is explained by fewer code Yellow calls being classified as P1 by dispatchers, resulting in a general de-prioritisation of code Yellow calls. Code Green response times do not differ significantly from their values in the base scenario. See Table B.5 for the tabular version of Figure 5.10.

Improving Prioritisation, Classification and Re-routing

The final set of scenarios in this section explore the effects on response times of increasing the dispatcher classification accuracy, while also increasing the rerouting probability and improving the prioritisation of P1 calls, relative to the base scenario. In these scenarios, prioritisation was improved by increasing the P1 prioritisation probabilities, and also by increasing the rerouting probability. The classification correction probability was set to 0.1, 0.3 and 0.5, in order to explore a range of moderate improvements to dispatcher classifications. Full parameter combinations used for these sub-scenarios are shown in Table 5.3. The effects on response times of each scenario are shown in Figure 5.11.

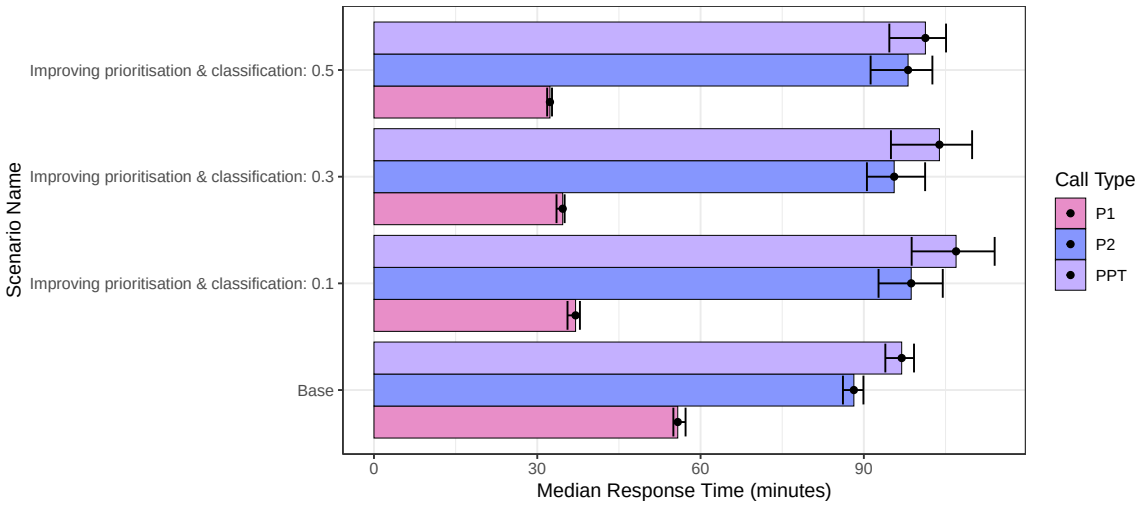


Figure 5.11: Median response time, by scenario and call type, for scenarios that improve prioritisation. Error bars show 95% bootstrapped BCa confidence intervals. The number in each scenario name is the dispatcher correction probability.

The increased probability of vehicles re-routing to P1 calls, and the increased prioritisation of P1 calls, cause notable decreases in P1 response times in all scenarios. Even in the *Improving prioritisation & classification: 0.1* scenario, where the dispatcher correction probability is only 0.1, median P1 response time is 37.04 minutes (95% CI: 35.56, 37.85). However, response times for other call types do increase substantially as a result of increased prioritisation of P1 calls, as noted previously.

When prioritisation of P1 calls is heightened, increasing the dispatcher correction probability does not decrease P1 response times much further, a sign of diminishing returns. For example, median P1 response time in the *Improving prioritisation & classification: 0.5* scenario is 32.60 minutes (95% CI: 31.82, 32.71), which is similar to the results in Figure 5.11.

5.4.2 Adding resources

Four sets of scenarios investigate the effects of augmenting the system's existing resources. These include adding ambulances (AMBs), patient transport vehicles (PTVs), independent practitioners (IPs) and supervised practitioners (SPs) to each of the five bases. The aim of each of these sets of scenarios is to reduce response times for P1 calls by increasing resources. Each sub-scenario involves increasing the number of staff or vehicles of the relevant type assigned to a given base by one unit (e.g. increasing the number of PTVs assigned to the base in Uitenhage by 1 vehicle).

Increasing AMBs

The first five sub-scenarios investigate the effects of increasing the number of ambulances available to each base, in turn, by one vehicle. Median response times for each sub-scenario are displayed in Figure 5.12.

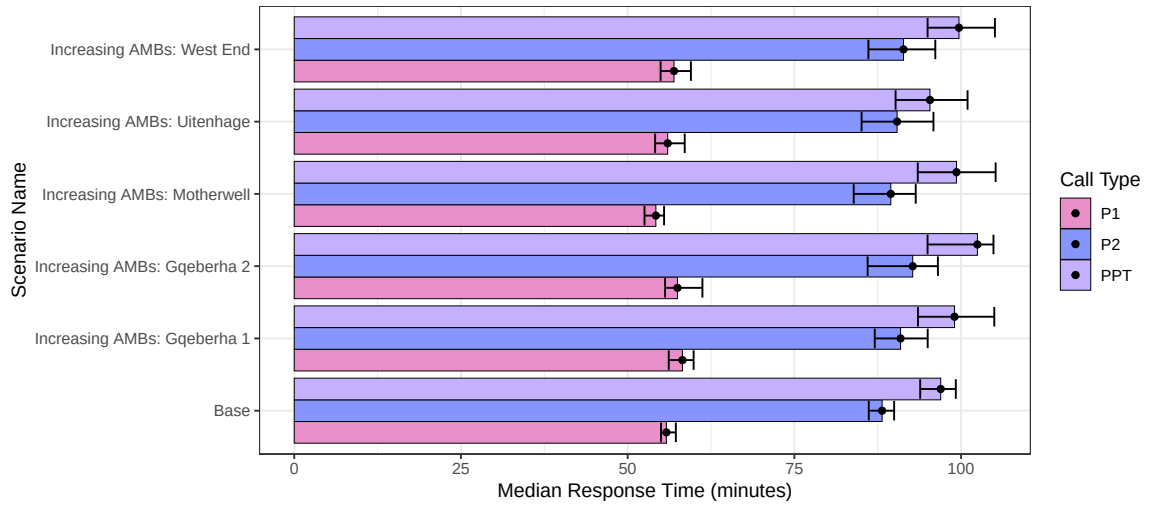


Figure 5.12: Median response time, by scenario and call type, for sub-scenarios that increase AMBs. Error bars show 95% bootstrapped BCa confidence intervals.

Only the addition of one ambulance was examined because it was expected that staff, not ambulances, are the primary bottleneck for decreasing response times. As expected, median response times for all call types remain approximately unchanged in all above scenarios relative to the base scenario. The *Increasing AMBs: Motherwell* sub-scenario shows a small decrease in median P1 response times relative to the base scenario, with a median response time of 54.23 minutes (95% CI: 52.53, 55.46). The Motherwell EMS base does have a 2:1 ratio of staff to vehicles in the base scenario, with one AMB and one PTV assigned. These results indicate that there could be too few vehicles relative to staff in the Motherwell EMS base.

Increasing PTVs

The effects on response time of increasing the number of PTVs available to each base are displayed in Figure 5.13.

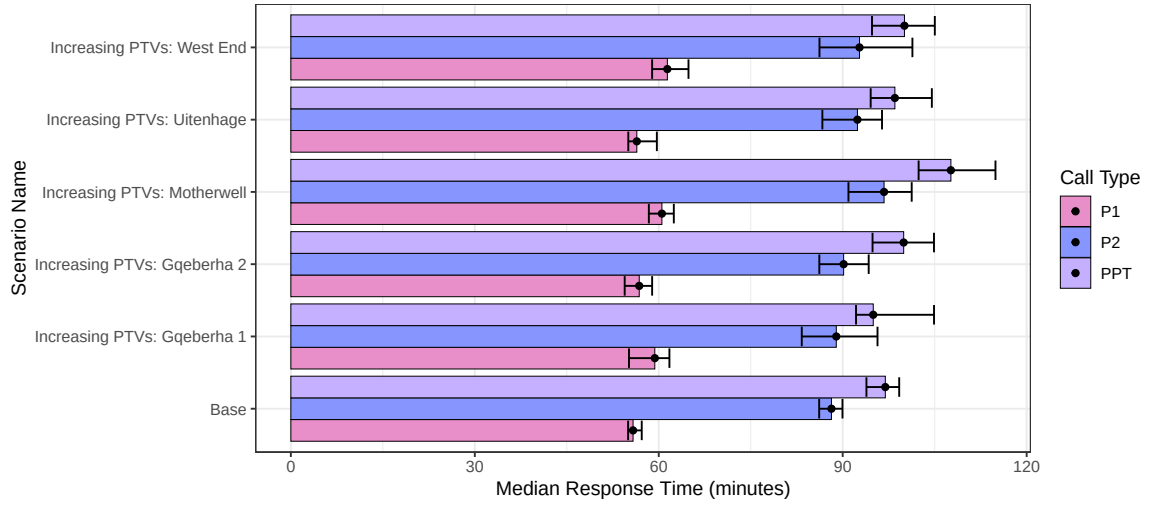


Figure 5.13: Median response time, by scenario and call type, for sub-scenarios that increase PTVs. Error bars show 95% bootstrapped BCa confidence intervals.

Response times for all call types are approximately unchanged in all the above sub-scenarios relative to the base scenario. By contrast to the previous scenario, here there are some small increases in response times relative to the base scenario, in some cases. For example, median P1

response times are slightly higher in the *West End* and *Motherwell* sub-scenarios, with times of 61.41 minutes (95% CI: 58.92, 64.85) and 60.51 minutes (95% CI: 58.39, 62.46) respectively.

One possible explanation for this effect is as follows: PTVs typically do not respond to P1 and P2 calls, which collectively make up the majority of calls. When additional PTVs are added to a base, the ratio of PTVs to AMBs increases, making it more likely that a PTV will be staffed and dispatched into the system while an AMB remains idle, due to insufficient staffing. When this happens, the system's capacity to respond to P1 and P2 calls is slightly reduced. This has the effect of increasing response times for all call types. Response times also increase for PPT calls because, while only the system's capacity to respond to P1 and P2 calls is reduced, these call types are also prioritised over PPT calls. The net increase is a slight increase in response times for all call types.

Increasing IPs

The effects on response time of increasing the number of IPs available to each base by one staff member are displayed in Figure 5.14. The scenarios that increase the number of staff available by one staff member increase the number available at a given ward at all times. Because no staff member can work 24/7, adding one staff member at all times means employing an additional four to five staff members, given a 40 hour work week.

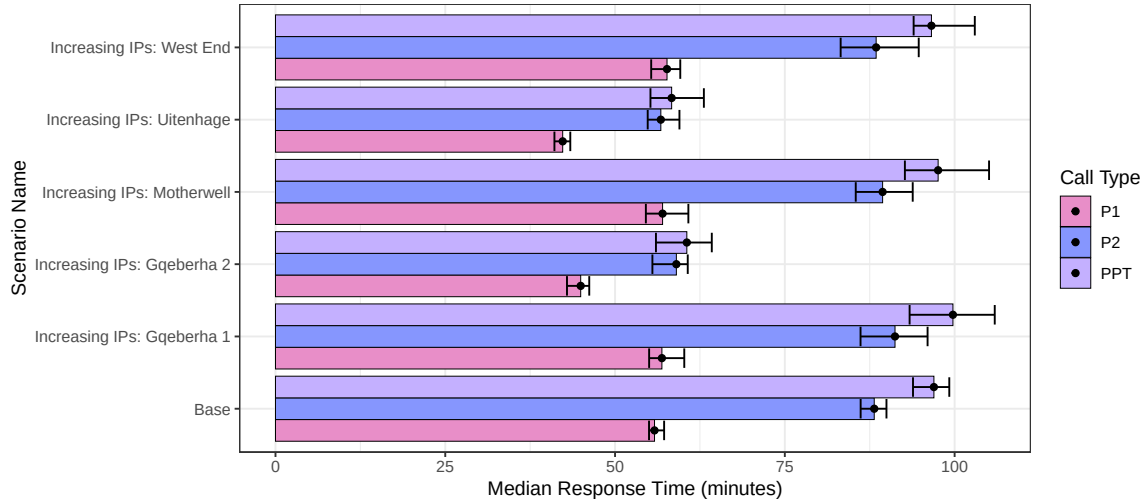


Figure 5.14: Median response time, by scenario and call type, for sub-scenarios that increase IPs. Error bars show 95% bootstrapped BCa confidence intervals.

For some wards, adding IPs significantly decreases median response times for all call types. In sub-scenarios involving adding IPs to bases Uitenhage and Gqeberha 2, median response times decrease significantly relative to the base scenario. In the *Uitenhage* and *Gqeberha 2* sub-scenarios, median P1 response times times are 42.29 minutues (95% CI: 41.07, 43.41) and 44.94 minutes (95% CI: 42.92, 46.20) respectively. For other bases, where there are insufficient vehicles available to sufficiently utilise the additional staff members, the increased IPs do not reduce response times.

Increasing SPs

The effects on response time of increasing the number of SPs available to each base by one staff member are displayed in Figure 5.15.

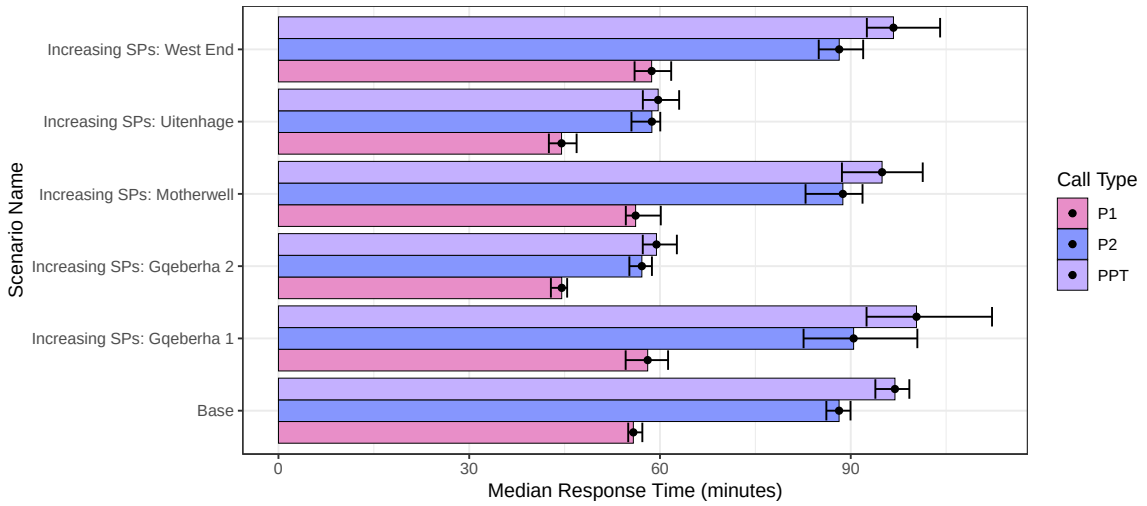


Figure 5.15: Median response time, by scenario and call type, for sub-scenarios that increase SPs. Error bars show 95% bootstrapped BCa confidence intervals.

Similar results are observed for this set of sub-scenarios as for the previous set. In the *Uitenhage* and *Gqeberha 2* sub-scenarios, median response times decrease significantly relative to the base scenario, for all types. In the *Uitenhage* and *Gqeberha 2* sub-scenarios, median P1 response times times are 44.52 minutes (95% CI: 42.53, 46.89) and 44.55 (95% CI: 42.84, 45.40) respectively. In other sub-scenarios, no decreases in response times are observed, for reasons discussed previously.

Adding one SP and one IP

A final set of scenarios involves adding one IP and one SP to each base. The effects on median response time of each of these sub-scenarios are displayed in Figure 5.16.

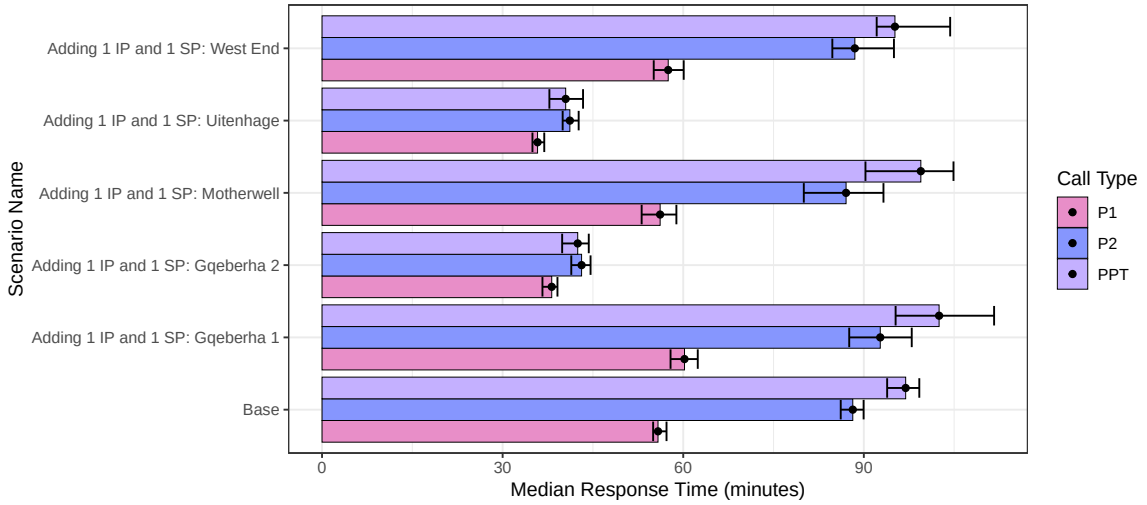


Figure 5.16: Median response time, by scenario and call type, for sub-scenarios that add one IP and one SP to each base. Error bars show 95% bootstrapped BCa confidence intervals.

