Integrating artificial neural networks, simulation and optimisation techniques in improving public emergency ambulance preparedness for heterogeneous regions under stochastic environments

By

Tichaona Wilbert Mapuwei

A thesis submitted in Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Applied Statistics
in the School of Mathematics, Statistics and Computer Science
University of KwaZulu-Natal

September 2021

UNIVERSITY OF KWAZULU-NATAL
INYUVESI YAKWAZULU-NATALI

Declaration

I, Tichaona Wilbert Mapuwei declare that the work presented under this thesis titled 'Integrating artificial neural networks, simulation and optimisation techniques in improving public emergency ambulance preparedness for heterogeneous regions under stochastic environments' is my own. I confirm that:

- The dissertation is my original research undertaken for the candidature of a doctor of philosophy degree at the university.
- The thesis has not been submitted for any degree or qualification at any other university.
- I have duly acknowledged the published work of other researchers.

[redacted] 06/04/2022

Mr. Tichaona W. Mapuwei Date

[redacted] 6/04/2022

Dr. Oliver Bodhlyera Date

Prof Henry G. Mwambi Date

Acknowledgments

I would like to express my gratitude to my supervisors Dr. Oliver Bodhlyera and Prof. Henry Mwambi for their guidance, patience, encouragement and support during the period of my study. I would also want to appreciate the support rendered to me by the entire Mathematics, Statistics and Computer science staff and postgraduate students.

I sincerely want to thank my family for their unwavering emotional and financial support throughout the period of my study. I would also want to thank God for granting me the opportunity, capacity and ability to study during the difficult but exciting moments of my career.

Abstract

The Bulawayo Emergency Medical Services (BEMS) department continues to rely on judgemental methods with limited use of historical data for future predictions, strategic, tactical and operational level decision making. The rural to urban migration trend has seen the sprouting of new residential areas, and this has put pressure to the limited health, housing and education resources. It is expected that as population increases, there is subsequent increase in demand for public emergency services. However, public emergency ambulance demand trends has been decreasing in Bulawayo over the years. This trend is a sign of limited capacity of the service rather than demand itself. The situation demanded for consolidated efforts across all sectors including research, to restore confidence among residents, reduce health risk and loss of lives.

The key objective was to develope a framework that would assist in integrating forecasting, simulation and optimisation techniques for ambulance deployment to predefined locations with heterogeneous demand patterns under stochastic environments, using multiple performance indicators. Secondary data from the Bulawayo Municipality archives from 2010 to 2018 was used for model building and validation. A combination of methods based on mathematics, statistics, operations research and computer science were used for data analysis, model building, sensitivity analysis and numerical experiments.

Results indicate that feed forward neural network (FFNN) models are superior to traditional SARIMA models in predicting ambulance demand, over a short-term forecasting horizon. The FFNN model is more inclined to value estimation as compared to SARIMA model, which is directional as depicted by the linear pattern over time. An ANN model with a 7-(4)-1 architecture was selected to forecast 2019 public emergency ambulance demand (PEAD). Peak PEAD is expected in January, March, September and December whilst lower demand is expected for April, June and July 2019.

Simulation models developed mimicked the prevailing levels of service for BEMS with six(6) operational ambulances. However. the average response times were well above 15 minutes, with significantly high average queuing times and number of ambulances queuing for service. These performance outcomes were highly undesirable as they pose a great threat to human based outcomes of safety and satisfaction with regards to service delivery.

Optimisation for simulation was conducted by simultaneously minimising the average response time and average queuing time, while maximising throughput ratios. Increasing the number of ambulances influenced the average response time below a certain threshold, beyond this threshold, the average response time remained constant rather than decreasing gradually. Ambulance utilisation inversely varied to increase in the fleet size. Numerical experiments revealed that reducing the response time results in the reduction in number of ambulances required for optimal ambulance deployment. It is imperative to simultaneously consider multiple performance indicators in ambulance deployment as it balances resource allocation and capacity utilisation, while avoiding idleness of essential equipment and human resources. Management should lobby for de-congestion and resurfacing of old and dilapidated roads to increase access and speed when responding to emergency calls.

Future research should investigate the influence of varying service time on optimum deployment plans and consider operational costs, wages and other budgetary constraints that influence the allocation of critical but scarce resources such as personnel, equipment and emergency ambulance response vehicles.

List of Acronyms

ANN: Artificial Neural Network.

AUR: Average Utility Ratio Expressed As a Percentage.

AVNRQ: Average Number of Calls In Response Queue.

AVQT: Average Queuing Time Per Call (minutes).

AVRT: Average Response Time Per Entity (minutes).

AVTIS: Average Total Duration in System (minutes).

BEMS: Bulawayo Emergency Medical Services.

EMS: Emergency Medical Services.

FFNN: Feed Forward Neural Network.

NEC: Non-emergency Calls.

NOA: Number of Ambulances.

PEAD: Public Emergency Ambulance Demand.

RTD: Response Time Distributions

TPR: Throughput Ratio Expressed As a Percentage.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Despite the ever- increasing rural to urban migration which has seen the population of urban areas in Zimbabwe increasing, Bulawayo City included, health services delivery remains low in terms of efficiency, effectiveness and equality. This migration trend has seen the sprouting of new residential areas (formal and informal) and has put pressure on the limited resources in terms of housing, health and education among many other critical human social amenities (Potts, 2000, 2006; Tawodzera, 2011). It is expected that as the population increases, there is subsequent increase in demand for public emergency services. This has not been the case with the demand trends for public ambulance emergencies in Bulawayo. One would assume that if service delivery is poor, an affected individual will turn to other alternatives which might as well be expensive or the individual might even live to meet his or her fate.

Despite reaching the highest annual public ambulance service demand total of 66630 in 1998, the trend has been decreasing up-until 2008 (6595) for Bulawayo City (Fig. 1.1). A significant increase was experienced between 2009 and 2011. This could have been influenced by the use of a multi-currency currency system which included the use of United States Dollars (USD) among a mixed bag of currencies. However, there has been a decreasing trend between 2012 up to 2018. The situation therefore, calls for consolidated efforts across all sectors including research in order to restore confidence among residents whilst reducing health risk and minimising the loss of lives. Planning into the future to restore a balanced fire and ambulance level of preparedness are vital in order to prevent the loss of lives and property in isolated incidences and let alone a disaster such as cholera outbreaks, fire or major road accidents.

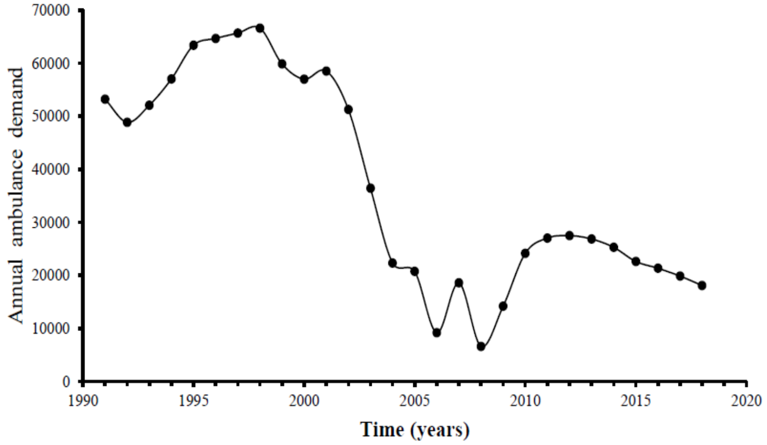


Figure 1.1: Annual PEAD, Bulawayo, Zimbabwe (1991-2018).

Zimbabwe has not been spared with a declining economy characterised by hyperinflation, drought and political violence that has seen the increase in poor service delivery across all economic, social and political divides. However, there has been renewed hope with the new dispensation after the November, 2017 events which led to the removal of the long-serving president who presided during this period of mixed fortunes in Zimbabwe. There has been a call for public and private investment partnership in all sectors of the economy with the health sector being one of them. The new government has since adopted the new mantra dubbed “Zimbabwe is Open for Business”. A blue print entitled “Transitional Stabilisation Programme (TSP)” was adopted by the Zimbabwean government with the objective of creating a conducive environment to stimulate the Zimbabwean economy. The research is envisaged to have been conducted at an opportune time as Zimbabwe is going through a new political and economic dispensation where the government has also made pronouncements for foreign direct investment.

Zimbabwe being a member of the United Nations is obliged also to focus on the sustainable development goals (SDGs) and the health sector being one of the critical areas. Goal number three (3) is focused on the good health and well-being of citizens (Qasim, 2013). Zimbabwe is also a signatory to the Abuja Declaration of April 2001, which compels African Union governments to allocate at least 15% of their total annual budget towards the health sector. The issue of public emergency ambulance service provision plays a critical role in ensuring the good health and well-being of the general populace.

The government of Zimbabwe after realising the importance of this sector undertook a national status survey on all fire and ambulance services in the country through the Ministry of Local Government, Public Works and National Housing. The objective of the audit was to assess the extent of preparedness in the provision of Fire and Ambulance services by Local Authorities.

1.1 Roles and Responsibilities of Fire and Ambulance Services

According to the Local Authorities Fire and Ambulance Emergency Services Operational Procedure Manual (Page 2-3), the functions can be divided into four broad but interdependent categories.

1.1.1 Command and Administration Unit

The functions of this unit includes: ensuring the overall safety of the Local Authority and its residents through the maintenance of a well-equipped and well-trained Fire Department, recruitment, training, and supervision of an appropriate number of Fire-fighters, Ambulance personnel, support staff and their development, procurement of appropriate equipment, guided by current trends, research, need, and experience, advising Council on matters relating to Public Safety, and ensuring a high-level of cooperation with other Council Departments, and the formulation of policies and the general maintenance of emergency vehicles, plant and equipment.

1.1.2 Fire Fighting and Rescue Unit

This unit of a Local Authority is responsible for: immediate dispatch of fire engines, ambulances and personnel to any emergency involving fire, or any other situation where human life is at risk, rescuing of affected or injured persons from the situation and the initiation of appropriate medical care and investigating the cause of the situation leading to identifying factors that could be changed to prevent a similar occurrence.

1.1.3 Emergency Medical Services Unit

Key functions of this unit include to assessing any call for help required, dispatching appropriate human and material resources and to implement appropriate medical care at the scene and transportation of the patient(s) to the most appropriate facility.

1.1.4 Fire Prevention Unit

This unit has the mandate of maintaining a high standard of fire prevention and public safety in all public buildings, thereby reducing to the lowest level possible, the risk of fire. It is also mandated to conduct regular inspections of all factories, commercial buildings, places where flammable liquids are stored, shops and places of public gathering, to ensure compliance with fire prevention by-laws and building codes.

1.2 Laws Governing Operations of Ambulance Services in Local Authorities

Section 200 (1) of the Urban Councils Act [Chapter 29:15] and First Schedule Section 35 of the Rural District Councils Act [Chapter 29:13] state that the Ambulance Services division provide emergency services in the form of conveying patients and performing first aid on patients from their home or incident scene to the hospital. The ambulance crew treats the patient during the rescue operation and performs en-route treatment of patients on their way to the hospital. Incidents are prioritised according to their demand for attention as life threatening, serious or minor. Life-threatening incidents require special medical attention other than just conveying the patient to hospital. These include maternity emergencies and medical emergencies like asthma, diabetes, choking, heart attack, stroke and collapsing cases.

According to the Operational Procedures Manual, categories of emergencies are set after considering incidents where life is in immediate danger and these are given a higher priority than the non-emergency request for transport. The control room attendant gives instructions to callers on what to do whilst the ambulance is on its way. Maternal conveyances and medical conveyances do not require special medical attention but just convey the patient to the hospital. When a patient has been conveyed and handed over to the hospital, the ambulance crew in the event that the patient has cash, receives payment and issues a receipt in accordance with the fees schedule. If the patient does not have cash on hand they prepare the client's bill which is supposed to be settled at the Local Authority's Treasury Department.

The laundry, cleaning and disinfection of vehicles are done using equipment available at the Local Authorities fire premises.

1.2.1 Fire Tenders and Ambulances Response Time

The International Standards of Fire Cover (2.14) states that the location of Fire stations should allow the responding team to reach the scene of the incident within five minutes after receiving the call in Cities, Municipalities and Town Councils and within twenty minutes in Rural District Councils after receiving the call. Where life is in danger, the fire tenders are supposed to be accompanied by an ambulance. The fire stations are supposed to have fire tenders and ambulances. The quantity of fire tenders and ambulances required by Local Authorities is dependent upon whether it is a City Council, Municipal Council, Town Council, Local Board or Rural District Council. According to Fire-by laws, Local Authorities are supposed to charge fire levy to ratepayers for capacitating the fire and ambulance services.

1.2.2 Funding of Fire and Ambulance Services in Local Authorities

The Fire and Ambulance services in Local Authorities are revenue collection units and their finances are controlled centrally by the Treasury Departments. The Treasury Department's source of funds is mainly from ratepayers in the form of rates and tariff charges. The Treasury Departments provide and manage funds of the Fire and Ambulance services in accordance with the approved budgets.

1.3 Operations of the BEMS Department

The Bulawayo Emergency Ambulance Services (BEMS) is a non-profit organisation that is managed by the Bulawayo City Council. The headquarters are at Famona fire station and the other substations are Northend, Nketa and Nkulumane. Each station was assigned to cover particular geographical zones in Bulawayo city. The control room is manned from The Tower Block in the city centre. This is where all calls are received and managed through a fully computerised emergency communication and mobilisation system manned by two (2) officers twenty-four (24) hours a day and three hundred and sixty-five (365) days a year. Details of the call received in the Control Room are recorded and the nearest station mobilised to respond to a reported emergency.

The BEMS department uses the manual system to capture information on every call that is received. This information enables the BEMS to prepare monthly reports on how they are performing. These reports have indicated that the organisation is finding it more and more difficult to meet its service targets.

When a call arrives at BEMS headquarters members of staff in the control room identify an available ambulance that is either idle at its base station or returning from its previously assigned job and dispatch this vehicle to the scene. After initial treatment at the scene, the ambulance crew transports the patient to a hospital, performs a hand over and take-over procedure to hospital staff, and then returns to its base station. If further transport services are not required, the ambulance returns directly to its base from the scene. The vehicle is considered available to receive calls as soon as it begins returning to base.

1.3.1 Operational Challenges being faced by BEMS Department in Recent Years

According to a report in response to the National Status Survey of 2017 on all fire and ambulance services in the country, Bulawayo City continues to face operational challenges in its service delivery. Bulawayo is ill-equipped in terms of aged equipment, obsolete command vehicles and ambulances. The authorized establishment of 176 fire and 114 ambulance crew staff has not been met over the years, Bulawayo operates with a meagre number of personnel currently at ±50 percent for both fire and rescue ambulance services.

The Ministry of Local Government, Public Works and National Housing imposed a freeze on the recruitment of fire-fighters and ambulance staff and this has affected the Bulawayo Emergency Services levels of preparedness in conducting their business. There is critical shortage of personnel. The planned outcome of the job evaluation and rationalization exercise has seen the division being downsized to 93 fire fighters to cover a three- (3) shift system. The absence of adequate personnel results in ill-preparedness of the emergency services to deal with major incidents or disasters.

Bulawayo City has four (4) fire stations strategically located to respond to any part of the City's ±483 square kilometres in ten (10) minutes. The City has grown exponentially without a corresponding increase in the number of fire stations and personnel to reduce the response time from ten (10) to five (5) minutes. The current economic performance and other factors affecting the financial position of the Local Authority is having a direct bearing on the number of stations that can be established and operated to reduce the response time to five (5) minutes.

Only six (6) ambulances are operated per shift to service the entire City and beyond owing to lack of Ambulance Technicians/Emergency Medical Technicians. Though old, the available ambulance fleet allows the division to operate ±11 ambulances per shift. However, this is against a stipulated requirement of thirty-two ambulances.

Suburbs such as Mahatshula, Woodville, Cow-dray Park, Waterford and others, takes between 20 and 30 minutes to reach them a situation that increases damage to property and loss of lives, above all dissatisfaction from the public and the affected.

The City of Bulawayo Fire and Ambulance Services has a fully computerised emergency communication and mobilisation system manned by two (2) officers 24/7, 365 days a year. Details of the call received in the Control Room are recorded and the nearest station mobilised to respond to a reported emergency. Notable deficiencies are the absence of a redundancy system to support continuity of communication in the event of a sudden breakdown of the system. A number of emergency vehicles lack mobile radios to ensure communication between responding teams and the emergency control room.

1.3.2 Persistent Technical Challenges faced by Ambulance Services Department

Though every call is recorded as it is received, this is still being done manually and the final consolidation of the information is done on a monthly basis. The summary of ambulance demand calls is done by category and on monthly basis. However, the daily demand is overlooked which could be used effectively to forecast either daily, weekly or monthly demand.

There has been no full utilisation of historical data to assist in forecasting demand for ambulance services in the near to distant future. The organisation relies on judgemental methods based on knowledge and qualified guesses of present and future events.

Though there is deliberate effort to achieve the 5 minutes' response time as an international practice, this has continued to elope the efforts by the responsible authorities. This failure to meet such standards has been compounded by a shortage of staff, ambulance vehicles and the ever-increasing residential housing boundaries and the population of the city.

Response time is defined as the time interval between receiving a call to the time that an ambulance first arrives at the scene is a key performance measure.

Calls are still being attended on “first come first serve” (FIFO) though priority is being given to Road Traffic Accidents (RTA). The speed of response depends not on the condition of the patient but on the volume of calls being responded to and how far away the nearest ambulance might be. This has serious consequences on the patient’s life and property. This motivates the researcher to also focus on issues of efficiency in terms of service delivery of the Bulawayo Fire Brigade Services.

Those in the fire and ambulance professions are concerned with rescuing lives and saving property. The question that arises is, is Bulawayo prepared for a major incident, let alone a disaster? The main objective in terms of ambulance logistics is to ensure that the customer who is the patient in need of ambulance services is satisfied. To achieve this, the whole process starts in the general planning by management, which has a strong bearing on the quality of service delivery. The issue of response time is a critical aspect of emergency response and has to be considered seriously, as every second counts when life or property is at risk.

The Bulawayo ambulance services continues to face problems, which are common to any ambulance service provider all over the world today. The ambulance services department faces a host of difficult policy questions related to the operation of the service. Henderson and Mason, (2004) propounded key sample questions that every other ambulance service provider should consider namely: how many ambulances should be employed and where should they be stationed, how can one distribute a limited number of ambulances between a high-demand residential area and a low-demand residential area, what procedures and policies must be adhered to as calls for assistance are received in order to ensure rapid response to calls while obtaining complete and quality information to allow for correct dispatching, should ambulances be used for non-urgent patient transfers without compromising the usual emergency response function, how dispatching decisions should be made when multiple vehicles are available for emergency dispatch and how can geographical information systems (GIS) be incorporated in the assignments and allocation of ambulances?

1.4 Problem Statement

Demand for public emergency ambulance services has been on the decrease for a prolonged period. Since efforts are gradually being made to restore sanity to the service delivery, there is a need to increase the confidence of clients with limited resources. Two important bold steps must be taken in order to address this fundamental issue. The first is to optimally use the current resources in order to restore confidence among the residents of Bulawayo whilst reducing the health risk and minimising the loss of lives. Secondly, there is a need to build capacity over the years to come and this can be attained by embarking on strategic and rigorous scientific research to meet the required international standards.

There is need therefore to develop robust and smart planning methods to ensure a balance between the number of ambulances available and skilled personnel manning these in line with the financial resources available and the corresponding demand within the demand zones.

Operational challenges arising from the depleted number of ambulances, increasing population, expansion of residential boundaries and the diminishing staff establishment of the respective stations has indirectly compromised issues such as equality, effectiveness and efficiency in the provision of public emergency ambulance services. There is need to seriously consider the use of forecasting, simulation and optimisation tools to forecast demand and track performance measurements that would assist in the allocation of these scarce resources.

Henderson and Mason (2005) found that statistical information, outlining when and where calls are likely to occur, is rarely used to support decision-making.. This is mainly due to the inability of the technical systems used by ambulance organisations (including BEMS) to supply relevant statistical data. Methods for utilising the data have to be developed and explored in this research. To effectively direct efforts into solving the challenges being faced by the BEMS, the researcher will focus on how neural networks, simulation modelling and optimisation techniques can be integrated and adopted in forecasting and capacity planning in addressing issues of equality, effectiveness and efficiency.

1.5 The Motivation of the Study

There has been a lot of work done in literature that focused on capacity and staffing of base stations, bases location, scheduling of ambulance crews whose focus was mainly based on the known deterministic demand and supply side of the ambulance service logistic chain (Batta et al., 1989; Başar et al., 2012; Bélanger et al., 2019). The deterministic demand component is the number of emergency ambulance calls received, and the deterministic supply component is the availability of ambulance personnel and ambulance vehicles (Bélanger et al., 2019). However, little has been done to assess the impact of the stochastic nature of the input data (emergency ambulance calls), the response time, service time and the overall utility of the ambulance vehicles which has a strong influence in addressing logistical issues such as equality, efficiency and effectiveness in the provision of public emergency services.

The situation calls for consolidated efforts across all sectors including research, in order to restore confidence among residents, reduce health risk and loss of lives. Planning into the future to restore a balanced fire and ambulance level of preparedness is vital in order to prevent the loss of lives in isolated incidences or even a disaster such as the cholera outbreaks that have been experienced in recent years (Chronicle, 2018), fire or major road accidents.

1.6 Research Objectives

The objectives of the study are to:

(i) Identify the operational and technical challenges being faced by the organisation in terms of public emergency ambulance service delivery.

(ii) Validate the forecasting performance of neural networks against traditional forecasting techniques in forecasting ambulance demand.

(iii) Forecast annual demand for resource planning and mobilisation.

(iv) Develop a framework for integrating forecasting, simulation and optimisation techniques for heterogeneous regions to improve levels of preparedness, equality, efficiency and effectiveness in emergency ambulance services.

1.7 Research Questions

The researcher is motivated by important questions stated below:

(i) What are the current operational and technical challenges being faced by the organisation in terms of public emergency ambulance service delivery?

(ii) Is there a comparative advantage of using neural networks in predicting near to distant deterministic demand to help improve public emergency ambulance preparedness?

(iii) How can simulation modelling be integrated with forecasting using artificial neural networks and optimisation techniques in decision making processes focused on improving levels of preparedness, equality, efficiency and effectiveness in emergency ambulance services?

1.8 Expected Benefits of the Study

The forecasted demand would assist in resource planning which is not limited to the allocation of ambulances to the respective stations but include staff training, ambulance servicing, community awareness campaigns, scheduling of annual leave days for members of staff, hiring of equipment or manpower and even routine checks on the alertness of the ambulance team as a whole.

The use of simulation techniques will help to measure the capacity of achieving set performance measurements as guided by international standards whilst directly or indirectly addressing important aspects of equality, efficiency and effectiveness in the provision of service to the citizens. The optimisation would ensure the development of optimal or near-optimal ambulance deployment plans under multiple performance indicators to ensure equality, efficiency and effectiveness in EMS resource mobilisation, allocation and utilisation. Additionally, the research will contribute to the body of literature by demonstrating how artificial neural networks (future demand) work. The integration of simulation modeling and optimization in decision making can prove to be effective and provide a number of scenarios for consideration prior to committing resources for emergency medical services systems. The integrated strategy can further be adopted for decision making to similar cases that involve a server-to-customer operating environment such as taxis and vehicle recovery machinery.

1.9 Delimitation of Study

The research will be confined to identifying appropriate neural network models, simulation and optimisation techniques that can be integrated and utilised in improving public ambulance emergency preparedness. The research will focus on the period spanning from January 1991 to December 2019. Bulawayo City Council's Department of Fire and Ambulance Services will be used as a case study.

1.10 Limitations of the Study

The study relied on secondary data which was largely available on hard copies. It implied manual data capturing as the current database is operating on partial automation. However, thorough data mining was performed to ensure completeness and accuracy of source data. The advent of COVID-19 posed serious challenges to the execution of the study as it minimised levels of engagement with relevant stakeholders. However, alternative channels of engagement were adopted to ensure the progress of the study. Though Bulawayo was used as a case study, some generalisations can be made for other public emergency service providers.

1.11 Organisation of the Study

The study comprises eight chapters, the first chapter covers the introduction, background, problem statement, objectives, research questions, delimitations as well as the limitations of the study. Chapter 2 covers the literature review, empirical evidence as well as the conceptual framework of the study. Chapter 3 provides the methodology while Chapter 4 covers the development of forecasting models using artificial neural networks (ANN) and seasonal autoregressive integrated moving average models (SARIMA). Chapter 5 focuses on simulation modelling of the emergency medical services (EMS) for the heterogeneous regions. Chapter 6 covers sensitivity analysis by adopting the optimisation for simulation approach. Chapter 7 focuses on numerical experiments to explore the influence of standardising response time on optimal ambulance deployment plans. Chapter 8, the final chapter presents the discussion and conclusion of the study.

Chapter 2

Literature Review

2.1 Introduction

Literature review is an assessment of a body of research that addresses the research questions. The literature review helps in identifying what is known about the specific study area and ultimately will help in the identification of issues that the body of research does not address. This helps to create a case for why further research has to be conducted. The objective of the review of literature is to describe, summarise and clarify literature in order to build a conceptual framework for the research.

This section will review literature on time series analysis, forecasting, simulation and optimisation and how these can be integrated in order to help improve the levels of preparedness for public ambulance emergency response through an informed decision-making process. In principle it will assess the current state of research on the topic, identify key questions that need further research and the determination of methodologies used in similar past studies. The literature review also provides some background knowledge of the mathematical concepts that are incorporated in the analysis.

2.2 Forecasting

Henderson and Mason (2005) highlighted that quantitative decision processes are becoming increasingly important in providing public accountability for the resource decisions that have to be made and that any solution to such problems require careful balancing of political, economic and medical objectives. There is need therefore, to strike a balance between numbers of ambulances available and skilled personnel manning these in line with the financial resources available through smart and robust planning.

Eldabi and Young (2007) argued that the provision of efficient and fair service to citizens or the population is a key goal of any emergency medical service and that this can be achieved by exploiting the use and integration of operations research, quantitative techniques and simulation techniques to provide decision making tools to emergency service management. They further argued that as populations are dynamic and evolving with regards to the demographics and their expectations, there is a corresponding increase in demand for better service to more people in a uniform manner.

To achieve this, future predictions have to be made using historical time series data. According to Tealab (2018), time series is a general problem of great practical interest in many disciplines and it allows you to discover, with some margin of error, the future values of a series from its past values.

2.3 The Nature of Time Series Data

A time series is a sequence of time-ordered data values that are measurements of some physical, biological, economical, or environmental phenomena of interest. Time series is defined by Shumway and Stoffer (2000) as a collection of random variables indexed according to the order they are obtained in time. A time series can also be defined as a sequence of numerical observations naturally ordered in time (Wilson et al., 2015).

Here x_t denotes the value taken by the time series at time $t = 0, 1, 2, ...$ A common time series variable used widely in time series analysis is the white noise. It is a collection of uncorrelated random variables w_t, with mean zero and a finite variance σ_w^2 (Shumway and Stoffer, 2000). Common components in time series data include trends, cyclical, irregular and seasonal variations. When analysing such data for forecasting purposes, two questions of paramount importance are that: 1) does the data exhibit a discernible pattern? and 2) can this pattern be exploited to make a meaningful forecast? The approach by which a researcher tries to answer such mathematical and statistical questions posed by these time correlations is referred to as time series analysis.

According to Chatfield and Weigend (1994) time series analysis has three goals: forecasting, modelling and characterisation. The aim of forecasting (predicting) is to accurately predict the short-term evolution of the system; the goal of modelling is to find the description that accurately capture features of the long-term behaviour of the system whilst characterisation is an attempt with little or no prior knowledge to determine important properties such as randomness of a system. The research will primarily focus on all three goals of time series analysis.

Time series analysis can be broadly classified into univariate time series and multivariate time series analysis. In univariate time-series data, the statistical relationship is unidirectional in that the forecast variable is influenced by its lags or the lags of other predictor variables (or features). A good example of univariate time series data is given by equation 2.1:

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t. \tag{2.1}$$

In multivariate time series data, the forecast variable is influenced by a co-movement of another time-series data. A good example of multivariate time series data is given by equation 2.2:

$$y_t = \mu + \phi_i y_{t-1} + \psi_j x_{t-1} + \epsilon_t, \tag{2.2}$$

where y_t and x_t are stationary time series and ϵ_t is the white noise. The mean μ and model parameters ϕ_i and ψ_j are also known model constants.

2.4 Time Series Forecasting Models

The demand for ambulance services in a zone, or the probability that an ambulance will be needed in a zone can be estimated in different ways. Common methods for forecasting demand include causal methods, time series models and judgemental methods. In causal methods, the forecasted demand is expressed as a function of independent variables such as for example, population and employment in the zone. Kamenetzky et al. (1982) described the development of a causal model for estimating the demand for pre-hospital care in Southwestern Pennsylvania.

Time series models are based on historical data, which is analysed to find out how the demand varies with time. However, the method to be used also depends on a variety of characteristics of the time series. It might depend on whether the time series data is univariate or multivariate time series data. Seasonal variations will be evident for ambulance health care as the demand in a zone might vary depending on the time of the day, which day of the week it is, or which month it is. In the study by Repede and Bernardo (1994), the demand inter-arrival rate for a district was found to depend on the day of the week and the hour of the day.

Judgemental methods rely on qualified guesses of the present and future events. The knowledge that a political rally, football tournament, marathon race, a big cultural or business exhibition event will take place in a zone is likely to increase the expected demand for ambulances. A combination of the three types of methods may be required in order to obtain a balanced estimation of the ambulance health care demand in a zone.

2.5 Conventional Time Series Forecasting Techniques

Linear time series forecasting techniques such as integrated autoregressive moving average models, moving averages, exponential smoothing have been used for time series analysis for decades in different areas of science, engineering, industry and commerce. Other various non-linear techniques have been proposed and applied to real-world problems, but like the linear techniques, they have a huge drawback as they require specific model assumptions (Alpaslan et al., 2012). Such non-linear models include Vector Autoregressive Models (VAR), Autoregressive Conditional Heteroskedastic (ARCH), Generalised Autoregressive Conditional Heteroskedastic (GARCH), and Threshold Autoregressive (TAR) models.

Real-world problems in the modern-day are characterised by non-linear patterns and behaviours which demand the use of non-linear techniques. There has been a shift of focus with the growing use and application of artificial neural networks because of their flexibility and remarkable features (Aras and Kocakoç, 2016). A neural network is defined by Hastie et al. (2009) as a non-linear statistical model which could be a two-stage regression or classification model, typically represented by a network diagram and good at modelling any complex function where the relationship is unknown.

2.6 Nonlinear Time Series Models

2.6.1 Artificial Neural Networks

Mitrea et al. (2009) defines an artificial neural network (ANN) as an information processing system that has been developed as generalizations of mathematical models of human neural biology. The use of artificial neural networks has grown over the years and has been applied to different fields in various forms due to their flexibility. According to Rather et al. (2015), these non-linear models overcome the limitation of linear models as they are able to capture the non-linear pattern of data, thus improving their prediction performance.

Over the years there has been a substantial use and comparison between neural networks and traditional time series forecasting techniques in different areas of applications such as health (Süt and Şenocak, 2007), geophysics (Goutorbe et al., 2006), geomechanics (Sadiq et al., 2003), stock markets (Herliansyah et al., 2017), chemical engineering (Noor et al., 2010; Prasad Y and Bhagwat, 2002), electrical engineering (Karami, 2010), global logistics (Bowersox et al., 2003), construction engineering (Zichun et al., 2012), financial business support (Tsai, 2008) and in insurance (Sha, 2008). Mitrea et al. (2009), made a comparison between neural networks and traditional forecasting methods in inventory management. The results showed that forecasting with neural networks offers better performance in comparison with traditional methods. Aish et al. (2015)used artificial neural networks to predict the efficiency of a reverse osmosis desalination plant in the Gaza Strip. A statistical model was compared with the developed models and it was found that ANN provided better predictions than conventional methods. However, a few researchers such as Raeesi et al. (2014) and Adebiyi et al. (2014) have made an attempt to apply neural networks on univariate data and make comparisons with the traditional methods. Moreover, such comparisons have not been applied to a case of demand for public emergency ambulance services. Raeesi et al. (2014) used feed-forward neural networks for short-term prediction using univariate data of the first

300 days of the year and values of the 301st days were predicted. The data was transformed into lags that would help to meet the multivariate conditions required by the neural network architect. Herliansyah et al. (2017) in their research study on the application of feed-forward neural networks (FFNN) for forecasting the Indonesia Exchange Composite Index concluded that with fewer data observations (54 training data points, 13 model validation data points), FFNN was still able to produce accurate predictions, react to sensitive changes and model the trend series well.

Despite these studies indicating that neural networks have superiority over traditional techniques, there are other studies reporting that traditional techniques outperform neural networks (Aras and Kocakoç, 2016). According to Anderson (1995) in a standard picture of the neuron presented in Fig. 2.1, dendrites receive inputs from other cells, the soma (cell body) and dendrites process and integrate the inputs, and information is transmitted along the axon to the synapses, whose outputs provide input to other neurons or effect organs.

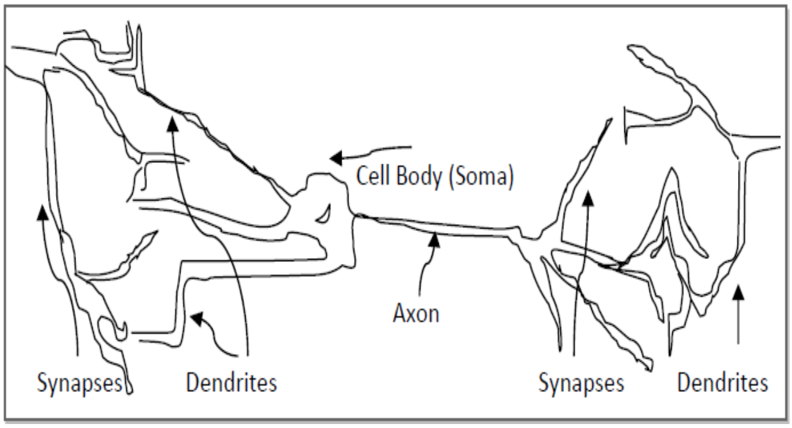


Figure 2.1: A Dendritic Tree of Several Neurons (Anderson J.A., 1995).

The structure of the artificial neural network (ANN) model is regarded as simpler than the immensely complex structure of the human neural system, in practice they can produce interesting and significant results in various real-life applications such as classification, image processing and forecasting (Alpaslan et al., 2012). In classification, the neural networks are able to carry pattern recognition, feature extraction, image extraction. Neural networks are also able to reduce white noise by recognising patterns in the inputs and producing noiseless outputs and are also capable to extrapolate historical data. Neural networks have the ability to learn and figure out how to perform their function on their own and are able to determine their function based only on sample units. They are also able to produce reasonable outputs for inputs it has not been taught how to deal with before.