Similar results are observed for this set of sub-scenarios as for the previous two sets. However, median response times are lower than previously achieved in the *Uitenhage* and *Gqeberha 2* sub-scenarios. Median response times in these sub-scenarios are considerably lower relative to the base scenario for all call types, indicating that these regions are bottlenecked by insufficient staff levels relative to vehicles. In the *Uitenhage* and *Gqeberha 2* sub-scenarios, median P1

response times times are 35.80 minutes (95% CI: 34.95, 36.91) and 38.16 minutes (95% CI: 36.61, 39.11) respectively.

As observed previously, adding staff members to bases which do not have a surplus of vehicles does not result in decreased median response times.

5.4.3 Comparing all scenarios

In Figure 5.17, the median reduction in P1 response time relative to the base scenario is compared for each scenario and sub-scenario:

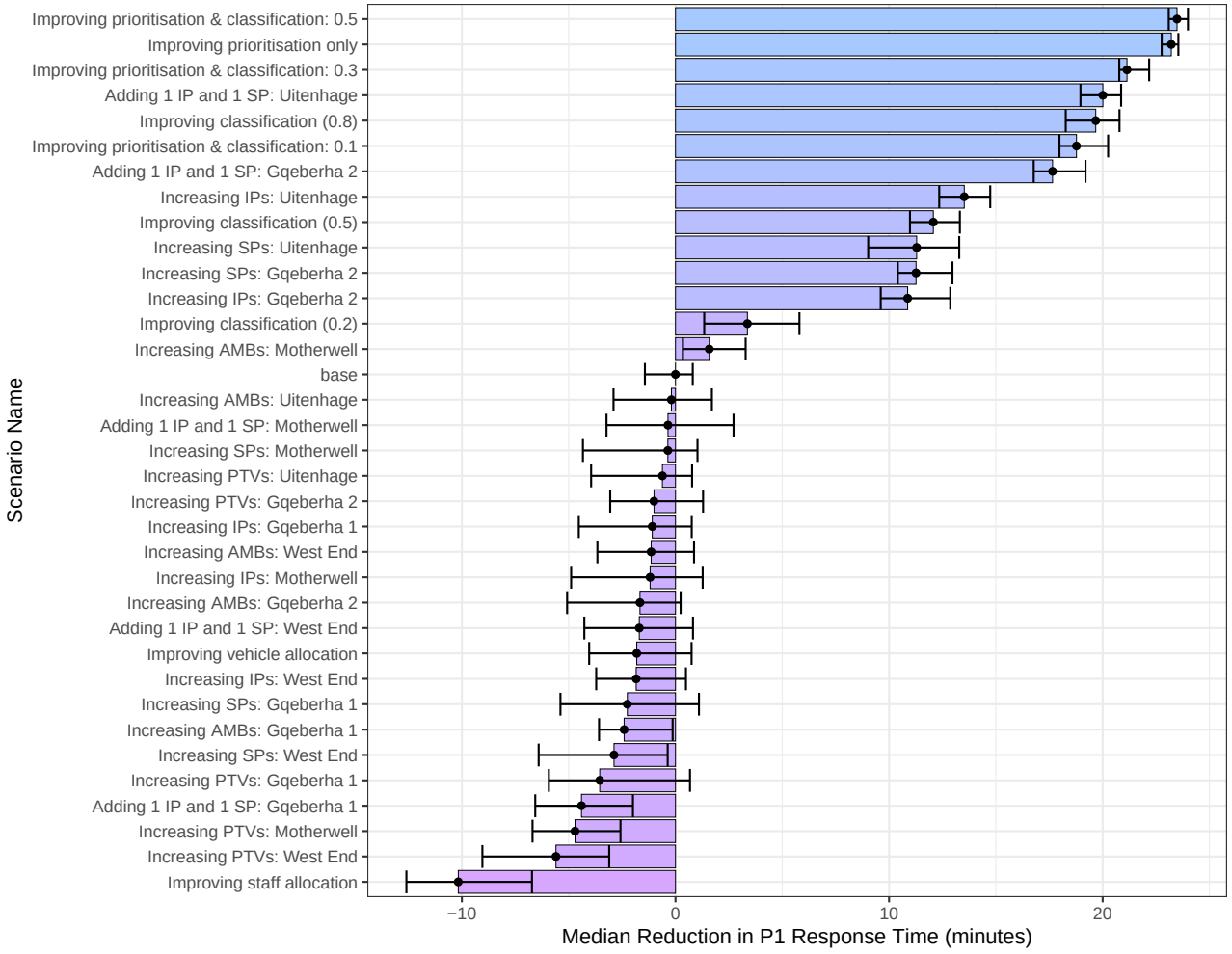


Figure 5.17: Comparison of all scenarios of single intervention types to base scenario, with respect to reduction in P1 response time. Error bars show 95% bootstrapped BCa confidence intervals.

The largest reductions in median P1 response times are seen in the scenarios involving changes to prioritisation and classification. Significant reductions in median P1 response are also observed in the *Adding 1 IP and 1 SP: Uitenhage* sub-scenario. Significant increases in median P1 response times are observed in the *Improving staff allocation* scenario, as we constrain which staff mixtures are capable of responding to P1 and P2 calls.

Each scenario has its benefits and drawbacks. For example, increasing the prioritisation of P1 and P2 calls is effective at reducing P1 and P2 response times, but this comes at the expense of PPT response times. Increasing the number of IPs is also effective at reducing response times,

in some cases, but each additional staff member employed incurs a financial cost to the ECDoH, which must be considered. Constraining the staff mixtures that can respond to P1 and P2 calls is effective at improving the call type, and staff mixture joint distribution, but comes at the expense of increased response times.

5.4.4 Combining scenarios

This section explores the effects of three further sets of scenarios, each at a different cost level. Each scenario combines combines multiple scenarios outlined above. The idea is that the drawbacks of some scenarios may be offset by the benefits of others, when they are implemented simultaneously. Increases to vehicles were not particularly effective at reducing response times in any of the scenarios (due to the main bottleneck being staff, not vehicles). Therefore, only increases to the numbers of staff members assigned to bases are considered in the combined scenarios.

These are separated into two low cost scenarios, two medium cost scenarios and two high cost scenarios. Cost is based purely on the number of additional staff members involved in the scenario, and does not include the cost associated with training interventions[1].

In the Low Cost scenarios, no staff members or vehicles are added; in the Medium Cost scenarios, one staff member is added, and in the High Cost scenarios two staff members are added. Since it was found in the previous scenarios (see Figure 5.17) that increasing the numbers of staff assigned to the Uitenhage base were particularly effective at reducing response times, the medium and high cost scenarios explore the effects of increasing the number of staff members assigned to the Uitenhage ward only. For full details on the parameters used in each of these scenarios, see Table 5.3. In all these combined scenarios, the P1 and P2 prioritisation probabilities, the priority correction probability, and the re-route probability were increased. In the *Low Cost 2*, *Medium Cost 2* and *High Cost 2* sub-scenarios, staff and vehicle allocations were also targeted. This was achieved be increasing the P1 AMB probability, and the P1 IP probability, and by decreasing the PPT AMB probability and PPT IP probability.

The effects of each of these combined scenarios on median response times is examined in Figure 5.18.

[1]The Low Cost 1, Medium Cost 1 and High Cost 1 scenarios all have the same parameter values, except that the higher cost scenarios have additional staff members. The same is true of the Low Cost 2, Medium Cost 2 and High Cost 2 scenarios. Implementing the Medium Cost and High Cost scenarios therefore has the same expected training costs as their lower cost counterparts, but with additional staff costs.

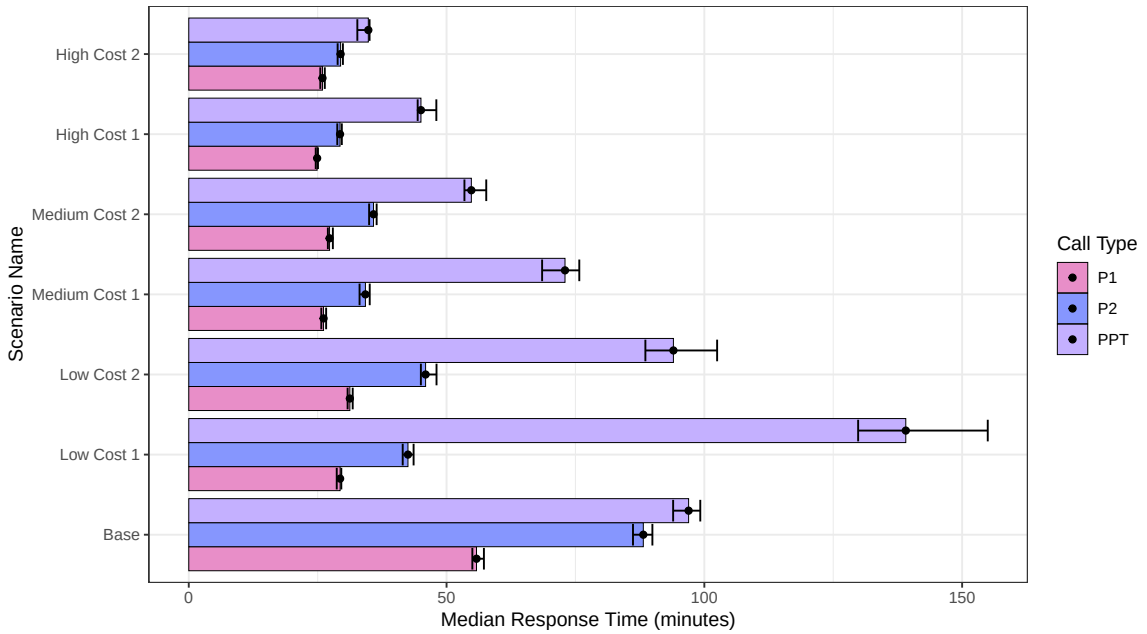


Figure 5.18: Median response time, by scenario and call type, for high and low cost scenarios. Error bars show 95% bootstrapped BCa confidence intervals.

Median P1 and P2 response times are considerably lower in all the above scenarios relative to the base scenario. There also appear to be diminishing returns to decreases in response times as cost levels increase. P1 response times are dramatically lower in the *Low Cost 1* scenario compared to the base scenario, but the incremental reduction achieved in *High Cost 1* is not large: P1 response times are 29.36 minutes (95% CI: 28.73, 29.59) in *Low Cost 1* and 24.90 minutes (95% CI: 24.63, 25.07) in *High Cost 1*.

Median PPT response times also exhibit diminishing returns as the scenarios progress from *Low Cost 1* to *High Cost 2*. However, there are still notable declines in response times with each progression between scenarios. It is high initially, in the Low Cost scenarios, because the increased prioritisation of P1 and P2 calls causes PPT calls to experience longer queue times. This is offset as the increased staff members in the Medium Cost and High Cost scenarios progressively decrease median response times. Within each cost level, median PPT response times also decline when moving from the first to the second sub-scenario within each cost level. For example, median PPT response time is 139.07 minutes (95% CI: 129.83, 154.84) in the *Low Cost 1* scenario, and only 94.00 minutes (95% CI: 88.57, 102.48) in the *Low Cost 2* scenario. See Table B.4.

This is likely because the probabilities of requiring IPs or AMBs to respond to PPT are both decreased to 0 in the *Low Cost 2*, *Medium Cost 2* and *High Cost 2* sub-scenarios. These changes make the resources available to respond to PPT calls slightly less constrained, since the model never requires PPT calls to be responded to by IPs or AMBs in these scenarios. The resulting effect is a reduction in PPT response times.

Some of the above scenarios aim to improve response times only, while the *Low Cost 2*, *Medium Cost 2* and *High Cost 2* scenarios also aim to improve staff and vehicle allocations. Figure 5.19 shows the distributions of vehicle types for each scenario, disaggregated by call types:

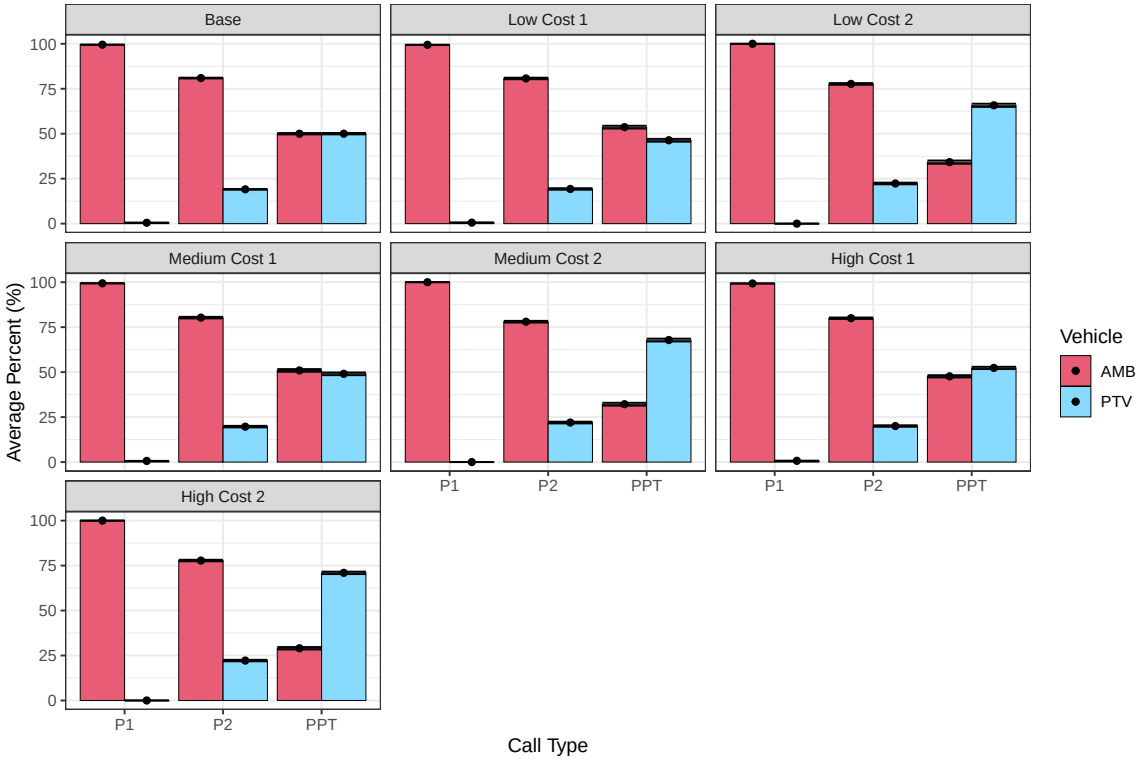


Figure 5.19: Joint distributions of vehicle types and call types, by scenario. Error bars show 95% bootstrapped BCa confidence intervals.

As in the *Improving vehicle allocation* scenario, the percentage of PPT calls responded to by AMBs declines in all *Low Cost 2*, *Medium Cost 2* and *High Cost 2* relative to the base scenario. In the base scenario 50.0% (95% CI: 49.6%, 50.4%) of PPT calls are responded to by ambulances. In all of the above-mentioned scenarios this figure declines significantly, to 34.2% (95% CI: 33.3%, 35.1%) for *Low Cost 2*, 32.2% (95% CI: 31.4%, 33.0%) for *Medium Cost 2*, and 29.0% (95% CI: 28.4%, 29.7%) for *High Cost 2*. See Table B.7. Additionally, the percentage of P1 calls responded to by PTVs in all the above-mentioned scenarios is 0%, while this figure is non-zero in the base scenario.

Figure 5.20 shows the effects on the staff mixture for each call type, in each scenario:

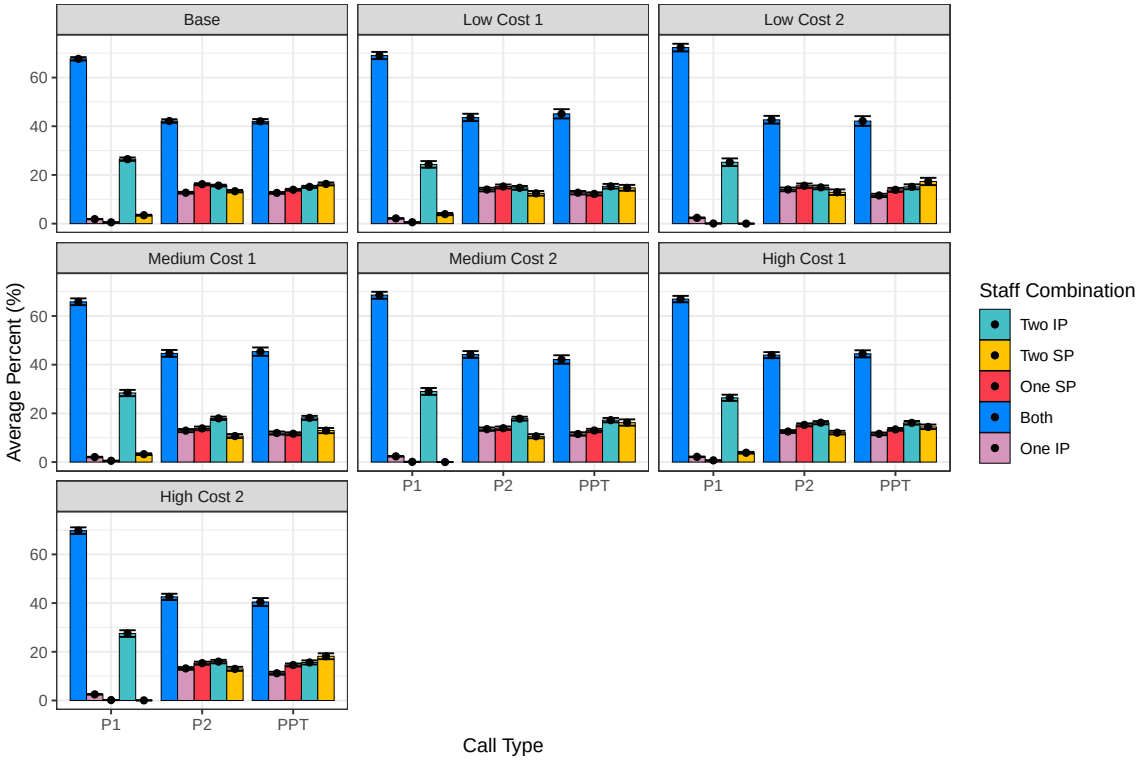


Figure 5.20: Joint distributions of staff mixture and call types, by scenario. Error bars show 95% bootstrapped BCa confidence intervals.

Most notably, the percentage of P1 calls responded to by One SP is 0% in the *Low Cost 2*, *Medium Cost 2* and *High Cost 2* scenarios. The percentage of P1 calls responded to by Two SPs is also 0% in all of these scenarios, implying that P1 calls are always responded to by at least one IP.

5.5 Sensitivity Analysis

In this section, the sensitivity analysis primarily explores the nature of the relationships between key input parameters and the Priority 1 response times. It is organised as follows:

First, the specifications of the Latin Hypercube Sampling, which was used to derive the parameter sets for the sensitivity analysis are outlined in Subsection 5.5.1. Then, the specifications of the Random Forest, which was used as the metamodel for the sensitivity analysis, are outlined in Subsection 5.5.2. Partial dependence plots are then used to show the modelled relationship between a subset of the model parameters and the response, holding the other parameters at their average values (B. M. Greenwell 2017). These plots are shown in Subsections 5.5.3 and 5.5.4, for univariate and bivariate cases respectively. Finally, the variable importance were extracted from the Random Forest, and these results are shown in Subsection 5.5.5.

5.5.1 Latin Hypercube Sampling

An R Shiny app, developed by Norman (2022) from MASHA, was used to generate the 3390 LHS parameter sets. Triangular distributions were used for all parameters. For discrete parameters, such as the number of IPs assigned to a particular EMS base, the continuous values produced by LHS were rounded to the nearest integer.

Parameter distributions were chosen in order to cover the range of reasonable values for each parameter. For example, with the probability parameters, this meant either constraining the distributions to lie within the range [0, 0.5], [0, 0.6], [0.5, 1] or allowing them to take on the full range [0, 1]. In the case of the Require AMB Probability, the range was set to [0.5, 1] for both P1 and P2 calls, and [0, 0.5] for PPT calls. This is because there are no reasonable scenarios in which the vehicle dispatching behaviour is such that the require AMB probability for PPT calls is greater than 0.5. For the discrete parameters, such as the number of AMBs assigned to West End, the values in the baseline scenario were used as the distribution's mean (μ), with the range set to $[\max(0, \mu - 1), \mu + 1]$. The ranges of the Reroute, Prioritisation and Priority Correction Probability parameters were set to [0, 1], to explore the relationships between these parameters and the outcome over their whole domains.

The distributions used for each parameter are specified in Table B.1, and the histograms of values obtained from LHS are displayed in Figure A.10, both in the Appendix.

5.5.2 Random Forest

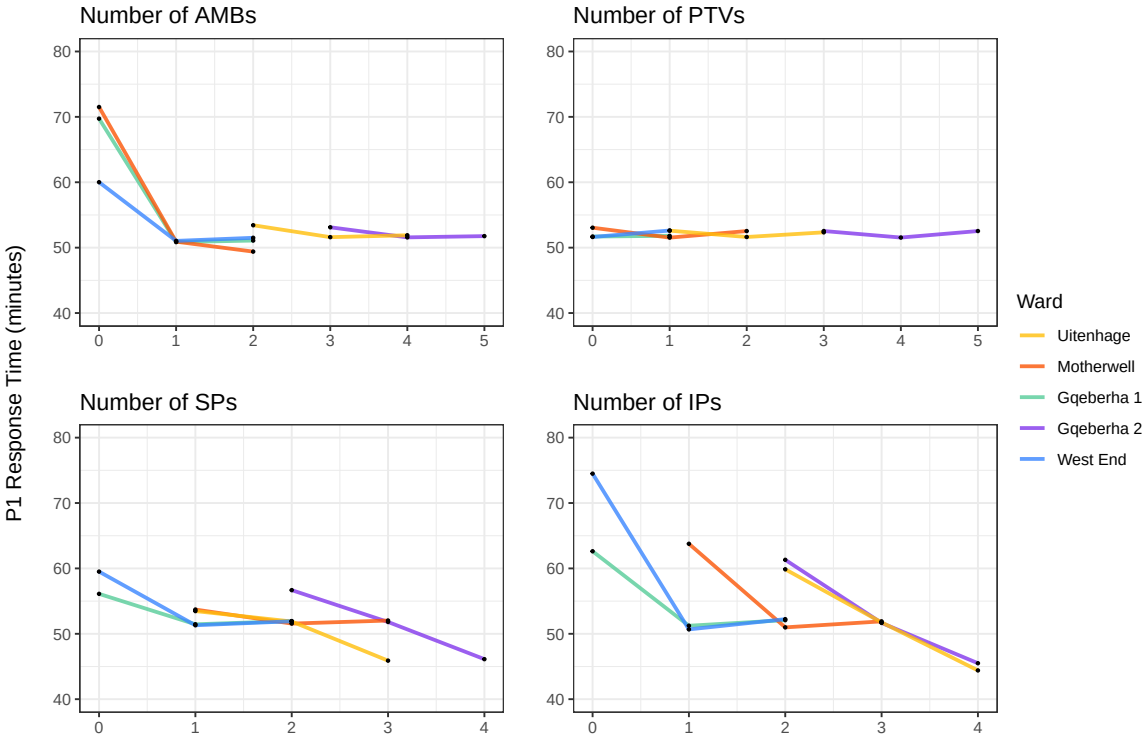
To fit the random forest, the response variable was median P1 response time, with values obtained from running the model using each parameter set generated with LHS. These values, and descriptions of what each hyperparameter means, can be found in Table 5.4. The explanatory variables were all the parameters in this set. The Mtry and Splitrule hyperparameters were tuned using a grid search (results of this grid search can be found in Figure B.3 in the Appendix). The values of other hyperparameters were set to the default values used by Ranger for regression problems. This approach, where only the Mtry and Splitrule are tuned is the default fitting procedure used by the Caret library in R, when fitting a random forest using the Ranger library. The decision to use the default fitting procedure was made because the random forest was only used for conducting the sensitivity analysis. The objective was to gain a better understanding of the relative importance of model parameters, and the nature of the relationship between the parameters and the outcome of interest, not to optimise the model's predictive performance. Therefore, extensive and computationally expensive tuning of additional hyperparameters was not considered necessary.

Hyperparameter	Value	Description
Mtry	**42**	Number of variables to consider for splitting at each node.
Number of trees	**500**	Total number of trees used in the model.
Target node size	**5**	The stopping criterion: the smallest number of observations remaining after partitioning the dataset with decision trees.
Importance mode	**impurity**	The "total decrease in node impurities from splitting on the variable, averaged over all trees." (Liaw & Wiener 2002) In this case, impurity is the residual variance.
Splitrule	**extratrees**	The method used to calculate the splits at each decision tree (Boehmke & B. Greenwell 2019). Extratrees is an implementation of Geurts *et al.* (2006)
N. random splits	**1**	Number of splitting values to consider for each variable considered.
R squared (OOB)	**0.25**	Out-of-bag R-squared.

Table 5.4: Random Forest hyperparameter values, and descriptions of each hyperparameter. Unless otherwise specified, the source for each description is the Ranger package documentation (Wright & Ziegler 2017).

5.5.3 One-way Partial Dependence Plots

First, partial univariate dependence plots are considered to examine the effect of changing each parameter while holding all others at their average values. Figure 5.21 shows the predicted median response time for P1 calls as the number of AMBs, PTVs, SPs and IPs are varied:



ed, and the

Figure 5.21: Univariate partial dependence plots, for number of AMBs, number of PTVs, number of SPs and number of IPs. X axes display the values of the varied parameters, and Y axes display the median P1 response time.

Holding other parameters constant, P1 RT declines as as the number of ambulances increases from 0 to 1 in Motherwell, West End and Gqeberha 1. Beyond this point, there are no further improvements in response times, where the bottleneck is staff rather than vehicles. Increases in the number of PTVs do not decrease P1 RT, holding other parameters constant.

As the number of SPs and IPs increase, there is a general trend towards lower P1 RT. As observed with increasing vehicles, there are points beyond which additional staff members do not decrease response times, where the bottleneck is vehicles rather than staff.

Figure 5.22 shows how the four wait time parameters affect P1 RT when varied individually:

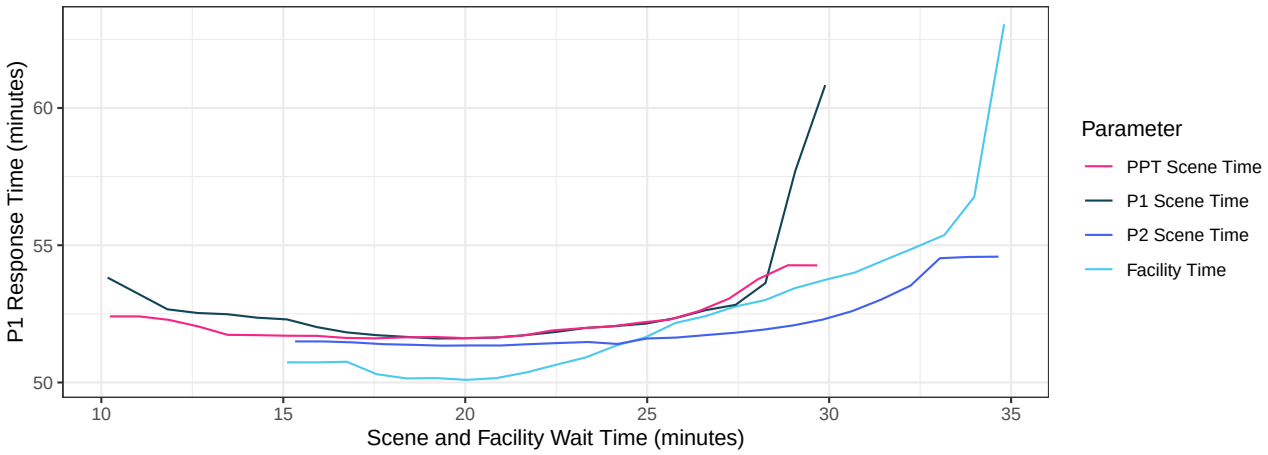


Figure 5.22: Univariate partial dependence plot for wait time parameters.

As wait times reach a threshold of around 2.5 ticks (corresponding to 25 minutes), P1 RT increases substantially.

Figure shows how the three prioritisation parameters affect P1 RT when varied individually:

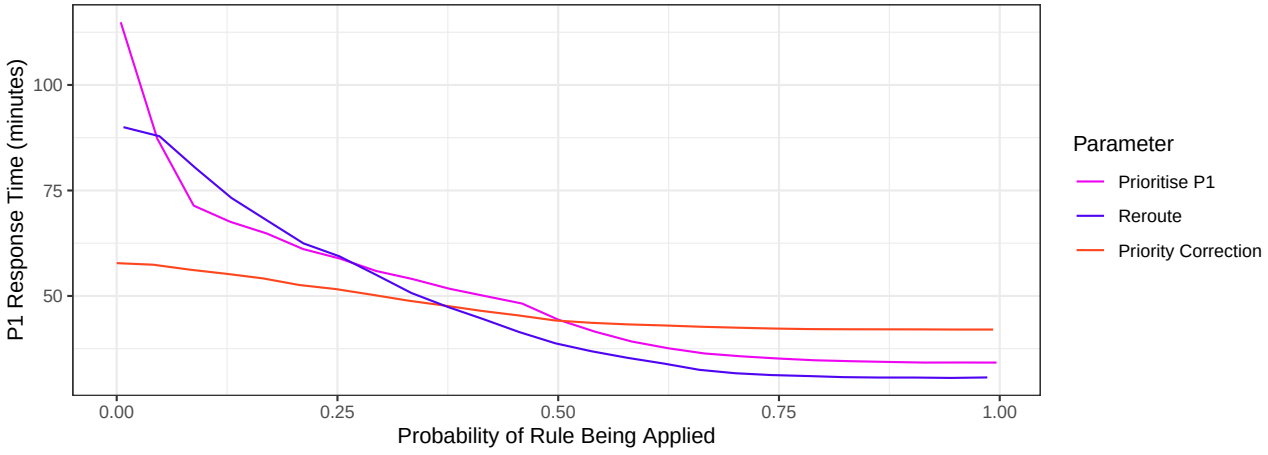


Figure 5.23: Univariate partial dependence plot for prioritisation parameters.

All three prioritisation parameters are inversely related to P1 RT, implying they are effective in lowering it. The priority correction parameter does not appear to affect P1 RT as strongly as the other two parameters.

5.5.4 Two-way Partial Dependence Plots

To investigate any interaction effects, bivariate partial dependence plots are used. Figure 5.24 shows two-way partial dependence plots for three sets of likely interactions: (1) Priority Correction Probability and P1 Prioritisation Probability, (2) Reroute Probability and P1 Prioritisation Probability and (3) Avg. Facility Time and Avg. Scene Time.

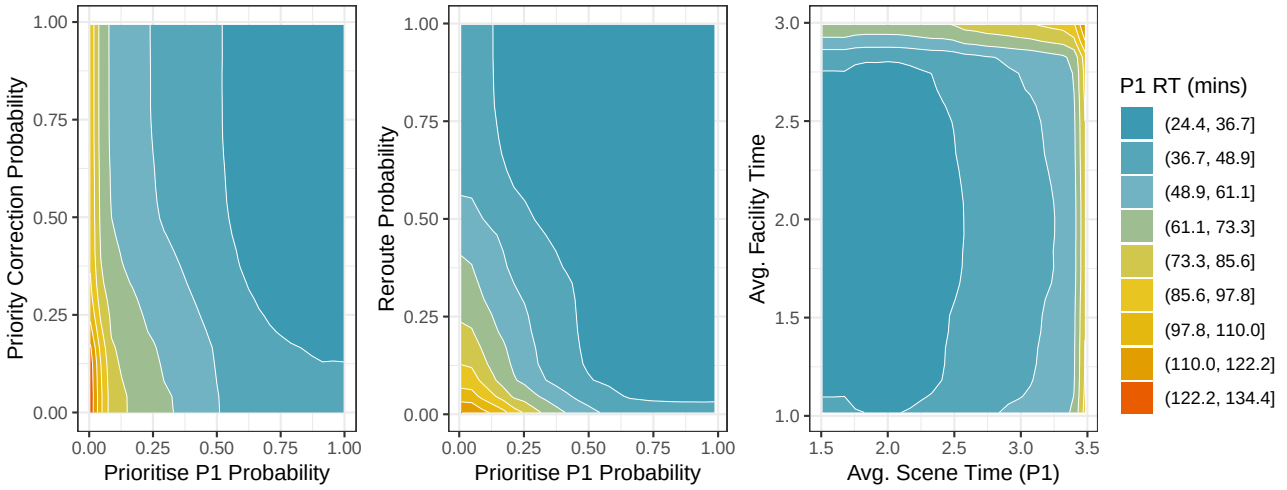


Figure 5.24: Bivariate partial dependence plots for number of IPs and AMBs assigned to each ward.

There are interactions present for the all three sets of parameters:

- Median P1 response time is highest when Prioritise P1 Probability is lowest and Priority Correction Probability is lowest. When Prioritise P1 Probability is low, median P1 response time remains high regardless of the value of Prioritise P1 Probability.
- Similarly, median P1 response time is highest when Prioritise P1 Probability is lowest and Reroute Probability is lowest. However, when either Reroute Probability or Prioritise P1 probability is high, median P1 response time is relatively low, regardless of the value of the other parameter.
- Finally, Avg. Scene Time and Avg. Facility Time appear to have have a weak interaction. Median P1 response time appears to be correlated with both time parameters, and is highest when both parameters are highest. When Avg. Facility Time is lowest, median P1 response time remains low until Avg. Scene Time approaches 35 minutes.