2.6.2 The Generic Neural Network Neuron

The generic neural network unit has two parts. The first part inputs from the synapses are added together, with individual synaptic contributions combining independently and adding algebraically, giving a resultant activity level (Anderson, 1995). The second part is used to generate the final output activity of the neuron by using the activity as the input to a nonlinear function relating activity level to output a value. The generic neural network neuron is presented in Fig. 2.2.

The single neuron has its response a non-linear function of the weighted sum of all input signals. A constant bias signal must be applied to all neurons in a feed-forward neural network. All weights that are used to weight the individual input signals are grouped in a weight vector W (Herliansyah et al., 2017). $f(*)$ is the non-linear sigmoid function (activation function) that is applied to the weighted sum of all input signals. The response of the single neuron is then given by equation 2.3:

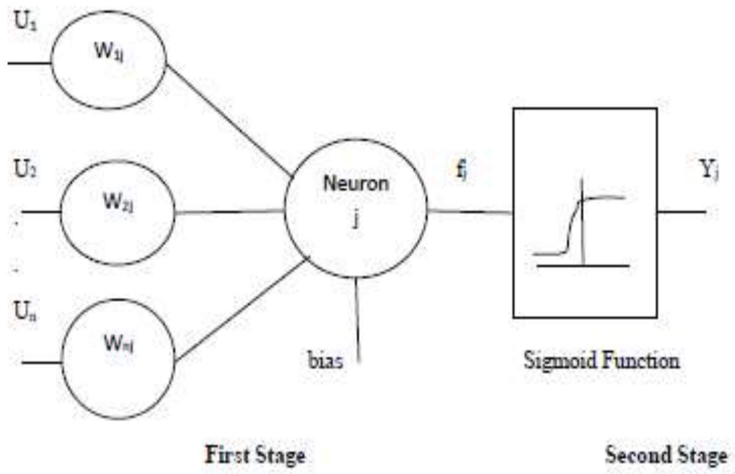


Figure 2.2: The Generic Neural Network Neuron.

$$Output = f(u_1 w_{1j} + u_2 w_{2j} + u_3 w_{3j} + ... + u_n w_{nj} + bias). \tag{2.3}$$

This can be generalised by the equation 2.4 given by:

$$Y_j = f(\sum_{n=1}^{N} W_n U_n + bias) = f(WU) + bias, \tag{2.4}$$

where W_n is the weight matrix , U_n is the input value matrix and bias is the error of approximation.

A widely used NN structure among modellers is the Feed Forward Neural Network (FFNN), also known as the multilayer perceptron (MLP) (Herliansyah et al., 2017). According to Mitrea et al. (2009) many varieties of neural network techniques including Multilayer Feed-forward NN, Recurrent NN, Time delay NN and Nonlinear Autoregressive Exogenous NN have been proposed, investigated, and successfully applied to time series prediction and causal prediction. The different architectures are shown in Fig. 2.3.

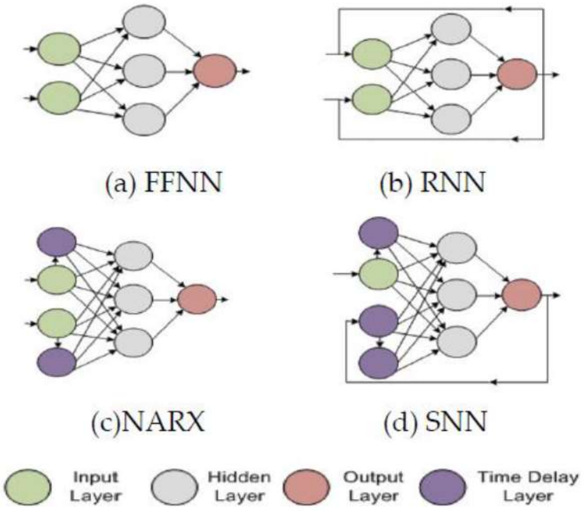


Figure 2.3: (a) Multilayer Feed Forward NN, (b) Recurrent NN, (c) Time Delay NN and (d) Nonlinear Autoregressive Exogenous NN.

2.6.3 Feed Forward Neural Networks

A feed-forward neural network consists of three basic layers, the input layer, hidden layer(s) and the output(s) as shown in Fig. 2.4. The network is called feed-forward because information flows only from the input to the output and there are no recurrent or backward connections. Each layer consists of neurons and there is no connection between neurons that are in the same layer. The activation functions are used for neurons in the hidden and output layers only; the output layer might contain one or more outputs depending on the phenomenon under investigation. However, literature has shown that the majority of studies use a single output for forecasting purposes (Alpaslan et al., 2012). A three-layer FFNN structure is shown in Fig. 2.4.

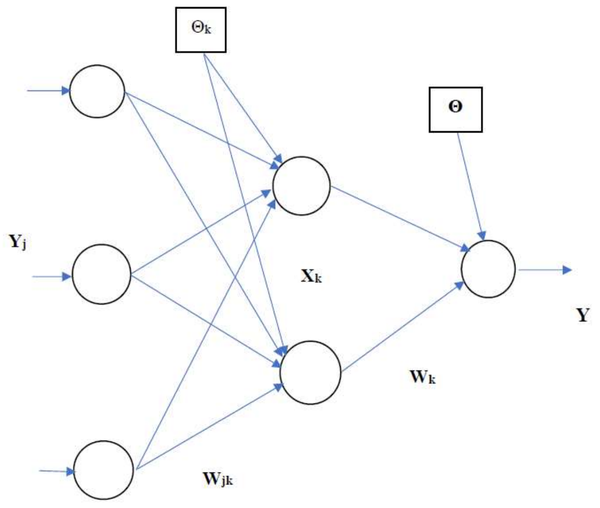


Figure 2.4: Feed-forward Neural Network For Time Series Forecasting.

The input vector is represented by Y_j denoted by $Y_j = \{y_1, y_2, y_3, ..., y_n\}$; $W_{jk}(j = 1, 2, 3, ..., n; k = 1, 2, 3, ..., m)$ is the connection weight vector of the j nodes of the input layer to the k nodes of the hidden layer; $X_k(k = 1, 2, 3, ..., m)$ is the vector of k neurons in the hidden layer; $W_k(k = 1, 2, 3, ..., m)$ is the connection weights of the k nodes of the hidden layer to the output layer; and Y is the unit output vector for the neural network with one output neuron. $\Theta_k(k = 1, 2, 3, ..., k)$ is the bias value of the hidden layer nodes and Θ is the bias value of the output layer. The output of the hidden layer is determined by the formula:

$$X_k = f(\Sigma_{j=1}^{n} W_{jk} Y_j + \Theta_k). \tag{2.5}$$

The output of the output layer is determined by the formula:

$$Y = f(\Sigma_{k=1}^{m} W_k X_k + \Theta), \tag{2.6}$$

where f is the activation function for the hidden and output layers.

2.6.4 Generation of Neural Network Weights

The weights in a neural network are the most important factor in determining its function. Training is the process of presenting the network with some sample data and modifying the weights to better approximate the desired function. The two main types of training are supervised training and unsupervised training. In supervised training, the neural network is supplied with inputs and the desired outputs. The response of the network to the inputs is measured and the weights are modified to reduce the difference between the actual and the desired outputs. A schematic representation of supervised neural network training is summarised in Fig. 2.5.

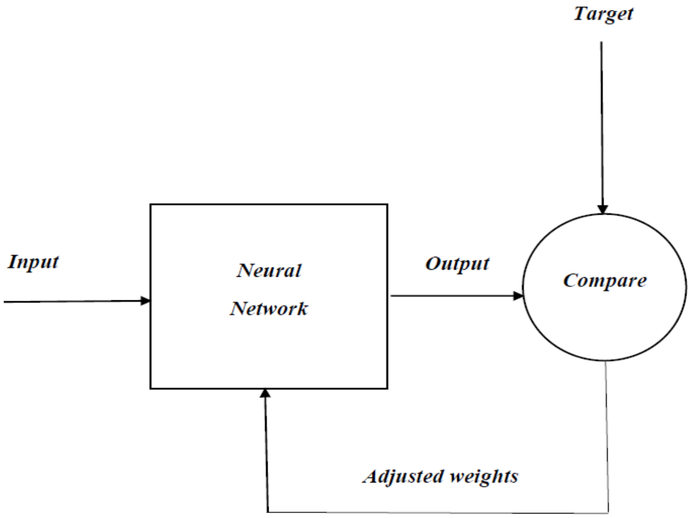


Figure 2.5: A Schematic Diagram for a Neural Network Supervised Training.

In unsupervised training, the neural network adjusts its own weights so that similar inputs generate similar outputs. The network identifies the differences in the inputs and patterns without any external assistance. According to Ciaburro and Venkateswaran (2017), an epoch is a single iteration through the process of providing the network with input and updating the network's weights. Several epochs are required to train the neural network efficiently. The process of adjusting the weights of a neural network uses an algorithm called backpropagation in order to minimise the error of estimating the model during the training of the neural network.

2.6.5 Activation Functions

Activation functions are applied to the weighted sum of the inputs of a neuron to produce the output. Widely used activation functions have a very large number of shapes and the activation function family include the logistic function, hyperbolic tangent, step function, linear function, rectified linear function, and arctangent functions Key desirable properties of an activation function are that; it must be differentiable, simple and fast in processing and should not be zero centred (Ciaburro and Venkateswaran, 2017).

The exact nature of the function has little effect on the abilities of the neural networks. Weights in a neural network are the most important factor in determining its function. The activation functions are used for neurons in the hidden and output layers. Activation functions provide the non-linear mapping, hence activation functions which that are nonlinear are desired in the hidden layer. Both linear and non-linear functions can be adopted for the output neuron(s).

The Sigmoid Function

A sigmoid function is smooth, continuous and monotonically increasing implying that the derivative of the sigmoid function is always positive. It has a bounded range but never reaches a maximum or minimum. The most commonly used sigmoid function is the logistic function as represented in Fig. 2.6 and is represented by a generalised equation 2.7:

$$f(x) = \frac{1}{1 + e^{-x}}. \tag{2.7}$$

The calculation of derivatives is important for neural networks and the logistic function has a generalised derivative given by equation 2.8:

$$f^{'}(x) = f(x)(1 - f(x)). \tag{2.8}$$

When a logistic activation function is used, values will be transformed into the interval $(0, 1)$.

The Hyperbolic Tangent Function

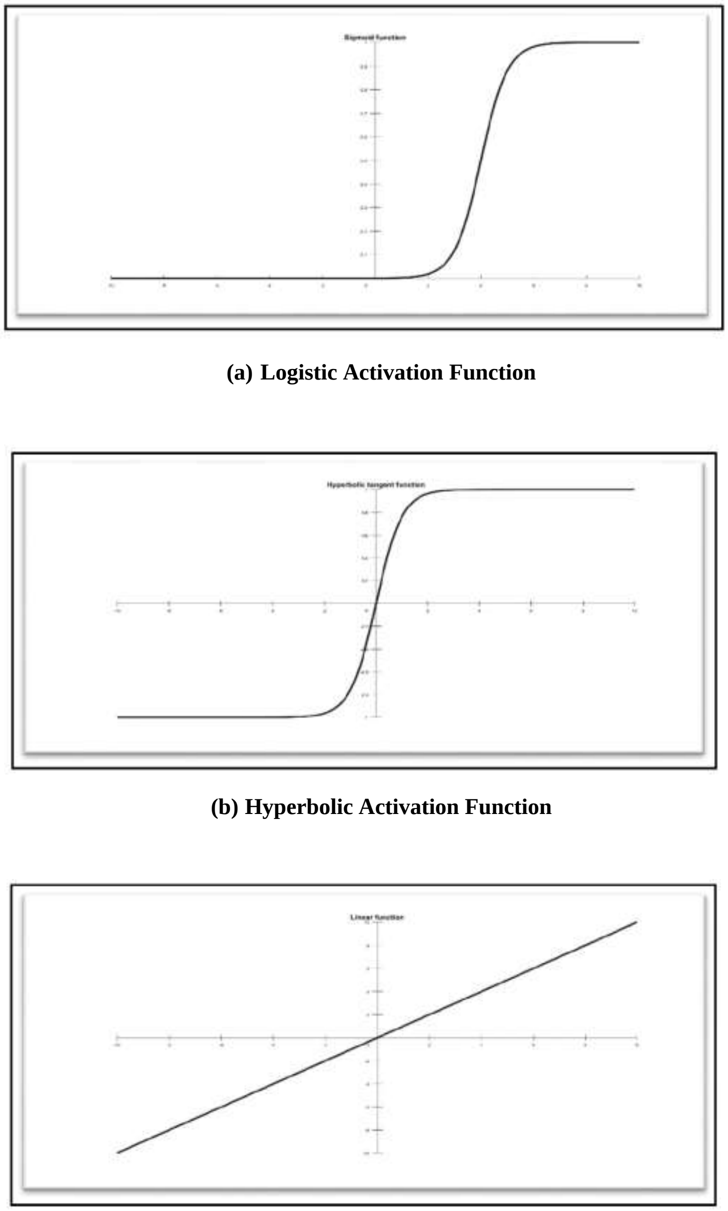
The hyperbolic tangent function (Fig. 2.6) is also the most commonly used activation scaled sigmoid function (Goutorbe et al., 2006) that transforms values into the interval $(-1, 1)$ by using the formula given by equation 2.9:

$$f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \tag{2.9}$$

The Linear Function

The linear function (Fig. 2.6) is mostly used in literature in the output layer where the desired outputs are real values from negative infinity to positive infinity. It is in the form of a polynomial of one degree which generates a straight line without any curves. The generalised formula is given by equation 2.10:

$$y = f(x) = x. \tag{2.10}$$



(a) Logistic Activation Function

(b) Hyperbolic Activation Function

(c) Linear Activation Function

Figure 2.6: A Collage of Sigmoid Activation Functions.

2.6.6 Neural Network Model Building Process

Aish et al. (2015) stated that there are a number of issues that need to be addressed in setting-up and training a multilayer neural network and these include: selecting how many hidden layers to use in the network, deciding on how many neurons to use in each hidden layer, finding a comprehensively optimal solution that avoids local minima, converging to an optimal solution in a reasonable period of time and validating the neural network for over fitting if it might have occurred during the training process. Seven (7) steps can be adopted whilst addressing the above-mentioned issues or concerns in the neural network model building and forecasting process. The seven key steps are summarised in Fig. 2.7.

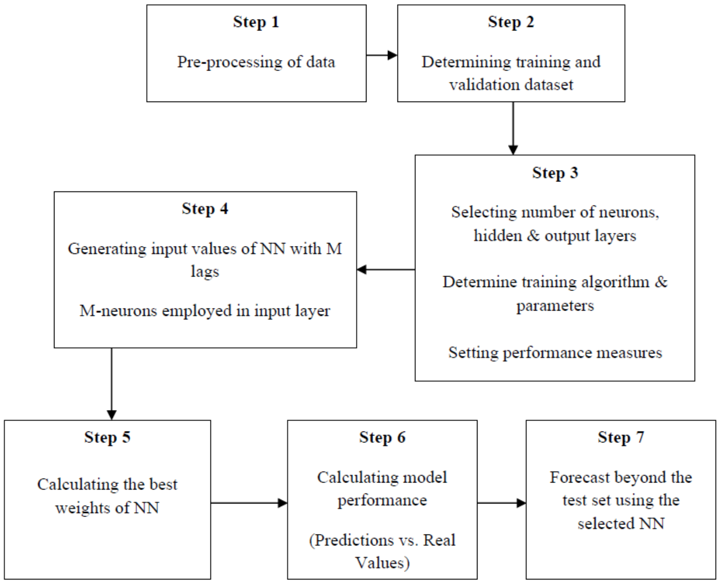

Step 1
Pre-processing of data
Step 2
Determining training and validation dataset
Step 3
Selecting number of neurons, hidden & output layers
Determine training algorithm & parameters
Setting performance measures
Step 4
Generating input values of NN with M lags
M-neurons employed in input layer
Step 5
Calculating the best weights of NN
Step 6
Calculating model performance
(Predictions vs. Real Values)
Step 7
Forecast beyond the test set using the selected NN

Figure 2.7: The Seven Step Neural Network Model Building and Forecasting Process.

Step 1 involves the pre-processing of data and choosing the activation functions for both the hidden and output layers. According to Mitrea et al. (2009) pre-processing represents data coding, enrichment and cleaning which involves accounting for noise and dealing with missing information. Pre-processing also involves the normalisation of data (scaling down) to an interval $[0, 1]$ in order to prevent saturation of hidden nodes before feeding into the neural network (Kitapcı et al., 2014). Normalisation also ensures that none of the variables over dominates the other variables and it assists in achieving quick convergence during training of the neural network. In data pre-processing, there are three basic normalisation techniques available: Min-Max normalisation, Z-score normalisation, and sigmoid normalisation (Kheirkhah et al., 2013). The min-max normalisation is the most commonly used technique and the formula can be generalised by equation 2.11:

$$y_i^{'} = \frac{y_i - Min(y_i)}{Max(y_i) - Min(y_i)}, \tag{2.11}$$

where $y_i^{'}$ is the normalised value at the time i, y_i, $Max(y_i)$ and $Min(y_i)$ represents the observed values, maximum and minimum values of observations respectively, of variable y_i over the range of data (Alpaslan et al., 2012).

Step 2 involves the determination of the lengths of training and the testing sets. The first set, called the 'model-building' set or training set, is used to develop the network model. The second data set, called the testing or prediction set is used to evaluate the forecasting ability of the selected model. Higher percentages are usually assigned to the training set and lower percentages to the testing set. The length of the testing set is implemented in various neural network architectures in literature at 10%, 15%, 20% and 30%. In a study by Alpaslan et al. (2012), they concluded that the smaller the testing set length, the more accurate the obtained forecasts are.

Step 3 involves the modelling aspects of the neural network. The architecture of the neural network is determined by the number of hidden and output layers, the number of neurons in all of the layers, training algorithm parameters and the performance measure. There is no general rule for the selection of the appropriate number of hidden layers and the most popular approach in finding the optimal number of the hidden layer is by trial and error (Güler and Übeyli, 2005).

It is important to note that neural networks without hidden units are equivalent to linear statistical forecasting methods (Cheng and Titterington, 1994). Hidden units perform the mapping between the input and output variables as well as to provide the non-linearity feature to neural networks, in addition to find out patterns in the dataset (Aras and Kocakoç, 2016). It is expected that the number of the parameters and the degree of non-linearity of the neural network increases as the number of hidden units (neurons) increases. A balance, therefore, has to be made as the use of an excessive number of hidden units can lead to overfitting of training data by the neural network. In general, a small number of hidden units are recommended to avoid the neural network from overfitting the training dataset. Wanas et al. (1998) after an empirical study, proposed that the number of hidden units is given by log (N) where N is the number of training samples. These can be increased during training of the neural network however; a general rule is that there must not exceed two-thirds of the input neurons.

Performance measures are used for comparing the performances of selected neural network models on validation data. The mean square error (MSE) and the mean absolute error (MAE) are the two classical accuracy measurements used in literature though other variants such as root mean squared error (RMSE) and mean absolute percentage error (MAPE) were implemented. Despite both the MSE and the MAE being measures of accuracy and the degree of spread, the former has been the most popular of the two due to its theoretical relevance in statistical modelling (Aras and Kocakoç, 2016). The MAE gives equal weight to all errors while the MSE assigns more weight to large errors. The MSE is more sensitive to outliers than the MAE. The formulations of the MSE, RMSE and MAE are given by:

$$MSE = \frac{1}{N}\Sigma_{t=1}^{N}(y_t - \hat{y_t})^2, \tag{2.12}$$

$$RMSE = \sqrt{\frac{1}{N}\Sigma_{t=1}^{N}(y_t - \hat{y_t})^2}, \tag{2.13}$$

$$MAE = \frac{1}{N}\Sigma_{t=1}^{N}|(y_t - \hat{y_t})|, \tag{2.14}$$

where y_t is the actual observation for the period t, $\hat{y_t}$ is the forecast for the same period, and N is the length of the test set. The MSE and the MAE are both measures of accuracy and the degree of spread of data points (Aras and Kocakoç, 2016). The MAE is a measurement of how close forecasts are to the actual data points; the average of the absolute errors (Herliansyah et al., 2017). Values predicted from the training sample and values in the test set were utilised to calculate the performance measures. Forecasts for the year 2018 were calculated by using the best weight values obtained in the selected model.

Step 4 involves the generation of input values of the neural network. The input values are lagged time series. If for a time series Y_t has m lagged time series $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-m}$, then m neurons are employed in the input layer. The selection of the number of inputs used in the model is based on trial and error (Belayneh et al., 2014).

Step 5 involves the calculation of the best weight values of the neural network through the use of a training algorithm. Training a network involves the adjusting of weights of the network using different learning algorithms. Süt and Şenocak (2007) used several training algorithms in their research that are available for use and these included the backpropagation (BP), quick propagation (QP), delta-bar-delta (DBD), the extended delta-bar-delta (EDBD) and Levenberg-Marquardt algorithms.

Backpropagation, also called backward error propagation or backprop, is the most popular and widely used network learning algorithm (Rumelhart et al., 1985; Anderson, 1995; Duda et al., 2012). There are used in a multilayer architecture, where in addition to the input and output layers, there are also one or more hidden layers that are neither input nor output. Backpropagation is a rule that generalises the gradient descent method as a way of changing the weights in the hidden layer of a FFNN. It gives the change $\Delta w_{jk}(i)$ in the weight of the connection between neurons j and k at iteration i as:

$$\Delta w_{jk}(i) = -\alpha \frac{\partial E}{\partial w_{jk}(i)} + \mu \Delta w_{jk}(i-1), \tag{2.15}$$

where α is called the learning coefficient, $\Delta w_{jk}(i-1)$ the weight change in the immediately preceding iteration and μ the momentum coefficient (Güler and Übeyli, 2005).

The learning coefficient ensures a maximum decrease of the error function thus increasing the convergence speed. If it is too small, convergence will be extremely slow and if too large, the error function will not converge. The momentum coefficient tends to aid convergence as it works as a low-pass filter by reducing rapid fluctuations. It applies smoothed averaging to the change in weights whilst also avoiding local minima (Goutorbe et al., 2006).

Step 6 involves the calculation of the performance measure. A performance measure is based on the difference between predictions from training (in sample) and testing set (real values) values. The MSE and the MAE are the commonly used performance measures and the model with the least value is selected for forecasting.

Step 7 is the forecasting stage where forecasts beyond the validation set are calculated using the best weight values obtained in step 5. According to Günay (2016), the assumption here is that if the model was able to predict the demand in this time interval with a good accuracy, it could be used for forecasting the future demand with high reliability.

2.7 Linear Time Series Analysis Methods

There are two main methods of forecasting linear time series data namely: The regression method and the Box-Jenkins method. The two methods differ in that a regression model is a structural or causal forecasting method that requires the forecaster to know in advance at least some of the determinants of the response variable whereas the Box-Jenkins technique can be used when these determinants of the variable are not readily available. Other linear models that can be used for prediction include the exponential smoothing models, generalised autoregressive conditional heteroscedasticity and stochastic volatility models, which are predominantly used in predicting stock returns (Rather et al., 2015). It is the Box-Jenkins methodology that is going to be used for forecasting purposes in this research. Jere et al. (2017) applied the same methodology in a study to forecast the second-hand car importation in Zambia and the methodology was found to be superior as compared to exponential smoothing models for univariate time series analysis.

2.7.1 Theoretical Properties of the ARIMA Models

In Box-Jenkins methodology, we start with the observed time series itself (with no explanatory variables) and examine its characteristics in order to get an idea of what black box (simulation model) we might use to transform the series into white noise (residuals). We begin by trying the most likely of many black boxes and, if we get white noise, we assume that this is the "correct" model to use in generating forecasts of the series. White noise has two characteristics: there is no relationship between consecutively observed values, and the previous values do not help in predicting future values.

2.7.2 Stationarity Assumption

When applying ARIMA models to time series data, it is assumed that the data is stationary. The stationarity assumption requires that the mean, variance and autocorrelation of the time series data does not change over time. Non-stationarity is common when components of time-series data such as trend, seasonality, cyclicity and irregularity are not accounted for in the model building process. When time-series data is not stationary, differencing both the ordinary and seasonal components is performed to ensure that stationarity is achieved.

2.7.3 Basic ARIMA Models

When choosing the best time series model, there are three basic types of models to be examined: Many variations within each of these three types can be developed depending on how the time series data is structured. The three basic models are: moving average (MA(q)) models, autoregressive (AR(p)) models and the mixed autoregressive/moving average models known as autoregressive integrated moving averages (ARIMA (p, d, q)) (Shumway and Stoffer, 2000).

Moving Average MA(q) Model

A moving average MA(q) model of order q predicts Y_t as a function of the past forecast errors in predicting Y_t (Shumway and Stoffer, 2000). Consider w_t to be the white noise series; a moving average would then take the form presented in equation 2.16:

$$Y_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \ldots + \theta_q w_{t-q}, \tag{2.16}$$

where:

- w_t is the value at time t of white noise.

- Y_t is the observed value at time t.
- $\theta_1, \theta_2, ...\theta_q$ are the coefficients (or ”weights”).
- $w_{t-1}, ..., w_{t-q}$ are the previous values of the white noise series.

If the linear combination of values is then applied with the backshift operator given by;

$$B^k w_t = w_{t-k}, \tag{2.17}$$

where

$$Bw_t = w_{t-1}, B^2 w_t = w_{t-2}, B^3 w_t = w_{t-3}, \ldots, B^k w_t = w_{t-k}. \tag{2.18}$$

The MA(q) model equation becomes:

$$Y_t = (1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q) w_t. \tag{2.19}$$

If $\theta(B)$ is the moving average operator given by;

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q. \tag{2.20}$$

Then the MA(q) model can be written in the equivalent form

$$Y_t = \theta(B) w_t. \tag{2.21}$$

Autoregressive AR(p) Models

The equation for the autoregressive model AR(p) of order p is similar to the MA(q) model except that the dependent variable Y_t depends on its own previous values rather than white noise series of residuals (Shumway and Stoffer, 2000). It takes the form presented in equation 2.22:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + w_t, \tag{2.22}$$

where;

- w_t is the value at time t of white noise.
- Y_t is the observed value at time t.
- $\phi_1, \phi_2, ...\phi_p$ are the coefficients (or "weights").
- $Y_{t-1}, ..., Y_{t-q}$ are the previous values of the time series at time t.

Since the regression of the AR(p) model is of the same series, the model adopts the name autoregression. The linear combinations can also be written using the backshift operator such that;

$$Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - ... - \phi_p Y_{t-p} = w_t, \tag{2.23}$$

$$(1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p) Y_t = w_t. \tag{2.24}$$

If $\phi(B)$ the autoregressive operator is introduced and is given by;

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p. \tag{2.25}$$

The AR(p) model can be written in the equivalent form

$$\phi(B) Y_t = w_t. \tag{2.26}$$

Autoregressive Moving Average ARMA (p, q) Models

The combination of an AR and an MA model is called ARMA, which stands for autoregressive moving average model (Shumway and Stoffer, 2000). It can be represented as follows:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + ... + \theta_q w_{t-q}. \tag{2.27}$$

The equation above defines a mixed autoregressive moving average model of order p, q; and is written as ARMA (p, q). When differencing is used to make a time series stationary, it is common to refer to the resulting model as an ARIMA (p, d, q) type model. The d inside the parenthesis refers to the degree of difference. An ARIMA model is then categorically referred to as an autoregressive integrated moving average model. The ARIMA (p, d, q) model can be concisely represented by an equation given by;

$$\phi(B)(1 - B)^d Y_t = \theta(B) w_t, \tag{2.28}$$

or equivalently as equation 2.29.

$$\phi(B) \bigtriangledown^d Y_t = \theta(B) w_t. \tag{2.29}$$

The higher-order differencing is represented by equation 2.30;

$$(1 - B)^d Y_t = \bigtriangledown^d Y_t. \tag{2.30}$$

Multiplicative Seasonal Autoregressive Integrated Moving Average (SARIMA) model

Components of time series data namely; seasonality, trend, cyclicity and irregularity, are key to model development. Seasonality refers to a consistent shape in the series that recurs with some periodic regularity. It can also be referred to as regular up and down fluctuation or short-term variations due to seasonal factors. Monthly ambulance calls data is likely to have a strong yearly component occurring at lags that are multiples of s = 12, because of strong relations of activities to the calendar year. This might lead to the adoption of autoregressive and moving average polynomials that identify with seasonal lags called multiplicative seasonal autoregressive integrated moving average models (Shumway and Stoffer, 2000) given by ARIMA (p, q, d) x (P, D, Q)s. The model can be written in the general form:

$$\Phi_p(B^s)\phi(B)\nabla_s^D\nabla^d Y_t = \Theta_Q(B^s)\theta(B)w_t, \tag{2.31}$$

where:

- w_t is the value at time t of white noise.
- Y_t is the observed value at time t.
- $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p$ is the ordinary autoregressive component of order p.
- $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q$ is the ordinary moving average component of order q.
- $\nabla^d Y_t = (1 - B)^d Y_t$ is the ordinary differenced component of order d.
- $\nabla_s^D = (1 - B^s)^D$ is the seasonal difference of order D at lag s.

- $\Phi_p(B^s)$ and $\Theta_Q(B^s)$ are the seasonal autoregressive and moving average difference of orders P and Q at lag s.

2.7.4 The Box-Jenkins Identification Process

The four basic steps of the Box-Jenkins identification process are outlined in Fig. 2.8. These include: Identifying the tentative model, estimation of model parameters, diagnostic check for the adequacy of the model and forecasting with the chosen model.

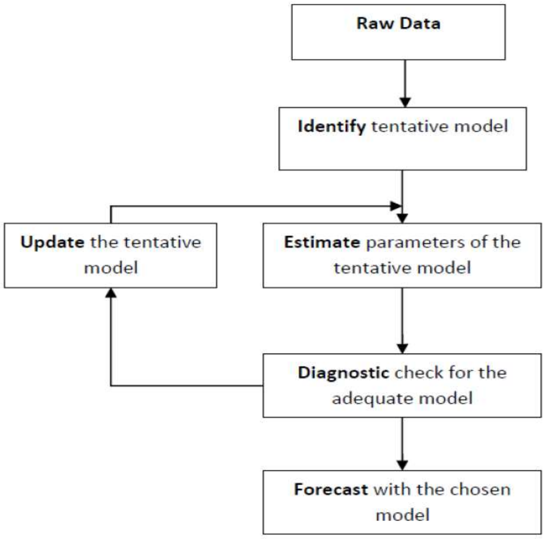

Raw Data
Identify tentative model
Update the tentative model
Estimate parameters of the tentative model
Diagnostic check for the adequate model
Forecast with the chosen model

Figure 2.8: The Box-Jenkins Identification Process.

2.7.5 Identifying the Tentative Model

As a first step, the raw series is examined to identify which of the many available models can be selected as the best representation of the series. The time series plot; the autocorrelation and partial autocorrelation functions of the time series in question can be used to accomplish this. If the raw series is not stationary, it will be necessary to modify the original series by transforming it. The main methods used to transform non-stationary time series data are differencing and Box-Cox transformations.

Differencing

Differencing is a special type of filtering, which is particularly useful for removing a trend by simply differencing a given series until it becomes stationary. This method is an integral part of the procedures advocated by Box and Jenkins, (1970) as highlighted by Anderson (1977). For non-seasonal data, first differencing is usually sufficient to attain apparent stationarity, so that the new series does not have the linear trend. Depending on the nature of time-series data, one can use second-order differencing in order to eliminate a quadratic trend so that a single long cycle not eliminated by the de-trending procedure will be addressed.

Box-Cox Transformation

The algorithmic and square root transformation are special cases of the class of transformations called the Box-Cox transformation and are applied to make a time series stationary. Given an observed time series x_t and a transformation parameter λ, the transformed series is given by equation 2.32;

$$g(x) = \begin{cases} \frac{x^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log(x), & \lambda = 0 \end{cases} \quad \text{..}(2.32)$$

This is effectively just a power transformation when $\lambda \neq 0$. If $\lambda = 0$, it becomes a logarithmic transformation. A Box-Cox power transformation on the dependent variable is a useful method to alleviate heteroscedasticity when the distribution of the dependent variable is not known. It is important to note that for the transformation to be applicable, the series has to be strictly positive (Proietti and Lütkepohl, 2013).

2.7.6 Estimation of Model Parameters

The second step in the process after a tentative model has been identified is the estimation of the parameters of the model. The general rules to be used in identifying a model if the time series data is found to be stationary at this stage can be summarised as:

- If the autocorrelation function abruptly stops at some point, say after q spikes, then the appropriate model is an MA (q) type.
- If the partial autocorrelation function abruptly stops at some point, say after p spikes, then the appropriate model is an AR (p) type.
- If neither function falls off abruptly, but does decline toward zero in some fashion, the appropriate model is an ARMA (p, q).

Autocorrelation Function (ACF)

A correlation of a variable with itself at different times is known as autocorrelation. Autocorrelation measures the linear predictability of the series at time t, say x_t using only the values x_s. It can be shown that $-1 \leq \rho(s,t) \geq 1$ using the Cauchy-Schwarz inequality (Shumway and Stoffer, 2017). The ACF which is a measure of association has a characteristic equation given by;

$$\rho_x(s,t) = \frac{\gamma_x(s,t)}{\sqrt{\gamma_x(s,s)\gamma_x(t,t)}}, \tag{2.33}$$

where:

- $\rho_x(s,t)$ is the autocorrelation function for a series x, at time s and t.

- $\gamma_x(s,t)$ is the autocovariance function for a series x, at time s and t.

- $\gamma_x(s,s)$ is the variance for a series x, at time s.

- $\gamma_x(t,t)$ is the variance for a series x, at time t.

The autocovariance function is defined by the second-moment product as:

$$\gamma_x(s,t) = cov(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)] \forall s, t. \tag{2.34}$$

The autocovariance of the time series at time s of the time series is given by:

$$\gamma_x(s,s) = cov(x_s, x_s) = E[(x_s - \mu_s)^2] = \sigma_x^2. \tag{2.35}$$

If a time series model is second order stationary, (stationary in both mean and variance): $\mu_t = \mu$ and $\sigma_x^2 = \sigma^2$ for all t, then an autocovariance function can be expressed as a function only of the time lag k:

$$\gamma_k = E[(x_t - \mu)(x_{t+k} - \mu)]. \tag{2.36}$$

The autocorrelation function ACF can now be defined as

$$\rho_k = \frac{\gamma_k}{\sigma^2}, \tag{2.37}$$

where $\sigma^2 = \gamma_0$.Thus, the autocorrelation is now only a function of lag k. In simple terms, ACF tells you how correlated points are with each other, based on how many time steps they are separated by.

Partial Autocorrelation Function (PACF)

The Partial Autocorrelation Function is a second function that expresses information useful in determining the order of an ARIMA model. This function is constructed by calculating the partial correlation between X_t and X_{t-4} after statistically removing the influence of X_{t-1}, X_{t-2}, and X_{t-3} from both X_t and X_{t-4}. The autoregressive order, p, is estimated as the lag of the last large partial autocorrelation. The values of PACF range from -1 to 1.