Figure 5.25 investigates any interactions between number of IPs and number of AMBs assigned to each ward, with bivariate partial dependence plots.

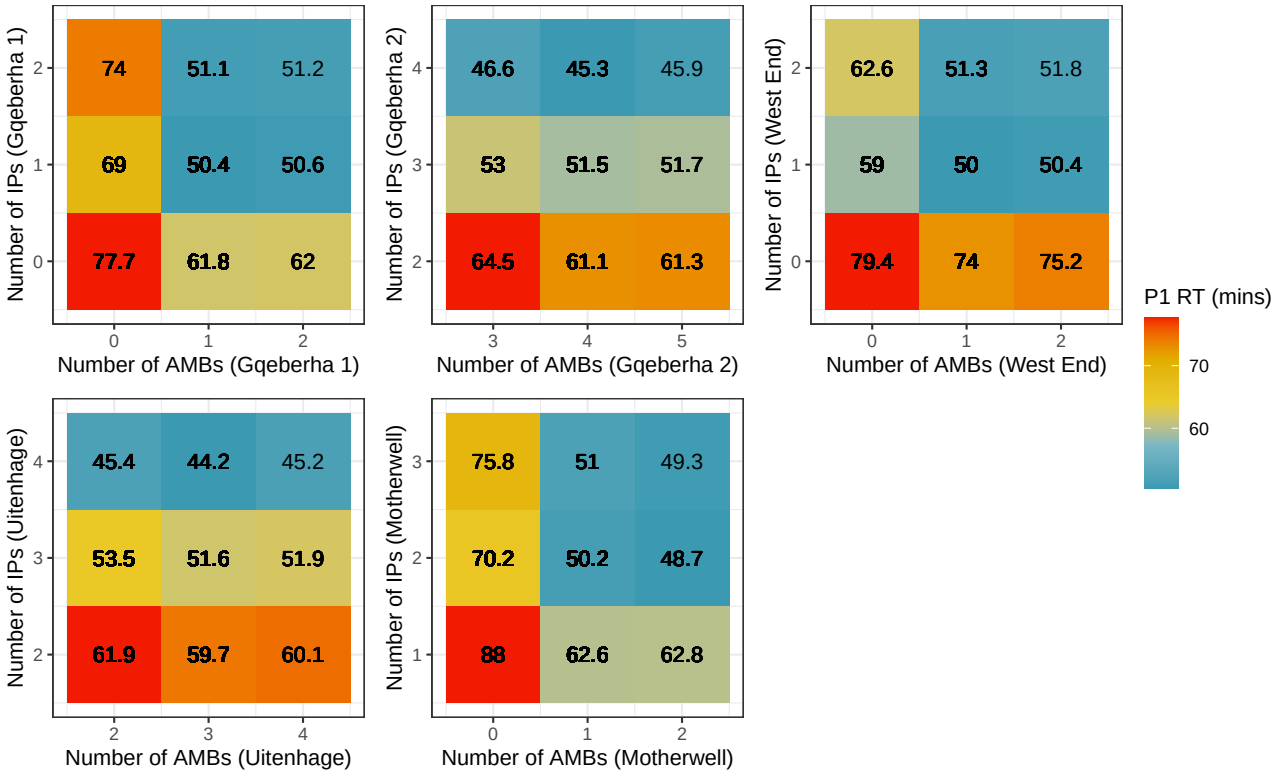


Figure 5.25: Bivariate partial dependence plots for number of IPs and AMBs assigned to each ward.

There appear to be interactions in Motherwell, West End and Gqeberha 1, where the reduction in median P1 response time associated with additional IPs depends on the number of AMBs available: when more AMBs are available, additional IPs have a greater reducing effect on response time. A similar, but less dramatic, effect is observed for Uitenhage. There does not appear to be any interaction in Gqeberha 2.

5.5.5 Variable Importance

Random Forests allow for the generation of variable importance plots, which measure the total reduction in residual variance obtained from splitting on each given variable, averaged over all decision trees (Liaw & Wiener 2002). Figure 5.26 is used to determine which parameters have the greatest effect on median P1 response time.

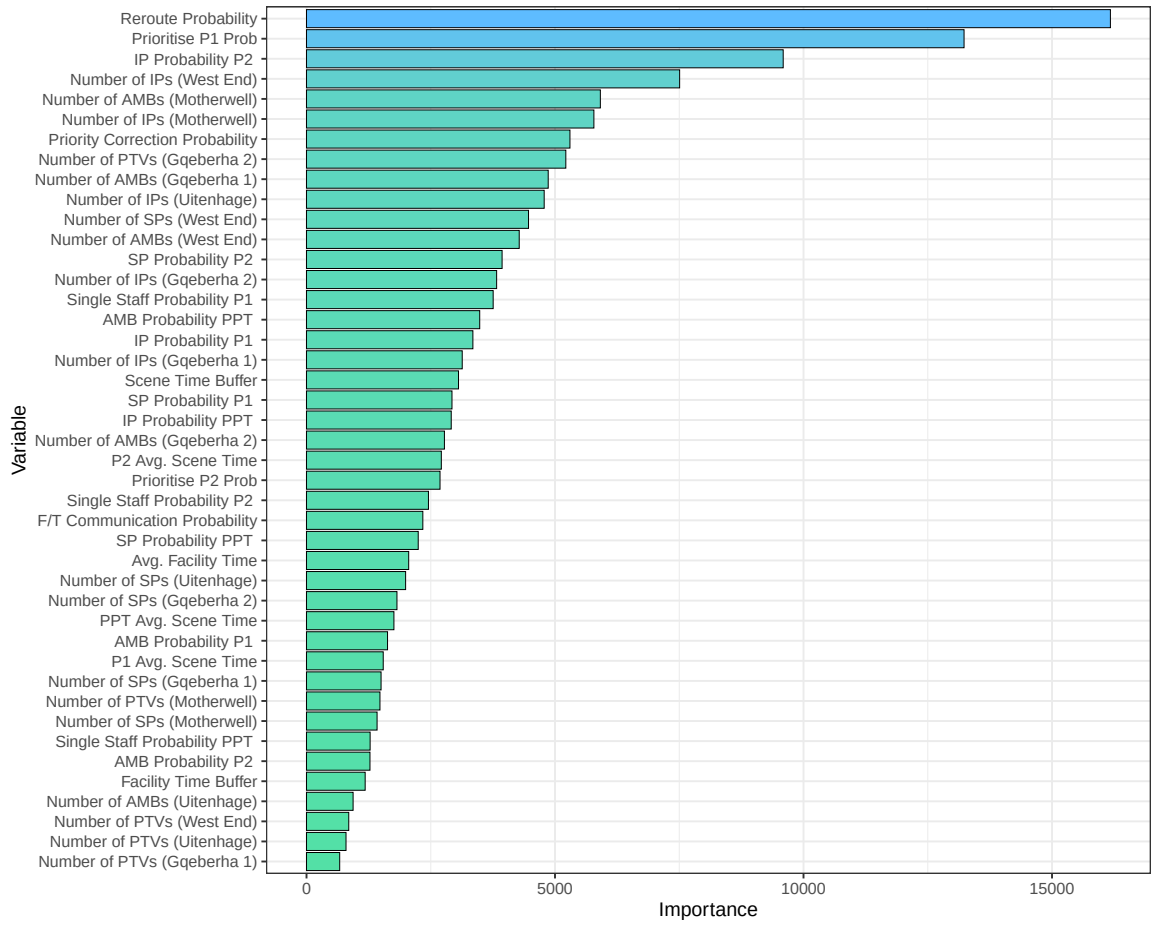


Figure 5.26: Random Forest variable importance. The total reduction in residual variance obtained from splitting on each given variable, averaged over all decision trees (Liaw & Wiener 2002).

The two most useful parameters for predicting P1 response times are Reroute Probability and Prioritise P1 Probability, both of which control the ways in which calls are prioritised in the model. While these parameters reduce P1 response time, the reductions come from focusing resources more on P1 calls, at the expense of potentially neglecting other call types. With high Reroute and Prioritise P1 probabilities, P2 and PPT calls experience increased response times, and these increases are at times unacceptably large, as explored in section 5.4.1.

The effects of the Reroute and Prioritise P1 probability parameters were explored in a number of scenarios. Reroute probability was changed in the *Improving prioritisation & classification* scenario as well as all *Low Cost 1* to *High Cost 2* scenarios. Prioritise P1 probability was also changed in the above scenarios, and also in the *Improving prioritisation* scenario.

It was previously found that increasing the number of IPs assigned to Uitenhage resulted in the largest reduction in P1 response time, compared with increasing the number of IPs assigned to other locations, albeit by a small margin. However, Figure 5.26 shows that the number of IPs assigned to West End is a stronger predictor of P1 response time than the number of IPs assigned to Uitenhage. This is likely because decreasing the number of IPs assigned to West End from its current value of 1 to 0 results in particularly large increases in response times (see Figure 5.25 and Figure 5.21). This makes the parameter particularly useful for predicting P1 response time, which explains its high variable importance score.

Chapter 6

Discussion

This chapter contains five sections. It begins with section 6.1, which summarises the main results. Section 6.2 then situates this research within the broader context of the EMS modelling literature, and discusses its novel contributions. Section 6.3 then uses these results to outline a set of recommendations to policy-makers. Finally, section 6.4 discusses the limitations of this research, and the chapter concludes with section 6.5 which reflects on these limitations to arrive at a set of ways in which this work could potentially be extended or improved in the future.

6.1 Results

The main objective of this research has been to identify ways to improve median response times for the most urgent calls in the Nelson Mandela Bay region. The most urgent calls are those that are classified as code Red or Blue by EMS personnel at the scene, and a good but imperfect proxy of call urgency is the priority assigned to each call by dispatchers. A second objective is to identify ways of increasing the probability that appropriate staff mixtures and vehicle types respond to calls. This is because it is not only essential that urgent calls are responded to in a timely manner, but also that the staff and vehicles that arrive are capable of providing a high quality of care.

The main results of this research are discussed in terms of these objectives:

1. Reducing median Priority 1 (P1) response times to below 30 minutes can be achieved through implementing changes to dispatching, rerouting and prioritisation behaviour alone, without increasing the number of staff or vehicles. However, these improvements come at the expense of increased median response times of Planned Patient Transport (PPT) calls, which in some cases almost double.

2. Reducing median P1 response time to below 30 minutes, while keeping median response times for P2 and PPT calls below their current values, can be achieved by adding one Independent Practitioner (IP) to the Uitenhage base, increasing prioritisation of P1 and P2 calls, increasing re-routing, and improving dispatcher call classifications.

3. The effects of increasing the number of staff or vehicles assigned to a particular base depend on the base in question. In cases where the number of staff is the bottleneck, and there is a surplus of vehicles, increasing the number of vehicles does not improve response times. And when vehicles are the bottleneck, increasing the number of staff assigned to a base is similarly ineffectual. In addition, the current allocation of vehicles as well as the spatial distribution of demand both affect how and whether response times decrease when the number of staff or vehicles assigned to a particular base are increased.

4. Increasing the number of IPs available to the Uitenhage or Gqeberha 2 (also referred to as Central) bases by one staff member results in the largest reductions in P1 response times, compared with increasing the number of IPs assigned to other bases. Increasing the number of SPs assigned to these bases results in a slightly smaller reduction in P1 response times.

5. Increasing the number of ambulances assigned to the Motherwell base results in only slight reductions in P1 response times. Increasing the number of ambulances or patient transport vehicles assigned to other bases does not significantly affect response times.

6. Improvements to staff and vehicle allocations can be achieved by ensuring that 1) P1 calls are always responded to by an ambulance and not fewer than two EMS personnel, at least one of whom is an Independent Practitioner, 2) single staff members are discouraged from responding to P2 calls, and 3) PPT calls are always responded to by patient transport vehicles. However, these improvements in staff and vehicle allocations come at the expense of increases in P1, P2 and PPT response times.

7. The effect described in (6) can be counteracted by increasing the number of staff employed.

8. Re-routing and prioritisation of P1 calls both reduce P1 response times, but these reductions come at the expense of increased response times for P2 and PPT calls, and requires ongoing training of staff.

9. If dispatcher call classifications are improved to better match triage classifications, the most urgent calls are more likely to be classified as P1, and responded to sooner. As a result, improvements to dispatcher classifications lead to reductions in response times for the most urgent code Red or Blue calls. Also, higher dispatcher classification accuracy results in fewer calls being classified as P1[1]. This means that when dispatcher classifications are improved, response times for the most urgent calls can be reduced without considerable increases in P2 and PPT response times.

6.2 Contributions to the Literature

Since the only other ground EMS simulation modelling study conducted in South Africa to date is the work of Stein *et al.* (2015), this research is the first such model of the Eastern Cape's EMS system, and the second of any ground EMS system in South Africa.

In addition, since there have been very few applications of simulation models to EMS systems in LMICs, this research has encountered methodological challenges that are not commonly seen in the literature. Addressing these challenges has included using fuzzy matching to standardise manually-entered locations, and using Google Maps to convert these text locations into approximate coordinates, rather than relying on precise location data collected by Computer Aided Dispatch (CAD) systems which are commonly used in higher income countries, as in the work of Yang *et al.* (2019). Also, because road quality and its impact on travel times was highlighted as an important consideration early on, Google Maps was used to estimate the travel time matrix, which includes road quality in its estimates (Lau 2020).

[1]This is because approximately one third of calls are classified as P1 by dispatchers, but only 8.9% of calls are classified as code Red or Blue by EMS personnel on the scene. As the dispatcher classifications converge to the triage classifications, the percentage of P1 calls declines.

6.3 Recommendations

At the time of writing (October 2023), a new Computer Aided Dispatch (CAD) system will soon be rolled out in the Nelson Mandela Bay region. This will likely result in improved data quality and system efficiency. The transition to this new system also presents an opportunity to re-think how the system operates, and address any inefficiencies.

These findings of this research are used to arrive at three sets of recommendations. These relate to changing how dispatchers operate, adding staff, and adding vehicles.

1. **Dispatching:**
   A number of improvements can be made to how calls are classified, and how vehicles and staff are dispatched. These changes would likely result in improved P1 response times, and more appropriate allocations of staff and vehicles, without the need for additional staff or vehicles. In this regard, training programmes targeting at the following areas could be carried out:

   (a) Since dispatchers frequently assign higher priorities to calls than do the EMS personnel on the scene, a training programme aimed at improving dispatcher classifications is recommended. If such a programme were effective, the number of calls misclassified as P1 would be reduced, leaving a smaller pool of P1 calls which would be easier to prioritise. Additionally, Alshehri *et al.* (2020) highlight the importance of dispatchers being medically trained, and able to ask relevant questions of patients, in order to assign the correct priorities to calls.

   (b) More frequently re-routing vehicles to P1 calls improves P1 response times. Similarly, prioritisation of calls by queueing non-P1 calls until all P1 calls have been responded to is also effective at reducing P1 response times. Therefore, if dispatchers more frequently re-route vehicles and queue non-P1 calls until all P1 calls have been responded to, P1 response times will likely reduce. These are both changes to dispatching behaviour that should be emphasised in the coming transition to a Computer Aided Dispatch (CAD) system. As it is essential to know the current locations of vehicles in order to re-route effectively, this will certainly be made easier by the introduction of a CAD system that provides real-time information about the current locations of vehicles to dispatchers.

   (c) It is recommended that the staff and vehicle rules outlined in section 6.1 be followed more closely, which can possibly be achieved through additional dispatcher training. However, because these rules result in increased response times, they are only recommended in combination with simultaneously increasing the number of EMS personnel in the system.

2. **Adding staff:**
   Because staff are a bottleneck in the Uitenhage and Gqeberha 2 bases, and increasing the numbers of IPs or SPs in these bases is particularly beneficial, it is recommended that at least one staff member is added to the Uitenhage or Gqeberha 2 bases at all times. For the largest decreases in response times, increasing the number of IPs is recommended. However, increasing the number of SPs also results in considerable reductions in response times for all call types. Recall that increasing the number of staff members assigned to a particular base in the simulation model by one means that one additional staff member is available at all times. In practice, having one additional staff member available at all times means employing four to five additional staff members per month, given 40 hours worked per week.

3. **Adding vehicles:**
   Because vehicles are not the bottleneck at any bases in the system, and neither additions to ambulances or patient transport vehicles resulted in noteworthy reductions in response times, the addition of vehicles is not recommended at this stage.

4. **Capturing of patient outcomes:**
   Finally, it is recommended that patient outcomes are linked to the CAD system's database, so that future researchers can more rigorously assess the relationship between the performance of the EMS system and the well-being of the patients it serves.

6.4 Limitations

The limitations of this research are either the result of data quality or simplifying modelling assumptions. In terms of data quality, because the main datasets were manually entered, it is likely that there are some capturing errors present in the data used to build and validate the model. The Computer Aided Dispatch system, due to automated data collection, should eliminate this issue in the future. Limitations introduced from modelling assumptions are listed below:

1. Travel times are assumed to be constant, regardless of the time of day. While traffic does typically yield to EMS vehicles, there is still likely some variation in travel times depending on the time of day and day of week.

2. Vehicles are assumed to be capable of holding a maximum of one patient. In reality, Patient Transport Vehicles do at times transport multiple patients.

3. The number of staff members assigned to each facility is assumed to be time-invariant.

6.5 Future Research

The model developed in this research can be applied to other urban EMS systems with relative ease, given the required input data. However, some extensions and improvements can be made to the model, and this section includes suggestions for future research.

I modelled the urban area of Nelson Mandela Bay due to the availability of electronically-captured data. However, rural districts in the Eastern Cape suffer from particularly long response times, and could benefit from evidence-based recommendations aimed at improving efficiency. Meents & Boyles (2010) estimated average response times for cases arriving at the rural Ngcwanguba Health Care Centre at 'almost 4 hours.' The high response times in the rural Eastern Cape are partly due to sparse population coverage and exceptionally poor road

quality (Meents & Boyles 2010). As road quality at times necessitates the use of 4×4 vehicles, adaptations to the model used in this research for use in rural settings would need to include 4×4 vehicle types and model road quality explicitly.

Other suggestions for refinements to the model include:

1. Using multiple travel time matrices depending on the time of day and day of week. This can be achieved using Google Maps.
2. Using an automatically-generated input dataset from the new Computer Aided Dispatch (CAD) system, in order to improve the robustness of the model results.
3. Extending the Patient Transport Vehicle aspect of the model to allow multiple patients to be transported in the same vehicle.
4. Extending the evaluations of each scenario to measure survival probability, which would enable estimates of lives saved.
5. Adding a cost model to allow for comparisons between scenarios with respect to more exact estimates of financial implementation costs.

Chapter 7

Conclusion

I began this research with the intention of identifying a number of evidence-based recommendations for improving the operations of the EMS system in Nelson Mandela Bay.

A set of potential interventions was formulated, in partnership with CHAI and the ECDoH. For estimating the real-world effects of each intervention, an agent-based simulation model of the system was developed. This model was built and validated using multiple sources of data, including a large historical dataset of EMS call metadata from the ECDoH, and a dataset of precise response times that I collected.

I found that the median response time of Priority 1 (P1) calls can be reduced to below the 30 minute target by improving dispatching, rerouting and prioritisation behaviour alone, without scaling up resources. Similarly, improvements to staff and vehicle allocations can also be achieved without increasing resources. However, the changes required to make these improvements rely to a large extent on increased prioritisation of P1 calls, which draws resources away from lower priority calls. This has the undesirable effect of considerably increasing response times for Priority 2 (P2) and Planned Patient Transport (PPT) calls. I also found that moderately increasing the number of staff employed while simultaneously improving dispatching, rerouting and prioritisation behaviour, reduces median P1 response time to below the 30 minute target, while keeping median P2 and PPT response times below their current values.

These results were used to derive a set of proposed interventions for policy-makers in the ECDoH. If implemented, these recommendations are likely to increase the operational efficiency and capacity of Nelson Mandela Bay's EMS system, decrease response times for the most urgent calls, improve allocations of staff and vehicles, and ultimately contribute to saving lives.

Bibliography

1. Aboueljinane, L., Sahin, E. & Jemai, Z. A review on simulation models applied to emergency medical service operations. *Computers &amp Industrial Engineering* **66,** 734–750. https://doi.org/10.1016/j.cie.2013.09.017 (Dec. 2013).
2. Alshehri, M., Pigoga, J. & Wallis, L. A mixed methods investigation of emergency communications centre triage in the Government Emergency Medical Services System, Cape Town, South Africa. *African Journal of Emergency Medicine* **10,** S12–S17. https://doi.org/10.1016/j.afjem.2020.02.004 (2020).
3. Aringhieri, R. *Ambulance location through optimization and simulation: the case of Milano urban area* Nov. 2007.
4. Batta, R., Dolan, J. M. & Krishnamurthy, N. N. The Maximal Expected Covering Location Problem: Revisited. *Transportation Science* **23,** 277–287. ISSN: 00411655, 15265447. http://www.jstor.org/stable/25768396 (2023) (1989).
5. Berg, P. L. V. D. & Essen, J. T. V. Comparison of static ambulance location models. *International Journal of Logistics Systems and Management* **32,** 292. https://doi.org/10.1504/ijlsm.2019.098321 (2019).
6. Berlin, G. N. & Liebman, J. C. Mathematical analysis of emergency ambulance location. *Socio-Economic Planning Sciences* **8,** 323–328. https://doi.org/10.1016/0038-0121(74)90036-6 (Dec. 1974).
7. Boehmke, B. & Greenwell, B. *Hands-On Machine Learning with R* https://doi.org/10.1201/9780367816377 (Chapman and Hall/CRC, Nov. 2019).
8. Borgonovo, E., Pangallo, M., Rivkin, J., Rizzo, L. & Siggelkow, N. Sensitivity analysis of agent-based models: a new protocol. *Computational and Mathematical Organization Theory* **28,** 52–94. https://doi.org/10.1007/s10588-021-09358-5 (Jan. 2022).
9. Boutilier, J. J. & Chan, T. C. Y. Ambulance Emergency Response Optimization in Developing Countries. *Operations Research* **68,** 1315–1334. eprint: https://doi.org/10.1287/opre.2019.1969. https://doi.org/10.1287/opre.2019.1969 (2020).
10. Brailsford, S. C., Sykes, J. & Harper, P. R. *Incorporating human behavior in healthcare simulation models* in *Proceedings of the 2006 Winter Simulation Conference* (2006), 466–472.
11. Cheema, B. & Twomey, M. *The South African Triage Scale* https://emssa.org.za/wp-content/uploads/2011/04/SATS-Manual-A5-LR-spreads.pdf (Western Cape Department of Health, 2012).
12. Chen, S. H. *et al.* Application of Machine Learning Techniques to an Agent-Based Model of Pantoea. *Frontiers in Microbiology* **12.** https://doi.org/10.3389/fmicb.2021.726409 (Sept. 2021).
13. Church, R. & ReVelle, C. The maximal covering location problem. *Papers of the Regional Science Association* **32,** 101–118. https://doi.org/10.1007/bf01942293 (Dec. 1974).
14. Collins, A. J., Seiler, M. J., Gangel, M. & Croll, M. Applying Latin hypercube sampling to agent-based models. *International Journal of Housing Markets and Analysis* **6** (eds Yates, D. S. R. & Elaine) 422–437. https://doi.org/10.1108/ijhma-jul-2012-0027 (Sept. 2013).
15. Daskin, M. S. A Maximum Expected Covering Location Model: Formulation, Properties and Heuristic Solution. *Transportation Science* **17,** 48–70. https://doi.org/10.1287/trsc.17.1.48 (Feb. 1983).

16. Erdogan, G., Erkut, E., Ingolfsson, A. & Laporte, G. Scheduling ambulance crews for maximum coverage. *Journal of the Operational Research Society* **61,** 543–550 (Apr. 2010).
17. Erkut, E., Ingolfsson, A. & Erdoğan, G. Ambulance location for maximum survival. *Naval Research Logistics (NRL)* **55,** 42–58. https://doi.org/10.1002/nav.20267 (Dec. 2007).
18. Fujiwara, O., Makjamroen, T. & Gupta, K. K. Ambulance deployment analysis: A case study of Bangkok. *European Journal of Operational Research* **31,** 9–18 (1987).
19. Gendreau, M., Laporte, G. & Semet, F. Solving an ambulance location model by tabu search. *Location Science* **5,** 75–88. https://doi.org/10.1016/s0966-8349(97)00015-6 (Aug. 1997).
20. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Machine Learning* **63,** 3–42. https://doi.org/10.1007/s10994-006-6226-1 (Mar. 2006).
21. Granato, B. & Li-Jessen, N. *Sensitivity Analysis for Dimensionality Reduction in Agent-Based Modeling* in (June 2020).
22. Greenwell, B. M. pdp: An R Package for Constructing Partial Dependence Plots. *The R Journal* **9,** 421–436. https://doi.org/10.32614/RJ-2017-016 (2017).
23. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585,** 357–362. https://doi.org/10.1038/s41586-020-2649-2 (Sept. 2020).
24. Henderson, S. G. & Mason, A. J. Ambulance service planning: simulation and data visualisation. *Operations research and health care: a handbook of methods and applications,* 77–102 (2004).
25. Hogan, K. & Revelle, C. Concepts and Applications of Backup Coverage. *Management Science* **32,** 1434–1444. ISSN: 00251909, 15265501. http://www.jstor.org/stable/2631502 (2023) (1986).
26. Horvat, A. M. *et al.* Binary Programming Model for Rostering Ambulance Crew-Relevance for the Management and Business. *Mathematics* **9,** 64. https://doi.org/10.3390/math9010064 (Dec. 2020).
27. HSPCA. *Revised Board Ruling on Supervised Practice* Oct. 2019. https://www.hpcsa.co.za/Content/upload/professional_boards/emb/guidelines/Revised_Board_ruling_on_Supervised_Practice_Oct_2019.pdf.
28. Ingolfsson, A., Erkut, E. & Budge, S. Simulation of single start station for Edmonton EMS. *Journal of the Operational Research Society* **54,** 736–746. https://doi.org/10.1057/palgrave.jors.2601574 (July 2003).
29. Janosikova, L., Jankovic, P., Kvet, P. & Zajacova, F. Coverage versus response time objectives in ambulance location. *International Journal of Health Geographics* **20.** https://doi.org/10.1186/s12942-021-00285-x (July 2021).
30. Kamnqa, S. *R27m needed to get Eastern Cape EMS plan off the ground — spotlightnsp.co.za* [Accessed 29-08-2023]. 2023. https://www.spotlightnsp.co.za/2023/02/14/r27m-needed-to-get-eastern-cape-ems-plan-off-the-ground/#:~:text=Working%20to%20address%20EMS%20challenges&text=Over%20the%20past%20three%20quarters,of%20calls%20met%20these%20targets..
31. Klugl, F. Agent-based simulation engineering (2016).
32. Knight, V., Harper, P. & Smith, L. Ambulance allocation for maximal survival with heterogeneous outcome measures. *Omega* **40,** 918–926. https://doi.org/10.1016/j.omega.2012.02.003 (Dec. 2012).
33. Kuhn & Max. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **28,** 1–26. https://www.jstatsoft.org/index.php/jss/article/view/v028i05 (2008).
34. Larson, R. C. Approximating the Performance of Urban Emergency Service Systems. *Operations Research* **23,** 845–868. https://doi.org/10.1287/opre.23.5.845 (Oct. 1975).

35. Lau, J. *Google Maps 101: How AI helps predict traffic and determine routes* Sept. 2020. https://blog.google/products/maps/google-maps-101-how-ai-helps-predict-traffic-and-determine-routes/.
36. Lee, T., Cho, S.-H., Jang, H. & Turner, J. G. *A simulation-based iterative method for a trauma center &amp#x2014 Air ambulance location problem* in *Proceedings Title: Proceedings of the 2012 Winter Simulation Conference (WSC)* (IEEE, Dec. 2012). https://doi.org/10.1109/wsc.2012.6465042.
37. Li, X., Zhao, Z., Zhu, X. & Wyatt, T. Covering models and optimization techniques for emergency response facility location and planning: a review. *Mathematical Methods of Operations Research* **74,** 281–310. https://doi.org/10.1007/s00186-011-0363-4 (July 2011).
38. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2,** 18–22. https://CRAN.R-project.org/doc/Rnews/ (2002).
39. Lubicz, M. & Mielczarek, B. Simulation modelling of emergency medical services. *European Journal of Operational Research* **29,** 178–185. https://doi.org/10.1016/0377-2217(87)90107-x (May 1987).
40. Majid, M. A., Fakhreldin, M. & Zuhairi, K. Z. in *Lecture Notes in Computer Science* 510–522 (Springer International Publishing, 2016). https://doi.org/10.1007/978-3-319-39510-4_47.
41. Majid, M. A., Aickelin, U. & Siebers, P.-O. Comparing Simulation Output Accuracy of Discrete Event and Agent Based Models: A Quantitative Approach. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.2830304 (2009).
42. McKay, M. D., Beckman, R. J. & Conover, W. J. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics* **21,** 239. https://doi.org/10.2307/1268522 (May 1979).
43. McLay, L. A. A maximum expected covering location model with two types of servers. *IIE Transactions* **41,** 730–741. https://doi.org/10.1080/07408170802702138 (June 2009).
44. MDB. *South African Local Municipal Boundary* 2020. https://dataportal-mdb-sa.opendata.arcgis.com/datasets/27bbdd5b041b4ba6b5707dfed5aa3923/explore.
45. Meents, E. & Boyles, T. Emergency medical services - poor response time in the rural Eastern Cape. *South African Medical Journal* **100,** 790. https://doi.org/10.7196/samj.4374 (Nov. 2010).
46. Metaweb Technologies, Inc. *OpenRefine* version 3.5.2. 2022. https://openrefine.org.
47. Norman, J. 2022. https://github.com/uct-masha/LHSCalibration/.
48. OECD. *History of DAC lists of aid recipient countries - OECD* 2023. https://www.oecd.org/development/financing-sustainable-development/development-finance-standards/historyofdaclistsofaidrecipientcountries.htm.
49. Pandas development team, T. *pandas-dev/pandas: Pandas* version latest. Feb. 2020. https://doi.org/10.5281/zenodo.3509134.
50. Pereira, A. & Broed, R. *Methods for Uncertainty and Sensitivity Analysis : Review and recomendations for implementation in Ecolego* in (2006). https://api.semanticscholar.org/CorpusID:15235914.
51. Potgieter, A. *et al.* Modelling Representative Population Mobility for COVID-19 Spatial Transmission in South Africa. https://doi.org/10.20944/preprints202106.0211.v1 (June 2021).
52. Pullum, L. & Cui, X. Techniques and Issues in Agent-Based Modeling Validation (Jan. 2012).
53. Repede, J. F. & Bernardo, J. J. Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky. *European Journal of Opera-*

tional Research **75,** 567–581. https://doi.org/10.1016/0377-2217(94)90297-6 (June 1994).
54. Ridler, S., Mason, A. J. & Raith, A. A simulation and optimisation package for emergency medical services. *European Journal of Operational Research* **298,** 1101–1113. https://doi.org/10.1016/j.ejor.2021.07.038 (May 2022).
55. Savas, E. S. Simulation and Cost-Effectiveness Analysis of New York's Emergency Ambulance Service. *Management Science* **15,** B–608–B–627. https://doi.org/10.1287/mnsc.15.12.b608 (Aug. 1969).
56. Saydam, C., Repede, J. & Burwell, T. Accurate estimation of expected coverage: A comparative study. *Socio-Economic Planning Sciences* **28,** 113–120. https://doi.org/10.1016/0038-0121(94)90010-8 (Jan. 1994).
57. Al-Shaqsi, S. Models of International Emergency Medical Service (EMS) Systems. *Oman Medical Journal.* https://doi.org/10.5001/omj.2010.92 (Oct. 2010).
58. Silva, P. M. S. & Pinto, L. R. *Emergency medical systems analysis by simulation and optimization* in *Proceedings of the 2010 Winter Simulation Conference* (IEEE, Dec. 2010). https://doi.org/10.1109/wsc.2010.5678938.
59. Stassen, W., Tsegai, A. & Kurland, L. A Retrospective Geospatial Simulation Study of Helicopter Emergency Medical Services' Potential Time Benefit Over Ground Ambulance Transport in Northern South Africa. *Air Medical Journal* **42,** 440–444. ISSN: 1067-991X. http://dx.doi.org/10.1016/j.amj.2023.07.005 (Nov. 2023).
60. Stats SA. *Mid Year Population Estiates* 2022. https://www.statssa.gov.za/?page_id=1854&PPN=P0302&SCH=73305.
61. Stats SA. *Nelson Mandela Bay Stats SA Census* 2011. https://www.statssa.gov.za/?page_id=1021&id=nelson-mandela-bay-municipality.
62. Stein, C., Wallis, L. & Adetunji, O. Meeting national response time targets for priority 1 incidents in an urban emergency medical services system in South Africa: More ambulances won't help. *South African Medical Journal* **105,** 840. https://doi.org/10.7196/samjnew.8087 (Sept. 2015).
63. Thacker, B. *et al.* Concepts of Model Verification and Validation (Oct. 2004).
64. Toregas, C., Swain, R., ReVelle, C. & Bergman, L. The location of emergency service facilities. *Operations research* **19,** 1363–1373 (1971).
65. Toro-Diaz, H., Mayorga, M. E., Chanta, S. & McLay, L. A. Joint location and dispatching decisions for emergency medical services. *Computers & industrial engineering* **64,** 917–928 (2013).
66. Unluyurt, T. & Tuncer, Y. Estimating the performance of emergency medical service location models via discrete event simulation. *Computers &amp Industrial Engineering* **102,** 467–475. https://doi.org/10.1016/j.cie.2016.03.029 (Dec. 2016).
67. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual* ISBN: 1441412697 (CreateSpace, Scotts Valley, CA, 2009).
68. Van Buuren, M., van der Mei, R., Aardal, K. & Post, H. *Evaluating dynamic dispatch strategies for emergency medical services: TIFAR simulation tool* in *Proceedings Title: Proceedings of the 2012 Winter Simulation Conference (WSC)* (IEEE, Dec. 2012). https://doi.org/10.1109/wsc.2012.6465214.
69. Vermuyten, H., Rosa, J. N., Marques, I., Beliën, J. & Barbosa-Póvoa, A. Integrated staff scheduling at a medical emergency service: An optimisation approach. *Expert Systems with Applications* **112,** 62–76. https://doi.org/10.1016/j.eswa.2018.06.017 (Dec. 2018).
70. Vile, J., Gillard, J., Harper, P. & Knight, V. Time-dependent stochastic methods for managing and scheduling Emergency Medical Services. *Operations Research for Health Care* **8,** 42–52. https://doi.org/10.1016/j.orhc.2015.07.002 (Mar. 2016).