The Goodness of Fit Test

Shumway and Stoffer (2017) proposed Akaike's Information Criterion (AIC) as a good measure of goodness of fit. The general equation of the AIC is given by equation 2.38:

$$AIC = \ln\left(\hat{\sigma}_k^2\right) + \frac{2k}{n}. \tag{2.38}$$

The terms are defined as follows:

- k is the number of parameters in the model.
- σ_k^2 is the maximum likelihood estimator of variance.
- n is the sample size.

The idea behind the AIC is to minimise σ_k^2 together with a term proportional to the number of parameters in the model. The model yielding the minimum AIC specifies the best model. If two or more models are almost similar in their fit of the data; it is suggested that the simpler of the similar models be chosen for actual forecasting.

2.7.7 Diagnostic Check of Selected Model

The third step is to diagnose if the correct model has been chosen. When a model has been fitted to a time series, it is advisable to check that the model does provide an adequate description of the data. As with most statistical models, this is usually done by looking at the residuals, which are defined as: **residual = observation − fitted value**. The residuals should: be normally distributed, be independent of each other, and have a constant mean and variance.

Test for the Normality of Residuals

Test for normality can be achieved in two ways, by plotting a histogram or a dot plot of the residuals and construction of a normal probability curve. The plot of the histogram should resemble a sample from a normal distribution centred at zero. A normal probability plot will resemble a straight line otherwise the normality assumption is violated. In visualising the straight line, more emphasise should be on the central values of the plot than on the extremes (Montgomery and Runger, 2014).

Test of Independence of Residuals

Two methods for testing for the independence of residuals are to be implemented. The first method involves the plotting of residuals against the fitted values. Authors have argued that if the model is correct, the plot should be structure less (Mapuwei et al., 2016). Another important tool is to plot the ACF of the residuals and this should not show any significant terms even though we expect approximately $\frac{1}{20}$ to be above $\pm\frac{2}{\sqrt{n}}$.

Test of Constant Mean and Variance on the Residuals

Plotting the time series of the residuals of the chosen model is sufficient to test for the homogeneity of the mean and variance of the residuals. This is expected to show any changes in the mean and variance with time.

The Ljung-Box Pierce Statistic

The Ljung-Box Pierce statistic tests whether the residual autocorrelations as a set are significantly different from zero (test of independence) and that any random shocks are probably uncorrelated. If the residuals are significantly different from zero, the model should be reformulated (Shumway and Stoffer, 2000). A hypothesis can be developed as follows:

$H_0 : Q = 0$ (the residuals are not correlated).

$H_1 : Q \neq 0$ (the residuals are correlated).

The Chi-square statistic is calculated as:

$$Q = n(n+2)\sum_{h=1}^{H}\frac{\hat{\rho}_e^2(h)}{n-h}, \tag{2.39}$$

with $H - p - q$ degrees of freedom, where:

- Q is the Ljung-Box-Pierce test statistic (or Q statistic).
- n is the length of the time series.
- h is the first k correlations being checked.
- p is the number of AR terms.
- q is the number of MA terms.
- $\hat{\rho}_e$ is the sample correlation value at lag h.

Under the null hypothesis at of model adequacy, asymptotically ($n \to \infty$), $Q \sim \chi^2_{H-p-q}$. Thus we reject the null hypothesis at α if the value of Q exceeds the $(1-\alpha)$ quantile of the χ^2_{H-p-q} distribution (Shumway and Stoffer, 2000).

2.7.8 Cross-Validation of Forecasting Models

After the diagnostic testing and the actual forecasting are done, it is always imperative to validate the chosen model. Validation of the selected model in any model building process is the final step. Model validation involves the checking of the model against independent data. They are three basic methods of validating a forecasting model namely: use of holdout sample to check the model and its predictive ability, also known as data splitting, collection of new data to check the model and its forecasting ability and comparison of results with theoretical expectations earlier results.

Data Splitting Process

Data splitting is reasonably effective when the data set is large enough to split the data into two sets. The first set, called the 'model-building' set, is used to develop the model. The second data set, called the validation or prediction set is used to evaluate the forecasting ability of the selected model. If the quality of the forecast is not acceptable, another forecasting model is selected. The problem with this technique is that we are not using all the information available to us to fit the model

Collection of New Data

Neter et al. (1989) proposed that the best means of model validation is through the collection of new data and examining whether the model developed from earlier data is still applicable to the new data. If it is justifiable then one has the assurance about the applicability of the model to data beyond those on which it is based and or to forecast future data with a minimum of errors. Collection of new data sometimes is neither practical nor feasible due to constraints such as data availability and costs.

Use of Theoretical Expectations

This is also another alternative though it heavily relies on other studies done previously. Sometimes it can be handy especially when there is a time constraint depending on the nature of the research.

2.8 Simulation Modelling

Introduction

There has been a vast amount of academic literature developed with regard to emergency service stations (ESS). The locations of ESS includes components such as fire brigades, emergency service stations, hospitals, police and ambulances which are key components in addressing issues of delivering quality health care and a reliable response system. Strategic planning and location of these ESS is anticipated to reduce the fatalities and disabilities caused by events that require emergency services. Such events include accidents, pandemic diseases such as cholera and malaria, natural disasters, illness, and fires among many others.

El Sayed (2012), in a review paper highlighted that there is a growing need across the world for emergency medical services (EMS) to increase coordination in patient care and quality care at lower costs by continuously monitoring the system's overall performance and effectiveness of the different pre-hospital interventions. Here, an emergency medical service (EMS) can be generalised as a system that provides pre-hospital care to a specified population or citizens in need for emergency medical service. The Institute of Medicine (IOM) report of 2006, defined quality as the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current knowledge. The report further highlighted that the six dimensions of quality care are: care that is safe, effective, patient-centered, timely, efficient and equitable.

Equality

According to Aringhieri et al. (2017), equity is one of the most challenging concerns in the healthcare sector, especially in EMS systems as it evaluates the fairness of how resources specifically EMS vehicles are allocated to service the demand in different geographical areas or population groups. Equity is usually bench marked on the distance or response time travelled by EMS vehicles. However, equity is an aspect that can be embedded in other factors of EMS such as the workloads of the EMS vehicles and medical staff.

Inequalities that may exist in the ambulance services may be associated with: differences between economically poor and rich communities, differences in people who may not use ambulance services to the same extent as people from other geographical areas or population groups with similar health needs and by levels of difficulty in accessing services by people living in different geographical areas. Demand zones located in the same city or urban zone can be different from each other in some characteristics, which calls for differential treatment of these demand zones. Bélanger et al. (2019) argue that equity is sometimes more of a perception than a measure as in some cases a customer might feel dissatisfied even though receiving service of a high standard, and might feel accessing worse service than what other customers are receiving. Despite such challenges, the issue of equality must be addressed in a more strategic, scientific and holistic manner.

Efficiency

The Scottish Ambulance Service Board Report of the priority-based Dispatch Project (2002) defined efficiency as a means of generating more or better output (volume or quality of service) from the available resources or achieving the same volume or quality of output with fewer resources. One can therefore generalise efficiency as delivering an intended output for fewer resources. Efficiency in this research is focusing on matching the resource deployment to the needs of a geographical area.

Effectiveness

Effectiveness implies doing what is expected and it implies being adequate to complete a set purpose. In emergency ambulance logistics, effectiveness would imply that the ambulances are meeting the expected transportation of the medical equipment, staff and patients from one point to another in the logistic chain without causing any unnecessary delays. In essence it encompasses issues around capacity utilisation levels. In simple terms ambulances must be fully utilised and idleness is not the desired outcome, whilst overuse is also undesirable. Therefore, there is a need to strike a balance to ensure the smooth provision of emergency ambulance services. In cases where there is overuse of ambulances, there is increased tear and wear leading to unnecessary breakdowns. Idleness would imply that some areas where the ambulance is needed are actually being deprived.

2.8.1 The Focus of Simulation Modelling

It is against this background that the overall aim of this research is also to ensure equality, efficiency and effectiveness in the mobilisation and allocation of resources considering possible variations in demand across the city of Bulawayo through the use of robust modelling techniques to help achieve high levels of preparedness.

Several models have been proposed in the literature, with different objective functions, constraints and methods of solutions in trying to address issues of quality in EMS. Pinto et al. (2015) highlighted that even though these models share common design elements, often the emergency delivery systems are specifically designed to meet the needs of a certain context. Aringhieri et al. (2017) emphasized that in order to manage a comprehensive and reliable EMS system, relevant data should be forecasted, complex systems should be modelled, efficient solutions and accurate dispatching policies should be designed. It is this level of expected rigour that is going to be adopted by this research while maintaining a reasonable balance among the interacting components in addressing these challenges.

The focus of the research is on the deployment of ambulances to service stations. The ultimate goal of any emergency medical system is to respond in an organized and quick way using minimum resources as possible, to all requests requiring emergency prehospital care. According to Zhen et al. (2014) the rising costs of medical equipment, increasing call volumes, worsening traffic conditions in urban areas make emergency medical service control centers face increasing pressure to meet performance targets. Such endeavors have been hampered by the uneven distribution of the population in the city, distribution of health centers, medical response vehicles, technical staff and the ambulance service stations resulting in failing to meet the performance measures such as the response time. The ambulance control centers have the responsibility to allocate the optimal number of ambulances in service stations so that medical service is provided to patients in an efficient manner. However, challenges continue to be faced in deploying the ambulances to waiting bases due to the uncertainty of the demand calls, travel time and service time which are often random in nature.

2.8.2 The General EMS Vehicle Dispatch and Service Process

Bélanger et al. (2019) classified EMS into two main groups: Anglo-American and Franco-German. In the Anglo-American group which the Zimbabwean government has adopted, the EMS is separated from the medical system and instead offer sorely paramedic care. The main objective in this scenario is to allow the EMS team to respond to calls as soon as possible whilst ensuring a fast and safe transportation of the patient to the appropriate medical center. The Franco-German system is designed to provide mobile and on-site medical care. It is considered in this scenario that the speed in the transportation of the patient is less important as compared to the former. The later system is yet to be adopted by the government of Zimbabwe.

The Anglo-American approach consists of a systematic process that adheres to the following steps when responding to emergency medical calls: the arrival of an emergency call, call screening, vehicle dispatching, vehicle travelling from its current location to the emergency scene, on scene treatment and finally patient transportation to a health facility. Figure 2.9 is a resemblance of the generic ambulance emergency system applicable to two main dispatch approaches that are widely considered in literature namely; the static dispatch approach and the multi-location dispatch approach.

In a static dispatch model, ambulances are strictly dispatched from their fixed bases whilst in the case of the multi-location model, ambulances are dispatched wherever they are. The dynamic dispatch approach which is not represented in Fig.3.1 is applicable where ambulances can be dispatched and rerouted at any time in order to serve a high priority call. According to Pinto et al. (2015), the dynamic dispatch approach always requires the system to be capable of identifying the position of the ambulance and this requires the use of a specialized software called the geographical information system (GIS).

2.9 Location, Relocation and Dispatching Trends in EMS Research

There has been recent development of optimization models aimed at addressing the dynamic contexts of EMS in assisting decision-makers in improving pre-hospital care service delivery. Başar et al. (2012) in an attempt to design a reusable model in different contexts, developed a systematic classification that goes a long way in locating the direction of this research which mainly focuses on the sizing and siting of emergency medical ambulances. Leknes et al. (2017)

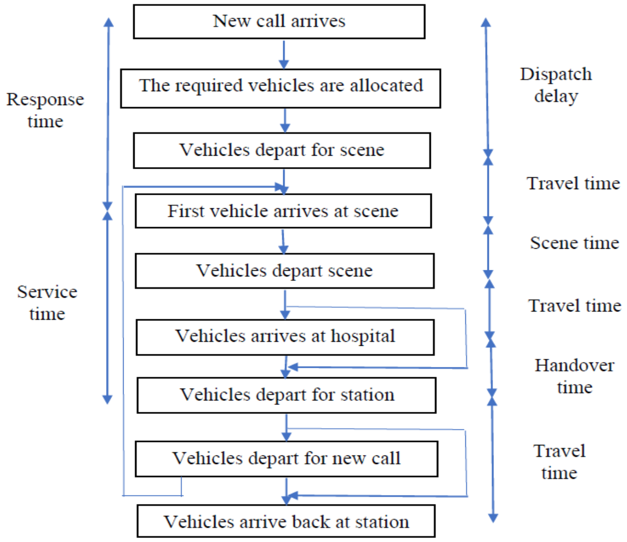

New call arrives
The required vehicles are allocated
Vehicles depart for scene
First vehicle arrives at scene
Vehicles depart scene
Vehicles arrives at hospital
Vehicles depart for station
Vehicles depart for new call
Vehicles arrive back at station
Response time
Service time
Dispatch delay
Travel time
Scene time
Travel time
Handover time
Travel time

Figure 2.9: EMS Vehicle Dispatch and Service Process (R. McCormack and G. Coates, 2015).

highlighted that the sizing of the fleet of ambulances and the allocation of ambulances to the ambulance stations are part of the tactical problems that any EMS provider has to address as part of strategic planning. Başar et al. (2012) categorized the optimization models into three main components namely: problem type, modeling and solution which culminated to a taxonomy of ESS (Table 2.1).

Table 2.1: Literature on the Taxonomy of the ESS Location Problem.

Section A	**Section B**
1. Problem type	2.2.5 Types of servers
1.1 Type of Emergency	2.2.5.1 Single
1.1.1 Fire	2.2.5.2 Multiple
1.1.2 Large scale emergency or disaster	2.2.6 Costs
1.1.3 Ambulance	2.2.7 Number of servers required
1.1.4 Hospital	2.2.8 Busy function
1.1.5 Police	2.2.8.1 System wide
1.1.6 General or not specified	2.2.8.2 Individual
1.2 Model Structure	2.2.9 Reliability
1.2.1 Deterministic	2.2.10 Queue size limit
1.2.2 Stochastic	**2.3 Decision variables**
1.2.2.1 Objective function	2.3.1 Coverage variables
1.2.2.2 Constraints	2.3.1.1 Single
1.3 Variation in Time	2.3.1.2 Multiple
1.3.1 Static	2.3.2 Server assignment variables
1.3.2 Dynamic	2.3.2.1 Continuous
1.3.2.1 Short-term (redeployment)	2.3.2.2 Binary/integer
1.3.2.2 Long term (strategic)	**2.4 Constraints**
1.4 Number of objectives	2.4.1 Coverage constraints
1.4.1 Single	2.4.1.1 At least once
1.4.2 Multiple	2.4.1.2 At least multiple
2. Modeling	2.4.1.3 Proportional
2.1 Objective Function	2.4.1.4 Mixed

Section A	**Section B**
2.1.1 Min. total no. of facilities	2.4.2 Server capacity constraints
2.1.2 Minimize the total distance or time to serve all calls	2.4.3 Priority of servers
2.1.3 Minimize the sum of costs	2.4.4 Max. no. of servers at each location
2.1.4 Minimize the maximum travel time or distance to any single call	2.4.5 Maximum distance constraints
2.1.5 Max. the covered area or demand once	2.4.6 Upper bound on no. of servers
2.1.6 Maximize the covered area or demand multiple times	2.5 **Type of model**
2.1.7 Maximize the total area or demand covered with a given probability	2.5.1 Integer programming
2.1.8 Maximize the expected area or demand covered	2.5.2 Dynamic programming
2.1.9 Max. the probability of service being available within a specified distance	2.5.3 Goal programming
2.2 Parameters	2.5.4 Fuzzy programming
2.2.1 Demand	2.5.5 Non-linear
2.2.1.1 Real demand	2.5.6 Simulation
2.2.1.2 Synthetic demand	**3. Solution**
2.2.2 Response data	**3.1 Optimal**
2.2.2.1 Travel time	**3.2 Heuristic**
2.2.2.2 Distance	**3.3 Metaheuristic**
2.2.3 Server capacity	3.3.1 Tabu search

Section A	**Section B**
2.2.3.1 Capacitated server	3.3.2 Ant colony
2.2.3.2 Uncapacitated server	3.3.3 Genetic algorithm
2.2.4 Server location	3.3.4 Simulated annealing
2.2.4.1 Discrete	3.3.5 Others
2.2.4.2 Continuous space	**3.4 Simulation**

Note. Table 2.1 adopted from Başar et al. (2012).

Aringhieri et al. (2017) in a review paper, classified the existing literature based on two key concepts: equity and uncertainty. However, the researcher observed that despite the difference in approach, there is convergence of ideas in the discussions by different authors with regards to the study of EMS systems. In a review paper by Bélanger et al. (2019), they proposed that in as much as a lot of research has been done on decision aspects of EMS, the decision levels can be broadly classified under strategic, tactical and operational levels as represented in Table 2.2.

The strategic decision level focuses on decisions that are static in nature such as the location of the ambulance stations, fleet dimensioning and staff hiring. The tactical level addresses decisions such as the location of standby sites, crew pairing and scheduling, and fleet management strategies. However, research efforts in the recent years have attempted to address the issue of uncertainty and dynamics involved in EMS and hence the emergence of models that focus more on the operational decisions. The main thrust at the operational level is to address short-term issues such as relocation and vehicle dispatching strategies as summarized in Table 2.2.

The complexity of contexts in which EMS manifest has led to the development of different models depending on the interaction between location, relocation and dispatch decisions over

Table 2.2: Classification of Decision Problems in EMS Management and Proposed Strategies and Models.

Decision Level	**Decisions**	**Strategies**	**Models**
Strategic	Ambulance station location Fleet dimensioning Staff hiring		
Tactical	Standby sites allocation Crew pairing and scheduling Fleet management strategies		
Operational	Ambulance location	Static location	Single coverage Multiple coverage Prob. and stochastic Stochastic and robust location-allocation Maximal survival Equity
		Relocation	Multi-period Real-time or online Compliance table or offline ADP-based
	Ambulance dispatching	Nearest vehicle Other rules	

Note. Table 2.2 adopted from Bélanger et al. (2019).

the years. Early studies that created the foundations of EMS research were based on the static ambulance location assumption. The static ambulance location problem assumes that after completing a mission, each ambulance will return to its designed standby site defined according to the predetermined location plan.

Bélanger et al. (2019) classified these early models into three main categories in their chronological order of evolution: single coverage deterministic models, multiple coverage deterministic models and probabilistic and stochastic models. Despite their classical assumption, static ambulance location models evolved over time by integrating more realistic features of the problem such as demand uncertainty, availability of vehicles, dispatch delays, distances travelled, traffic congestion and different types of response vehicles in the waiting bases.

2.9.1 Single Coverage Models

Early research on the formulation of static ambulance location problem was based on the idea of coverage. Toregas et al. (1971) proposed the location set covering problem (LSCP) based on the idea of coverage. The objective of the model was to minimize the number of vehicles ensuring that all zones are adequately covered. This class of optimization problems were classified as deterministic single coverage models. It assumed that a vehicle was always available upon arrival of a demand call.

However, the model had its own limitations related to the assumptions adopted for deterministic single coverage models. Firstly, the number of ambulances required to accomplish a complete coverage might be very high and unrealistic considering that the human and material resources are always scarce and costly. Secondly, decision makers are not merely concerned about coverage, instead they also consider seriously the best levels of utilisation of the allocated ambulance vehicles. The models also assumed that a vehicle is always available upon arrival of an emergency call which is always not the case in practice. A practical example arises when two consecutive emergency demand calls are received from the same zone covered by the same vehicle and there is little time between servicing the first call and the second call. A delay however has to occur in such instances.

2.9.2 Multiple Coverage Models

Church and ReVelle (1974) upon realizing these limitations, formulated the maximal covering location problem (MCLP) which sought to maximize the demand covered by a specific vehicle fleet size. Other researchers made follow ups to these early problems in carrying out empirical researches that would try to address different aspects of the problem by looking at scenarios where two different types of ambulances are applied Schilling et al. (1979). Despite their simplicity in formulation and limitations in application to real world EMS, deterministic single coverage models then opened up research space for other variants such as the deterministic multiple coverage models and probabilistic and stochastic models.

The deterministic multiple coverage models unlike the deterministic single coverage models improved in the aspect being more applicable to real-life situations. Researchers were focused on issues around the stochastic or randomness nature of the calls for emergency vehicles and their availability. They mainly focused on increasing the chances or likelihood of having a demand zone covered by one vehicle by increasing the number of vehicles available to cover the demand zone.

This led to the development of the hierarchical objective set covering model (HOSC) proposed by Daskin and Stern (1981) to cover the aspect of multiple coverage. The HOSC minimized the number of vehicles needed to ensure a complete coverage and maximize the number of vehicles that can cover a zone. The HOSC had its own limitations as the additional vehicle would directly increase the number of vehicles required to ensure complete zone coverage, The allocation of two vehicles to a demand zone might not necessarily ensure full vehicle utilization as there is high likelihood that the vehicles will not be busy simultaneously at the same time. Moreover, it is argued that since it does not consider the specific demand of a zone, there is likelihood that HOSC will group vehicles to zones that can be easily covered leaving harder to reach zones only to be covered once.

Several variants to already discussed optimization problems were developed by researchers such as Eaton et al. (1986), Gendreau et al. (2001), Doerner et al. (2005), Beaudry et al. (2010) and more recently Su et al. (2015). Each of the research conducted aimed at improving the limitations of one model or addressing other important aspects of the EMS problem.

Important aspects of the EMS that were addressed by such models included: putting limits to the response time, maximizing the demand covered twice given a specific number of vehicles to locate, minimizing both costs of delayed services and the operating costs, imposing limits to ambulance workloads and dispatching multiple types of vehicles under various priority levels.

A more significant contribution for static ambulance location models was the maximal-multiple location covering problem (MMLCP). Storbeck (1982) proposed a goal programming formulation named the maximal-multiple location covering problem (MMLCP). The aim of the MMLCP was to locate a fixed fleet of vehicles in order to simultaneously minimize the uncovered demand and maximize the number of demand zones covered by more than one vehicle. It is important to highlight that the later research was inclined at collecting empirical data that would then be transformed to solve realistic EMS situations.

2.9.3 Probabilistic and Stochastic Models

The study of EMS has a significant number of variables that are often characterized by uncertainty in their occurrence. This led to the development of probabilistic and stochastic models that rely on the calculation of expected values and are often referred to expected covering location models. These models generally seek to determine a set of vehicle locations that maximizes the expected coverage.

Daskin (1983) developed a maximum expected covering location problem (MEXCLP) whose objective was to locate a given number of ambulance vehicles in order to maximize the expected coverage. The expected coverage depended on what was referred to as the busy fraction. Bélanger et al. (2019) defined the busy fraction as the probability that a vehicle is unavailable to respond to an emergency call. In developing the MEXCLP it was assumed that the busy fraction is known and the same for all vehicles, the busy fraction is independent of the vehicle location and each vehicle operates independently. Despite the significant contribution of the MEXCLP in improving service delivery, it had its own limitations.

Several variants of probabilistic and stochastic models were developed to address different aspects of the EMS. These included the adjusted maximum expected covering location problem (AMEXCLP) which allowed the relaxation of the vehicle independence assumption by integrating a corrective factor in the objective function using the queuing theory. Of significance was the proposal by Batta et al. (1989) to use the hypercube model to estimate the expected coverage given a set location plan. However, the limitations of these models was that they assumed deterministic travel times which in practice varies between two locations due to factors such as roads, weather and traffic congestion.

Research by Ingolfsson et al. (2008), proposed a model in which the call delay time (chute time) and the dispatching priorities for ambulance vehicles are considered in the model. Restrepo et al. (2009) propounded a model that locate ambulances over a set of waiting bases in order to minimize calls that are expected to find no ambulance available in their waiting bases. The concept of queuing theory made a mark in the development of EMS optimization models. Galvão et al. (2005) proposed a model inspired by the maximum availability model (MALP) that uses queuing theory to represent a more realistic life situation of the EMS problem. The model incorporated a corrective factor in the probabilistic constraints that consider vehicle-specific busy fractions rather than zone-specific ones and the cooperation of vehicles in responding to emergency calls. The model embedded the the hypercube model in the local search methods.

2.9.4 Stochastic and Robust Location-allocation Models

The stochastic and robust location-allocation models endeavored to account for the randomness of the call arrivals. According to Bélanger et al. (2019), unlike the other discussed demand coverage maximization models, location-allocation inspired models aim to minimize costs under demand satisfaction constraints. Zhang and Li (2015) formulated an ambulance location-allocation model that they applied to a case of a city in China with a range of twenty to seventy stations. The model simultaneously minimized the ambulance operating and transportation costs and the demands not served on time. They incorporated chance-constraints to deal with demand uncertainty and these constraints ensured with a given probability that the number of vehicles located in a demand zone can satisfy the number of concurrent demands emanating in the area assigned to it.

Boujemaa et al. (2018) formulated a two-stage stochastic location-allocation model to design an EMS in Tunisia. The model was designed to simultaneously determine the location of ambulance stations, the number and type of ambulances to be deployed and the demand zones to be covered by each station. Despite the high level of formulation of stochastic and robust location-allocation models, they require vast computational turn around time in solving them and this has reduced their attractiveness for adoption.

2.9.5 Fuzzy Models

Aringhieri et al. (2017) indicated that this paradigm of fuzzy models is mostly applicable to deal with the uncertainty in the number of emergency calls. The fuzzy paradigm allows the use of qualitative data as well as expert-based knowledge by characterising them as linguistic terms. In essence, the fuzzy models are applied when the stochastic framework or the probabilistic paradigm cannot be used.

2.9.6 Human Outcome-based Models

Early research by Daskin and Stern (1981) and Daskin (1983), focused on performance-based indicators such as coverage and the proportion of emergency calls responded within a fixed time frame . Over the years, there has been growing need to satisfy the patient's outcomes such as safety and satisfaction whilst optimizing the use of scarce resources. This has seen the emergence of the maximal survival and equity models that are inclined to integrating patient outcomes into the decision-making process by EMS organizations.

Survival Models

Erkut et al. (2008) proposed the maximal survival location problem (MSLP) considering patient outcomes. The model considers the probability of survival of a patient by including it in the objective function that attempts to maximize the expected number of lives. When the model was applied to a case of Edmonton in Canada, the results indicated a significant increase in the number of survivors. Bélanger et al. (2019) argues that despite benefits through the consideration of patient survivability in location models, response time thresholds and coverage are still the important metrics in evaluating EMS performance.

Equity Models

Aringhieri et al. (2017) indicated that equity is one of the most challenges in the healthcare sector and specifically EMS as it evaluates the fairness of how resources are distributed to patients in heterogenous societies. Such disparities exist between urban and rural areas or between high density suburbs and low-density suburbs. If issues around equity are not addressed in such scenarios, it would imply that lives are valued differently in different areas. This has led to the development of two schools of thought namely: horizontal equity and vertical equity. In addressing horizontal equity, it is considered that all demand nodes are treated in an equal manner whilst in vertical equity demand nodes are treated differently. In general practice, EMS providers are deemed to be providing equitable services if they favor disadvantaged groups.

Bélanger et al. (2019) argues that equity is a complex phenomenon in the study of EMS and can be more meaningful when it considers the standing perspective of the key stakeholder(s). They considered the patient's perspective and the service provider's perspective as key stakeholders in EMS system. The patient is mainly concerned about fairness in the context of patient's outcomes and patient's waiting time. The stakeholder perspective mainly focuses on the issues around fairness in the distribution of workload as an example. This aspect of EMS affects directly the retention of skilled personnel and the level of attracting new employees to the organization.

2.10 Why Simulation Modelling?

Simulation modelling is the process of designing a model of a real system and conducting experiments with this model for the purposes of understanding the behavior of the system and or evaluating various strategies for the operation of the system.

Henderson and Mason (2005) asserted that even though the hypercube model remains a powerful modelling approach, it requires several assumptions with regards to the way ambulances are dispatched whilst posing a great threat in convincing decision makers to adopt its predictions due to model complexities. This is a common feature for most models that have been discussed so far in literature.

Zhen et al. (2014) highlighted that even though minimal covering models, maximum covering models and double standard models have been developed based on either integer programming or dynamic programming formulation methodologies, finding their solutions is time consuming as they need to solve an optimisation sub-model every time a decision has to be made. It is from this stand point that the research team adopted a simulation optimisation method that enabled them to evaluate operational performance of deployment plans through a detailed simulation model.

Pinto et al. (2015) argues that ambulance services have been examined in a variety of ways including queueing theory and hypercube models, mixed integer programming, stochastic optimization and dynamic programming. However, they purported that despite simulation modelling being seemingly costly in implementing, it allows a more detailed description of the system and allows the analysis of dynamic effects. Such uniqueness makes the simulation approach to system analysis a cut above the rest especially in the aspect of decision making. Pinto et al. (2015) argued further and stated that the high number of variables and the random nature of demand makes the analysis for decision-making a combinational problem with a higher number of alternatives and renders deterministic methods unattractive.

According to Eldabi and Young (2007), the pressure for better services, low availability of resources and need to assess the impact of changes before actual implementation has created huge opportunities of increasing modelling and simulation in healthcare. Thus, simulation modelling is an attractive alternative as it allows an analysis of different scenarios before the actual implementation. Pinto et al. (2015) acknowledged that there could be many dimensions of uncertainty in modelling an ambulance service, however, these dimensions can be portioned with relative ease and changed when moving from one ambulance service provider to another.

Simulation modelling techniques allows one to gain information and insight into the operation of the system without disturbing the system. Simulation also allows one to develop resource and operation policies to improve the system performance by observing different system scenarios through visualization and animation. In other cases, simulation modeling is used to test and validate new systems before their implementation. It has been however observed in literature that every other analytical method used in the discipline of operations research such as: linear programming, network analysis, meta heuristics, queuing theory, game theory and simulation has got its on merits and demerits when applying them to solve real life problems. Below are a summary of the advantages and disadvantages of using the simulation analytical method or technique.

2.10.1 Advantages of Simulation

Some of the advantages of using the simulation modelling technique are summarised as follows:

- The technique requires fewer assumptions in application and captures more of the actual or true characteristics of a system under study.

- The concept of simulation is much easier to comprehend, thus making it easier to have a buy-in or to justify to key stakeholders involved in the study.

- The technique allows one to gain information and insight on how the system under study works, its behaviour over time and not just the end result.

- The communication tools embedded in the technique such as visualisation and animation stimulates discussion about all aspects of the system under study.

- The technique allows one to use non-standard distributions, hence it provides the flexibility to describe events and timings as they occur.

- The technique allows for the assessment of different system scenarios using different performance measures without disturbing the actual system in operation.

2.10.2 Disadvantages of Simulation

Some of the disadvantages of using the simulation modelling technique are summarised as follows:

- The utility of the simulation study results depends on the skills of the modeller and the quality of the model which makes it subjective in terms of its application and adoption.
- The collection of reliable secondary and primary data for the development of a simulation model can be time consuming and this translates to be expensive.
- Simulation models do not produce an optimal solution, however they serve as a tool for analysis of the behaviour of a system under the stipulated conditions by the researcher.

Aringhieri et al. (2017) argued that a combination of appropriate approaches could help EMS managers to come up with more realistic models which could fully capture the complexity of the system and provide a better overview of the whole system.

2.10.3 Empirical Evidence of Simulation Modelling in EMS

There has been a wide range of research on emergency medical services using simulation modelling as a solution method of preference. In other cases, there has been a deliberate attempt to integrate different operations research techniques in order to improve the robustness of the results and analysis of the developed models. Silva and Pinto (2010) integrated simulation and optimisation techniques to analyse and evaluate the emergency medical system of the city of Belo Horizonte in Brazil. In their research, they focused on two critical aspects of service: how the system responded to an increased demand and the re-sizing of the ambulance fleet in order to significantly reduce the response time. Simulation in this case allowed different scenarios to be assessed without interfering with the actual EMS system.

Whereas the use of optimisation for simulation improved the search for optimal settings of the system. More recently, Pinto et al. (2015) designed a generic method to develop simulation models for ambulance systems which integrated simulation and optimisation techniques. The model was validated using a case study of Belo Hoizonte in Brazil and the UK system.

Aboueljinane et al. (2014) carried out a simulation study to improve the performance of the emergency medical service of the French Val-de-Marne department. They focused on five strategies namely: varying the number and workload of resources, improving the EMS team deployment, regionalising the response, multi-period redeployment and process improvement. In all these strategies, they employed the discrete event simulation (DES) model in assessing different scenarios whilst using coverage and the utilisation rate as the performance measurements. In the application of the multi-period redeployment strategy, where the assignment of EMS teams were adjusted to adequately cover changing demand patterns, Aboueljinane et al. (2014) managed to demonstrate on how the simulation optimisation can be incorporated in the DES model in order to handle the large number of possible redeployment plans. Results of this strategy indicated that the multi-period redeployment solution provided improved coverage and utilisation rates. Coverage here is considered as the percentage of calls for which the response time does not exceed a specific target time. An example could be 80% of calls less or equal to 20 minutes of response time. The human resources utilisation rate was defined as the total workload divided by the total operating time.

Zhen et al. (2014) developed a simulation optimisation framework for ambulance deployment and relocation and applied it to the metropolitan city of Shanghai in China. In this framework, they applied a simulation optimisation method to evaluate the operational performance of deployment plans through a detailed simulation model. Key considerations were made on the stochastic and dynamic nature of emergency call arrivals, complex traffic conditions and emergency fulfillment processes such as ambulance traveling and serving processes were made. Their research findings indicated that simulation optimisation is more appropriate for making deployment decisions when there are limited number of ambulances (scarce resources) in order to optimise the deployment instead of just allocating them in a balanced manner. Another important result from this research was that establishing more ambulance bases or involving more hospitals into the system does not guarantee the deployment of ambulances in a balanced way. The main reason proposed to influence this phenomenon was that the uneven ambulance deployment decision is directly influenced by the uneven distributions of population and uneven traffic densities and congestion among different areas.

2.11 A Framework for Developing Simulation Models

Simulation models for EMS are generally developed specific for each case. However, there are three core areas of any EMS namely: call generation, dispatch of ambulances and ambulance journey. In the system, ambulances are permanent entities and calls are temporary entities. Attributes have to be generated for both the ambulances and the calls at the beginning of a simulation run. The ambulance tasks include: travel, service at the scene, delivery at the hospital and restocking. However, all or a few of these ambulance tasks maybe performed due to the nature of the call. It is therefore imperative to discuss some of the salient features of the EMS in order to help in the formulation of the input parameters of the simulation model, the assumptions to be made, performance measures and the selection criteria for the optimum model.

2.11.1 Call Features

The main characteristic feature of calls is its level of uncertainty which manifest in the form of seasonal patterns and the associated randomness in their occurrence. Such trends are highly depended on the geographical distribution of the population, call arrival rate, type of emergency incident and the level of risk of the patient demanding for service. The calls have to be directed to the respective substations or waiting bases. These calls also need to be categorized depending on the severity of the case based on the judgement of the dispatchers at the call center.

Arrival Rates

The demand for emergency medical services is uncertain and random in nature. According to Matteson et al. (2011), variations can be observed with seasonal trends over a year, days of the month, week of the year and hours of the day. There is need to strike a balance to avoid either underestimating or overestimating the call arrival rates. Underestimating the call arrivals has implications such as under-staffing and excessively high response times. Over-estimating the call arrivals would imply over-staffing and high operational inefficiencies. Pinto et al. (2015) acknowledged that resources are always scarce and there is always need to operate a system within some cost constraints and this implies that there is need to maximize the utilization of resources.