71. Wajid, S. & Nezamuddin, N. A robust survival model for emergency medical services in Delhi, India. *Socio-Economic Planning Sciences* **83,** 101342. https://doi.org/10.1016/j.seps.2022.101342 (Oct. 2022).
72. Wears, R. L. & Winton, C. N. *Simulation modeling of prehospital trauma care* in *Proceedings of the 25th conference on Winter simulation* (1993), 1216–1224.
73. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* **77,** 1–17 (2017).
74. Xiang, X., Kennedy, R., Madey, G. & Cabaniss, S. *Verification and validation of agent-based scientific simulation models* in *Agent-directed simulation conference* **47** (2005), 55.
75. Yang, W. *et al.* Simulation modeling and optimization for ambulance allocation considering spatiotemporal stochastic demand. *Journal of Management Science and Engineering* **4,** 252–265. https://doi.org/10.1016/j.jmse.2020.01.004 (Dec. 2019).

Appendix A

Plots

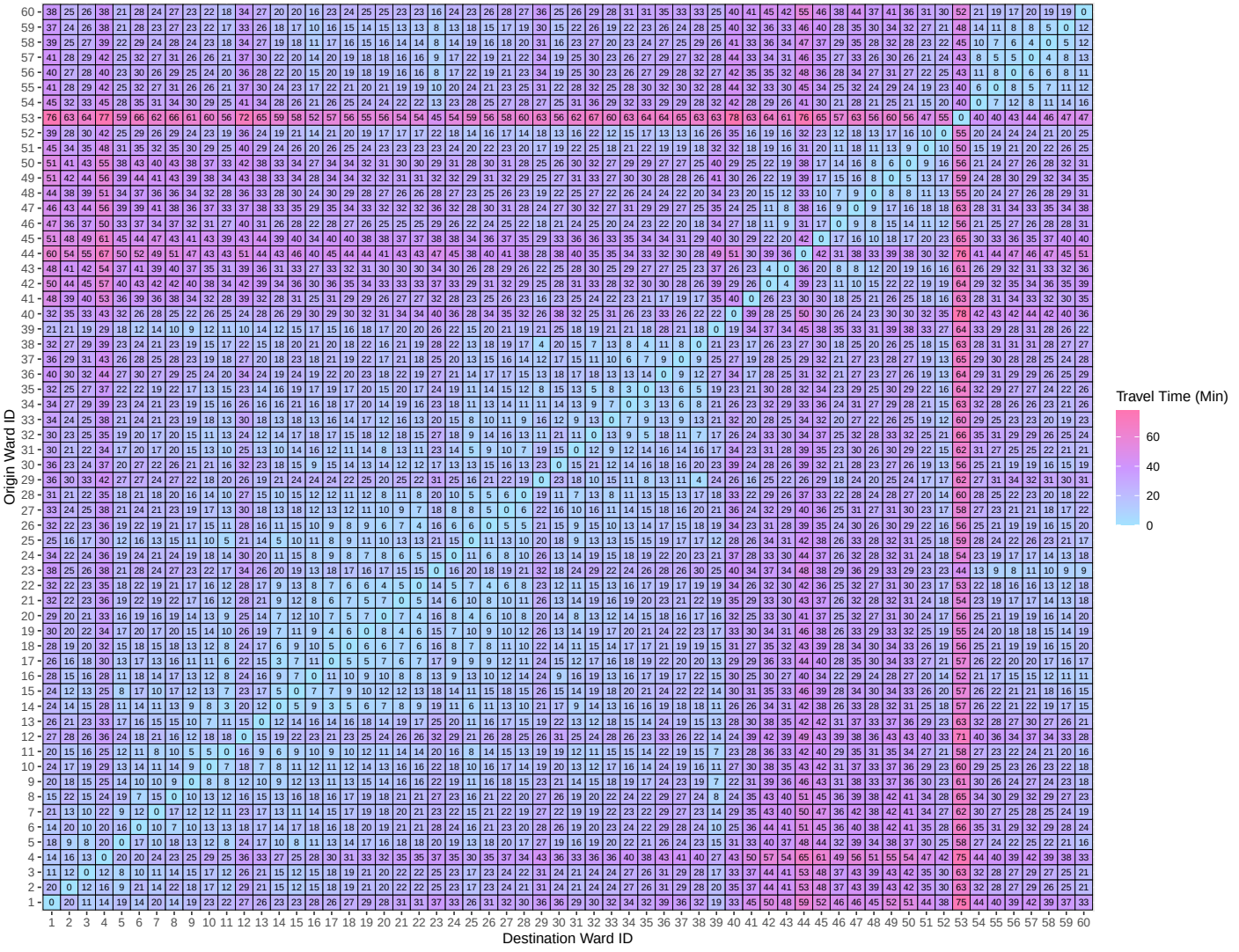


Figure A.1: Inter-ward Travel Time Matrix. Data from Google.

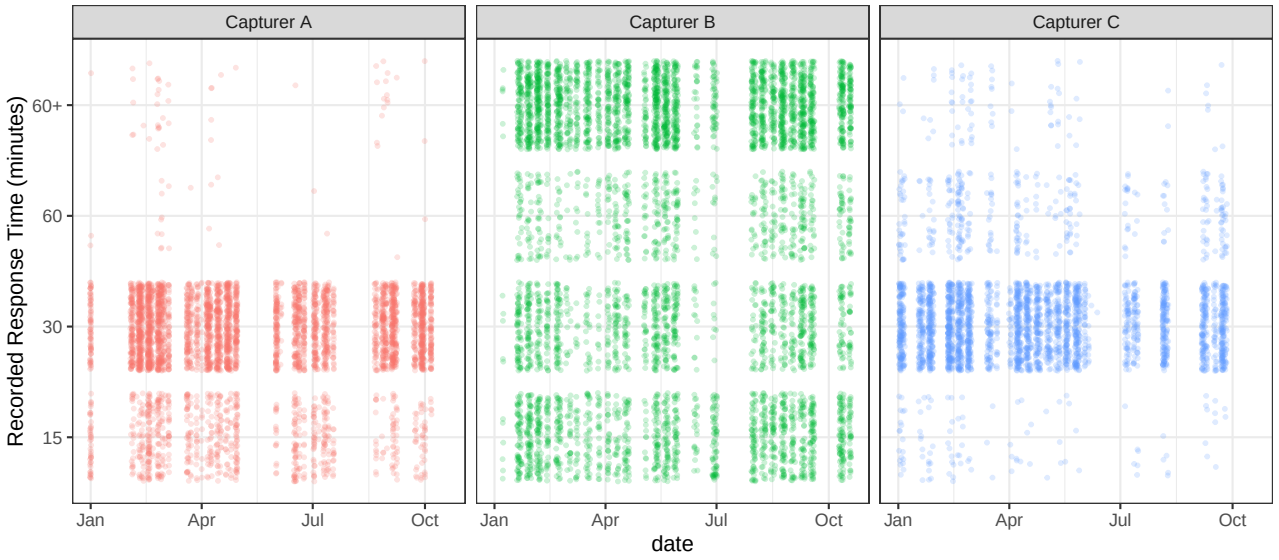


Figure A.2: Recorded Response Times by data capturer (2020).

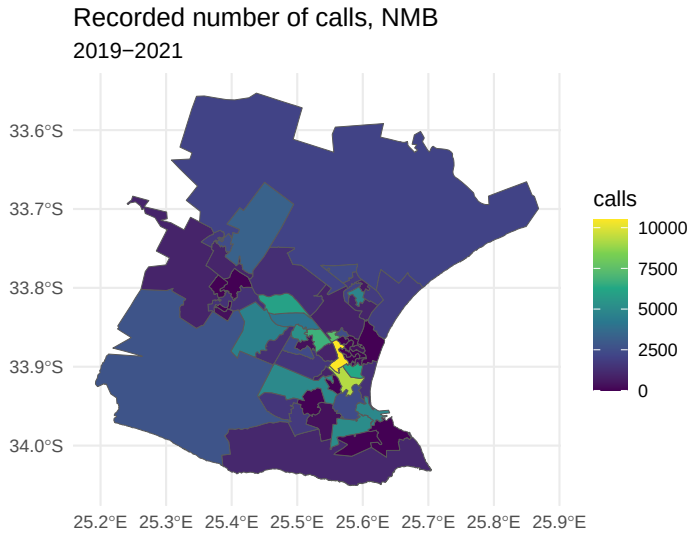
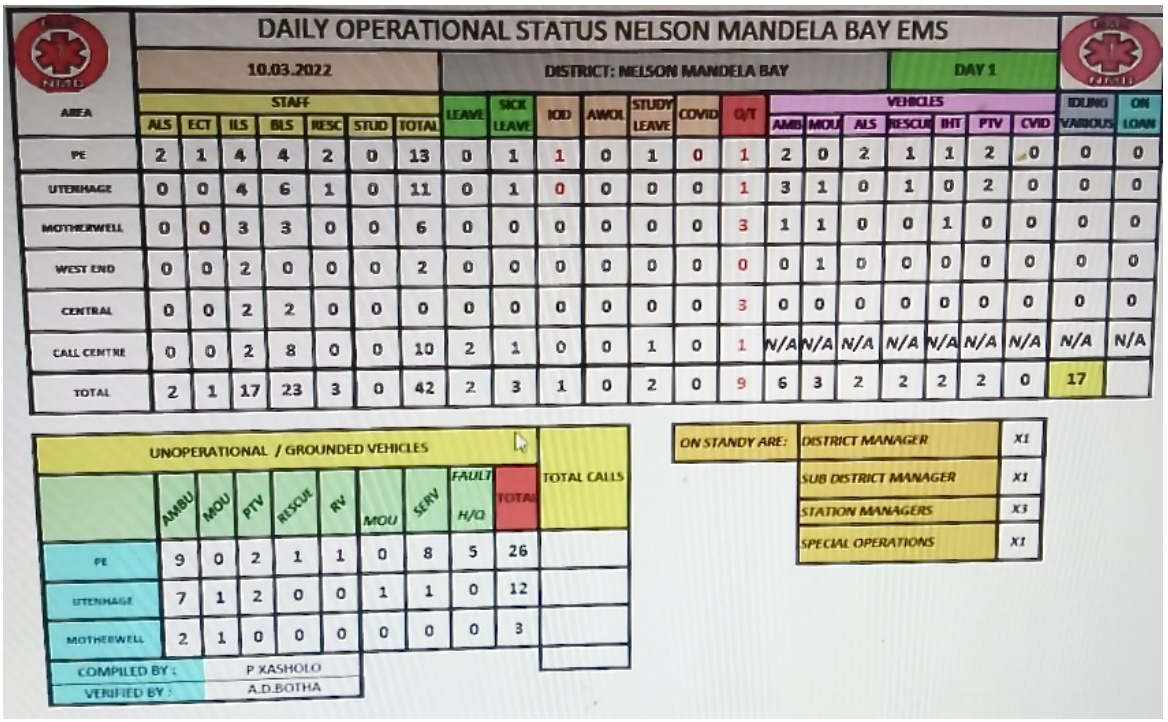


Figure A.3: Overall spatial EMS call distribution, by ward.



DAILY OPERATIONAL STATUS NELSON MANDELA BAY EMS

10.03.2022 | DISTRICT: NELSON MANDELA BAY | DAY 1

AREA	STAFF							LEAVE	SICK LEAVE	IOD	AWOL	STUDY LEAVE	COVID	Q/T	VEHICLES							IDLING VARIOUS	ON LOAN
	ALS	ECT	ILS	BLS	RESC	STUD	TOTAL								AMB	MOU	ALS	RESCUE	IHT	PTV	CVID		
PE	2	1	4	4	2	0	13	0	1	1	0	1	0	1	2	0	2	1	1	2	0	0	0
UTENHAGE	0	0	4	6	1	0	11	0	1	0	0	0	0	1	3	1	0	1	0	2	0	0	0
MOTHERWELL	0	0	3	3	0	0	6	0	0	0	0	0	0	3	1	1	0	0	1	0	0	0	0
WEST END	0	0	2	0	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
CENTRAL	0	0	2	2	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0
CALL CENTRE	0	0	2	8	0	0	10	2	1	0	0	1	0	1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
TOTAL	2	1	17	23	3	0	42	2	3	1	0	2	0	9	6	3	2	2	2	2	0	17	

UNOPERATIONAL / GROUNDED VEHICLES

	AMBU	MOU	PTV	RESCUE	RV	MOU	SERV	FAULT H/Q	TOTAL	TOTAL CALLS
PE	9	0	2	1	1	0	8	5	26	
UTENHAGE	7	1	2	0	0	1	1	0	12	
MOTHERWELL	2	1	0	0	0	0	0	0	3	

COMPILED BY : P XASHOLO

VERIFIED BY : A.D.BOTHA

ON STANDY ARE:		
	DISTRICT MANAGER	X1
	SUB DISTRICT MANAGER	X1
	STATION MANAGERS	X3
	SPECIAL OPERATIONS	X1

Figure A.4: Daily operational status spreadsheet, March 10th 2022.

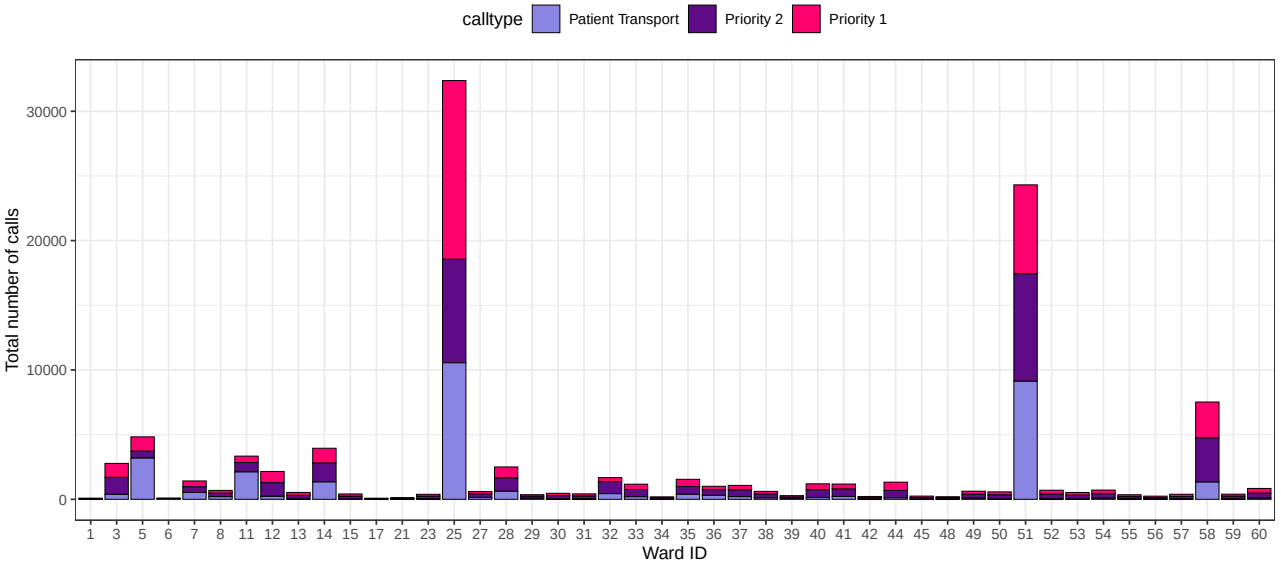


Figure A.5: Distribution of Destination Ward IDs, by Call Type.