Several methods in literature have been proposed in order to model the call arrival rates. Setzler et al. (2009) in a comparative study on emergency call demand used artificial neural networks and their approach offered a significant improvement at the hourly level. Channouf et al. (2007) made an attempt to model the daily call arrival rate as a Gaussian with fixed day-of-week, month-of-year, special day effects and fixed day month interactions. Despite the advent of several methods attempting to model call arrival rates, Matteson et al. (2011) emphasized that the standard and universal practice is that hourly call-arrival volume is a Poisson distribution.

Geographical Distribution of Calls

There is general consensus in literature that different zones generate different demand patterns resulting in non-uniform call distributions. The call-arrival also varies from time to time even within a time frame of a day. In a case study in Turkey by Swalehe and Aktas (2016), they recommended that more ambulances should be deployed near residential areas at night than workplaces and more should be deployed near workplaces during day times than at night to ensure that ambulances are closer to where they might be needed most. This is a clear indication that geographical distribution of calls plays a vital role in designing simulation models.

2.11.2 Types of Emergency and Ambulance Requirements

When calls are received by an operator at a centralized call center, the call operator determines the level of the emergency into three categories. Category 1 is classified as urgent and life threatening cases which include among them road traffic accidents, general accidents or emergencies such as cardiac arrests. Category 2 is classified as urgent but not life threatening cases. It comprises of maternity emergencies that would have been verified from clinics and are for referral transfers (maternity clinics). Category 2 also includes maternity cases from home and these are emergency calls related to maternity issues from homes that have not been verified by medical doctors (clinics from home). Lastly, Category 3 is classified as non-urgent calls which comprises of removals or transfers from one institution to another and might also include chronic cases, psychiatric patient transfers and blood transfers.

Ambulance Features

Different scenarios exist in literature as propounded by Aboueljinane et al. (2014) and Pinto et al. (2015), where different kinds of ambulances are available to respond to different kinds of emergency incidences. In other setups there is only one type of ambulance or with the same fitted features to respond to emergency calls. In other scenarios, two or more different types of vehicles are available for dispatch. In recent years there has been the use of helicopters for emergency rescue missions and the Scottish Ambulance Service is a good example. When designing analytical and simulation models, such variations if considered ensures the robustness of the model.

Ambulance Dispatch

Three main dispatch models that exist in literature are the static, multi-location and the dynamic dispatch models. In a static dispatch model, ambulances are strictly dispatched from their fixed bases. After rendering service at the scene, the ambulance either transfers the patient to the medical center or returns back to the base and stay there waiting for the next dispatch. In the case of the multi-location model, ambulances are dispatched wherever they are. The allocation to a new case can occur from the moment the ambulance delivers a patient at the hospital or soon after completing service at the scene of an incident. In the multi-location model, the ambulances do not necessarily have to return to their waiting stations before the next dispatch. The dynamic dispatch approach is applicable where ambulances can be dispatched and rerouted at any time in order to serve a high priority call. It implies that the ambulance remains available even after it has been dispatched.

Calls requiring emergency medical services arrive by phone and is attended to by a dispatcher. The dispatcher upon interrogating the caller for emergency services makes a decision by asking some predefined questions and determines the priority of the call (Zhen et al., 2014). Aringhieri et al. (2017) defines dispatching in the context of EMS as the act of choosing appropriate EMS vehicles to respond to emergency calls based on the nature and location of calls. Old EMS organisations used to employ the first come first serve (FIFO) criteria or the closest-idle criteria, which sends the closest available EMS vehicle to each call. Current trends have seen the development of priority rules which normally assigns high priorities to life threatening incidences.

Lee (2012) proposed that ambulance dispatching decisions can be categorised as either call-initiated or ambulance vehicle-initiated and these are highly dependent on the busyness of the EMS system. In the call-initiated based decision, the dispatcher is mandated to select one of the current idle ambulance vehicles to be dispatched after the arrival of an emergency call. In the vehicle-initiated approach, the dispatcher has to choose on the emergency calls queuing whenever an ambulance vehicle is freed to service another call. Call-initiated decisions are idle in scenarios where the EMS system load is low whereas the vehicle-initiated decisions are appropriate in high load EMS systems.

Ambulance Routing

In practice, it is possible that an ambulance can also be given a specific route to follow when responding to an emergency call. The concept of rerouting is applicable in the context of dynamic dispatch model. Aringhieri et al. (2017) defines routing as the act of defining the exact path a dispatched ambulance should follow to reach a patient or a specified hospital. When ambulances are being routed special consideration is given to essential parameters such as the road damage status, distance and level of congestion. Pinto et al. (2015) argued that dynamic dispatch models are more efficient as they require fewer ambulances. However, if re-routing does not occur frequently, the multi-location dispatch model is more appropriate as it is simple and does not require the use of the expensive GIS tools.

EMS Crew and Ambulance Journey

The three key elements of the ambulance journey along with the EMS crew are: provision of service at the scene, delivery of patient at the hospital and restocking of medical resources. The servicing process of the ambulance crew begins when the ambulance arrives at the scene.

There are three possibilities that could confront the ambulance crew: false and malicious alarm, false alarm with good intent and the true existence of an emergency case that needs attention of which service is rendered.

A call is classified as false and malicious if the call is made to the station and yet there is no genuine reason for making the call as there would not have been an emergency at all. Such calls are made by individuals with the intent to test the responsiveness and preparedness of the EMS. However, such is not desirable as it puts to waste the scarce resources such as fuel, human capital and time among many others. In the case of false alarm with good intent, the call will be genuine as there will be need for emergency services, however, the source might then find another alternative means of transport to transfer the patient to a medical centre or would have received the required attention before the arrival of the emergency medical crew.

After rendering service on the scene, the ambulance is likely to transfer the patient to the hospital. However, it is possible that the ambulance can be deployed back to the waiting base or may be dispatched to serve a new call in the case where there is no need to transfer the patient to a medical institution. There are cases when ambulances are diverted when they reach a hospital or medical institution and realise that they are fully occupied in terms of their capacity. Such ambulance diversions cause system delays. It is also a reality that an ambulance could be diverted to a specific medical center due to the nature or condition of the patient during transition from the scene to an intended medical center. Some models would also try and account for such system delays. Lastly, it is the responsibility of the ambulance crew to ensure that the medical aid materials are replenished time and again when they return to their waiting bases.

2.11.3 Simulation Model Formulation Framework

The simulation model development framework will adopt the one applied by Pinto et al. (2015) which focused on four fundamental aspects namely: Analysis of the main input parameters, development of the simulation model, analysis of the performance measures, and use of optimization for simulation to analyse the scenarios.

However, it is important to note that the model development is embedded in the general steps that has to be performed when conducting a simulation study. The steps are summarised in Figure 2.10.

Ingolfsson et al. (2008) presented an ambulance optimisation model that minimises the number of ambulances needed to provide a specific service level. Their model considered the randomness that occurs in the inter-arrival of calls, response time (delay and travell time) and service time (delay on scene and travel to health institution or waiting station). These stochastic aspects of service delivery indirectly incorporates the uncertainty in the ambulance availability in determining the response time. They argued that models that do not account for the uncertainty of one of the components of response time and ambulance availability is likely to overestimate the possible service level for a given number of ambulances or underestimate the number of ambulances needed to provide a specific service level. therefore, the simulation models to be developed will take cognisance of the uncertainty or stochastic nature of the inter-arrival of calls, response time and service time.

Due to the heterogeneous nature of the medical service demand, the problem is focused on how to deploy a given amount of ambulances to sub-stations (waiting bases) in the city in order to optimise the service delivery whilst addressing issues of equality, effectiveness and efficiency. The geographical location of a sub-station has an indirect influence on the overall performance of the service delivery in terms of the response time and service time. Leknes et al. (2017), emphasised that the service time highly depends on the travelling distances in

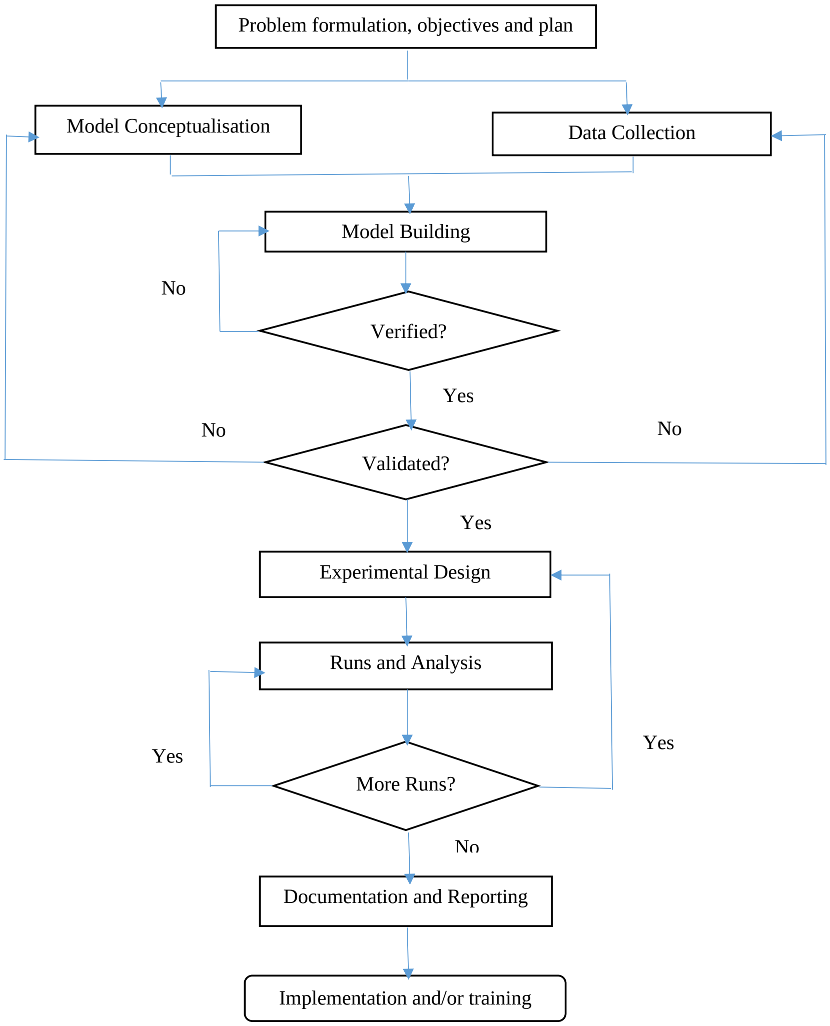


Figure 2.10: Fundamental Steps in a Simulation Study.

the area that the station covers, nature of roads and distance to the nearest hospital. Hence, sub-stations were considered separately to ensure robustness of the simulation models. The model also acknowledges that ambulances are not always available, hence the calls have to queue for the service.

In order to assess the performance of the simulation model for future demand forecasts from ANN models, specific multiple performance measures were adopted. The basis of using multiple performance measures is derived from Knight et al. (2012), who were able to demonstrate the benefits of using multiple performance measures rather than a single performance measure based on average response time. The performance measures to be considered for model building and model evaluation are namely: the average entity time in system, average response time, average response queue time, average number of calls in response queue and the ambulance utility levels.

The study will further explore some hypothetical situations using numerical experiments in order to determine some managerial implications that might arise due to unpredictable circumstances in future. The study also employed discrete event simulation that simulates the operational process levels of an EMS involving the call arrival, dispatch, ambulance travelling and servicing activities.

2.12 Optimisation for Simulation

Azadivar (1999) and Pinto et al. (2015) are in agreement that the optimisation for simulation model cannot be written in the standard form as in deterministic models as the objective function or constraints or both are embedded in the simulation model as decision parameters of the model and can only be evaluated by a computer simulation. Such optimisation problems can be solved using non-formal ways referred to as heuristics.

In computer science, artificial intelligence and mathematical optimisation, a heuristic is a technique used to solve a problem more quickly when classic methods are too slow. This approach to problem solving, learning, or discovery employs a practical method not guaranteed to be optimal or perfect, but sufficient for immediate goals. The process of determining the optimal solution for simulation by heuristics is shown in Fig. 2.11. The optimisation problem according to Fu (2002) can be formulated as follows:

$$Min_{\theta\varepsilon\Theta} J(\theta) = E\left[C(\theta, w)\right] \tag{2.40}$$

subject to:

$$R(\theta, w) \leq RT$$

$$N_i(\theta, w) \leq NLS \sim \forall i$$

where:

- θ is vector of input variables (the set of substations for allocation and quantity of ambulances at each base).
- $J(\theta)$ is objective function.
- w is number of replication.
- $E\left[C(\theta, w)\right]$ is expected value of $C(\theta, w)$.
- $R(\theta, w)$ is response times for the sample θ in replication w for ambulance units respectively.
- $N_i(\theta, w)$ is quantity of ambulance units allocated in base i at the sample θ and replication w.
- RT is upper response times of the ambulance units.

- NLS is upper bounds for total ambulance units in each base.

As suggested by Pinto et al. (2015), to solve the problem, an optimiser and a simulator must work together as shown in Fig. 2.11. The simulator evaluates the performance of each candidate solution, while the optimiser uses heuristics to seek the candidate solutions.

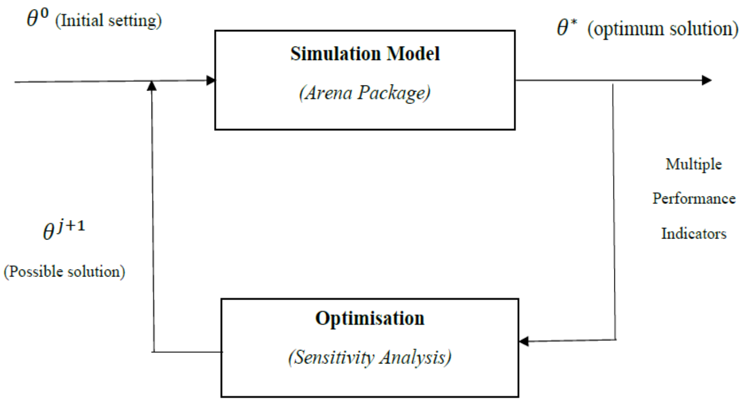


Figure 2.11: A Schematic Diagram of Optimisation for Simulation Approach (Pinto et al., 2015).

2.13 Chapter Conclusion

2.13.1 Relevance of Literature to Research Objectives

The literature review covered the area of time series analysis models (non-linear or linear) and how they can be used to predict future values that can be used for decision making at the strategic level. The emergency of artificial neural networks and its application to other critical and related sectors were discussed. The literature review has also discussed the early approaches applied in EMS studies in solving issues around location, relocation and dispatch of ambulances in an EMS system together with the associated challenges. The review of literature has also highlighted the growing advantages of using simulation modelling techniques in solving EMS system challenges of inequality, inefficiencies and ineffectiveness in the allocation of emergency ambulances.

Literature has also indicated that current research in EMS is seized in finding ways of addressing challenges at strategic, tactical and operational levels. The main strategic challenge is the location of ambulances and stations. Tactical challenges are the sizing of the fleet of ambulances and their allocation to the ambulance stations. Operational challenges involves decisions on which ambulance(s) to be dispatched and the reallocation of ambulances.

However, the available literature failed to locate the comparison of the use of emerging non-linear times series analysis techniques (ANN models) and the traditional linear time series analysis techniques (SARIMA) in predicting future demand for public emergency ambulances. Furthermore, the literature was unable to identify how the predicted static demand forecasts can be integrated by simulation modelling and optimisation techniques in addressing issues of resource mobilisation and allocation in order to increase levels of preparedness to respond to calls in an equitable, efficient and effective manner whilst assessing performance.

2.13.2 Conceptual Framework

The conceptual framework highlights how forecasting, simulation and optimisation techniques can be integrated in an attempt to improve public emergency ambulance preparedness whilst achieving important decision outcomes of resource mobilisation and allocation. Such a matrix is expected to address issues of equality, efficiency and effectiveness in the management of scarce resources whilst also being able to visualise and assess the performance of the EMS system. The research will follow the following conceptual framework as presented in Fig. 2.12.

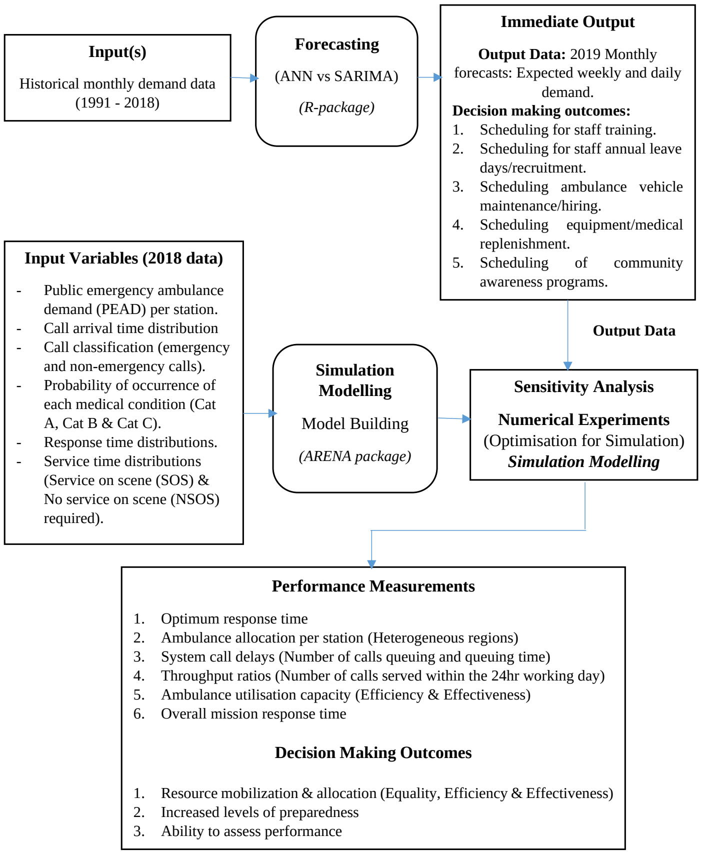

Input(s)
Historical monthly demand data (1991 - 2018)
Forecasting
(ANN vs SARIMA)
(R-package)
Immediate Output
Output Data: 2019 Monthly forecasts: Expected weekly and daily demand.
Decision making outcomes:
1. Scheduling for staff training.
2. Scheduling for staff annual leave days/recruitment.
3. Scheduling ambulance vehicle maintenance/hiring.
4. Scheduling equipment/medical replenishment.
5. Scheduling of community awareness programs.
Output Data
Input Variables (2018 data)
- Public emergency ambulance demand (PEAD) per station.
- Call arrival time distribution
- Call classification (emergency and non-emergency calls).
- Probability of occurrence of each medical condition (Cat A, Cat B & Cat C).
- Response time distributions.
- Service time distributions (Service on scene (SOS) & No service on scene (NSOS) required).
Simulation Modelling
Model Building
(ARENA package)
Sensitivity Analysis
Numerical Experiments
(Optimisation for Simulation)
Simulation Modelling
Performance Measurements
1. Optimum response time
2. Ambulance allocation per station (Heterogeneous regions)
3. System call delays (Number of calls queuing and queuing time)
4. Throughput ratios (Number of calls served within the 24hr working day)
5. Ambulance utilisation capacity (Efficiency & Effectiveness)
6. Overall mission response time
Decision Making Outcomes
1. Resource mobilization & allocation (Equality, Efficiency & Effectiveness)
2. Increased levels of preparedness
3. Ability to assess performance

Figure 2.12: The Conceptual Framework of the Study.

Chapter 3

Methodology

3.1 Introduction

This chapter focuses on the methodology which is principally guided by the conceptual framework. It focuses on the sources of data, analysis tools, software packages and methods of data analysis applied to the research study. The focus of the study was be divided into three broad sections namely: forecasting, simulation and optimisation for simulation as presented in the conceptual framework.

3.2 Forecasting

In this section, sources of data, time series analysis models and the model building processes are discussed in detail. Performance measures and selection criteria of the two models of interest; the FFNN and seasonal ARIMA model, will be discussed together with their mathematical formulation and application to time series analysis.

3.2.1 Model Input Data

Historical data on ambulance services for the Bulawayo city municipality, covering the period January 1991 to December 2018, were retrieved from the archives. For purposes of developing forecasting models, data for the period January 2010 to December 2018 was used because of the absence of cyclic and irregularity components in the time series data as compared to previous years. Each call for public ambulance emergency services is received and recorded at a control centre operated continuously everyday throughout the year. Monthly compilation is done and the information captured is stored in a database for billing and other management processes.

The data splitting approach was adopted in order to compare the two suggested models. Data from January 2010 to December 2017 was used for model building and the data for the year 2018 was used for model cross-validation and comparisons of the two modelling approaches. This translates to 96 observations for model building and 12 observations for model cross-validation. The flowchart of the methodology is presented in Fig. 3.1.

3.2.2 Feed-Forward Neural Networks

The feed-forward neural network architecture was trained by the "neuralnet" function of the R-package, which is a network training function that updates weights and bias values during training. The network is called feed-forward because information flows only from the input to the output and there are no recurrent or backward connections as presented in Fig. 2.4. Each layer consists of neurons and there is no connection between neurons that are in the same layer. A three-layer FFNN structure was also presented in Fig. 2.4.

The input data consisted of seven (7) inputs based on the lagged annual monthly demand for the years 2010 to 2016. The input and output data of the neural network is summarised in Table 3.1. Here, we assume that the demand for the next month is a function of the past

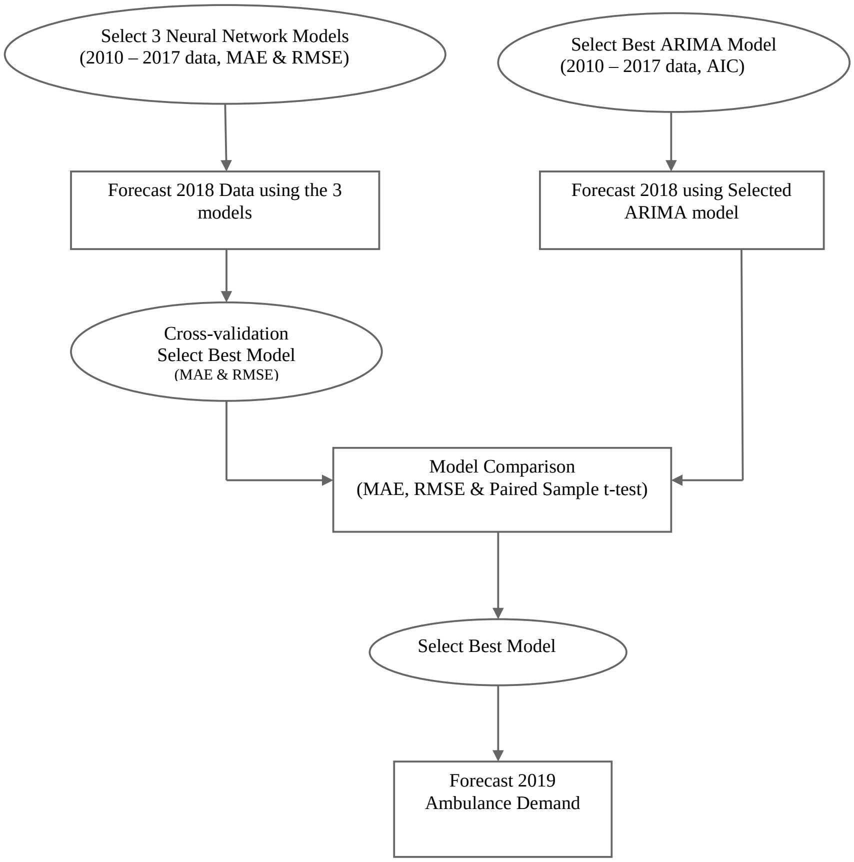

Select 3 Neural Network Models
(2010 – 2017 data, MAE & RMSE)
Select Best ARIMA Model
(2010 – 2017 data, AIC)
Forecast 2018 Data using the 3 models
Forecast 2018 using Selected ARIMA model
Cross-validation
Select Best Model
(MAE & RMSE)
Model Comparison
(MAE, RMSE & Paired Sample t-test)
Select Best Model
Forecast 2019
Ambulance Demand

Figure 3.1: Forecasting Model Building and Selection Procedure.

values of the previous months recorded at the same time. The model equation is represented by the generalised equation 3.1:

$$y_{t+12} = f(y_t, y_{t-12}, y_{t-24}, y_{t-36}, y_{t-48}, y_{t-60}, y_{t-72}) + \Theta_{t+12}, \tag{3.1}$$

where y_t is the observed demand value for the current month, y_{t-12} is the demand value of the previous month and so on, whilst y_{t+12} is the demand for the next month, and Θ_{t+12} is the associated bias. Günay (2016) implemented the same approach in his study of forecasting annual gross electricity demand. The selection of the optimal number of inputs is based on trial and error as mentioned in (Belayneh et al., 2014).

Table 3.1: Sample of Input and Output Data of the Neural Network.

Year	2010	2011	2012	2013	2014	2015	2016	2017
Month	y_{t-72}	y_{t-60}	y_{t-48}	y_{t-36}	y_{t-24}	y_{t-12}	y_t	y_{t+12}
January	1790	2152	2299	2288	2232	1954	1804	1724
February	1602	2095	2196	2238	2092	1853	1770	1603
March	1939	2238	2291	2413	2254	2003	1858	1753
April	1955	2214	1933	2073	2043	1813	1805	1637
May	2049	2215	2155	2261	2106	1832	1732	1674
June	1949	2253	2209	2246	2059	1758	1719	1554
July	2000	2274	2404	2093	2207	1869	1760	1724
August	2046	2255	2486	2214	2119	1834	1748	1650
September	2118	2419	2446	2247	2044	1818	1788	1583
October	2247	2011	2441	2289	2132	2033	1810	1649
November	2193	2446	2235	2197	1971	1898	1762	1626
December	2245	2353	2387	2289	2018	1952	1779	1674

Data Preprocessing

When designing ANN, data is pre-processed first before model fitting. Pre-processing represents data coding, enrichment and cleaning which involves accounting for noise and dealing with missing information. Pre-processing also involves the normalisation of data (scaling down) to an interval $[0, 1]$ in order to prevent saturation of hidden nodes before feeding into the neural network (Kitapcı et al., 2014). Three basic normalisation techniques that can be used in data preprocessing, namely; min-max normalisation, Z-score normalisation and the sigmoid normalisation technique according to Kheirkhah et al. (2013). The input values were first normalised using the minimum-maximum criteria in order to optimise the convergence rate of the neural network model.

Neural Network Model Training and Testing Sets

When the data has been processed, the network model-building process begins. The processed data is separated into two sets, namely; model-building set and testing set. The first set, called the 'model-building' set or training set, is used to develop the network model. The second set, called the testing or prediction set is used to evaluate the forecasting ability of the model. Higher percentages are usually assigned to the training set and lower percentages to the testing set (Tsai, 2008). In different neural network architectures in literature (Kitapcı et al., 2014; Alpaslan et al., 2012), 10%, 15%, 20%, and 30% have been implemented to time-series data as the length of the testing set. Seventy-two (72) and twenty-four (24) observations were used as training and testing sets respectively. This translates to 75% for the training set and 25% for the testing set. It has been observed in the literature that the smaller the test set length, the more accurate the forecasts are (Alpaslan et al., 2012).

Neural Network Model Architecture

The architecture of the neural network is determined by the number of hidden and output layers, the number of neurons in all of the layers, training algorithm parameters and the performance measure. There is no general rule for the selection of the appropriate number of hidden layers and the most popular approach in finding the optimal number of the hidden layer is by trial and error (Güler and Übeyli, 2005). The architecture of the neural network can be generalised by an expression 3.2:

$$I - (H_1, H_2, H_3, ..., H_n) - O, \tag{3.2}$$

where I represents the number of input nodes, H_n the number of neurons in hidden layer n, and O the number of neurons in the output layer. An example is an ANN with seven (7) input nodes, one hidden layer with three (3) neurons and one (1) output neuron can be represented as 7-(3)-1 respectively.

Neural Network Model Training

Training of a NN is a process of determining the weights of the neural network as well as the number of neurons in the different layers of the network. According to Prasad Y and Bhagwat (2002), this is an optimisation process whose aim is to find a set of optimum weights with which reasonable predictions can be made. Supervised training with resilient backpropagation as the training algorithm was used and demand calls of 2017 were set as the target values. Backpropagation, also called backward error propagation or backprop, is the most popular and widely used network learning algorithm (Sadiq et al., 2003; Rumelhart et al., 1985; Anderson, 1995; Duda et al., 2012). Backpropagation is a rule that generalises the gradient descent method as a way of changing the weights in the hidden layer of a FFNN. It gives the change $\Delta w_{jk}(i)$ in the weight of the connection between neurons j and k at iteration i as

$$\Delta w_{jk}(i) = -\alpha \frac{\partial E}{\partial w_{jk}(i)} + \mu \Delta w_{jk}(i-1), \tag{3.3}$$

where α is called the learning coefficient, $\Delta w_{jk}(i-1)$ the weight change in the immediately preceding iteration and μ the momentum coefficient (Güler and Übeyli, 2005).

The learning coefficient ensures a maximum decrease of the error function thus increasing the convergence speed. If it is too small, convergence will be extremely slow and if too large, the error function will not converge. The momentum coefficient tends to aid convergence as it works as a low-pass filter by reducing rapid fluctuations. It applies smoothed averaging to the change in weights whilst also avoiding local minima (Goutorbe et al., 2006).

The resilient backpropagation was used with the threshold value for the training data set at 0.01. The threshold or average error ensures that a network only converges only when the error reaches below the chosen threshold value (Rather et al., 2015). The learning rate factors were set at a minimum of 0.5 and maximum of 1.2 The momentum was not pre-set allowing the machine to adopt default values.

The selection of the initial number of hidden neurons was initially based on the logarithm of the training observations (Wanas et al., 1998), and the calculated value (log 84) was two. These were increased during training of the neural network however; a general rule is that there must not exceed two-thirds of the input neurons.

The logistic function and the linear function were implemented as the activation functions in the hidden and output layers respectively. Alpaslan et al. (2012) in their statistical approach in selecting activation functions recommended the use of the logistic activation function in the hidden layers and the linear function in the output layer. The results obtained accurate short-term forecasting horizons. A single output neuron with a linear activation function was applied to the neural network.

Neural Network Model Selection

The number of hidden layers and neurons in the selection process were systematically varied in order to obtain the most accurate models as proposed by Sadiq et al. (2003). It is important to note that neural networks without hidden units are equivalent to linear statistical forecasting methods (Cheng and Titterington, 1994). Hidden units perform the mapping between the input and output variables as well as to provide the non-linearity feature to neural networks, in addition to finding out patterns in the dataset (Aras and Kocakoç, 2016). The selection of the best three models among the neural networks was based on the performance measures namely: mean absolute error (MAE) and the residual mean square error (RMSE).

The MSE and the MAE are both measures of accuracy and the degree of spread of data points (Aras and Kocakoç, 2016). The MAE is a measurement of how close forecasts are to the actual data points; the average of the absolute errors (Herliansyah et al., 2017). Values predicted from the training sample and values in the test set were utilised to calculate the performance measures. Forecasts for the year 2018 were calculated by using the best weight values obtained in the selected model.

3.2.3 Multiplicative Seasonal Autoregressive Integrated Moving Average(SARIMA) Model

According to Rather et al. (2015), various linear models can be used for prediction including exponential smoothing models, generalised autoregressive conditional heteroscedasticity and stochastic volatility models, which are predominantly used in predicting stock returns. However, Jere et al. (2017) applied the Box-Jenkins methodology using autoregressive integrated moving average models (ARIMA) on a univariate time series in a study to forecast the second-hand car importation in Zambia and the methodology was found to be superior to exponential smoothing models. The Box-Jenkins methodology was adopted in this study.

ARIMA Model Building

When developing an ARIMA model using the Box-Jenkins methodology, four stages have to be followed: identification, estimation, diagnosis and forecasting. Components of time series data namely; seasonality, trend, cyclicity and irregularity, are key to model development. Seasonality refers to a consistent shape in the series that recurs with some periodic regularity. It can also be referred to as regular up and down fluctuation or short-term variations due to seasonal factors. Monthly ambulance calls data is likely to have a strong yearly component occurring at lags that are multiples of s = 12, because of strong relations of activities to the calendar year. This might lead to the adoption of autoregressive and moving average polynomials that identify with seasonal lags called multiplicative seasonal autoregressive integrated moving average models as proposed by Shumway and Stoffer (2000) given by ARIMA (p, q, d) x (P, D, Q)s.

Stationarity Assumption

When applying ARIMA models to time series data, it is assumed that the data is stationary. The stationarity assumption requires that the mean, variance and autocorrelation of the time series data does not change over time. Non-stationarity is common when components of time series data such as trend, seasonality, cyclicity and irregularity are not accounted for in the model building process. When time-series data is not stationary, differencing both the ordinary and seasonal components as presented in equation (2.31) is performed to ensure that stationarity is achieved.

Model Selection and Diagnosis

The Akaike Information Criterion (AIC) proposed by Shumway and Stoffer (2000) for time series data, was used to select the best model considering that the time series is stationary and that the model attains a minimum AIC value. Model diagnostic checks were implemented to check whether the residuals of the model resembled white noise. The plot of residuals against time was also applied to check for constant mean and variance assumption. A normal probability plot and the histogram plot were applied to test for the normality of the residuals.

The ACF and the PACF of the residuals were plotted to check for the independence of residuals. To complement the visual approaches, the Ljung-Box Pierce statistic as proposed by Shumway and Stoffer (2000) was used to further ascertain the independence of residuals.

3.2.4 Comparison of the Neural Network and SARIMA Models

A cross-validation study was conducted by selecting the best model from each of the FFNN and SARIMA models. The predicted values and the actual values were used to calculate the performance of each model using the mean absolute error (MAE) and the root mean square error (RMSE) as the performance measures. MAE gives equal weighting to differences from actual values whilst the RMSE apportions a huge penalty in these differences and is more appropriate in identifying outliers (Aras and Kocakoç, 2016). This makes the MAE a more appropriate tool for testing goodness of fit as compared to the RMSE. The paired sample t-test (at 5% level of significance) was carried out to validate any significant differences between the actual values and predicted values of each of the two models using the Minitab statistical package. For a model to be of good fit, there should not be significant differences between the actual and predicted values ($p > 0.05$). Using the selected superior model, forecasts for 2019 were carried out and inferential deductions made.

3.3 Simulation

The simulation model development will focus on four fundamental aspects namely: derivation of the main input parameters, development of the simulation model, analysis of the performance measures, and use of optimization for simulation to analyse the scenarios.

3.3.1 Description of the Bulawayo EMS Response System

The Bulawayo Emergency Medical Services (BEMS) adopted the regionalised response strategy where EMS teams are assigned to serve a pre-specified area or region. In this strategy, it is assumed that if the assigned EMS team(s) is busy, the closest team must perform the mission. The main advantage of this strategy is to minimise travel times due to the reduced size of the geographic area that the EMS teams need to travel between call locations.

Bulawayo is divided into two broad regions, the eastern and western regions, for emergency response purposes. The eastern region covers the low-density suburbs characterised by low population densities as compared to the western region characterised by high population densities. Both the Eastern and Western regions are further split into two subregions to which an ambulance base station is assigned. Currently there are four base stations, two in each region namely: Famona and Northend (Eastern region), Nketa and Nkulumane (Western region). The regions and their respective suburbs are summarised in Table 3.2.

Call Features

The study considered the geographical distribution of emergency calls of the four stations: Famona, Northend, Nketa and Nkulumane. The study will determine the distributions of arrival rates of calls for each substation using historical data of 2018 using the ARENA package.

Types of Emergency and Ambulance Requirements

The types of emergencies are broadly categorised into five(5) distinct categories with assigned unique codes for tracking, rescue team deployment and reporting purposes. These are summarised in Table 3.3. For simulation modelling purposess, the model will adopt the codes Cat A, Cat B and Cat C for distinguishing the different emergency response categories

Table 3.2: Regions, Stations and Areas of Coverage for BEMS.

Region	Station	Suburbs Covered
Eastern	Famona	City Centre, Hillside, Ascot, Willsgrove, Glencoe, Lochview Riverside, Buena Vista, Manningdale, Sunnighill, Woodlands, Waterford, Matsheumhlope, Fortune gate, Malindela, Bradford, Selbourne park, Whitecairns, Eloana, Fourwinds, Bradfield, Barham green, Hillcrest, Surburbs, Centenary park, Waterfalls, Morningside, Belmont, Ilanda, Montrose, Greenhill
	Northend	Pardonhust, Saurcetown, Tagela, Norwood, Richmond Northend, Windsor park, Glenville, Trenance, Highmount Parklands, Harrisvale, Quenspark West, Khumalo, Romney Park, Nortgate, Newmansford, Lobenvale, Kennilworth Sunnyside, Glengary, Woodville, Northlea, Quenspark East Makokoba, Thorngrove, Steeldale, Mzilikazi, Kingsdale Nguboyenja, Burombo flats, Barbourfields, Mahatshula
Western	Nketa	Matshobana, Mpopoma, Kelvin north, Entumbane, Njube, Lobengula, Iminyela, Mabutweni, Pelandaba, Luveve Enqotshweni, Enqameni, Gwabalanda, Pumula, Hyde park Magwegwe, Pumula, Cowdray Park, Emakhandeni
	Nkulumane	Nkulumane 1 to 12, Kelvin west and east, Tshabalala, Sizinda Tshabalala extension, Nketa 6 to 9, Emganwini, West Somerton Southworld, Sidojiwe hostel, Donnington, Newton west, Westgate,

as discussed in literature review.

Table 3.3: Description of Emergency Response Categories and Codes.

Simulation Code	Data Code	Description
Cat A: Urgent and life threatening	RTA	Road traffic accidents
	1A	Accident/Emergencies
Cat B: Urgent but not life threatening symptoms	1B	Maternity clinics
	2	Clinics from home
Cat C: Non-urgent calls	3	Removals/transfers

The BEMS is currently operating using the Anglo-American response strategy where the EMS is separated from the medical system as it offers only paramedic care. The BEMS uses different kinds of vehicles but fitted with the same equipment features to respond to emergency calls.

Ambulance dispatch, which is the act of choosing an appropriate EMS vehicle to respond based on the nature and location of call guided by set standard rules and guidelines is performed by a dispatcher upon receiving calls requiring EMS. Currently, BEMS is inclined to call-initiated dispatch decision-making strategy where the dispatcher is mandated to select one of the idle ambulance vehicles to be dispatched after the arrival of an emergency call. BEMS employs the first come first serve (FIFO) dispatch strategy with priority given to road traffic accidents in the case where waiting calls are in the response system. The BEMS assumed a multi-location dispatch model, where the ambulances may be dispatched from wherever they are.

When responding to calls, EMS crews are not given specific routes to follow as in the case of dynamic dispatch systems. In cases where an ambulance call is cancelled, it is recorded and normally such cases occur when there is a duplication of calls or the use of other emergency ambulance service providers by the caller. When responding, EMS medical crews can encounter: false and malicious calls (FAM), false alarm with good intent (FAGI) and true existence of a call. All these scenarios are reported and recorded during response. The EMS crew is expected to provide service at the scene, deliver a patient to a medical institution, perform hand-over and take-over procedures at medical centre, restocking and fueling of vehicle.

The study will assume a static ambulance deployment model which seeks to allocate a fixed number of ambulances to a set of known fixed base stations with the objective of ensuring that the best medical outcomes for patients are achieved. This will help determine the capacity and staffing of ambulance stations by optimising the number of ambulances needed to provide an efficient and effective service level for the set of existing ambulance stations with known locations. A logical presentation of the BEMS multi-location dispatch model is presented in Fig. 3.2.

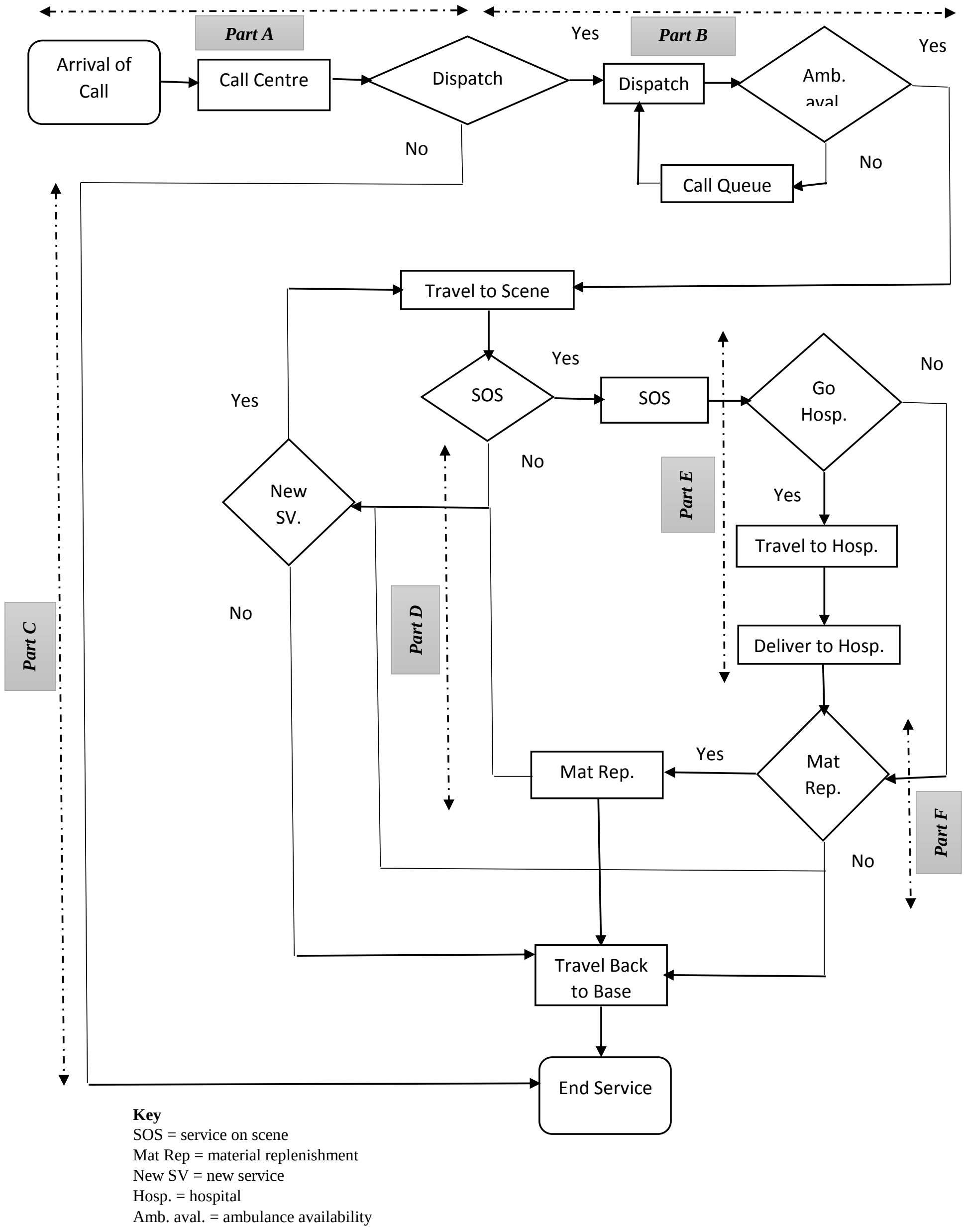


Figure 3.2: BEMS Multi-location Dispatch Model.

Part A of the dispatch model represents the call generation process whilst Part B represents the dispatch process. Part C represents cancelled calls which emanate from calls that do not require an ambulance response or occurs when the caller sort for another service provider or there has been a duplication of a call. Part D represents a case where service on the scene is not required. Normally these are calls recorded as a false malicious alarm (FAM) or a false alarm good intention (FAGI). Part E represents a case where service on the scene is required and the patient is transferred to the hospital. Part F represents the material replenishment process where a decision is made whether to replenish the medical resources or not. It also includes aspects of vehicle service or refueling.

3.3.2 Data Manipulation and Analysis

The simulation model to be developed will incorporate the randomness in call arrivals, response time and service time. The model will adhere to the assumptions listed in the next section which depicts the activities and protocols followed by the BEMS in responding to emergency calls as discussed.

Assumptions of the Simulation Model

- The arrival rate of calls may vary and is time-dependent. Calls are also related to the socio-economic conditions of the population.
- The calls are serviced as per first come first serve (FIFO).
- The ambulance could only serve one call at a time.
- The ambulances have the same capacity in terms of size and equipment and each ambulance team is made up of the driver and an attendant.
- Ambulances are to be allocated randomly.

- Response time is the time between the receipt of a call and to when the ambulance team arrive at the scene.

- Service time is the time between the arrival of the ambulance team at the scene until they have performed hand-over take-over at the medical centre and the vehicle is ready to depart for the station and available to perform another task.

- Total duration in the system is the time from when a call is received up to when the ambulance is ready to depart for the station ready to perform another task.

Estimation of Statistical Distributions of Simulation Model Parameters

The call inter-arrival time, response time, and service time distributions will be generated in the ARENA simulation package using the Input Analyser module on the 2018 historical data. Possible statistical distributions that can be fitted to the empirical distribution functions generated by these different sample data in ARENA include the Poisson, Normal, Exponential, Erlang, Gamma, Log-normal and Weibull distribution. The selection of the best distribution is based primarily on the square error (s.e) criterion given by equation 3.4.

$$s.e = \Sigma_{i=1}^{n}(f_i - f(x_i))^2, \tag{3.4}$$

where:

- n is the number of histogram intervals,

- f_i is the relative frequency of the data for the i^{th} interval, and

- $f(x_i)$ is the relative frequency of the fitted probability distribution function in the i^{th} interval.

Test for the goodness of fit is an additional feature of the Input Analyser module to ascertain the goodness of fit of the hypothetical distribution using both the Chi-square test and the Kolmogorov-Smirnov (KS) test. The Chi-square goodness of fit test is a non-parametric test used to find out how the observed value is significantly different from the expected value. The KS-test is also a non-parametric test that compares the sample data to a known distribution and assists in determining whether they have the same distribution. Both tests compare a known hypothetical (specified) probability distribution to the distribution generated by the sample data. For both the Chi-square and KS-test for the goodness of fit, in general, the higher the p-value the better the fit.

The Chi-square test for goodness of fit is guided by the following hypothesis:

H_0: The data comes from the specified distribution.

H_1: The data does not match the specified distribution.

The KS-test for goodness of fit is guided by the following hypothesis:

H_0: The data comes from the specified distribution.

H_1: At least one data value does not match the specified distribution.

Estimation of Proportions of Model Parameters

The monthly and daily occurrences of demand per station were computed from the forecasts data generated by the feed-forward neural network. Allocations to the different stations (Famona, Northend, Nketa and Nkulumane) will be based on proportions calculated from the historical data of 2018. The probability of occurrence of each medical condition/category (Cat A, Cat B and Cat C) shall be computed in Excel based on the 2018 annual historical data.

Simulation Model Development and Performance Measures

The simulation model will be developed using the ARENA simulation package. The performance measures considered for the simulation models are the average entity time in system, average response time, average response queue time, the average number of calls in the response queue, and capacity utilization of ambulances. These were also generated using the ARENA package.

3.3.3 Sensitivity Analysis

Sensitivity analysis entails looking at different scenarios and observing how they affect the overall model performance. It involves changes to model parameters and subsequently observing how these changes affect the general model performance. It helps to answer pertinent questions or concerns post-analysis. The research will explore the following scenarios as part of simulation model development sensitivity analysis.

- Optimum static ambulance deployment maintaining response time distributions (RTD).
- Optimum static ambulance deployment to predicted ANN demand, maintaining the RTD.

3.3.4 Numerical Experiments

Numerical experiments are focused on evaluating different scenarios and observing how they affect the overall model performance. It also involves changing model parameters and subsequently observing how these changes affect the general model performance. It helps to answer important questions and concerns post analysis and their managerial implications. The research will explore the numerical experiments as part of simulation model development.

- The influence of standardising the response time on the optimal ambulance deployment plan.

When analysing the different scenarios, special consideration will be given to the simulation model performance measures as proposed in literature.

3.4 Chapter Conclusion

Chapter 3 focused on the research methodology which was anchored on the development of forecasting and simulation models and subsequently on sensitivity analysis of developed models. The next chapter, Chapter 4 will focus on the development of forecasting models using both the artificial neural network and SARIMA models.

Chapter 4

Forecasting Models: ANN and SARIMA Models

4.1 Introduction

Chapter 4 focuses on the development of forecasting models for univariate time series of ambulance demand for short-term forecasting horizons using artificial neural networks (ANN) and seasonal autoregressive integrated moving average (SARIMA) models. A comparative analysis was conducted to select the best model with high predictive power of the public emergency ambulance demand (PEAD) using monthly historical data from 2010 to 2017 and compared against observed data of 2018. Based on the selected model, short-term annual PEAD forecasts for 2019 were predicted. Results of ANN models are presented first followed by results of the SARIMA models and finally the best model is selected and used for prediction.

4.2 The Demand Time Series Plot

A time series plot of the monthly public ambulance demand is summarized in Fig. 4.1. Chatfield, (1996) defined trend as a long-term change in the mean level and this trend is evident in the time series data. Differencing the time series data would help achieve stationarity when using ARIMA models. There are also seasonal patterns that recur after every multiple of lags of twelve (12), hence there is an expectation of including seasonality in the model. The random components are generally insignificant. The neural network models however, do not require any prior conditions such as linearity or stationarity of the time series data.

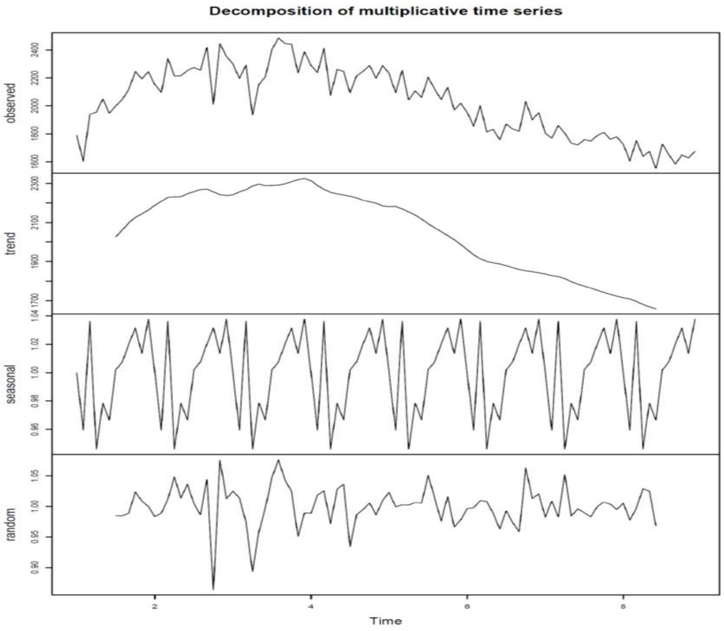


Figure 4.1: The Public Emergency Ambulance Time Series Plot.

4.2.1 Selecting the Best Neural Network Model

Several models of different architectures were systematically selected starting with two hidden units (log 84 = 2) in a hidden layer and gradually increasing them. The models were predominantly divided into two distinct sets, one with a single hidden layer and the other with two hidden layers respectively. The selection of the number of inputs in the model was based on trial and error (Belayneh et al., 2014). The mean square error (MSE) and mean absolute error (MAE) were used as the performance measures during training. Three models were selected and forecasts were generated for the year 2018 as a model cross-validation process. The RMSE and MAE were used as final performance measures for selecting a suitable model for the neural network and the results are summarized in Table 4.1 and Table 4.2 respectively.

Table 4.1: Feed Forward Neural Network Model Selection.

		Testing set	Testing set	Validation	Validation
Model	Structure	(MSE)	(MAE)	(RMSE)	(MAE)
1	7-(3)-1	268.14	5.29	165.28	114.54
2	7-(3,2)-1	169.41	3.26	138.20	108.08
3	7-(4)-1	402.18	6.26	137.19*	94.00*

Note:* is the minimum value of the performance measure across all models.

The architecture of the FFNN $(7-(4)-1)$ with seven input nodes, one hidden layer (4 neurons) and one output neuron was the best model with the lowest MAE of 94.0 and RMSE of 137.19. The neural network topology is presented in Figure 4.2.

4.2.2 Multiplicative SARIMA Model Selection

Several models were considered by the algorithm in the R-package but the $ARIMA(0,1,1)(0,0,2)_{12}$ model was the best according to the Akaike Information Criteria (AIC) with a value $AIC = 1162.57$ as presented in Table 4.3.

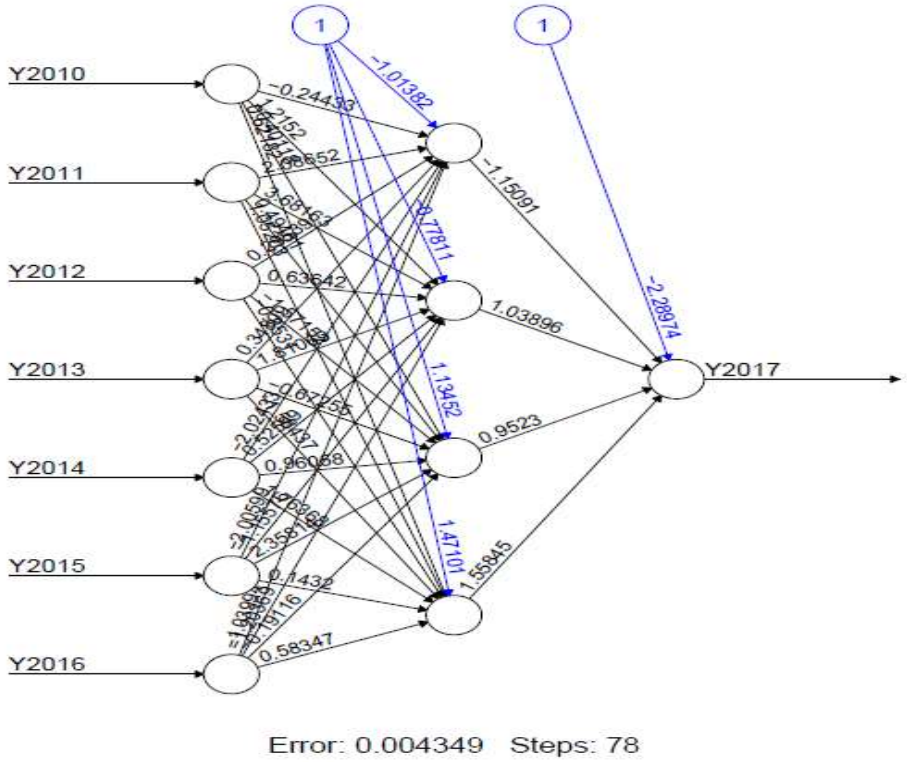


Figure 4.2: Visualisation of the Architecture of the FFNN (7-(4)-1).

Table 4.2: Forecast Data Generated by the Three Neural Network Models.

	Actual	7-(3)-1	7-(3,2)-1	7-(4)-1
Month	(2018)	FFNN Model 1	FFNN Model 2	FFNN Model 3
January	1569	1613	1613	1623
February	1452	1481	1498	1500
March	1684	1659	1656	1668
April	1477	1499	1409	1407
May	1440	1291	1266	1417
June	1470	1098	1162	1268
July	1612	1563	1547	1588
August	1515	1444	1476	1462
September	1543	1563	1569	1539
October	1330	1673	1696	1686
November	1427	1608	1594	1617
December	1564	1633	1650	1652
	MAE	114.54	108.08	94.00*
	RMSE	165.28	138.20	137.19*

Note:* is the minimum value of the performance measure across all models.

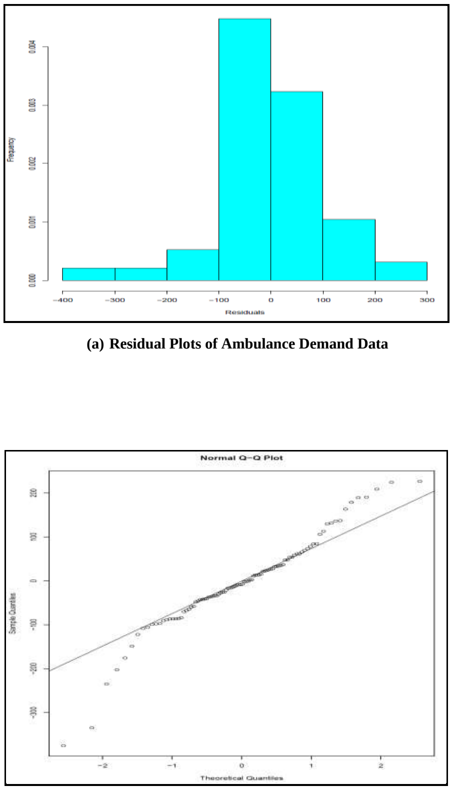
Table 4.3: Summary of $ARIMA(0,1,1)(0,0,2)_{12}$ Model Estimation Results.

Parameter	MA 1	SMA 1	SMA 2	Variance	Log. likelihood	AIC
Coefficient	-0.5767	0.1953	0.4011	10900	-577.28	1162.57
s.e.	0.0871	0.1023	0.1447			

SARIMA Model Diagnosis

The histogram of the residuals (Fig. 4.3) showed that the data is normally distributed with a mean zero. The normal probability Q-Q plot of residuals (Fig. 4.3) showed a linear trend to confirm that the residuals are normally distributed. When visualizing the straight line, more emphasis was given to the central values of the plot than on the extremes (Montgomery and Runger, 2014). The results are sufficient to affirm that the normality assumption is satisfied.

The plots of the ACF and PACF of residuals resembled an insignificant number of the spikes being significantly different from zero (Fig. 4.4), as it is expected that approximately $\frac{1}{20}$ to be above $\pm\frac{2}{\sqrt{n}}$. The Ljung-Box test statistic (at 5% level of significance) value of



(a) Residual Plots of Ambulance Demand Data

(b) The Normal Q-Q plot of Residuals

Figure 4.3: Test for Normality of Residuals of SARIMA Model.

$p = 0.6118 (> 0.05)$ calls for the failure to reject the null hypothesis and hence, conclude that the residuals are not correlated. This validates the requirement that the residuals must be independent.

Plotting residuals against time (order of the data) does the test for homogeneity of the mean and variance of the residuals. This is expected to show any changes in the mean and variance with time. The plot of the residuals against order of the data (Fig. 4.5) shows that the mean of the residuals varies closely to zero with a relatively constant variance as expected.

4.2.3 Comparison of the Neural Network and SARIMA Models

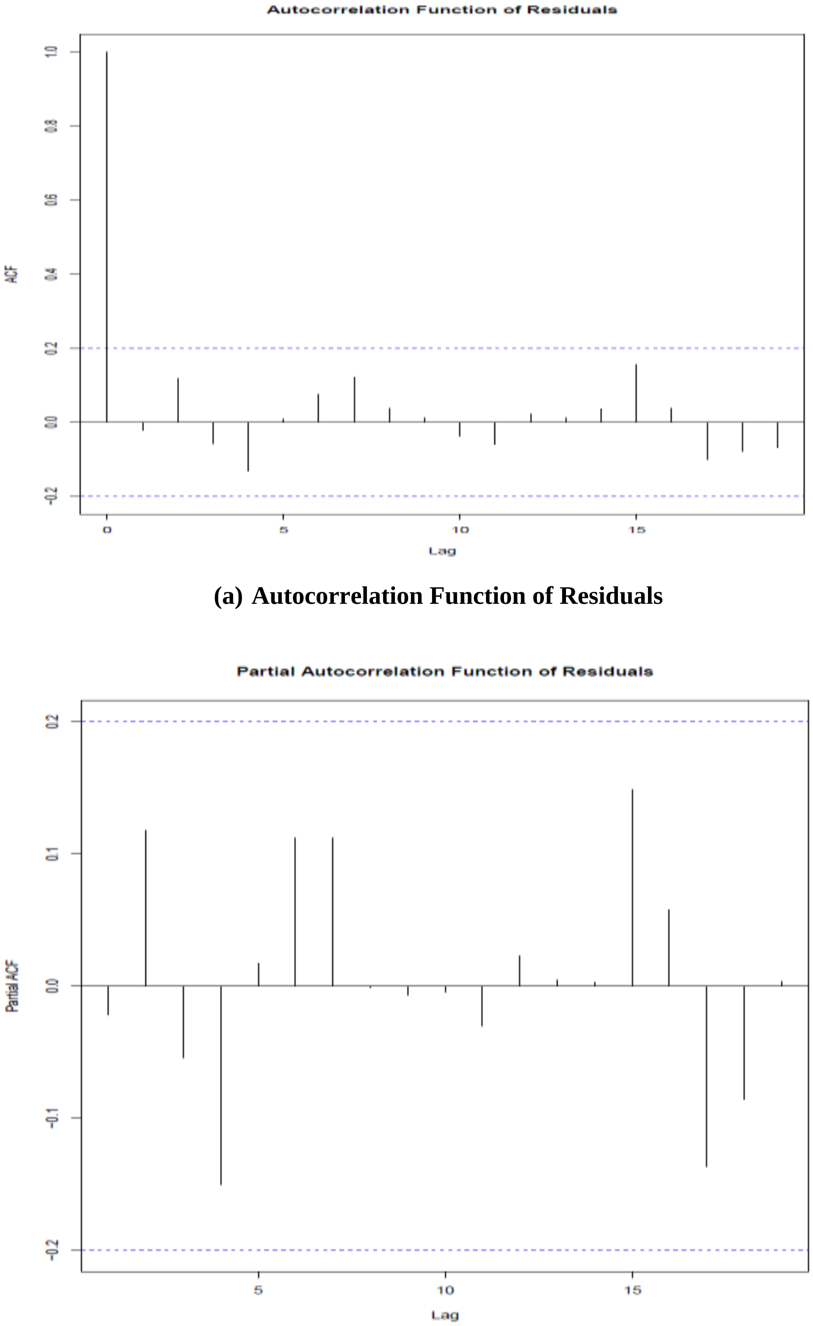
The actual monthly ambulance demand data for 2018 was compared with the forecasts using the two selected models as presented in Table 4.4.

Table 4.4: Forecasts of Ambulance Demand by FFNN and SARIMA Models.

Month	Actual (2018)	SARIMA	FFNN
January	1569	1605	1623
February	1452	1595	1500
March	1684	1624	1668
April	1477	1601	1407
May	1440	1593	1417
June	1470	1580	1268
July	1612	1587	1588
August	1515	1596	1462
September	1543	1605	1539
October	1330	1600	1686
November	1427	1589	1617
December	1564	1606	1652
	MAE	105.71	94.00∗
	RMSE	125.28∗	137.19

Note:* is the minimum value of the performance measure across all models.

It can be observed in Fig. 4.6 that the pattern of the FFNN model is inclined to value



(a) Autocorrelation Function of Residuals

(b) Partial Autocorrelation Function of Residuals

Figure 4.4: Test for Independence of Residuals of SARIMA Model.

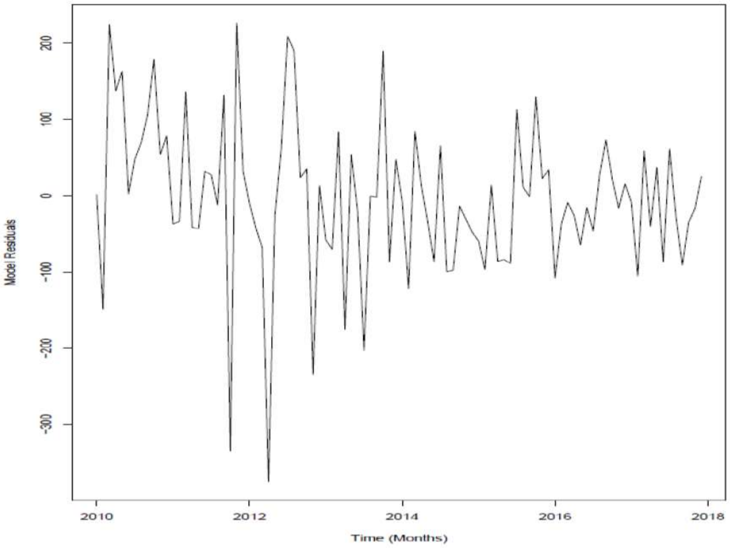


Figure 4.5: Homogeneity Test for the Mean and Variance of Residuals of SARIMA Model.

estimation as compared to SARIMA which is directional as depicted by the linear pattern over time. The FFNN was able to detect the hidden patterns of the time series data. These findings concur with findings from a study carried out by Adebiyi et al. (2014) for stock price predictions.

In terms of the MAE the FFNN is superior to the SARIMA model. When considering the RMSE, the SARIMA model is superior to the FFNN. However, it must be noted from literature (Aras and Kocakoç, 2016), that the MAE gives equal weighting to differences from actual values whilst the RMSE apportions a huge penalty in these differences and is more appropriate in identifying outliers. This makes the MAE an appropriate tool for the goodness of fit.

Paired sample t-tests at 5% level of significance for the actual values and predicted values were applied to both models. The calculated p-value for FFNN was 0.493(> 0.05) (Table

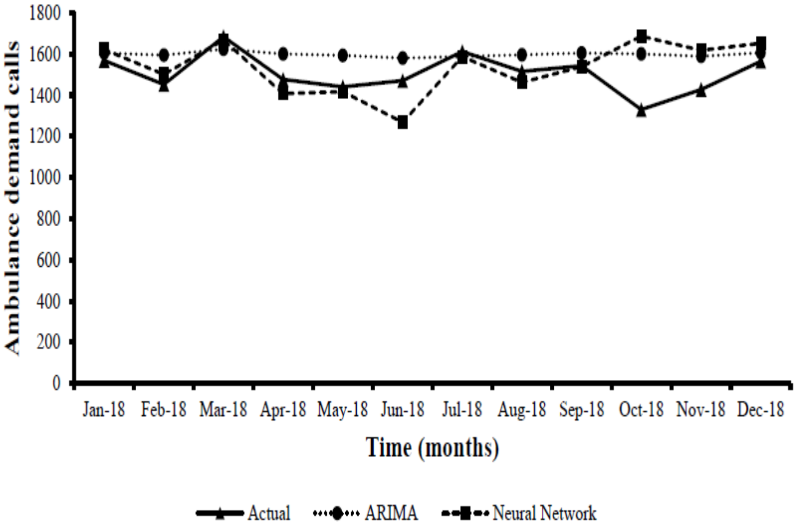


Figure 4.6: Comparison of Actual and Predicted Values of Time Series Models.

4.5) and we conclude that there is no significant difference between the forecast and actual values of public emergency ambulance demand. The calculated p-value for the SARIMA model was 0.005(< 0.05) (Table 4.6) and we conclude that there is significant difference between the forecast and actual values of public emergency ambulance demand. The results resonate with findings in the literature that propounded that neural network models are superior to ARIMA models in time series prediction (Adebiyi et al., 2014).

Table 4.5: Paired Sample T-test Results: Actual versus FFNN Forecast of 2018 Demand.

	N	Mean	s.d.	S.E. Mean	t- value	p- value	95% C.I.
Actual	12	1506.92	94.54	27.29	-0.71	0.493	(-177.67,60.33)
FFNN	12	1535.58	127.78	36.89			
Difference	12	-28.67	140.07	40.44			

Table 4.6: Paired Sample T-test Results: Actual versus SARIMA Forecast of 2018 Demand.

	N	Mean	s.d.	S.E. Mean	t-value	p-value	95% C.I.
Actual	12	1506.92	94.54	27.29	-3.55	0.005	(-148.23,-34.77)
SARIMA	12	1598.42	11.33	3.27			
Difference	12	-91.5	89.29	25.78			

4.2.4 Public Emergency Ambulance Demand Forecast for 2019

The selected neural network model was used to forecast the public emergency ambulance demand for 2019 as shown in Figure 4.7. The demand is expected to peak in January, March, September and December 2019. Lower demand is projected in April, June and July.

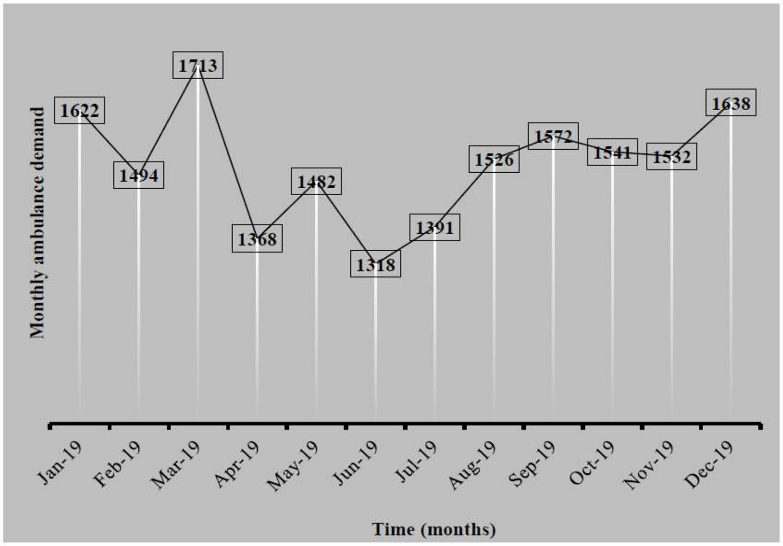


Figure 4.7: Public Emergency Ambulance Demand Forecast for 2019 Using FFNN.

Even though models that forecast weekly, daily, and hourly demands were not developed at this level, such important quantitative measures can be derived from such forecasts and be fully utilised for strategic planning purposes to ensure that there is adequate preparedness to respond to ambulance calls and a summary is given in Table 4.7 and Fig. 4.8.

Table 4.7: Monthly, Weekly and Daily Projected Number of Calls for 2019.