Reg No	New Category	Site	District
GGW796EC	CATEGORY 1: Sedan 1400-1600cc	UITENHAGE	Nelson Mandela
GGW816EC	CATEGORY 1: Sedan 1400-1600cc	TSOLO	OR Tambo
GGW310EC	CATEGORY 1: Sedan 1400-1600cc	WILLOWMORE	Sarah Baartman
GGG365EC	CATEGORY 1: Sedan 1400-1600cc	Port Elizabeth	EMS College
GGG359EC	CATEGORY 1: Sedan 1400-1600cc	Bhisho	Head Office
GGW311EC	CATEGORY 1: Sedan 1400-1600cc	PE	Nelson Mandela
GGX024EC	CATEGORY 10: LDV Rescue	BURGERSDORP	Joe Gqabi
GGW853EC	CATEGORY 10: LDV Rescue	MTHATHA	OR Tambo
GGV470EC	CATEGORY 10: LDV Rescue	EAST LONDON	Buffalo City
GGW856EC	CATEGORY 10: LDV Rescue	CRADOCK	Chris Hani
GGW852EC	CATEGORY 10: LDV Rescue	GRAAFF REINET	Sarah Baartman
GGX215EC	CATEGORY 10: LDV Rescue	MQANDULI	OR Tambo
GGW946EC	CATEGORY 10: LDV Rescue	NGQELENI	OR Tambo
GGW948EC	CATEGORY 10: LDV Rescue	LUSIKISIKI	OR Tambo
GGX025EC	CATEGORY 10: LDV Rescue	MATATIELE	Alfred Nzo
GGS195EC	CATEGORY 10: LDV Rescue	LABORIA HOUSE	Sarah Baartman
GGW947EC	CATEGORY 10: LDV Rescue	Mzamba	Alfred Nzo
GGX020EC	CATEGORY 10: LDV Rescue	LADY FRERE	Chris Hani
GGS187EC	CATEGORY 10: LDV Rescue	MIDDELBURG	Chris Hani
GGR151EC	CATEGORY 10: LDV Rescue	Mzamba	Alfred Nzo
GGR156EC	CATEGORY 10: LDV Rescue	Alice	Amathole
GGR153EC	CATEGORY 10: LDV Rescue	Mthatha	OR Tambo
GGR150EC	CATEGORY 10: LDV Rescue	King Williams Town	Buffalo City
GGR155EC	CATEGORY 10: LDV Rescue	SOMERSET EAST	Sarah Baartman
GGR154EC	CATEGORY 10: LDV Rescue	MACLEAR	Joe Gqabi
GGR152EC	CATEGORY 10: LDV Rescue	Queenstown	Chris Hani
GGV619EC	CATEGORY 10: LDV Rescue	PE	Nelson Mandela
GGX206EC	CATEGORY 10: LDV Rescue	MOTHERWELL	Nelson Mandela
GGW849EC	CATEGORY 10: LDV Rescue	PE	Nelson Mandela
GGX214EC	CATEGORY 10: LDV Rescue	PE	Nelson Mandela

Figure A.6: Snapshot of EMS fleet register.

CASE NO	CODE	P1	P2	P3	G	Y	R	B	B	15M	30	60M	60+	F/T	HOSP/FROM	TO	REASON	CALLTAKER	DISPATCHER	AMB	CREWS	CREWS
1	22	1				1				1					GELVANDALE	L/H	KNOWN CARDIAC C/O SEVERE CHEST PAIN	COETZER	BRAINEERS	A282	WELCOME	JACOBS
2	28		1											1	EZINYOKA	F/T	CLINIC CASE	MALONI	HLONGWANA	A294	MAZIKO	MILISI
3	15	1				1				1					JOE SLOVO	WEMOU	PIMIGRAVIDA C/O LABOUR PAINS	MALONI	HLONGWANA	G186	VAN NIEKERK	COMMONS
4	31		1			1						1			L/H	PEPH	RIGHT KIDNEY INJURY	VAYO	HLONGWANA	A138	VAN WYK	MABONA
5	31		1			1					1				LIV	DNH	PNEUMONIA	COETZER	HLONGWANA	A138	VAN WYK	MABONA
6	29		1			1				1					KWAMAGXAKI	DNH	SHORTNESS OF BREATH	COETZER	BRAINEERS	G024	BOOYSEN	KWANELE
7	15	1				1				1					CHATTY 12	DNH	G6 C/O LABOUR PAINS	COETZER	HLONGWANA	A294	MAZIKO	MILISI
8	15	1				1				1				1	SOWETO	F/T	NO PT NO ESCORT	COETZER	HLONGWANA	A294	MAZIKO	MILISI
9	15	1				1				1					BARCELONA	WEMOU	PRIMIGRAVIDA C/O LABOUR PAINS	MALONI	HLONGWANA	G186	VAN NIEKERK	COMMONS
10	29		1											1	NEW BRIGHTN	F/T	NO RESPONSE	MALONI	NOT DISP	NOT DISP	NOT DISP	NOT DISP
11	29		1			1							1		S/S	DNH	VOMITTING,WEAKNESS	COETZER	BRAINEERS	A141	VAN HEEREN	MQWEBEDU
12	29		1											1	KWAZAKHELE	F/T	CANCELLED BY CALLER	MALONI	NOT DISP	NOT DISP	NOT DISP	NOT DISP
13	32		1			1					1				MORNING SIDE	LIV	FOR DIALYSIS	KATU	KATU	A091	PEDRO	NO PARTNER
14	32		1			1					1				CAPE ROAD	LIV	FOR DIALYSIS	KATU	HLONGWANA	A091	PEDRO	NO PARTNER
15	32		1			1					1				CAPE ROAD	LIV	FOR DIALYSIS	KATU	HLONGWANA	A091	PEDRO	NO PARTNER
16	32		1			1					1				ALGOA PARK	DES-CHEM	FOR DIALYSIS	GEORGE	HLONGWANA	A091	PEDRO	NO PARTNER
17	15		1			1					1				SOWETO	DNH	PT IN LABOUR	VAYO	HLONGWANA	G186	V.NIEKERK	COMMONS
18	29		1			1					1				S/S	DNH	FITS	BRAINERS	BRAINEERS	A203	HUMAN	QWABE
19	29		1			1					1				EZINYOKA	DNH	V+ D -WEAKNESS	COETZER	BRAINEERS	A310	BHELWANA	KWANELE
20	29	1							1		1			1	SALT LAKE	F/T	BOA OUT	VAYO	BRAINEERS	G024	BOOYSEN	KWANELE
21	32		1		1					1					NEW BRIGHTN	L/H R.U	HEAMODIALYSIS	KATU	HLONGWANA	T131	PHONGOLO	NO PARTNER
22	32		1		1					1					NEW BRIGHTN	L/H R.U	HEAMODIALYSIS	KATU	HLONGWANA	T131	PHONGOLO	NO PARTNER
23	32		1		1					1					NEW BRIGHTN	L/H R.U	HEAMODIALYSIS	KATU	HLONGWANA	T131	PHONGOLO	NO PARTNER
24	33		1			1					1				PALMRIDGE	F/T	NO RESPONSE	MULLER	SVR	T091	MULLER	NO PARTNER
25	33		1		1					1					FITCHARD CORNER	LIV	HEAMODIALYSIS	MULLER	SVR	T091	MULLER	NO PARTNER
26	33		1		1					1					GREENBUSHES	LIV	HEAMODIALYSIS	MULLER	SVR	T091	MULLER	NO PARTNER
27	33		1		1					1					WALMER	LIV	HEAMODIALYSIS	MULLER	SVR	T091	MULLER	NO PARTNER

Figure A.7: Snapshot of raw call centre spreadsheet, for June 2021.

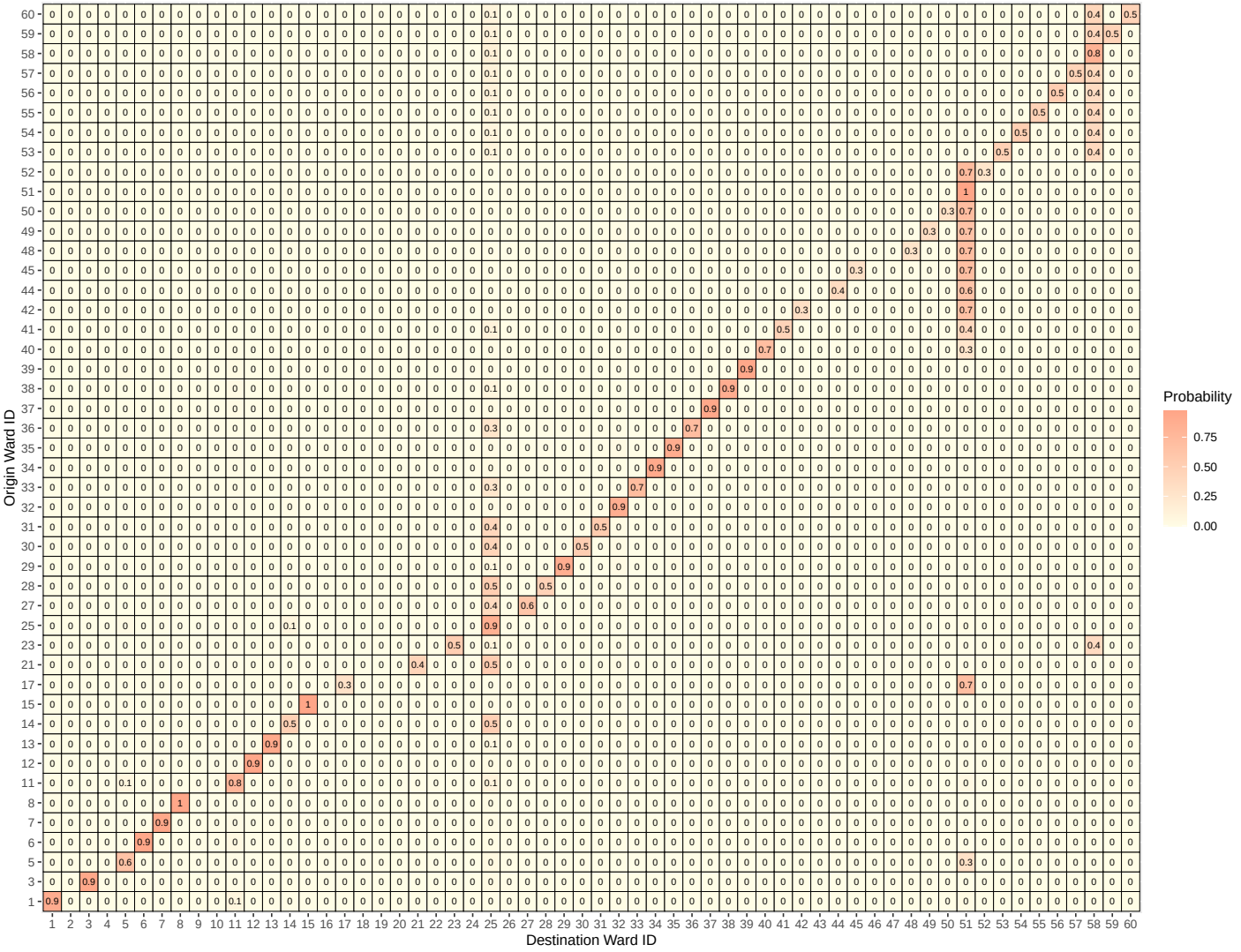


Figure A.8: Priority 2 Destination Ward Probability Matrix. Each row provides the probability distribution over destination wards, conditional on a given origin ward.

Note: Rows for wards without any observed demand are omitted. In this figure, some rows do not sum to 1 due to rounding.

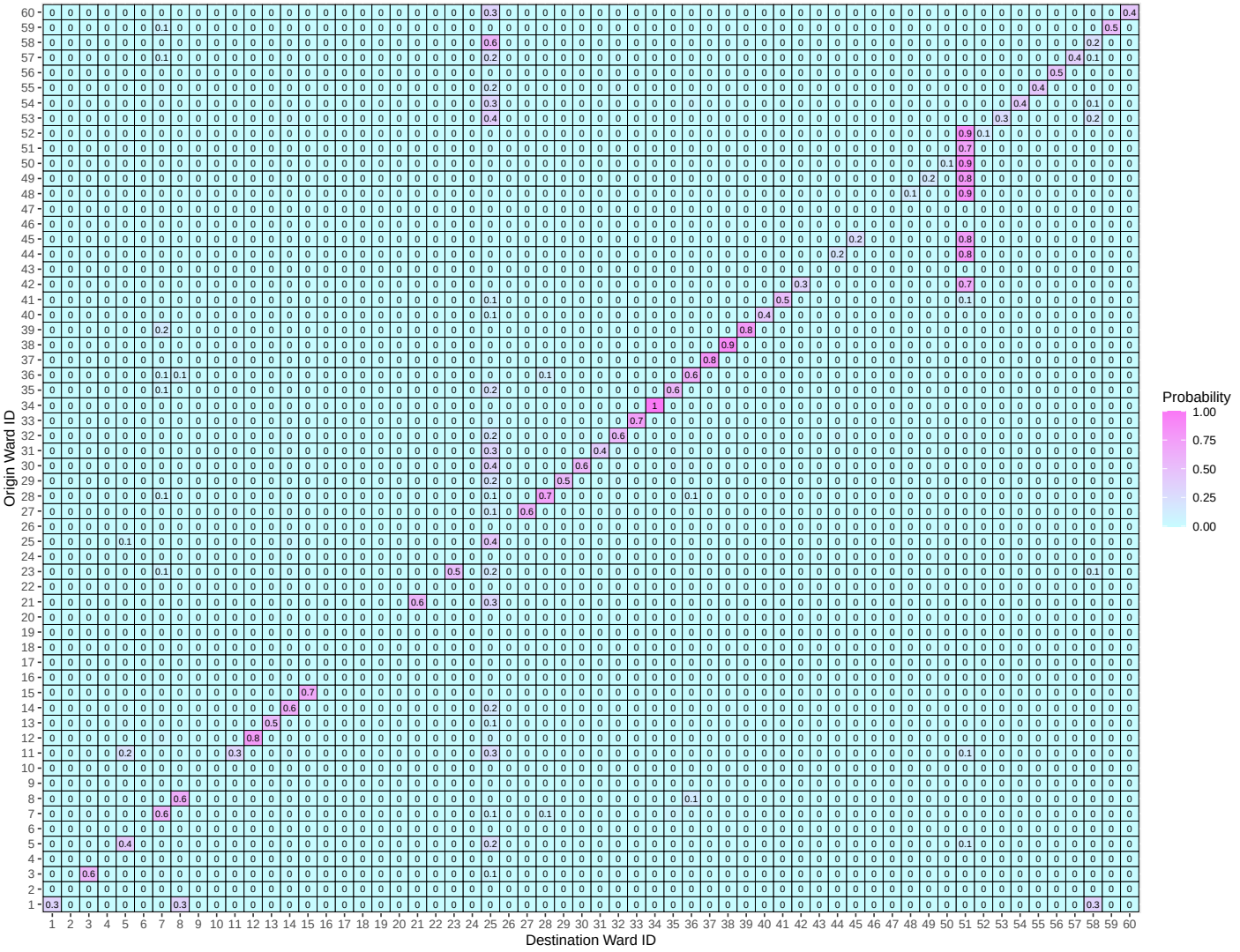


Figure A.9: Planned Patient Transport Destination Ward Probability Matrix. Each row provides the probability distribution over destination wards, conditional on a given origin ward.

Note: Rows for wards without any observed demand are omitted. In this figure, some rows do not sum to 1 due to rounding.

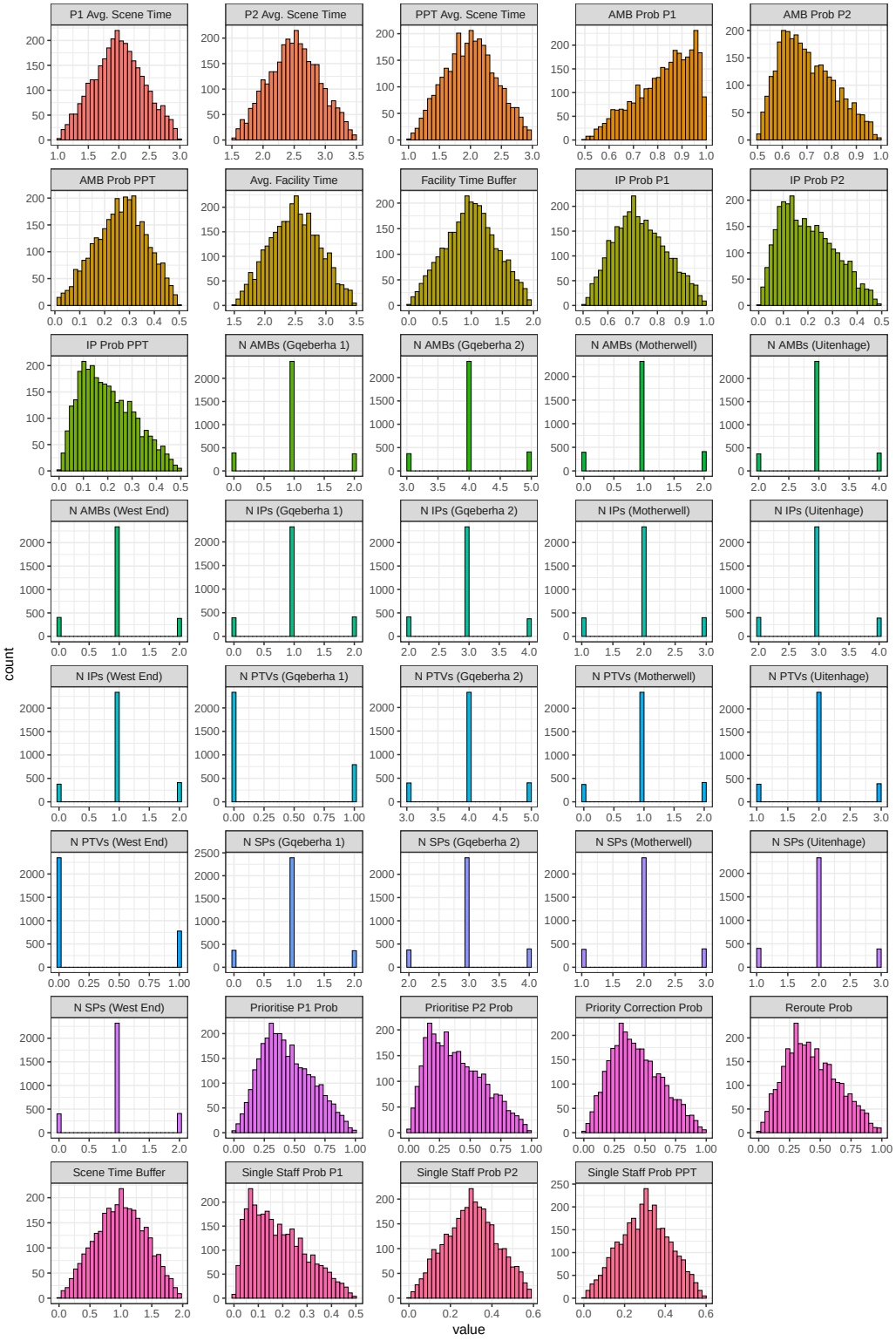


Figure A.10: Histograms of parameter distributions, formed by LHS.