Year (2019)	Monthly number of calls	Number of days in a month	Weekly demand	Daily demand
January	1622	31	406	53
February	1494	28	374	54
March	1713	31	429	56
April	1368	30	342	46
May	1482	31	371	48
June	1318	30	330	44
July	1391	31	348	45
August	1526	31	382	50
September	1572	30	393	53
October	1541	31	386	50
November	1532	30	383	52
December	1638	31	410	53

4.3 Chapter Conclusion

Performance measures; MAE and the paired sample t-test indicate that the FFNN models are superior to traditional SARIMA models in time series prediction of ambulance demand in the city of Bulawayo, over a short-term forecasting horizon. It was found that the FFNN model is more inclined to value estimation as compared to the SARIMA model, which is directional as depicted by the linear pattern over time.

Therefore, the FFNN model derives its model prediction accuracy from this unique characteristic. The FFNN model building process used 96 observations, where seventy-two (72) and twenty (24) observations were used as training and testing sets respectively. With such small sample data, it can be concluded that the FFNN model is able to accomplish accurate predictions as it was able to detect the hidden patterns of the time series better than the SARIMA model. The SARIMA model required the assumption of stationarity to be

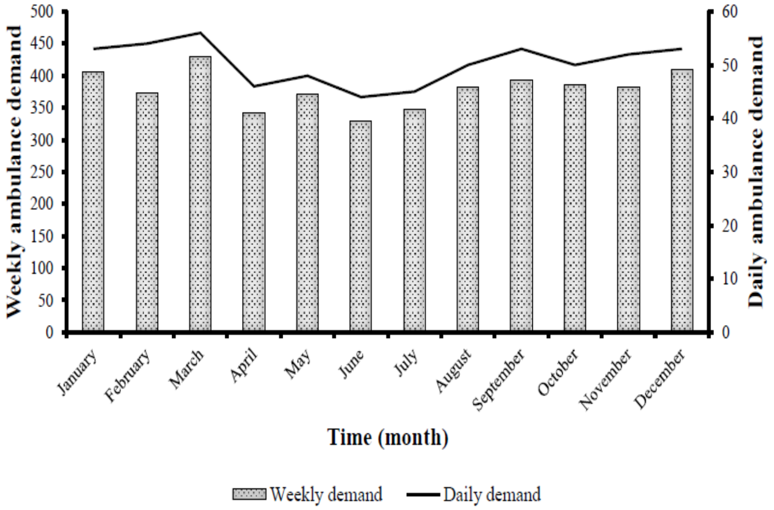


Figure 4.8: Projected Weekly and Daily Number of Ambulance Calls for 2019.

satisfied before model building and also to be verified post model development. However, such assumptions are not required in the development of FFNN. Therefore the FFNN is a parsimonious, simple model with no assumptions and requires fewer variables in the model building but with greater explanatory power as compared to the SARIMA model.

It was observed that using the architecture of one hidden layer produced more accurate results than those obtained from architectures with two hidden layers for short-term forecasting horizons. Therefore, the use of a single hidden layer is adequate in developing FFNN for short-term forecasting horizons. Reducing the number of input neurons did not improve the model accuracy. The researchers recommend that Bulawayo City Council should deliberately adopt and integrate such forecasting tools to assist them in their strategic resource planning activities.

Based on the performance measures, the parsimonious FFNN model was selected to predict short-term annual ambulance demand. Demand forecasts with FFNN for 2019 reflected the expected general trends in Bulawayo. The forecasts indicate high demand during the months of January, March, September and December. These four non-consecutive months are often characterised by public holidays and wet weather conditions resulting in high demand for public ambulance service demand. Key ambulance logistic activities such as vehicle servicing, replenishment of essential equipment and drugs, staff training, leave days scheduling and mock drills need to be planned for April, June and July when low demand is anticipated. This deliberate planning strategy would avoid a dire situation whereby ambulances are available but without adequate staff, essential drugs and equipment to respond to public emergency calls (or vice versa).

Chapter 5

Simulation Modelling for Heterogeneous Regions

5.1 Introduction

In this chapter, simulation modelling is performed to determine the appropriate model for static ambulance deployment for the heterogeneous regions using the prevailing service delivery levels. No attempt was made to find an optimal solution which will be covered in the next chapter using sensitivity analysis and optimisation techniques. The chapter focuses on the development of simulation models that incorporate the randomness that occurs in the inter-arrival of calls, the response time (delay and travel time) and service time (delay on scene and travel to health institution or base station). These random aspects of EMS parameters indirectly incorporate the uncertainty in the ambulance availability in determining the overall response time. According to Ingolfsson et al. (2008), models that do not consider and incorporate the uncertainty of response time, service time and ambulance availability are likely to overestimate or underestimate the number of possible ambulances required to provide specific service levels. In cases where the ambulance is not available, it usually manifests as queuing delays as the ambulances will be busy responding to other calls. It was a common feature that a backup is offered from the nearest sub-station or that fire fighting vehicle or in other cases a supervisors' vehicle(s) are used to respond due to unavailability of ambulances at the specific sub-station offering EMS. Random and non-random model input parameters required for simulation model development included among them, the response time distributions, service time distributions and proportion of occurrence of different emergency categories. Performance measures such as the average entity time in the system, average response time, average response queue time, the average number of calls in the response queue, throughput ratio and the ambulance utility levels were used to evaluate the models.

The chapter also considered that the geographical location of a sub-station has an indirect influence on the overall performance of service delivery as influenced by the randomness nature of inter-arrivals of calls, response time, service time and the prevalence of different levels of the urgency of calls. The service time depends on the traveling distances in the geographical area that the sub-station covers, the nature and density of roads, and the distance to the nearest hospital. Hence, sub-stations were considered separately to ensure the robustness of the simulation models. Leknes et al. (2017) presented almost a similar EMS problem which focused on an ambulance station location and allocation problem which they referred to as the Maximum Expected Location Problem for Heterogeneous Regions (MEPLP-HR). In their case they formulated a mixed-integer linear program to solve the problem. In our case we adopted simulation model techniques.

5.2 Estimation of Simulation Model Input Parameters

Call inter-arrival time, response time, service on scene delay time (SOS) and no-service on scene required delay time (NSOS) distributions were generated in the ARENA simulation package using the 2018 historical data. A summary of the results is presented in Table 5.1.

Table 5.1: Simulation Model Parameter Distributions for the Heterogeneous Sub-stations.

Famona	*Inter-arrival time*	*Response time*
	0.999+WEIB(180;1.17)	2+GAMM(22;1.48)
	Service on scene delay	*No-service on scene delay*
	-0.001+164*BETA(2.7;6.47)	-0.5+72*BETA(0.606;1.2)
Northend	*Inter-arrival time*	*Response time*
	3+GAMM(143;1.33)	2+GAMM(23.9;1.36)
	Service on scene delay	*No-service on scene delay*
	N(51.6;23.7)	-0.001+WEIB(25.9;0.834)
Nketa	*Inter-arrival time*	*Response time*
	-0.001+WEIB(64;1.06)	-0.001+ERLA(18.5;2)
	Service on scene delay	*No-service on scene delay*
	2+201*BETA(3.81;10.9)	-0.001+EXPO(23.6)
Nkulumane	*Inter-arrival time*	*Response time*
	0.999+GAMM(93.6;1.73)	0.999+GAMM(21.8;1.62)
	Service on scene delay	*No-service on scene delay*
	N(53;19.8)	-0.5+63*BETA(0.484;0.79)

It was also necessary to determine the proportion of emergency calls and non-emergency calls. The emergency calls are those that required the dispatch of an ambulance after being assessed by the dispatcher in the call centre. The non-emergency calls included cancelled calls and those that were attended to by other private emergency service providers. Global values of these parameters were calculated for all the four sub-stations as presented in Table 5.2.

Table 5.2: Summary of Model Input Parameters for the Heterogeneous Sub-stations.

Item	Parameter	Frequency	Proportion	% Frequency
Call filter	Emergency calls	16648	0.92	92%
	Non-emergency calls	1435	0.08	8%
	Total	18083	1.00	100%

Calls required to be categorised into three categories namely: category A (Road traffic accidents/emergencies), category B (Maternity clinics/maternity clinics from home) and category C (Removals/transfers) together with their probability of occurrences. It was observed from the data that these vary from one sub-station to another and hence were computed separately for each sub-station. The service on scene delay (SOS) and no service on scene required delay (NSOS) proportions of occurrence was also computed and the statistics are summarised in Table 5.3. The NSOS proportions were derived from the false alarm good intent (FAGI) and false alarm malicious (FAM) calls as these usually result in fewer time delays as compared to cases where service on the scene is required and rendered to the patient(s).

Table 5.3: Nature of Service and Call Category Classification Proportions by Sub-station.

Station	SOS	NSOS	Total	Cat A	Cat B	Cat C	Total
Famona	0.84	0.16	1.0	0.69	0.26	0.05	1.0
Northend	0.84	0.16	1.0	0.58	0.34	0.08	1.0
Nketa	0.93	0.07	1.0	0.56	0.37	0.07	1.0
Nkulumane	0.94	0.06	1.0	0.62	0.36	0.02	1.0

The NSOS emergencies emanate from the FAM and FAGI where the general service time is smaller as compared to cases where SOS is required and rendered. Proportions of the SOS emergencies are seemingly higher in the western suburbs (Nketa and Nkulumane) as compared to the eastern suburbs (Famona and Northend).

5.3 Developing Simulation Models

Simulation models are to be developed taking consideration of the model parameters developed in the last section whilst performance measurements are recorded from the ARENA simulation package. In developing the simulation model, the number of ambulances was incremented from one (1) to the stipulated number of allocated ambulances to each sub-station whilst changes in performance measures namely: average time of a call in system, average response time, average number of calls in response queue, average queue time, throughput ratio and ambulance utility levels were being observed.

The throughput ratio, represents the number of emergency ambulance calls that are served divided by the calls generated for the 24 hour day period and is expressed as a fraction. A fraction given by $\frac{10}{15}$, would imply that of the fifteen(15) received emergency calls, ten(10) were served up to the point where a patient was delivered at a health institution during the 24 hour working period. The remaining five(5) calls were still within the system awaiting some form of service. According to the official reports from the department of the fire brigade, Famona was allocated one(1), Northend one(1), Nketa three(3) and Nkulumane one(1) ambulance(s) respectively. A schematic diagram of the simulation model in ARENA is presented in Figure 5.1.

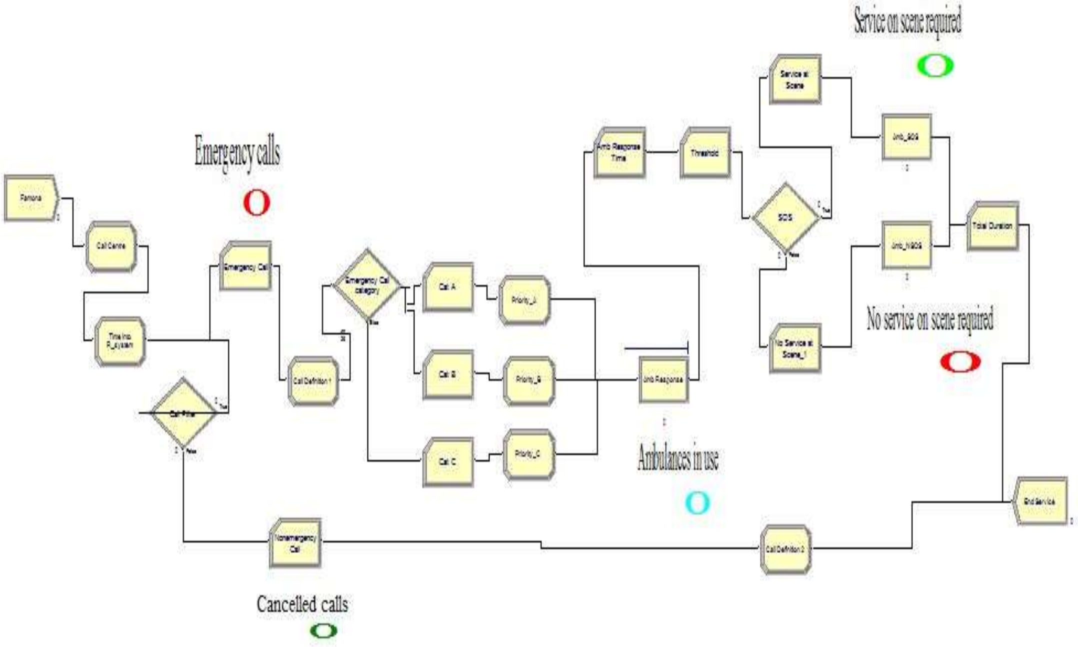


Figure 5.1: A Schematic Diagram of a Simulation Model in ARENA.

5.3.1 Simulation Model Building for Famona Station

Famona station was allocated one(1) ambulance to service its allocated geographical area. A summary of simulation modelling results is presented in Table 5.4.

Table 5.4: Simulation Model Performance Measures for Famona Station.

Number of Ambulances	1
Average time in system (minutes)	86.15
Average response time (minutes)	40.51
Average number in response queue	0.04
Average queue time (minutes)	6.58
Throughput ratio	$\frac{8}{9}$=0.89
Ambulance utility (%)	48

On average a single emergency call is queuing for an ambulance for an average time of approximately seven(7) minutes. The ambulance response time is averaging at approximately 41 minutes, while it takes approximately 86 minutes to completely service an emergency call. The single ambulance allocated has an utility of 48% of its allocated day-time which implies that it is busy 48% of the day. The throughput ratio stands at 89%.

5.3.2 Simulation Model Building for Northend Station

Northend station was allocated one(1) ambulance to service its allocated geographical area. A summary of simulation modelling results is presented in Table 5.5.

Table 5.5: Simulation Model Performance Measures for Northend Station.

Number of Ambulances	1
Average time in system (minutes)	102.74
Average response time (minutes)	58.70
Average number in response queue	0.19
Average queue time (minutes)	24.99
Throughput ratio	$\frac{11}{11}$=1.00
Ambulance utility (%)	60

On average an emergency call in Northend station's geographical area of service is queuing for approximately twenty-five(25) minutes. The ambulance response time is averaging at approximately 59 minutes and it takes approximately 103 minutes to completely service an emergency call. The single ambulance allocated has a utility level of 60% which implies that it is busy 60% of the day. The throughput ratio stands at 100%.

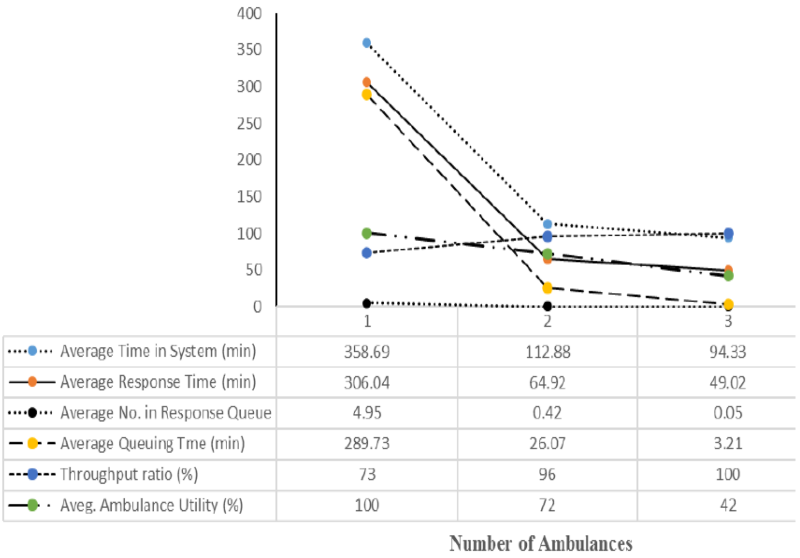
5.3.3 Simulation Model Building for Nketa Station

According to the official reports from the department of fire brigade, Nketa station was allocated three(3) ambulances to service its allocated geographical area. A summary of simulation modelling results is presented in Table 5.6.

Table 5.6: Simulation Model Performance Measures for Nketa Station.

Number of Ambulances	1	2	3
Average time in system (minutes)	358.69	112.88	94.33
Average response time (minutes)	306.04	64.92	49.02
Average number in response queue	4.95	0.42	0.05
Average queue time (minutes)	289.73	26.07	3.21
Throughput ratio	$\frac{16}{22}$=0.73	$\frac{22}{23}$=0.96	$\frac{24}{24}$=1.00
Ambulance utility (%)			
Ambulance 1	100	78	54
Ambulance 2		66	45
Ambulance 3			53
Average ambulance utility (%)	100	72	42

Results for the Nketa station indicate that as the number of ambulances increases, there is a corresponding improvement in the performance measures as submitted in Table 5.6. The average time that an emergency call spends in the system decreases as the number of allocated ambulances increases. The response time, the number of calls in the response queue and the corresponding average time in the queue also decreases as the number of ambulance allocation increases. The throughput ratio, increases with an increase in allocated ambulances. However, the ambulance utilisation levels decrease as the number of allocations increases. These general insights are presented in Figure 5.2.



	1	2	3
Average Time in System (min)	358.69	112.88	94.33
Average Response Time (min)	306.04	64.92	49.02
Average No. in Response Queue	4.95	0.42	0.05
Average Queuing Tme (min)	289.73	26.07	3.21
Throughput ratio (%)	73	96	100
Aveg. Ambulance Utility (%)	100	72	42

Figure 5.2: Influence of Varying Fleet Size on Performance Measures.

For the three(3) allocated ambulances for Nketa station, there is an average response time of approximately 49 minutes with an average queue size of one(1) ambulance. The resulting average queuing time of approximately three(3) minutes was observed. With the three allocated ambulances, the ambulance utilisation stands at 42% with a throughput ratio of 100% implying that all calls are responded to and attended to a point where appropriate medical attention is provided for the 1440 minutes (24 hours) day.

5.3.4 Simulation Model Building for Nkulumane Station

Nkulumane station was allocated one(1) ambulance to service its allocated geographical area. A summary of simulation modelling results is presented in Table 5.7.

An emergency call in the Nkulumane station's geographical area of service with a single

Table 5.7: Simulation Model Performance Measures for Nkulumane Station.

Number of Ambulances	1
Average time in system (minutes)	97.81
Average response time (minutes)	43.21
Average number in response queue	0.02
Average queue time (minutes)	3.67
Throughput ratio	$\frac{7}{7}$=1.00
Ambulance utility (%)	46

ambulance allocated has an average queuing time of approximately 4 minutes. The ambulance response time is averaging at approximately 43 minutes and it takes approximately 98 minutes to completely service an emergency call. The single ambulance allocated has a utility level of 46% which implies that it is busy 46% of the day. The throughput ratio stands at 100%.

5.4 Chapter Conclusion

The chapter focused on developing simulation models for the four sub-stations established to service-specific geographical areas. Model input parameters such as inter-arrival time distributions, response time distributions, service time distributions, proportions of occurrence of emergency calls and emergency categories were computed using historical data. The models developed are adequately mimicking the prevailing EMS process for Bulawayo city. It was observed that the number of false alarm malicious (FAM) and false alarm good intent (FAGI) calls are more prevalent in the eastern suburbs (Famona and Northend) as compared to their counterparts in the western suburbs (Nketa and Nkulumane). This might imply that eastern suburb residents find themselves with a wide range of alternatives for health emergencies resulting in more cases of false alarm with good intent (FAGI). Model performance measures were observed and recorded, and among them included average time a call spends in the system, average response time per entity, average response queue time, average number of calls in response queue and ambulance utility.

Results from Nketa station which was allocated more than one ambulance indicate that as the number of ambulance sizes increases, there is a corresponding improvement in the performance measures. The average time that an emergency call spends in the system decreases as the number of allocated ambulances increases. The response time, the number of calls in the response queue and the corresponding average time in the queue also decrease as the number of ambulance sizes increases. The throughput ratio, increases with an increase in allocated ambulances. The ambulance utilisation levels decrease as the ambulance fleet size increases.

The average response times are relatively high in comparison to international standards of 5 to 10 minutes as stipulated in the literature. Average queuing times and number of ambulances queuing are significantly high and undesirable in terms of service delivery as they have a negative bearing on human-based outcomes of safety and satisfaction. Safety in terms of the chances of survival and satisfaction in terms of quality of service delivery across the EMS response cycle. The general expectation is that no call should queue for service. Hence, there is a need to determine the optimum ambulance deployment models that minimises the number of ambulances needed to provide a specific service level. The next chapter seeks to address the issue by adopting optimisation for simulation through the use of sensitivity analysis.

Chapter 6

Sensitivity Analysis: Optimisation for Simulation

6.1 Introduction

This chapter focuses on finding optimal fleet size by applying the optimisation for simulation technique to analyse the different model scenarios based on the performance measures. Sensitivity analysis on the developed models was conducted in order to investigate various scenarios in the build-up to public emergency preparedness. The chapter also focuses on integrating the ANN public emergency ambulance demand (PEAD) forecasts, whilst adjusting the ambulance fleet sizes in order to optimise the levels of preparedness. The objective is to evaluate whether the adopted deployment plan is adequate to meet future demand levels as predicted by ANN. Multiple performance measures such as average entity time in the system, average response time, average response queue time, the average number of calls in the response queue, throughput ratios, and the ambulance utility levels will be considered to evaluate the models.

6.2 Optimum Static Ambulance Deployment Maintaining RTD

This section employs sensitivity analysis to explore the influence of increasing the ambulance fleet size whilst maintaining the same response time distributions for all the sub-stations. Six performance measures namely: average entity time in the system, average response time, average response queue time, the average number of calls in the response queue, throughput ratios, and the ambulance utility levels were used to identify the optimal static ambulance deployment.

6.2.1 Ambulance Deployment Plan for Famona Station

Famona station was initially allocated one(1) ambulance to service its allocated geographical area. Increasing the ambulance fleet size from one(1) to two(2) resulted in positive significant improvements of all the performance measures except for the utilization levels which dropped from 48% to 24% as presented in Table 6.1. Further increase in fleet size from two(2) to three(3) did not result in changes of all the performance measures except for the utilisation levels which decreased from 24% to 15% respectively. Therefore two(2) ambulances result as the optimal static ambulance deployment when maintaining the system predetermined response time distribution.

This optimal static deployment of two(2) ambulances for Famona station will result in an average response time of 32 minutes, with no queuing calls and an average call time in the system of 73 minutes. The two ambulances are expected to be busy 24% of the 24 hour shift time. Eighty percent of the calls received in the shift will be serviced right up to the process end depending on the needs of the patient.

Table 6.1: Simulation Model Performance Measures for Famona Station.

Number of Ambulances	1	**2**	3
Average time in system (minutes)	86.15	**73.06**	73.06
Average response time (minutes)	40.51	**32.29**	32.29
Average number in response queue	0.04	**0.0**	0.0
Average queue time (minutes)	6.58	**0.0**	0.0
Throughput ratio	$\frac{8}{9}$=0.89	$\frac{8}{10}$=**0.80**	$\frac{8}{10}$=0.80
Ambulance utility (%)			
Ambulance 1	48	**31**	17
Ambulance 2		**16**	19
Ambulance 3			10
Average ambulance utility (%)	48	**24**	15

6.2.2 Ambulance Deployment Plan for Northend Station

Northend station was initially allocated one(1) ambulance to service its allocated geographical area. Increasing the ambulance fleet size from one(1) to two(2) resulted in positive significant improvements of all the performance measures except for the utilization levels which decreased from 60% to 19% as presented in Table 6.2. Further increase in fleet size from two(2) to three(3) did not result in changes of all the performance measures except for the utilization levels which decreased from 19% to 13% respectively. Therefore two(2) ambulances result as the optimal static ambulance deployment when maintaining the system predetermined response time distribution.

This optimal static deployment of two(2) ambulances for Northend station will result in an average response time of 44 minutes, with no queuing calls and an average call time in the system of 79 minutes. The two ambulances are expected to be busy 19% of the 24 hour shift time. All of the calls received in the 24-hour shift will be serviced right up to the process end depending on the needs of the patient.

Table 6.2: Simulation Model Performance Measures for Northend Station.

Number of Ambulances	1	**2**	3
Average time in system (minutes)	102.74	**78.89**	78.89
Average response time (minutes)	58.70	**44.39**	44.39
Average number in response queue	0.19	**0.0**	0.0
Average queue time (minutes)	24.99	**0.0**	0.0
Throughput ratio	$\frac{11}{11}$=1.00	$\frac{7}{7}$=**1.00**	$\frac{7}{7}$=1.00
Ambulance utility (%)			
Ambulance 1	60	**31**	26
Ambulance 2		**7**	13
Ambulance 3			0.0
Average ambulance utility (%)	60	**19**	13

6.2.3 Ambulance Deployment Plan for Nketa Station

Nketa station was officially allocated three(3) ambulances to service its geographical zone. As the number of ambulances was increased systematically from three(3) to five(5), there was a corresponding significant improvement of all the performance measures except for the utilisation levels with observed decreasing trends. The utilisation levels decreased from 42% to 30% as presented in Table 6.3.

Table 6.3: Simulation Model Performance Measures for Nketa Station.

Number of Ambulances	1	2	3	4	**5**	6
Aveg. time in system (min.)	358.69	112.88	94.33	95.74	**87.43**	87.43
Aveg. response time (min.)	306.04	64.92	49.02	41.5	**40.68**	40.68
Aveg. no. in response queue	4.95	0.42	0.05	0.01	**0.0**	0.0
Aveg, queue time (min.)	289.73	26.07	3.21	0.81	**0.0**	0.0
Throughput ratio	$\frac{16}{22}$ = 0.73	$\frac{22}{23}$=0.96	$\frac{24}{24}$=1.0	$\frac{22}{24}$=0.92	$\frac{25}{25}$=**1.0**	$\frac{25}{25}$=1.0
Ambulance utility (%)						
Ambulance 1	100	78	54	55	**48**	39
Ambulance 2		66	45	35	**28**	20
Ambulance 3			53	28	**29**	27
Ambulance 4				30	**11**	12
Ambulance 5					**35**	30
Ambulance 6						24
Aveg. ambulance utility (%)	100	72	42	37	**30**	25

A further attempt to increase the ambulance fleet size from five(5) to six(6) did not yield any changes in the performance measures except for the ambulance utility levels which decreased from 30% to 25% respectively. Therefore five(5) ambulances result as the optimal static ambulance deployment when maintaining the system predetermined response time distribution.

This optimal static deployment of five(5) ambulances to Nketa station will result in an average response time of 41 minutes, with no queuing calls and an average call time in the system of 87 minutes. The five ambulances are expected to be busy 30% of the 24-hour time shift. All of the calls received in the 24-hour shift will be serviced right up to the process end depending on the needs of the patient.

6.2.4 Ambulance Deployment Plan for Nkulumane Station

Nkulumane station was officially allocated a single ambulance to service the allocated geographical zone. When the ambulance fleet size was increased from one(1) to two(2), all the performance measures significantly increased positively except for the ambulance utilization levels which decreased from 29% to 24% as presented in Table 6.4. An attempt to increase the fleet size further from two(2) to three(3) ambulances did not yield any changes for the performance measures except for the utilization levels which dropped from 24% to 16% as expected. Therefore two(2) ambulances result as the optimal static ambulance deployment when maintaining the system predetermined response time distribution for Nkulumane station.

This optimal static deployment of two(2) ambulances for the Nkulumane station will result in an average response time of 44 minutes, with no queuing calls and an average call time in the system of 83 minutes. The two ambulances are expected to be busy 24% of the 24 hour shift time. Eighty-six percent of the calls received in the 24-hour shift will be

Table 6.4: Simulation Model Performance Measures for Nkulumane Station.

Number of Ambulances	1	**2**	3
Average time in system (minutes)	97.81	**83.04**	83.04
Average response time (minutes)	43.21	**43.95**	43.95
Average number in response queue	0.02	**0.0**	0.0
Average queue time (minutes)	3.67	**0.0**	0.0
Throughput ratio	$\frac{7}{7}$=1.00	$\frac{6}{7}$=**0.86**	$\frac{6}{7}$=0.86
Ambulance utility (%)			
Ambulance 1	29	**32**	32
Ambulance 2		**15**	0.0
Ambulance 3			15
Average ambulance utility (%)	29	**24**	16

serviced right up to the process end depending on the needs of the patient.

6.2.5 Summary Statistics for Static Ambulance Deployment Plan

This section explored the concept of optimising the fleet sizes whilst maintaining constant input parameters such as response time, service time and category occurrence proportions without specifying the actual number of ambulances being serviced on a specific 24 hour (1440 minutes) working day. Table 6.5 is a summary of performance statistics for the optimum static ambulance deployment for the four sub-stations. The abbreviations used in Table 6.5 are defined as:

- AVTIS is the average total duration in the system (minutes).
- AVRT is the average response time per entity (minutes).
- AVNRQ is the average number of calls in the response queue.
- AVQT is the average queuing time per call (minutes).
- NEC is non-emergency calls
- TPR is the throughput ratio expressed as a percentage.

- AUR is the average utility ratio expressed as a percentage.

- NOA is the number of ambulances.

Table 6.5: Summary Statistics for Static Ambulance Deployment by Sub-station.

Sub-station	AVTIS (min.)	AVRT (min.)	AVNRQ	AVQT (min.)	TPR (%)	AUR (%)	Optimal size	Current size	Deficit
Famona	73.06	32.29	0	0	80	24	2	1	1
Northend	78.89	44.39	0	0	100	29	2	1	1
Nketa	87.43	40.68	0	0	100	30	5	3	2
Nkulumane	83.04	43.95	0	0	86	24	2	1	1
Overall	80.61	40.33	0	0	91.5	24.25	11	6	5

It can be observed that even though optimum fleet sizes have been determined for each sub-station, the response times remain relatively high ranging from 32.29 minutes (Famona) and 44.39 minutes (Northend). Results indicate that the average total duration time in the system for each call is high in the western suburbs serviced by Nketa and Nkulumane sub-stations as compared to the eastern suburbs covered by Famona and Northend sub-stations respectively. Across all the sub-stations, it is taking an average of more than an hour (60 minutes) to completely service an emergency ambulance call.

Under the prevailing conditions, there is a deficit of five (5) ambulances to maintain the optimal fleet size where queues and queuing time for an ambulance are reduced to zero. However, there is a need to look into the future and assess the performance of the models using forecasts from the ANN. This will help in the strategic, tactical and operational planning in the EMS delivery process.

6.3 Integrating ANN PEAD Forecasts in Ambulance Deployment

The objective is to integrate the ANN public emergency ambulance demand (PEAD) forecasts and the optimal models developed in the prior section. The optimum ambulance allocations determined in the previous section were: Famona (2), Northend (2), Nketa (5) and Nkulumane (2) ambulances respectively. The inter-arrival time, service time and response time distributions will be maintained in the sensitivity analysis.

6.3.1 Computation of Expected Daily PEAD from ANN Forecasts

In order to apportion the predicted public emergency ambulance demand(PEAD) by ANN to the different sub-stations, proportions of occurrence of demand calls at each station were computed using historical data of 2018. A summary of the calculations of proportions is presented in Table 6.6. The calculated proportions were as follows: Famona (18.1%), Northend (17.6%), Nketa (47.1%) and Nkulumane (17.2%) respectively.

These proportions were then used to apportion the 2019 ANN forecasts as an integration of the forecasting and simulation concepts for future EMS preparedness. A summary of the expected monthly public ambulance demand calls per station is presented in Table 6.7.

The expected average daily calls per station were further calculated and results are summarised in Table 6.8. These calculated expected daily calls will be incorporated in the determination of the optimal number of ambulances to be allocated in each station every month of 2019.

Table 6.6: Proportions of PEAD Occurrences in Sub-stations.

Month	Famona	Northend	Nketa	Nkulumane	Total
January	261	254	679	248	1442
February	244	237	633	232	1346
March	273	266	711	260	1510
April	250	243	650	238	1381
May	245	238	636	233	1352
June	251	244	652	238	1384
July	267	260	694	254	1475
August	259	252	675	246	1433
September	247	240	643	235	1365
October	224	218	585	213	1240
November	238	231	620	226	1315
December	254	247	662	242	1405
Station Total	3013	2930	7840	2865	16648
Calculated Ratio	0.181	0.176	0.471	0.172	1
% ratio	18.1%	17.6%	47.1%	17.2%	100%

6.3.2 Optimum Deployment Plan for Famona Station

The ANN forecasts indicated that over the whole year, across the different 12 months would assume values of 8, 9 and 10 as the expected daily number of calls for Famona station. Reference is made to Table 6.8. The ambulance fleet sizes were incremented from one(1) to three(3) whilst monitoring the performance measures.

In all cases with expected daily demands of 8, 9 or 10 calls respectively, increasing the number of ambulances beyond two(2) ambulances did not significantly improve any of the performance measures (Table 6.9). The deployment of two(2) ambulances would therefore ensure that no ambulance is queuing with throughput ratios above 89% recorded. It can be observed that ambulance utility levels decrease with an increase in ambulance fleet size. Thus increasing the ambulance fleet size beyond the threshold of two(2) ambulances did not bring any significant changes in the performance measures. Thus two(2) ambulances are the optimum ambulance allocation throughout the calendar year of 2019 as they were adequate to cover the expected annual daily demands of 8, 9, and 10 for the Famona sub-station.

Table 6.7: Expected Monthly Ambulance Demand Per Sub-station.

Station	ANN Forecast (2019)	Famona (18.1%)	Northend (17.6%)	Nketa (47.1%)	Nkulumane (17.2%)
January	1622	294	285	764	279
February	1494	270	263	704	257
March	1713	310	301	807	295
April	1368	248	241	644	235
May	1482	268	261	698	255
June	1318	239	232	620	227
July	1391	252	245	655	239
August	1526	276	269	719	262
September	1572	285	277	740	270
October	1541	279	271	726	265
November	1532	277	269	722	264
December	1638	296	289	771	282
Total	18197	3294	3203	8570	3130

Table 6.8: Expected Daily Ambulance Demand Per Sub-station.

Station	Days in the Month	Famona	Northend	Nketa	Nkulumane
January	31	9	9	25	9
February	28	10	9	25	9
March	31	10	10	26	10
April	30	8	8	21	8
May	31	9	8	23	8
June	30	8	8	21	8
July	31	8	8	21	8
August	31	9	9	23	8
September	30	10	9	25	9
October	31	9	9	23	9
November	30	9	9	24	9
December	31	10	9	25	9
	Overall average	9	9	24	9
	Maximum	10	10	25	10
	Minimum	8	8	21	8

6.3.3 Optimum Deployment Plan for Northend Station

The ANN forecasts indicated that over the whole year, across the different 12 months would assume values of 8, 9, and 10 as the expected daily number of calls for Northend station.

Table 6.9: Optimum Deployment Plan for Famona Station.