Appendix B

Tables

Parameter	Mean	Min	Max
P1 Avg. Scene Time	2.00	1	3
P2 Avg. Scene Time	2.50	1.50	3.50
PPT Avg. Scene Time	2.00	1.00	3.00
AMB Prob P1	0.80	0.50	1.00
AMB Prob P2	0.70	0.50	1.00
AMB Prob PPT	0.30	0.00	0.50
Avg. Facility Time	2.50	1.50	3.50
Facility Time Buffer	1.00	0.00	2.00
IP Prob P1	0.70	0.50	1.00
IP Prob P2	0.20	0.00	0.50
IP Prob PPT	0.20	0.00	0.50
N AMBs (Gqeberha 1)	1.00	0.00	2.00
N AMBs (Gqeberha 2)	4.00	3.00	5.00
N AMBs (Motherwell)	1.00	0.00	2.00
N AMBs (Uitenhage)	3.00	2.00	4.00
N AMBs (West End)	1.00	0.00	2.00
N IPs (Gqeberha 1)	1.00	0.00	2.00
N IPs (Gqeberha 2)	3.00	2.00	4.00
N IPs (Motherwell)	2.00	1.00	3.00
N IPs (Uitenhage)	3.00	2.00	4.00
N IPs (West End)	1.01	0.00	2.00
N PTVs (Gqeberha 1)	0.25	0.00	1.00
N PTVs (Gqeberha 2)	4.00	3.00	5.00
N PTVs (Motherwell)	1.00	0.00	2.00
N PTVs (Uitenhage)	2.00	1.00	3.00
N PTVs (West End)	0.25	0.00	1.00
N SPs (Gqeberha 1)	1.00	0.00	2.00
N SPs (Gqeberha 2)	3.00	2.00	4.00
N SPs (Motherwell)	2.00	1.00	3.00
N SPs (Uitenhage)	2.00	1.00	3.00
N SPs (West End)	1.00	0.00	2.00
Prioritise P1 Prob	0.45	0.00	1.00
Prioritise P2 Prob	0.45	0.00	1.00
Priority Correction Prob	0.45	0.00	1.00
Reroute Prob	0.45	0.00	1.00
Scene Time Buffer	1.00	0.00	2.00
Single Staff Prob P1	0.20	0.00	0.50
Single Staff Prob P2	0.30	0.00	0.60
Single Staff Prob PPT	0.30	0.00	0.60

Table B.1: Triangular distributional parameters used for LHS, for each model parameter.

Scenario	P1			P2			PPT		
	Median	Lower	Upper	Median	Lower	Upper	Median	Lower	Upper
Base	55.81	55.00	57.23	88.16	86.16	89.97	96.95	93.87	99.21
Adding 1 IP and 1 SP: Gqeberha 1	60.21	57.89	62.41	92.73	87.57	97.96	102.50	95.28	111.64
Adding 1 IP and 1 SP: Gqeberha 2	38.16	36.61	39.11	43.11	41.40	44.60	42.47	39.89	44.31
Adding 1 IP and 1 SP: Motherwell	56.16	53.09	58.86	87.05	80.03	93.26	99.47	90.27	104.90
Adding 1 IP and 1 SP: Uitenhage	35.80	34.95	36.91	41.17	39.97	42.63	40.47	37.77	43.36
Adding 1 IP and 1 SP: West End	57.51	55.10	60.08	88.51	84.77	95.00	95.17	92.14	104.34
Improving classification (0.2)	52.44	50.01	54.47	94.99	90.22	99.64	98.45	93.62	107.18
Improving classification (0.5)	43.74	42.50	44.83	90.13	83.28	97.24	94.86	84.81	104.48
Improving classification (0.8)	36.13	35.04	37.71	86.15	75.83	95.09	85.79	80.73	94.61
Improving prioritisation & classification: 0.1	37.04	35.56	37.85	98.70	92.70	104.59	106.94	98.35	113.89
Improving prioritisation & classification: 0.3	34.67	33.63	35.03	95.57	90.60	101.46	103.88	94.96	109.99
Improving prioritisation & classification: 0.5	32.33	31.82	32.71	98.12	91.24	101.81	101.31	94.68	105.10
Improving prioritisation only	32.60	32.26	33.05	42.12	40.73	43.28	156.10	142.04	171.62
Improving staff allocation	65.98	62.71	68.84	121.49	112.66	131.77	125.14	115.03	134.98
Improving vehicle allocation	57.62	55.05	59.90	79.82	77.26	84.69	74.57	69.58	78.91
Increasing AMBs: Gqeberha 1	58.22	56.16	59.89	90.93	87.06	94.99	99.02	93.54	104.98
Increasing AMBs: Gqeberha 2	57.47	55.61	61.21	92.75	85.99	96.53	102.45	94.97	104.86
Increasing AMBs: Motherwell	54.23	52.53	55.46	89.47	83.90	93.20	99.31	93.51	105.19
Increasing AMBs: Uitenhage	56.00	54.12	58.56	90.41	85.08	95.86	95.34	90.18	100.98
Increasing AMBs: West End	56.95	54.94	59.48	91.37	86.10	96.14	99.69	94.98	105.08
Increasing IPs: Gqeberha 1	56.90	55.02	60.19	91.22	86.15	96.02	99.75	93.37	105.91
Increasing IPs: Gqeberha 2	44.94	42.92	46.20	59.03	55.50	60.70	60.57	56.02	64.25
Increasing IPs: Motherwell	57.00	54.54	60.80	89.40	85.44	93.83	97.57	92.68	105.07
Increasing IPs: Uitenhage	42.29	41.07	43.41	56.74	54.81	59.48	58.33	55.21	63.08
Increasing IPs: West End	57.65	55.32	59.60	88.44	83.22	94.72	96.59	93.97	102.97
Increasing PTVs: Gqeberha 1	59.36	55.14	61.74	88.96	83.32	95.68	94.98	92.18	104.89
Increasing PTVs: Gqeberha 2	56.82	54.45	58.92	90.15	86.19	94.24	99.95	94.87	104.89
Increasing PTVs: Motherwell	60.51	58.39	62.46	96.75	90.94	101.25	107.65	102.38	114.92
Increasing PTVs: Uitenhage	56.42	55.03	59.69	92.40	86.68	96.41	98.51	94.56	104.54
Increasing PTVs: West End	61.41	58.92	64.85	92.74	86.21	101.37	100.07	94.78	105.02
Increasing SPs: Gqeberha 1	58.07	54.60	61.27	90.45	82.58	100.47	100.34	92.49	112.22
Increasing SPs: Gqeberha 2	44.55	42.84	45.40	57.15	55.19	58.73	59.45	57.29	62.66
Increasing SPs: Motherwell	56.17	54.63	60.13	88.76	82.89	91.85	94.93	88.61	101.32
Increasing SPs: Uitenhage	44.52	42.53	46.89	58.72	55.51	60.04	59.71	57.31	63.01
Increasing SPs: West End	58.69	56.00	61.76	88.19	84.97	91.95	96.72	92.53	104.06

Table B.2: Bootstrapped median response times and 95% BCa confidence intervals by scenario and for each call type

mtry	min.node.size	splitrule	RMSE	Rsquared	MAE	RsquaredSD
2	5.00	variance	4.68	0.18	1.56	0.04
2	5.00	extratrees	4.73	0.17	1.56	0.05
22	5.00	variance	4.60	0.20	1.46	0.04
22	5.00	extratrees	4.43	0.26	1.41	0.05
42	5.00	variance	4.79	0.16	1.49	0.05
42	5.00	extratrees	4.38	0.27	1.41	0.04

Table B.3: Random Forest grid search results.

Scenario	P1			P2			PPT		
	Median	**Lower**	**Upper**	**Median**	**Lower**	**Upper**	**Median**	**Lower**	**Upper**
Base	55.81	55.00	57.23	88.16	86.16	89.97	96.95	93.87	99.21
High Cost 1	24.90	24.63	25.07	29.34	28.81	29.68	45.03	44.40	47.90
High Cost 2	25.93	25.51	26.38	29.45	28.88	29.90	34.83	32.73	35.07
Medium Cost 1	26.14	25.71	26.64	34.26	33.15	35.08	72.98	68.54	75.75
Medium Cost 2	27.31	26.98	27.95	35.84	34.99	36.45	54.81	53.46	57.70
Low Cost 1	29.36	28.73	29.59	42.51	41.52	43.61	139.07	129.83	154.84
Low Cost 2	31.25	30.79	31.80	45.94	45.02	47.89	94.00	88.57	102.48

Table B.4: Bootstrapped median response times and 95% BCa confidence intervals by scenario and for each call type, for combined scenarios

Scenario	**Red**			**Yellow**			**Green**		
	Median	Lower	Upper	Median	Lower	Upper	Median	Lower	
Base	58.99	57.81	60.11	69.43	67.83	70.41	91.54	89.37	94.26
Improving classification (0.2)	55.08	52.42	57.35	73.04	68.36	76.66	94.79	88.84	100.97
Improving classification (0.5)	45.10	43.45	46.44	72.78	68.07	76.97	91.38	82.78	100.74
Improving classification (0.8)	36.68	35.19	38.53	75.16	69.46	84.07	84.86	76.97	95.01

Table B.5: Bootstrapped median response times and 95% BCa confidence intervals by scenario and for each triage classification

Variable	Type	Stat	Base	Staff Sc.	HC 1	HC 2	MC 1	MC 2	LC 1	LC 2
Two IP	P1	Mean	26.5	28.9	26.4	27.5	28.4	29.0	24.3	25.2
		Lower	25.8	27.2	25.1	26.2	27.1	27.6	23.0	23.6
		Upper	27.2	30.7	27.6	28.9	29.7	30.5	25.7	26.8
	P2	Mean	15.6	18.2	16.2	16.0	17.9	17.8	14.7	14.9
		Lower	15.3	17.2	15.5	15.3	17.2	17.0	13.9	14.1
		Upper	16.0	19.2	16.8	16.7	18.6	18.6	15.5	15.7
	PPT	Mean	15.1	13.9	16.1	15.6	18.2	17.2	15.3	15.1
		Lower	14.7	13.0	15.4	14.8	17.3	16.3	14.4	14.1
		Upper	15.5	14.9	16.8	16.5	19.0	18.1	16.3	16.2
Two SP	P1	Mean	3.4	0.0	3.8	0.0	3.2	0.0	3.9	0.0
		Lower	3.3	0.0	3.5	0.0	2.9	0.0	3.5	0.0
		Upper	3.7	0.0	4.2	0.1	3.6	0.0	4.4	0.0
	P2	Mean	13.3	15.8	12.1	13.0	10.7	10.6	12.4	12.9
		Lower	12.8	14.3	11.4	12.1	9.9	9.8	11.5	11.7
		Upper	13.8	17.4	12.8	13.8	11.5	11.4	13.4	14.0
	PPT	Mean	16.3	18.4	14.4	18.2	12.9	16.2	14.7	17.3
		Lower	15.7	16.6	13.5	17.0	11.9	14.9	13.5	15.8
		Upper	16.9	20.2	15.4	19.4	14.0	17.5	15.9	18.8
IP & SP	P1	Mean	67.7	71.0	66.9	69.8	65.8	68.6	69.0	72.3
		Lower	67.0	69.3	65.6	68.4	64.4	67.1	67.6	70.7
		Upper	68.4	72.8	68.3	71.1	67.2	70.0	70.5	73.9
	P2	Mean	42.1	48.0	43.9	42.6	44.6	44.2	43.6	42.6
		Lower	41.4	46.1	42.7	41.3	43.2	42.8	42.1	41.1
		Upper	42.9	50.0	45.2	43.8	46.1	45.5	45.1	44.2
	PPT	Mean	42.0	39.8	44.5	40.5	45.3	42.1	45.1	42.1
		Lower	41.2	37.7	43.0	38.8	43.6	40.4	43.3	40.1
		Upper	42.9	41.9	45.9	42.1	47.1	43.8	47.0	44.1
One SP	P1	Mean	0.5	0.0	0.7	0.2	0.5	0.1	0.6	0.1
		Lower	0.5	0.0	0.6	0.1	0.5	0.1	0.5	0.1
		Upper	0.6	0.0	0.8	0.2	0.6	0.1	0.6	0.1
	P2	Mean	16.2	10.3	15.3	15.3	13.8	13.9	15.3	15.6
		Lower	15.8	9.6	14.7	14.7	13.1	13.2	14.5	14.7
		Upper	16.6	11.0	15.9	16.0	14.6	14.6	16.1	16.4
	PPT	Mean	13.9	16.4	13.4	14.6	11.6	13.0	12.2	13.9
		Lower	13.6	15.3	12.9	14.0	11.0	12.3	11.6	13.1
		Upper	14.3	17.5	14.0	15.2	12.2	13.6	12.8	14.7
One IP	P1	Mean	1.9	0.0	2.2	2.5	2.1	2.4	2.1	2.4
		Lower	1.8	0.0	2.0	2.4	1.9	2.2	2.0	2.2
		Upper	1.9	0.0	2.3	2.6	2.2	2.5	2.3	2.5
	P2	Mean	12.7	7.6	12.5	13.2	12.9	13.5	14.0	14.0
		Lower	12.3	6.9	12.0	12.6	12.3	12.8	13.3	13.3
		Upper	13.1	8.3	13.1	13.7	13.6	14.2	14.7	14.8
	PPT	Mean	12.6	11.6	11.6	11.2	11.9	11.5	12.7	11.6
		Lower	12.3	10.6	11.0	10.6	11.3	10.9	12.0	10.9
		Upper	13.0	12.5	12.1	11.8	12.6	12.2	13.4	12.3

Table B.6: Bootstrapped staff mixture percentages and 95% BCa confidence intervals by scenario and for each call type. Full column names are, respectively: Variable, Type, Statistic, Base Scenario, Improved Staff Allocation Scenario, High Cost 1, High Cost 2, Medium Cost 1, Medium Cost 2, Low Cost 1, Low Cost 2

Vehicle	Type	Stat	Base	Vehicle Sc.	HC 1	HC 2	MC 1	MC 2	LC 1	LC 2
AMB	P1	Mean	99.5	100.0	99.3	100.0	99.4	100.0	99.4	100.0
		Lower	99.5	100.0	99.2	100.0	99.3	100.0	99.4	100.0
		Upper	99.5	100.0	99.3	100.0	99.4	100.0	99.5	100.0
	P2	Mean	80.9	79.3	80.0	77.8	80.3	78.0	80.7	77.7
		Lower	80.7	78.7	79.6	77.4	79.9	77.5	80.3	77.2
		Upper	81.2	79.9	80.4	78.2	80.8	78.5	81.1	78.1
	PPT	Mean	50.0	31.6	47.7	29.0	50.9	32.2	53.6	34.2
		Lower	49.6	30.5	47.0	28.4	50.2	31.4	52.9	33.3
		Upper	50.4	32.6	48.3	29.7	51.7	33.0	54.4	35.1
PTV	P1	Mean	0.5	0.0	0.7	0.0	0.6	0.0	0.6	0.0
		Lower	0.5	0.0	0.7	0.0	0.6	0.0	0.5	0.0
		Upper	0.5	0.0	0.8	0.0	0.7	0.0	0.6	0.0
	P2	Mean	19.1	20.7	20.0	22.2	19.7	22.0	19.3	22.3
		Lower	18.8	20.2	19.6	21.8	19.2	21.5	18.9	21.9
		Upper	19.3	21.4	20.4	22.6	20.1	22.5	19.7	22.8
	PPT	Mean	50.0	68.4	52.3	71.0	49.1	67.8	46.4	65.8
		Lower	49.6	67.4	51.7	70.3	48.3	67.0	45.6	64.9
		Upper	50.4	69.5	53.0	71.7	49.8	68.6	47.1	66.7

Table B.7: Bootstrapped vehicle type percentages and 95% BCa confidence intervals by scenario and for each call type. Full column names are, respectively: Vehicle, Type, Statistic, Base Scenario, Improved Vehicle Allocation Scenario, High Cost 1, High Cost 2, Medium Cost 1, Medium Cost 2, Low Cost 1, Low Cost 2