Calls (N)	NOA	AVTIS (min.)	AVRT (min.)	AVNRQ	AVQT (min.)	TPR (%)	AUR (%)	NSOS (min.)	SOS (min.)
N=8	1	86.15	38.80	0.03	5.32	8/8	0.45	16.44	65.9
	2	**76.06**	**33.07**	**0.0**	**0.0**	**8/8**	**0.21**	**15.62**	**64.37**
	3	73.06	33.07	0.0	0.0	8/8	0.14	15.62	64.37
N=9	1	86.15	40.51	0.04	6.58	8/9	0.48	16.44	65.9
	2	**73.06**	**32.29**	**0.0**	**0.0**	**8/9**	**0.23**	**15.62**	**64.37**
	3	73.06	32.29	0.0	0.0	8/9	0.15	15.62	64.37
N=10	1	86.15	40.51	0.04	6.58	8/9	0.48	16.44	65.9
	2	**73.06**	**32.29**	**0.0**	**0.0**	**8/9**	**0.24**	**15.62**	**64.37**
	3	73.06	32.29	0.0	0.0	8/9	0.15	15.62	64.37

Reference is made to Table 6.8. The ambulance fleet sizes were incremented from one(1) to three(3) whilst monitoring the performance measures.

In all the cases with expected daily demands of 8, 9 or 10 calls respectively, increasing the ambulance fleet size beyond two(2)ambulances did not significantly improve any of the performance measures (Table 6.10). The deployment of two(2) ambulances would therefore ensure that no ambulance is queuing with throughput ratios of 100% recorded. It was also observed that ambulance utility levels decrease with an increase in ambulance fleet size. Thus increasing the ambulance fleet size beyond the threshold of two(2) ambulances did not bring any significant changes in the performance measures. Thus two(2) ambulances are the optimum ambulance allocation throughout the calendar year of 2019 as they were adequate to cover the expected annual daily demands of 8, 9, and 10 for Northend station.

6.3.4 Optimum Deployment Plan for Nketa Station

The ANN forecasts indicated that over the whole year, across the different 12 months would assume values of 21, 23, 24, 25 and 26 as expected daily number of calls. Reference is made to Table 6.8. The ambulance fleet sizes were incremented from two(2) up to six(6) ambulances whilst recording the performance measures.

Table 6.10: Optimum Deployment Plan for Northend Station.

Calls (N)	NOA	AVTIS (min.)	AVRT (min.)	AVNRQ	AVQT (min.)	TPR (%)	AUR (%)	NSOS (min.)	SOS (min.)
N=8	1	93.87	47.63	0.09	17.91	7/7	0.37	31.82	52.0
	2	**94.34**	**49.26**	**0.0**	**0.0**	**6/6**	**0.20**	**34.47**	**55.68**
	3	94.34	49.26	0.0	0.0	6/6	0.13	34.47	55.68
N=9	1	93.79	46.67	0.09	15.67	8/8	0.43	31.82	52.22
	2	**85.65**	**49.26**	**0.0**	**0.0**	**6/6**	**0.18**	**27.78**	**53.63**
	3	85.65	49.26	0.0	0.0	6/6	0.13	27.78	53.63
N=10	1	103.01	54.85	0.14	22.61	9/9	0.50	31.82	52.84
	2	**78.89**	**44.40**	**0.0**	**0.0**	**7/7**	**0.19**	**26.84**	**53.63**
	3	78.89	44.40	0.0	0.0	7/7	0.13	26.84	53.63

In all the cases with expected daily demands of 21, 23, 24, 25 and 26 calls respectively, increasing the ambulance fleet size beyond five(5) ambulances did not significantly improve any of the performance measures (Table 6.11). The deployment of five(5) ambulances would therefore ensure that no ambulance is queuing with throughput ratios of 100% recorded. It was also observed that ambulance utility levels decreased with an increase in ambulance fleet size. Thus increasing the ambulance fleet size beyond the threshold of five(5) ambulances did not bring any significant changes in the performance measures. Thus five(5) ambulances are the optimum ambulance allocation throughout the calendar year of 2019 as they were adequate to cover the expected annual daily demands of 21, 23, 24, 25 and 26 for Nketa station.

6.3.5 Optimum Deployment Plan for Nkulumane Station

The ANN forecasts indicated that over the whole year, across the different 12 months would assume values of 8, 9, and 10 as the expected daily number of calls for the Nkulumane station. Reference is made to Table 6.8. The ambulance fleet sizes were incremented from one(1) to three(3) whilst monitoring the performance measures.

In all the cases with expected daily demands of 8, 9 or 10 calls respectively, increasing

Table 6.11: Optimum Deployment Plan for Nketa Station.

Calls	NOA	AVTIS	AVRT	AVNRQ	AVQT	TPR	AUR	NSOS	SOS
(N)		(min.)	(min.)		(min.)	(%)	(%)	(min.)	(min.)
N=21	2	120.75	66.38	0.30	24.20	18/18	0.61	53.18	54.61
	3	101.16	51.82	0.05	4.05	19/19	0.43	0.0	49.34
	4	93.72	44.02	0.01	0.97	20/20	0.32	11.84	51.7
	5	**92.21**	**41.75**	**0.0**	**0.0**	**20/20**	**0.26**	**11.84**	**52.49**
	6	92.21	41.75	0.0	0.0	20/20	0.26	11.84	52.49
N=23	2	113.3	61.06	0.3	22.48	19/19	0.60	53.18	52.06
	3	102.59	53.74	0.05	3.66	21/21	0.48	0.0	48.84
	4	94.47	42.95	0.01	0.88	22/22	0.36	11.84	53.41
	5	**89.23**	**39.71**	**0.0**	**0.0**	**22/22**	**0.27**	**11.84**	**51.32**
	6	89.23	39.71	0.0	0.0	22/22	0.23	11.84	51.32
N=24	2	112.16	59	0.3	21.36	20/20	0.63	53.18	53.16
	3	98.97	51.62	0.05	3.5	22/22	0.48	23.42	49.75
	4	95.74	42.95	0.01	0.88	22/22	0.37	11.84	54.74
	5	**89.73**	**40.7**	**0.0**	**0.0**	**23/23**	**0.29**	**19.56**	**53.45**
	6	89.73	40.7	0.0	0.0	23/23	0.24	19.56	53.45
N=25	2	115.24	59.58	0.31	20.98	21/21	0.69	53.18	56.08
	3	96.87	49.79	0.05	3.35	23/23	0.50	23.42	49.33
	4	95.74	41.49	0.01	0.84	22/23	0.39	11.84	54.74
	5	**88.97**	**41.07**	**0.0**	**0.0**	**24/24**	**0.30**	**20.14**	**53.45**
	6	88.97	41.07	0.0	0.0	24/24	0.25	20.14	53.45
N=26	2	113.51	62.25	0.35	23.01	22/22	0.69	43.77	52.93
	3	94.33	49.02	0.05	3.21	24/24	0.51	17.09	49.33
	4	95.74	41.49	0.01	0.81	22/23	0.37	11.84	54.74
	5	**87.43**	**40.68**	**0.0**	**0.0**	**25/25**	**0.30**	**15.78**	**52.65**
	6	87.43	40.68	0.0	0.0	25/25	0.25	15.78	52.65

the ambulance fleet size beyond two(2)ambulances did not significantly improve any of the performance measures (Table 6.12). The deployment of two(2) ambulances would therefore ensure that no ambulance is queuing with throughput ratios of 86% recorded. It was also observed that ambulance utility levels decrease with an increase in ambulance fleet size. Thus increasing the ambulance fleet size beyond the threshold of two(2) ambulances did not bring any significant changes in the performance measures. Thus two(2) ambulances are the optimum ambulance allocation throughout the calendar year of 2019 as they were adequate

Table 6.12: Optimum Deployment Plan for Nkulumane Station.

Calls	NOA	AVTIS	AVRT	AVNRQ	AVQT	TPR	AUR	NSOS	SOS
(N)		(min.)	(min.)		(min.)	(%)	(%)	(min.)	(min.)
N=8	1	97.81	43.21	0.02	3.67	7/7	0.46	50.86	56.06
	2	**83.04**	**43.95**	**0.0**	**0.0**	**6/7**	**0.20**	**34.5**	**53.54**
	3	83.04	43.95	0.0	0.0	6/7	0.13	34.5	53.54
N=9	1	97.81	43.21	0.02	3.67	7/7	0.46	50.86	56.09
	2	**83.04**	**43.95**	**0.0**	**0.0**	**6/7**	**0.20**	**34.5**	**53.54**
	3	83.04	43.95	0.0	0.0	6/7	0.13	34.5	53.54
N=10	1	97.81	43.21	0.02	3.67	7/7	0.46	50.86	56.09
	2	**83.04**	**43.95**	**0.0**	**0.0**	**6/7**	**0.20**	**34.5**	**53.54**
	3	83.04	43.95	0.0	0.0	6/7	0.13	34.5	53.54

to cover the expected annual daily demands of 8, 9, and 10 for the Nkulumane station.

6.3.6 Computation of Fleet Sizes for Expected Daily Calls from ANN Forecasts

The Optimum ambulance deployment plan results when integrating ANN forecasts for 2019 for all the sub-stations will not change (Table 6.13).

Table 6.13: Computation of Fleet Sizes on Expected Daily Calls (N) from ANN Forecasts.

		Expected Daily Calls (N)							
Region	Station	N=8	N=9	N=10	N=21	N=23	N=24	N=25	N=26
Eastern	Famona	2	2	2	-	-	-	-	-
Suburbs	Northend	2	2	2	-	-	-	-	-
Western	Nketa	-	-	-	5	5	5	5	5
Suburbs	Nkulumane	2	2	2	-	-	-	-	-

A summary of the annual deployment plan is presented in Table(6.14). Indications are that the initial optimum deployment plan will not change when considering the annual ANN expected daily demand forecasts.

The results therefore, imply that the initially determined optimum deployment plan of

Table 6.14: Optimal Fleet Size For ANN Expected Daily Ambulance Demand Forecasts by Sub-station.

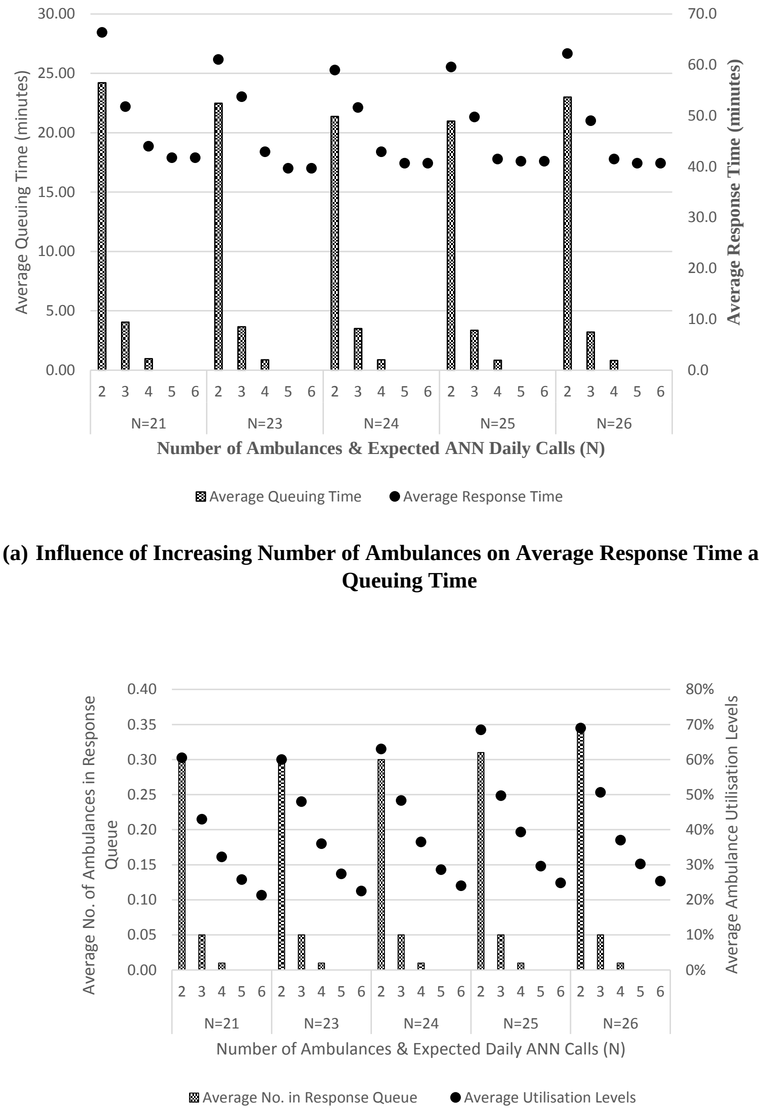
Station	Famona	Northend	Nketa	Nkulumane	Fleet Size
January	2	2	5	2	11
February	2	2	5	2	11
March	2	2	5	2	11
April	2	2	5	2	11
May	2	2	5	2	11
June	2	2	5	2	11
July	2	2	5	2	11
August	2	2	5	2	11
September	2	2	5	2	11
October	2	2	5	2	11
November	2	2	5	2	11
December	2	2	5	2	11

eleven(11) ambulances is adequate to meet future demand as predicted by ANN.

6.4 Chapter Conclusion

Chapter six (6) focused on developing optimal ambulance deployment plans by incorporating ANN forecasts and varying the fleet sizes while observing the multiple performance measures. Key insights can be derived from Nketa Station with a wide variation in the expected number of calls and ambulance fleet size required for deployment in achieving the optimal solution. A summary of key indicators was tracked as fleet sizes were incremented to optimal levels as presented in Figure 6.1. As the fleet size is increased there is a corresponding decrease in the average response time to a specific threshold (5 ambulances) and beyond this optimal fleet size, the average response time remains constant. It was also observed that as the fleet size increases to a specific threshold, the average queuing time also decreases to zero. The same applies to the number of calls queuing in the response time. However, the ambulance utilisation levels vary inversely with an increase in fleet size. The optimum deployment plans still have a high average response time which is far above the recommended international standards, hence the need to adjust accordingly through the numerical experiments.

From the results, it was observed that increasing the number of ambulances influences the average response time below a certain threshold. When the fleet size is increased beyond this threshold, no significant changes occur in the performance measures. It was observed that as the fleet size is increased, the ambulance utilisation levels decreased. Hence, there is a need to balance resource allocation and capacity utilisation to avoid the idleness of essential equipment and human resources. This has a significant direct influence on additional costs that have to be incurred with increased number of emergency ambulances and personnel, whom in some cases will be idle. The integration of ANN forecasts did not bring any significant changes to the initial optimum deployment plan: Famona (2), Northend (2), Nketa (5) and Nkulumane (2) ambulances respectively, throughout the the year of 2019. The results imply that the initial deployment plan was adequate to meet future demand for



(a) Influence of Increasing Number of Ambulances on Average Response Time and Queuing Time

(b) Influence of Increasing Number of Ambulances on Number of Ambulances in Response Queue and Ambulance Utilisation Levels

Figure 6.1: Implications of Varying Fleet Sizes on Performance Indicators.

ambulance services. Under the prevailing conditions, there is a deficit of five (5) ambulances to maintain the optimal fleet size where queues and queuing time for an ambulance are reduced to zero. Average response times across the four heterogeneous regions remain high far beyond the 5 to 10 minutes international standard. Chapter seven (7) will investigate the influence of changing the response time distributions on the average response time and optimum deployment plans in order to meet international prescribed standards.

Chapter 7

Numerical Experiments

7.1 Introduction

In this chapter, numerical experiments are conducted in order to investigate various scenarios in the build-up to public emergency preparedness. The chapter will integrate the ANN public emergency ambulance demand (PEAD) forecasts, whilst adjusting the ambulance fleet sizes in order to optimise the levels of preparedness. Special consideration would involve the determination of an optimal static ambulance deployment plan by varying the response time to international standards against the predicted ANN values. Performance measures such as the average entity time in the system, average response time, average response queue time, the average number of calls in the response queue, throughput ratio, and the ambulance utility levels will be considered to evaluate the models.

7.2 Empirical Evidence

Several researchers have made similar attempts in conducting numerical experiments to specific areas across the world involving emergency medical services. Ingolfsson et al. (2008)

focused on the allocation of ambulance vehicles to a set of (existing or planned) ambulance stations with known locations and alluded that the action to reduce response time due to delays (pre-trip delay and queuing delay) is far easier and less costly to reduce than travel times. Pre-trip delays emanate from call delay or chute delay. A call delay is the time spent on taking a call, establishing the severity of the call and dispatching an ambulance crew. Chute delay is the time that elapses from when a crew is dispatched until the vehicle starts moving. Queuing delays occur when no ambulance(s) are available either busy attending to other calls and is often attributed to system congestion. The study indicated that reducing the travel times usually requires adding ambulance stations or hospitals which is costly as the municipality is currently financially under-resourced. Their emphasis on response time was that reducing the time by 5 seconds, is actually 5 seconds saved and it does not matter which component of response time these savings have come from. Here, the expectation is that reducing the response time has a huge bearing in improving service delivery, survival rates and patient satisfaction.

Leknes et al. (2017) presented almost similar research of an EMS problem which focused on ambulance station location and allocation problem which they referred to as the Maximum Expected Location Problem for Heterogeneous Regions (MEPLP-HR). Its main objective was to give the population of Sor-Trondelag County in Norway the best possible EMS according to a set of selected performance measures. They highlighted that the probability of an available ambulance depends on the arrival rate of calls at a station, number of ambulances allocated to the station, the service time of the ambulances and the priority given to calls (urgency levels) as determined by the EMS provider. In their research they were able to demonstrate that as the response time decreases, there is a corresponding increase in the probability of survival of a patient. They further demonstrated that as the service rate (calls/hour) increases, the probability of no available ambulance decreases. They were also able to demonstrate that as the arrival rate increases, the probability of no available ambulances increases as it translates to an increase in demand for EMS provision.

Zhen et al. (2014) applied the simulation optimisation framework for ambulance deployment and relocation to the city of Shanghai in China. In their case, they also carried out some numerical experiments to determine the influence of parameters on the response time and the ambulance deployment plan. They were able to observe that the number of ambulances, the number of ambulance bases, and the number of hospitals had an impact on the average response time.

The focus of this chapter is therefore to explore the influence of varying the response time distribution on the optimum deployment plan to international standards of 5 to 10 minutes by adopting a uniform distribution $U(5;10)$. Both $U(10;15)$ and $U(5;10)$ distributions were investigated in order to track any trends in the influence of varying response time distributions on deployment plans. This was strongly motivated by the fact that it is easier, cheaper and feasible for management to control some of the processes that are directly linked to the response time. Implications to managerial decision-making as a result of the findings will be discussed in detail.

7.3 Influence of Standardising Response Time

This section focuses on adjusting fleet size by varying the response time distributions (RTD) to a uniform distribution with parameters $U(10;15)$ and $U(5;10)$ whilst incorporating the ANN forecasts. The uniform distribution of the response time would allow the response time to vary between 10 and 15 minutes and, 5 and 10 minutes respectively. The simulation model parameters such as the inter-arrival of calls and service time distributions would be maintained whilst the performance measures such as the overall total duration of a call in system, an average number of calls in response queue, the average time in response queue, throughput ratio and ambulance utilisation levels will be observed to evaluate the models.

This section explores the impact of reducing the response time on the overall deployment plan for future planning by including the ANN forecasts. Table 7.1 is a summary of key parameters to be considered in the model build-up process.

Table 7.1: Proposed Simulation Model Response Time Distributions and Parameters Per Sub-station.

Station	Initial RTD	Proposed RTD	Current Opt. Fleet Size	Expected ANN forecasts
Famona	2+GAMM(22;1.48)	U(10;15)/U(5;10)	2	8,9,10
Northend	2+GAMM(23.9;1.36)	U(10;15)/U(5;10)	2	8,9,10
Nketa	-0.001+ERLA(18.5;2)	U(10;15)/U(5;10)	5	21,23,24,25,26
Nkulumane	0.999+GAMM(21.8;1.62)	U(10;15)/U(5;10)	2	8,9,10

7.3.1 Comparison of Optimum Deployment Plans for Famona Station

The ANN forecasts indicated that over the whole year, across the different 12 months would assume values of 8, 9, and 10 as the expected daily number of calls for Famona Station (Table 7.1). The response time distribution is assumed to follow a uniform distribution given by $U(10;15)$. Results indicate that increasing the number of ambulances beyond the threshold of one(1) ambulance would not improve the performance measures for all the cases (N= 8, 8 and 10) as presented in Table 7.2. However, the average response time has significantly improved by standardising the response time distribution.

A comparison of the different performance changes due to the influence of the changes in response time distributions on the optimal deployment plan for Famona Station by adopting a $U(5,10)$ response distribution was conducted. A comparative summary of consolidated results of the optimum deployment plans is presented in Table 7.3.

The overall optimal deployment plan for Famona changes for all the expected daily forecasts (N=8, 9 and 10) from ANN. The optimum number of ambulances (NOA) decreases

Table 7.2: Optimum Deployment Plan for Famona Station: RTD $\sim U(10; 15)$.

ANN	NOA	AVTIS	AVRT	AVNRQ	AVQT	TPR	AUR	NSOS	SOS
Forecasts		(min.)	(min.)		(min.)		(%)	(min.)	(min.)
N=8	**1**	**52.92**	**12.92**	**0.0**	**0.0**	**8/8**	**0.29**	**15.62**	**64.37**
	2	52.92	12.92	0.0	0.0	8/8	0.15	15.62	64.37
	3	52.92	12.92	0.0	0.0	8/8	0.10	15.62	64.37
N=9	**1**	**52.92**	**12.89**	**0.0**	**0.0**	**8/9**	**0.33**	**15.62**	**64.37**
	2	52.92	12.89	0.0	0.0	8/9	0.17	15.62	64.37
	3	52.92	12.89	0.0	0.0	8/9	0.11	15.62	64.37
N=10	**1**	**52.92**	**12.79**	**0.0**	**0.0**	**8/10**	**0.33**	**15.62**	**64.37**
	2	52.92	12.29	0.0	0.0	8/10	0.18	15.62	64.37
	3	52.92	12.79	0.0	0.0	8/10	0.12	15.62	64.37

from two(2) to one(1) as the response time is set between 5 to 10 minutes using the uniform distribution $U(5; 10)$ for all the cases. Therefore reducing the response time impacts positively on the performance measures/indicators. The average response time decreases whilst the number of ambulances required to be deployed decreases without compromising service delivery. Notably, the average total time a call is reported to be in the system significantly decreases even though fewer ambulances would have been deployed across all the considered cases. Moreover, the average utilisation levels (AUR) of ambulances increased significantly with the reduced response time. Under these prevailing conditions, no emergency ambulance call is expected to queue for service. The throughput ratios remained constant across the optimal deployment plans.

7.3.2 Comparison of Optimum Deployment Plans for Northend Station

The ANN forecasts indicated that over the whole year, across the different 12 months would assume values of 8, 9, and 10 as the expected daily number of calls for Northend station (Table 7.1). The response time distribution is initially assumed to follow a uniform distribution given by $U(10; 15)$ and the summary of results is presented in Table 7.4.

Table 7.3: Comparison of Optimum Deployment Plans for Famona Station.

ANN Forecasts	Response Time Distribution	Opt. NOA	AVTIS (min.)	AVRT (min.)	AVNRQ	AVQT (min.)	NEC	TPR	AUR	NSOS (min.)	SOS (min.)
N=8	2+GAMM(22;1.48)	2	76.06	33.07	0	0	0	8/8	0.21	15.62	64.37
	U(10;15)	1	52.92	12.92	0	0	0	8/8	0.29	15.62	64.37
	U(5;10)	**1**	**47.92**	**7.92**	0	0	0	**8/8**	**0.27**	**15.62**	**64.37**
N=9	2+GAMM(22;1.48)	2	73.06	32.29	0	0	0	8/9	0.23	15.62	64.37
	U(10;15)	1	52.92	12.92	0	0	0	8/9	0.29	15.62	64.37
	U(5;10)	**1**	**47.92**	**7.89**	0	0	0	**8/9**	**0.30**	**15.62**	**64.37**
N=10	2+GAMM(22;1.48)	2	73.06	32.29	0	0	0	8/10	0.24	15.62	64.37
	U(10;15)	1	52.92	12.92	0	0	0	8/10	0.29	15.62	64.37
	U(5;10)	**2**	**47.92**	**7.79**	0	0	0	**8/10**	**0.30**	**15.62**	**64.37**

Table 7.4: Optimum Deployment Plan for Northend Station: RTD $\sim U(10; 15)$.

ANN	NOA	AVTIS	AVRT	AVNRQ	AVQT	TPR	AUR	NSOS	SOS
Forecasts		(min.)	(min.)		(min.)		(%)	(min.)	(min.)
N=8	1	53.71	16.25	0.02	4.14	6/6	0.21	23.78	51.14
	2	**56.61**	**11.83**	**0.0**	**0.0**	**6/6**	**0.12**	**34.47**	**55.68**
	3	56.61	11.83	0.0	0.0	6/6	0.08	34.47	55.68
N=9	1	51.28	15.87	0.02	3.55	7/7	0.23	23.61	51.14
	2	**58.04**	**11.63**	**0.0**	**0.0**	**7/7**	**0.14**	**34.47**	**55.36**
	3	58.04	11.63	0.0	0.0	7/7	0.09	34.47	55.36
N=10	1	57.45	15.54	0.02	3.11	8/8	0.30	36.38	51.14
	2	**58.90**	**11.63**	**0.0**	**0.0**	**8/8**	**0.17**	**34.47**	**54.95**
	3	58.90	11.63	0.0	0.0	8/8	0.11	34.47	54.95

A comparison of the performance changes due to the influence of changes in response time distribution on optimal deployment plan for Northend Station was performed incorporating results from adopting a $U(5, 10)$. A comparative summary of consolidated results is presented in Table 7.5.

The overall optimal deployment plan for Northend Station did not change for expected daily forecasts (N=8, 9 and 10) from ANN forecasts. The optimum number of ambulances (NOA) remains at two(2), however, significant decreases in the average response time (AVRT) and the average total duration time of call in the system (AVTIS) were recorded respectively. In all the cases discussed no emergency call is expected to queue for service. The total number of ambulances to be served within a day as depicted by the throughput ratios (TPR) of (6/6, 7/7 and 8/8) is expected to increase with reduced response time represented by $U(10; 15)$ and $U(5; 10)$ respectively.

7.3.3 Comparison of Optimum Deployment Plans for Nketa Station

The ANN forecasts indicated that over the whole year, across the different 12 months would assume values of 21, 23, 24, 25 and 26 as expected daily number of calls (Table 7.1). The

Table 7.5: Comparison of Optimum Deployment Plans for Northend Station.

ANN	Response Time	Opt.	AVTIS	AVRT	AVNRQ	AVQT	NEC	TPR	AUR	NSOS	SOS
Forecasts	Distribution	NOA	(min.)	(min.)		(min.)			(min.)	(min.)	
N=8	2+GAMM(23.9;1.36)	2	94.34	49.26	0	0	2	6/6	0.20	34.47	55.68
	U(10;15)	2	56.61	11.83	0	0	2	6/6	0.12	34.47	55.68
	U(5;10)	**2**	**51.61**	**6.53**	0	0	2	**6/6**	**0.11**	**34.47**	**55.68**
N=9	2+GAMM(23.9;1.36)	2	85.65	49.26	0	0	3	6/6	0.18	27.78	53.63
	U(10;15)	2	58.04	11.63	0	0	2	7/7	0.14	34.47	55.36
	U(5;10)	**2**	**53.04**	**6.63**	0	0	2	**7/7**	**0.13**	**34.47**	**55.36**
N=10	2+GAMM(23.9;1.36)	2	78.39	44.40	0	0	3	7/7	0.19	26.84	53.63
	U(10;15)	2	58.90	11.63	0	0	2	8/8	0.17	34.47	54.95
	U(5;10)	**2**	**53.90**	**6.63**	0	0	2	**8/8**	**0.15**	**34.47**	**54.95**

response time distribution is initially assumed to follow a uniform distribution given by $U(10;15)$ and results are summarised in Table 7.6.

Table 7.6: Optimum Deployment Plan for Nketa Station: RTD $\sim U(10,15)$.

ANN	NOA	AVTIS	AVRT	AVNRQ	AVQT	TPR	AUR	NSOS	SOS
Forecasts		(min.)	(min.)		(min.)		(%)	(min.)	(min.)
N=21	2	85.12	21.30	0.10	8.45	19/19	0.46	49.24	64.73
	3	**65.49**	**12.72**	**0.0**	**0.0**	**19/19**	**0.29**	**28.41**	**57.34**
	4	65.49	12.72	0.0	0.0	19/19	0.22	28.41	57.34
	5	65.49	12.72	0.0	0.0	19/19	0.17	28.41	57.34
	6	65.49	12.72	0.0	0.0	19/19	0.14	28.41	57.34
N=23	2	82.48	12.63	0.0	0.0	20/21	0.48	49.24	60.89
	3	**65.48**	**12.63**	**0.0**	**0.0**	**20/21**	**0.31**	**28.41**	**57.1**
	4	65.48	12.63	0.0	0.0	20/21	0.24	28.41	57.1
	5	65.48	12.63	0.0	0.0	20/21	0.19	28.41	57.1
	6	65.48	12.63	0.0	0.0	20/21	0.16	28.41	57.1
N=24	2	82.05	21.81	0.13	9.0	20/20	0.51	49.24	60.82
	3	**64.29**	**12.57**	**0.0**	**0.0**	**21/22**	**0.32**	**28.41**	**55.53**
	4	64.29	12.57	0.0	0.0	21/22	0.24	28.41	55.53
	5	64.29	12.57	0.0	0.0	21/22	0.19	28.41	55.53
	6	64.29	12.57	0.0	0.0	21/22	0.16	28.41	55.53
N=25	2	80.31	21.45	0.13	8.57	21/21	0.52	40.27	60.82
	3	**64.29**	**12.68**	**0.0**	**0.0**	**21/23**	**0.33**	**28.41**	**55.53**
	4	64.29	12.68	0.0	0.0	21/23	0.25	28.41	55.53
	5	64.29	12.68	0.0	0.0	21/23	0.20	28.41	55.53
	6	64.29	12.68	0.0	0.0	21/23	0.16	28.41	55.53
N=26	2	78.56	21.05	0.13	8.19	22/22	0.54	40.27	59.23
	3	**64.29**	**12.68**	**0.0**	**0.0**	**21/23**	**0.33**	**28.41**	**55.53**
	4	64.29	12.68	0.0	0.0	21/23	0.25	28.41	55.53
	5	64.29	12.68	0.0	0.0	21/23	0.20	28.41	55.53
	6	64.29	12.68	0.0	0.0	21/23	0.16	28.41	55.53

A comparison of the performance changes due to the influence of changes in response time distribution by considering $U(10;15)$ and $U(5;10)$ uniform distributions on optimal deployment plans for Nketa Station was performed. A comparative summary of consolidated results is presented in Table 7.7 for all cases (N=21, 23, 24, 25 and 26) respectively.

Results in Table 7.7 indicate that reducing the response time by adopting a uniform

Table 7.7: Comparison of Optimum Deployment Plans for Nketa Station

ANN (N)	Response Time Distribution	Opt. NOA	AVTIS (min.)	AVRT (min.)	AVNRQ	AVQT (min.)	NEC	TPR	AUR	NSOS (min.)	SOS (min.)
N=21	-0.001+ERLA(18.5;2)	5	92.21	41.75	0	0	1	20/20	0.26	11.84	52.49
	U(10;15)	3	65.49	12.72	0	0	2	19/19	0.29	28.41	57.34
	U(5;10)	**3**	**60.49**	**7.72**	0	0	2	**19/19**	**0.26**	**28.41**	**57.34**
N=23	-0.001+ERLA(18.5;2)	5	89.23	39.71	0	0	1	22/22	0.27	11.84	51.32
	U(10;15)	3	65.48	12.63	0	0	2	20/21	0.31	28.41	57.10
	U(5;10)	**3**	**60.48**	**7.63**	0	0	2	**20/21**	**0.29**	**28.41**	**57.10**
N=24	-0.001+ERLA(18.5;2)	5	89.73	40.7	0	0	1	23/23	0.29	19.56	53.45
	U(10;15)	3	64.29	12.57	0	0	2	21/22	0.32	28.41	55.53
	U(5;10)	**3**	**60.48**	**7.56**	0	0	2	**20/22**	**0.30**	**28.41**	**57.10**
N=25	-0.001+ERLA(18.5;2)	5	88.97	41.07	0	0	1	24/24	0.30	20.14	53.45
	U(10;15)	3	64.29	12.68	0	0	2	21/23	0.33	28.41	55.53
	U(5;10)	**3**	**60.48**	**7.56**	0	0	2	**20/22**	**0.30**	**28.41**	**57.10**
N=26	-0.001+ERLA(18.5;2)	5	87.43	40.68	0	0	1	25/25	0.30	15.78	52.65
	U(10;15)	3	64.29	12.68	0	0	2	21/23	0.33	28.41	55.53
	U(5;10)	**3**	**60.48**	**7.56**	0	0	2	**20/22**	**0.30**	**28.41**	**57.10**

distribution $U(5;10)$ resulted in the decrease of the optimal number of ambulances required to achieve an optimal deployment plan. For Nketa Station the adoption of $U(5;10)$ would result in a drop of the threshold fleet size from five(5) to three(3) ambulances across the different scenarios (N=21, 23, 24, 25 and 26). Utilisation capacity levels increased with a reduction of response time regardless of the decrease of the fleet size from five(5) to three(3) ambulances across all the cases. The throughput ratios remain relatively high despite the decrease in optimum fleet size where no emergency ambulance call is queuing for service. Hence, service provision is not compromised by the resulting influence of reducing response time and fleet size.

7.3.4 Comparison of Optimum Deployment Plans for Nkulumane Station

The ANN forecasts indicated that over the whole year, across the different 12 months would assume values of 8, 9, and 10 as the expected daily number of calls for Nkulumane station (Table 7.1). The response time distribution is initially assumed to follow a uniform distribution given by $U(10;15)$ and a summary of results is presented in Table 7.8.

Table 7.8: Optimum Deployment Plan for Nkulumane Station: RTD $\sim U(10;15)$.

ANN Forecasts	NOA	AVTIS (min.)	AVRT (min.)	AVNRQ	AVQT (min.)	TPR (%)	AUR (%)	NSOS (min.)	SOS (min.)
N=8	**1**	**57.64**	**12.71**	**0.0**	**0.0**	**7/7**	**0.28**	**34.5**	**52.76**
	2	57.64	12.71	0.0	0.0	7/7	0.14	34.5	52.76
	3	57.64	12.71	0.0	0.0	7/7	0.09	34.5	52.76
N=9	**1**	**57.64**	**12.71**	**0.0**	**0.0**	**7/7**	**0.28**	**34.5**	**52.76**
	2	57.64	12.71	0.0	0.0	7/7	0.14	34.5	52.76
	3	57.64	12.71	0.0	0.0	7/7	0.09	34.5	52.76
N=10	**1**	**57.64**	**12.71**	**0.0**	**0.0**	**7/7**	**0.28**	**34.5**	**52.76**
	2	57.64	12.71	0.0	0.0	7/7	0.14	34.5	52.76
	3	57.64	12.71	0.0	0.0	7/7	0.09	34.5	52.76

A comparison on the performance changes due to the influence of changes in response time distribution on optimal deployment plan for Nkulumane Station was performed by incorporating results from further adopting a $U(5;10)$ distribution as the response time distribution. A comparative summary of consolidated results is presented in Table 7.9 for all cases (N=8, 9 and 10) respectively.

In all the cases (N=8, 9 and 10), the threshold of one(1) ambulance was achieved and any increase beyond this fleet size would not positively influence changes in the performance measures (Table 7.9). Ambulance utility levels (AUR) increased as the optimal number of ambulances allocated decreased. Results indicate that reducing the response time between 5 and 10 minutes by adopting a $U(5;10)$ distribution resulted in the decrease of the optimal number of ambulances required to achieve an optimal deployment plan, without having calls queuing for ambulance response services. Throughput ratios were observed to increase with the variations in the response time distributions i.e. $U(5;10)$.

7.3.5 Optimal Fleet Sizes for ANN Forecasts and Selected RTD

A summary of the optimum deployment plans when integrating ANN forecast and the respective response time distributions are summarised in Table 7.10. Generally, the number of ambulances required is high in the Western suburbs as compared to the Eastern suburbs. The integration of the expected daily forecasts from ANN and the simulation modelling process resulted in an optimal deployment plan presented in Table 7.10. The deployment plan for the $U(5;10)$ indicates that eight(8) ambulances are required in February, March, September and December whilst seven(7) ambulances are required for January, April, May, June, July, August, September, October and November respectively.

Table 7.9: Comparison of Optimum Deployment Plans for Nkulumane Station.

ANN	Response Time	Opt.	AVTIS	AVRT	AVNRQ	AVQT	NEC	TPR	AUR	NSOS	SOS
(N)	Distribution	NOA	(min.)	(min.)		(min.)				(min.)	(min.)
N=8	0.999+GAMM(21.8;1.62)	2	83.04	43.95	0	0	0	6/7	0.20	34.5	53.54
	U(10;15)	1	57.64	12.71	0	0	0	7/7	0.28	34.5	52.76
	U(5;10)	**1**	**52.64**	**7.71**	0	0	0	**7/7**	**0.26**	**34.5**	**52.76**
N=9	0.999+GAMM(21.8;1.62))	2	83.04	43.95	0	0	0	6/7	0.20	34.50	53.54
	U(10;15)	1	57.64	12.71	0	0	0	7/7	0.28	34.5	52.76
	U(5;10)	**1**	**52.64**	**7.71**	0	0	0	**7/7**	**0.26**	**34.5**	**52.76**
N=10	0.999+GAMM(21.8;1.62)	2	83.04	43.95	0	0	0	6/7	0.20	34.50	53.54
	U(10;15)	1	57.64	12.71	0	0	0	7/7	0.28	34.5	52.76
	U(5;10)	**1**	**52.64**	**7.71**	0	0	0	**7/7**	**0.26**	**34.5**	**52.76**

7.4 Implications of The Numerical Experiments

Standardising the response time between 5 to 10 minutes by adopting a uniform distribution $U(5, 10)$) had a positive influence on the optimal deployment plan. It resulted in a significant decrease in the number of ambulances to be deployed. The decrease in ambulances deployed did not affect the overall performance of EMS provision as it resulted in the decreases of the average response time, average total duration of the call in the system and reduced queuing time to zero. The ambulance utilisation levels and the throughput ratios remained relatively high and in some cases were observed to be increasing, implying that reducing the response time impacts positively on the overall performance of the models. A comparison of the ambulance deployment plans before and after adjusting for the response time distributions is shown in Fig. 7.1.

To management, it is imperative to seriously consider ways of reducing the response time as it has significant bearing in reducing the required number of ambulances, a critical but scarce resource. A reduced number of ambulances, brings about reduced requirement of human capital and also a reduced workload as staff members would be able to be rotated more frequently as per international standards. One way is to focus on reducing the per-trip delays which are operationally possible. There is a need therefore to manage activities within the call center. This might include digitisation of switch boards in the call center, training of the paramedics and provision of proper and modern equipment to the response teams as this will go a long way in reducing the chute delay time. Provision of well serviced and equipped ambulances will translate to reducing the ambulance travel time to the scene where the emergency service is required. As an important department of Bulawayo City Council, the department of fire and ambulance services could also lobby for decongestion and resurfacing of old and dilapidated roads in order to increase access and speed when responding to emergency calls based on scientific evidence.

Table 7.10: Optimal Fleet Sizes for ANN Forecasts and The Response Time Distributions.

Station	RTD	Jan.	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
Famona	2+GAMM(22;1.48)	2	2	2	2	2	2	2	2	2	2	2	2
Northend	2+GAMM(23.9;1.36)	2	2	2	2	2	2	2	2	2	2	2	2
Nketa	-0.001+ERLA(18.5;2)	5	5	5	5	5	5	5	5	5	5	5	5
Nkulumane	0.999+GAMM(21.8;1.62)	2	2	2	2	2	2	2	2	2	2	2	2
Ambulance	Optimal Deployment	11	11	11	11	11	11	11	11	11	11	11	11
Deployment	Current Fleet Size	6	6	6	6	6	6	6	6	6	6	6	6
Status	Deficit	5	5	5	5	5	5	5	5	5	5	5	5
Famona	U(10;15)	1	2	2	1	1	1	1	1	2	1	1	2
Northend		2	2	2	2	2	2	2	2	2	2	2	2
Nketa		3	3	3	3	3	3	3	3	3	3	3	3
Nkulumane		1	1	1	1	1	1	1	1	1	1	1	1
Ambulance	Optimal Deployment	7	8	8	7	7	7	7	7	8	7	7	8
Deployment	Current Fleet Size	6	6	6	6	6	6	6	6	6	6	6	6
Status	Deficit	1	2	2	1	1	1	1	1	2	1	1	2
Famona	U(5;10)	1	2	2	1	1	1	1	1	2	1	1	2
Northend		2	2	2	2	2	2	2	2	2	2	2	2
Nketa		3	3	3	3	3	3	3	3	3	3	3	3
Nkulumane		1	1	1	1	1	1	1	1	1	1	1	1
Ambulance	Optimal Deployment	7	8	8	7	7	7	7	7	8	7	7	8
Deployment	Current Fleet Size	6	6	6	6	6	6	6	6	6	6	6	6
Status	Deficit	1	2	2	1	1	1	1	1	2	1	1	2

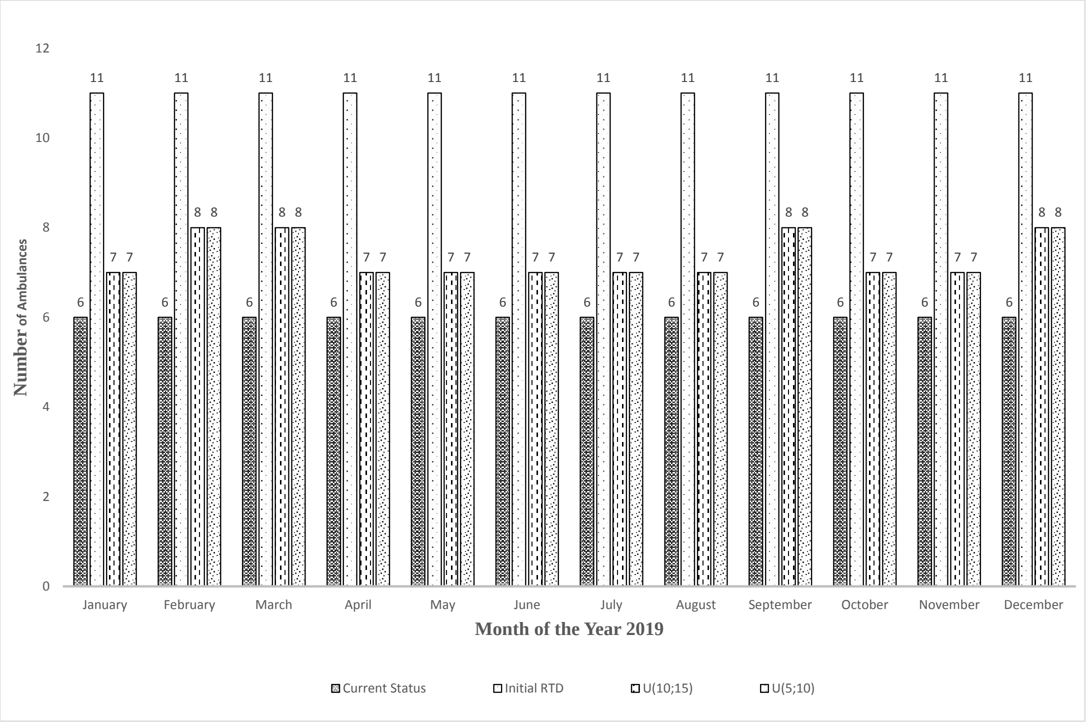


Figure 7.1: Comparison of Annual Optimum Deployment Plans for BEMS (2019).

7.5 Chapter Conclusion

In this chapter, numerical experiments were conducted by integrating the ANN public emergency ambulance demand (PEAD) forecasts, whilst adjusting the ambulance fleet sizes in order to optimise the levels of preparedness. The focus of this chapter was to explore the influence of varying the response time distribution on the optimum deployment plan to international standards of 5 to 10 minutes by adopting a uniform distribution given by $U(5; 10)$. This was strongly motivated by the fact that it is easier, cheaper and feasible for management to control some of the processes that are directly linked to the response time. Performance measures such as the average total entity time in the system, average response time, average response queue time, the average number of calls in the response queue, and the ambulance utility levels were considered in evaluating the models. Implications to managerial decision-making as a result of the findings were discussed in detail. A comparison of the deployment plans obtained from Chapter 6 and Chapter 7 were conducted in order to ascertain the influence of the response time as a parameter on the performance of the simulation models. It can be concluded from the results that for medical resources such as ambulances, it does not always translate that, the more resources deployed the better the performance. The increase in the number of ambulances does not always positively influence the average response time. When the number of ambulances exceeds a certain threshold, the average response time stays at a certain level rather than decreasing gradually. It is also imperative to consider simultaneously multiple performance indicators such as the average total duration of a call in the system, average number of calls in queue, the average call queuing time, throughput ratios and the ambulance utilisation levels to complement the average response time as a performance measure. It is imperative for management to seriously consider ways of reducing the response time as it has significant bearing in reducing the required number of ambulances, a critical but scarce resource. It would also translates to reduced workloads on the response teams. However, based on the scientific evidence, management could lobby for the decongestion and resurfacing of old and dilapidated roads in order to increase access and speed when responding to emergency calls. Training and provision of appropriate and modern equipment to the response teams will go a long way in reducing the chute delay time.

Chapter 8

Discussion and Conclusion

8.1 Introduction

The overall cross-cutting objective of the thesis was to develop and demonstrate the applicability of a strategy to integrate forecasting using artificial neural networks (ANN), simulation and optimisation techniques for ambulance deployment to predefined locations with heterogeneous demand patterns under stochastic environments, using multiple performance indicators. Chapter 8 forms the basis of formulating the conclusion of the thesis as it highlights the discourse from Chapter 1 to Chapter 7. It discusses the findings and makes conclusions to the study.

8.2 Discussion

The Bulawayo ambulance service department continues to rely on judgemental methods with limited use of historical data for future predictions. The organisation continues to face problems, which are common to any ambulance service provider all over the world today. Despite the ever-increasing rural to urban migration which has seen the population of urban areas in Zimbabwe increasing, Bulawayo City included, heath services delivery remains low in terms of efficiency, effectiveness and equality. This migration trend has seen the increase of unemployment in urban areas, the sprouting of new residential areas, both formal and informal and this has put pressure on the limited resources in terms of housing, health and education among many other critical human social amenities.

It is expected that as the population increases, there is a subsequent increase in demand for public emergency services. This has not been the case with the demand trends for public emergency ambulance in Bulawayo. However, the trend has been decreasing and this observed decrease might be a sign of limited capacity of the service rather than demand itself. The situation therefore, called for consolidated efforts across all sectors including research, in-order to restore confidence among residents, reduce health risks and loss of lives. Planning into the future to restore a balanced emergency ambulance level of preparedness is necessary in order to prevent the loss of lives in isolated incidences or even a disaster such as cholera outbreaks, fire or major road accidents that have been experienced in recent years in Bulawayo.

One of the key objectives of this thesis was to determine the operational and technical challenges being faced by Bulawayo Emergency Medical Services (BEMS) in terms of emergency medical service (EMS) delivery. The second objective was to make a comparative analysis on the predictive power between artificial neural networks (ANN) and traditional seasonal autoregressive integrated moving average models (SARIMA) in forecasting annual public emergency ambulance demand (PEAD). The other broad key objective was to integrate forecasting (using artificial neural networks), simulation and optimisation techniques in decision-making process in ambulance preparedness for heterogeneous regions using multiple performance indicators. This was envisaged to address issues around equality, effectiveness and efficiency in service delivery.

Chapter 1 focused on the introduction of the thesis by providing a background of the study, problem statement, motivation of the study, research objectives, research questions, expected benefits of the study and the delimitation of the study. Chapter 2 focused on the review of related literature. It explored both nonlinear and linear time series models with emphasis placed on artificial neural networks and SARIMA models suitable for forecasting PEAD for short-term forecasting horizons. Literature on forecasting, simulation modelling and optimisation techniques was explored considering the gaps that exists in the body of knowledge with regards to emergency medical services. The review of literature culminated in the development of a conceptual framework for the study.

Chapter 3 dwelt on the methodology of the study aimed at integrating forecasting, simulation and optimisation techniques for ambulance preparedness. Justifications and empirical evidence on the use of simulation for modelling the EMS were presented. The data used in this study was collected from the Bulawayo Municipality archives from 2010 to 2018 for model building and validation. A combination of methods based on mathematics, statistics, operations research and computer science were used for data manipulation. Various computing packages such as R, Minitab, Excel and Arena were employed and integrated for data analysis, model building and sensitivity analysis. Chapter 4 focused on a comparative analysis of the predictive power of artificial neural networks and SARIMA models in forecasting PEAD for 2019. The mean absolute error (MAE), root mean square error (RMSE) and the paired sample t-test were used as performance measures to select the best demand forecasting model.

Chapter 5 presented the development of simulation models for the heterogeneous regions of Bulawayo city using Arena. Chapter 6 presented the process of sensitivity analysis by integration simulation and optimisation techniques in determining optimum ambulance deployment plans under multiple performance indicators. The performance indicators included average response time, average total duration time of the call in the system, average response queue time, average number of calls in the response queue, throughput ratio and the ambulance utilisation levels. Further analysis was implemented in integrating ANN forecasts under multiple performance indicators whilst maintaining the response time distributions in determining optimal deployment plans for the four heterogeneous regions namely: Famona, Northend, Nketa and Nkulumane. Chapter 7 presented numerical experiments that explored the influence of varying the response time distributions to the international standard of between 5 to 10 minutes by adopting a uniform distribution $U(5;10)$ on the optimal deployment plan. However, in order to trace any trends, a uniform distribution given by $U(10;15)$ was also explored. Chapter 8 focused on the discussion of the results and conclusion of the thesis.

In the comparative study on modelling univariate time series ambulance demand, the performance measures; MAE and the paired sample t-test indicate that the feed-forward neural network (FFNN) models are superior to traditional SARIMA models in time series prediction of ambulance demand in the city of Bulawayo, over a short-term forecasting horizon. It was discovered that the FFNN model is more inclined to value estimation as compared to the SARIMA model, which is directional as depicted by the linear pattern over time. The FFNN model derives its model prediction accuracy from this unique characteristic. The FFNN model building process used 96 observations, where seventy-two (72) and twenty (24) observations were used as training and testing sets respectively. With such small sample data, the FFNN model was able to accomplish accurate predictions as it was able to detect the hidden patterns of the time series better than the SARIMA model. The SARIMA model required the assumption of stationarity to be satisfied before model building and also to be verified post-model development. However, such assumptions were not required in the development of FFNN. Therefore the FFNN is a parsimonious, simple model with no assumptions that requires fewer variables in the model building but with greater explanatory power as compared to the SARIMA model. It was observed that using the architecture of one hidden layer produced more accurate results than those obtained from architectures with two hidden layers for short-term forecasting horizons. Therefore, the use of a single hidden layer is adequate in developing FFNN for short-term forecasting horizons. Reducing the number of input neurons did not improve the model accuracy. An ANN model with a 7-(4)-1 architecture was selected to forecast the 2019 public emergency ambulance demand (PEAD). Peak PEAD is expected in January, March, September, and December whilst lower demand is expected for April, June, and July 2019.

Probabilistic and stochastic simulation model input parameters were developed using the 2018 data to capture the random or stochastic nature of the inter-arrival rates of calls, response time, service time, occurrences of emergency calls, and their levels of severity due to the heterogeneous demand zones namely: Famona, Northend, Nketa, and Nkulumane. The number of false alarm malicious (FAM) and false alarm good intent (FAGI) calls were prevalent in the eastern suburbs (Famona and Northend) as compared to the western suburbs (Nketa and Nkulumane). Implications are that eastern suburb residents find themselves with a wide range of alternatives for health emergencies resulting in more cases of FAGI. This however, justifies the need for equitable deployment of ambulance resources to meet the heterogeneous needs of the populace by ensuring that ambulances are deployed where they are needed most. Simulation models developed mimicked the prevailing levels of service for BEMS with six(6) operational ambulances. The general simulation models developed indicated that average response times are well above 15 minutes, with significantly high average queuing times and the number of ambulances queuing for service. These performance outcomes are highly undesirable as they pose a great threat to human-based outcomes of safety and satisfaction with regards to service delivery. The general expectation is that no call should queue for service. Hence, there was a need to determine the optimum ambulance deployment plans that minimises the response time whilst adjusting for the number of ambulances needed to provide a specific service level.

Optimisation for simulation was conducted by simultaneously minimising the average response time, average queuing time and maximizing throughput ratios and utililisation levels. Increasing the number of ambulances influenced the average response time below a certain threshold, beyond this threshold, the average response time stayed at a certain level rather than decreasing gradually and no significant changes occurred in other performance measures. Ambulance utilisation inversely varied to increase in the fleet size. A total of eleven(11) ambulances are required to meet the future demand of 2019. Under these prevailing conditions, there is a deficit of five(5) ambulances to maintain a balanced optimal fleet size where queues and queuing time for an ambulance are reduced to zero. However, the average response times remained high, above the recommended international standards.

The influence of varying the response time distributions on the optimum deployment plans to international standards of 5 to 10 minutes by adopting a uniform distribution given by $U(5;10)$ was explored using numerical experiments. The ANN public emergency ambulance demand (PEAD) forecasts were incorporated, whilst adjusting the ambulance fleet sizes in order to optimise the levels of preparedness. This was strongly motivated by the fact that it is easier, cheaper and feasible for management to control processes that are directly linked to the response time such as pre-trip delays, chute time and queuing time. The adoption of $U(5;10)$ resulted in a decrease in the total ambulance deployment from eleven(11) to eight(8) ambulances. This implies that reducing the response time results in the reduction in number of ambulances required for optimal ambulance deployment.

8.3 Conclusion

The researcher recommends that Bulawayo City Council should deliberately adopt and integrate artificial neural networks as key forecasting tools to assist in strategic resource planning activities due to its explanatory power. Key ambulance logistic activities such as vehicle servicing, replenishment of essential equipment and drugs, staff training, leave days scheduling and mock drills need to be planned for when low demand is predicted. This deliberate planning strategy would avoid a dire situation whereby ambulances are available but without adequate staff, essential drugs and equipment to respond to public emergency calls. It is imperative for EMS research to simultaneously consider multiple performance indicators to complement the average response time in ambulance deployment. This goes a long way in balancing resource allocation and capacity utilisation to avoid the idleness of essential equipment and human resources.

For medical resources such as ambulances, the more resources deployed does not always translate to better performance. Decision-makers in EMS must seriously consider ways of reducing the response time as it has significant bearing in reducing the required number of ambulances, a critical but scarce resource. Efforts must be directed towards digitisation of switch boards in the call center, training of the paramedics and provision of relevant modern equipment to the response teams. Training and provision of appropriate and modern equipment to the response teams will go a long way in reducing the call delay time, chute time and ultimately the response time. This also translates to reduced workloads on the response teams as fewer ambulances are required implying that a few response teams are required to meet the public emergency ambulance demand needs. Based on the scientific evidence, management could lobby for de-congestion and resurfacing of old and dilapidated roads in order to increase access and speed when responding to emergency calls.

An important contribution of this paper was to develop and demonstrate a framework for integrating forecasting, simulation and optimisation techniques for ambulance deployment in a heterogeneous region under multiple performance measures. The methodology removed several simplifying assumptions that are necessary for other models. The proposed strategy ensures that ambulances are deployed where they are needed most (equity), ambulances meet the set performance targets (effectiveness) and ensures that throughput ratios and utilisation levels remain significantly high (efficiency) to avoid idleness or excessive overloads. The simplified strategy is key for preparedness and can be adapted to operational environments that involve a server-to-customer relationship with relative ease. Such could involve the operation of a fleet of taxis or breakdown recovery vehicle service. Future research should endeavour to investigate the influence of varying service times on the optimum deployment plans and also consider an extension of the study to include operational costs, wages and other budgetary constraints that influence the allocation of critical but scarce resources such as personnel, equipment and emergency ambulance response vehicles.

Appendix A

Proof of Publications

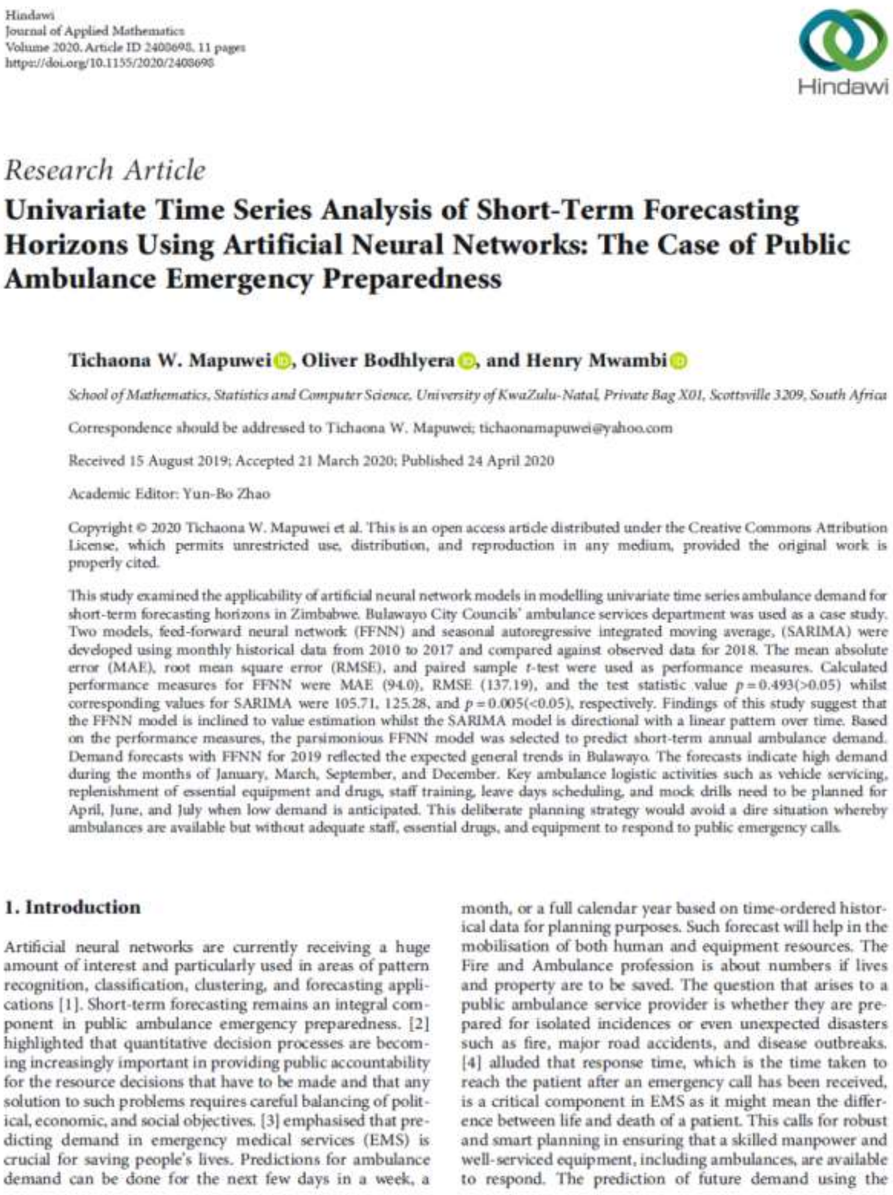
Three research papers are under consideration here. One(1) paper has been published and two(2) papers were submitted for reviewing purposes pending publication. The Title of the research papers, journals and authors are listed as follows:

- **Univariate Time Series Analysis of Short-term Forecasting Horizons Using Artificial Neural Networks: The Case of Public Ambulance Emergency Preparedness**. *Tichaona W. Mapuwei, Oliver Bodhylera and Henry Mwambi. Journal of Applied Mathematics - Hindawi. Volume 2020, https://doi.org/10.1155/2020/2408698.*

- **Integrating Artificial Neural Networks, Simulation and Optimisation Techniques in Ambulance Deployment for Heterogeneous Regions**. *Tichaona W. Mapuwei, Oliver Bodhylera and Henry Mwambi. Journal of Simulation - Taylor and Francis.*

- **Assessing the Impact of Varying Response Time Distributions on Ambulance Deployment Plans in Heterogeneous Regions Using Multiple Performance Indicators**. *Tichaona W. Mapuwei, Oliver Bodhylera and Henry Mwambi. Simulation Modelling Practice and Theory - Elservier.*

PUBLICATION NUMBER 1



Hindawi
Journal of Applied Mathematics
Volume 2020, Article ID 2408698, 11 pages
https://doi.org/10.1155/2020/2408698

Research Article

Univariate Time Series Analysis of Short-Term Forecasting Horizons Using Artificial Neural Networks: The Case of Public Ambulance Emergency Preparedness

Tichaona W. Mapuwei, Oliver Bodhlyera, and Henry Mwambi

School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Private Bag X01, Scottsville 3209, South Africa

Correspondence should be addressed to Tichaona W. Mapuwei; tichaonamapuwei@yahoo.com

Received 15 August 2019; Accepted 21 March 2020; Published 24 April 2020

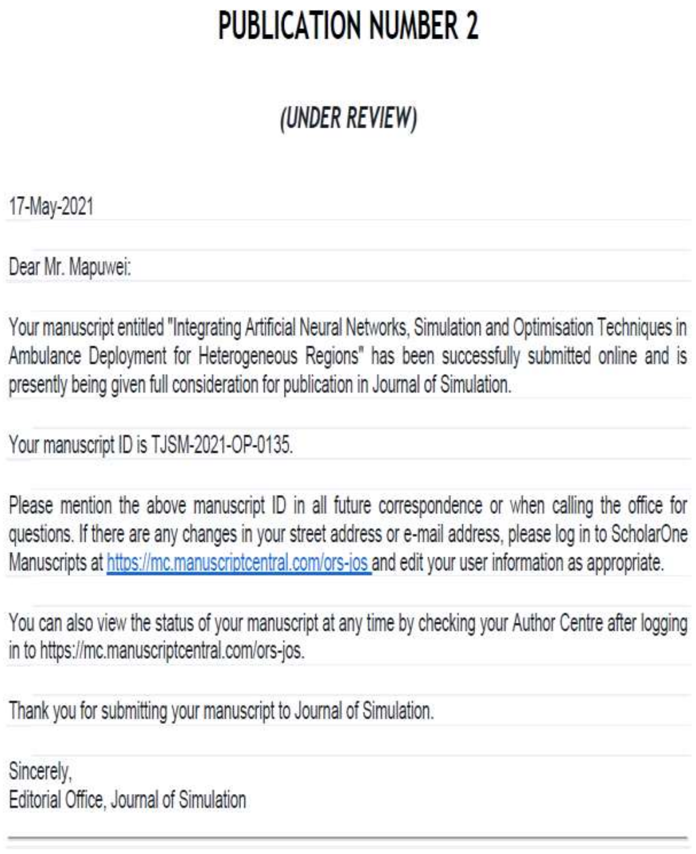
Academic Editor: Yun-Bo Zhao

This study examined the applicability of artificial neural network models in modelling univariate time series ambulance demand for short-term forecasting horizons in Zimbabwe. Bulawayo City Councils' ambulance services department was used as a case study. Two models, feed-forward neural network (FFNN) and seasonal autoregressive integrated moving average, (SARIMA) were developed using monthly historical data from 2010 to 2017 and compared against observed data for 2018. The mean absolute error (MAE), root mean square error (RMSE), and paired sample t-test were used as performance measures. Calculated performance measures for FFNN were MAE (94.0), RMSE (137.19), and the test statistic value $p = 0.493 (>0.05)$ whilst corresponding values for SARIMA were 105.71, 125.28, and $p = 0.005 (<0.05)$, respectively. Findings of this study suggest that the FFNN model is inclined to value estimation whilst the SARIMA model is directional with a linear pattern over time. Based on the performance measures, the parsimonious FFNN model was selected to predict short-term annual ambulance demand. Demand forecasts with FFNN for 2019 reflected the expected general trends in Bulawayo. The forecasts indicate high demand during the months of January, March, September, and December. Key ambulance logistic activities such as vehicle servicing, replenishment of essential equipment and drugs, staff training, leave days scheduling, and mock drills need to be planned for April, June, and July when low demand is anticipated. This deliberate planning strategy would avoid a dire situation whereby ambulances are available but without adequate staff, essential drugs, and equipment to respond to public emergency calls.

1. Introduction

Artificial neural networks are currently receiving a huge amount of interest and particularly used in areas of pattern recognition, classification, clustering, and forecasting applications [1]. Short-term forecasting remains an integral component in public ambulance emergency preparedness. [2] highlighted that quantitative decision processes are becoming increasingly important in providing public accountability for the resource decisions that have to be made and that any solution to such problems requires careful balancing of political, economic, and social objectives. [3] emphasised that predicting demand in emergency medical services (EMS) is crucial for saving people's lives. Predictions for ambulance demand can be done for the next few days in a week, a month, or a full calendar year based on time-ordered historical data for planning purposes. Such forecast will help in the mobilisation of both human and equipment resources. The Fire and Ambulance profession is about numbers if lives and property are to be saved. The question that arises to a public ambulance service provider is whether they are prepared for isolated incidences or even unexpected disasters such as fire, major road accidents, and disease outbreaks. [4] alluded that response time, which is the time taken to reach the patient after an emergency call has been received, is a critical component in EMS as it might mean the difference between life and death of a patient. This calls for robust and smart planning in ensuring that a skilled manpower and well-serviced equipment, including ambulances, are available to respond. The prediction of future demand using the

Figure A.1: Publication Number 1: Abstract



PUBLICATION NUMBER 2

(UNDER REVIEW)

17-May-2021

Dear Mr. Mapuwei:

Your manuscript entitled "Integrating Artificial Neural Networks, Simulation and Optimisation Techniques in Ambulance Deployment for Heterogeneous Regions" has been successfully submitted online and is presently being given full consideration for publication in Journal of Simulation.

Your manuscript ID is TJSM-2021-OP-0135.

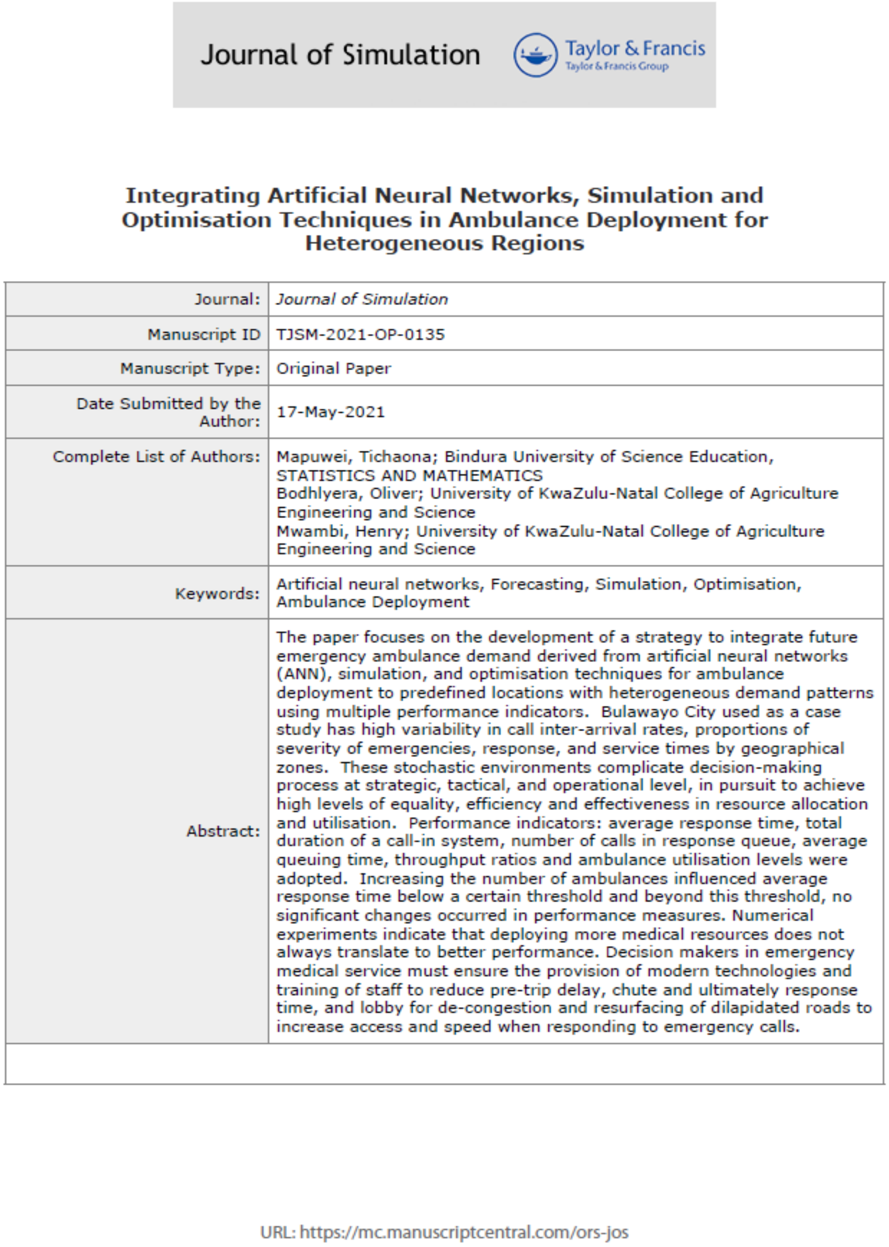
Please mention the above manuscript ID in all future correspondence or when calling the office for questions. If there are any changes in your street address or e-mail address, please log in to ScholarOne Manuscripts at https://mc.manuscriptcentral.com/ors-jos and edit your user information as appropriate.

You can also view the status of your manuscript at any time by checking your Author Centre after logging in to https://mc.manuscriptcentral.com/ors-jos.

Thank you for submitting your manuscript to Journal of Simulation.

Sincerely,
Editorial Office, Journal of Simulation

Figure A.2: Research Paper Number 2: Submission Communication Letter



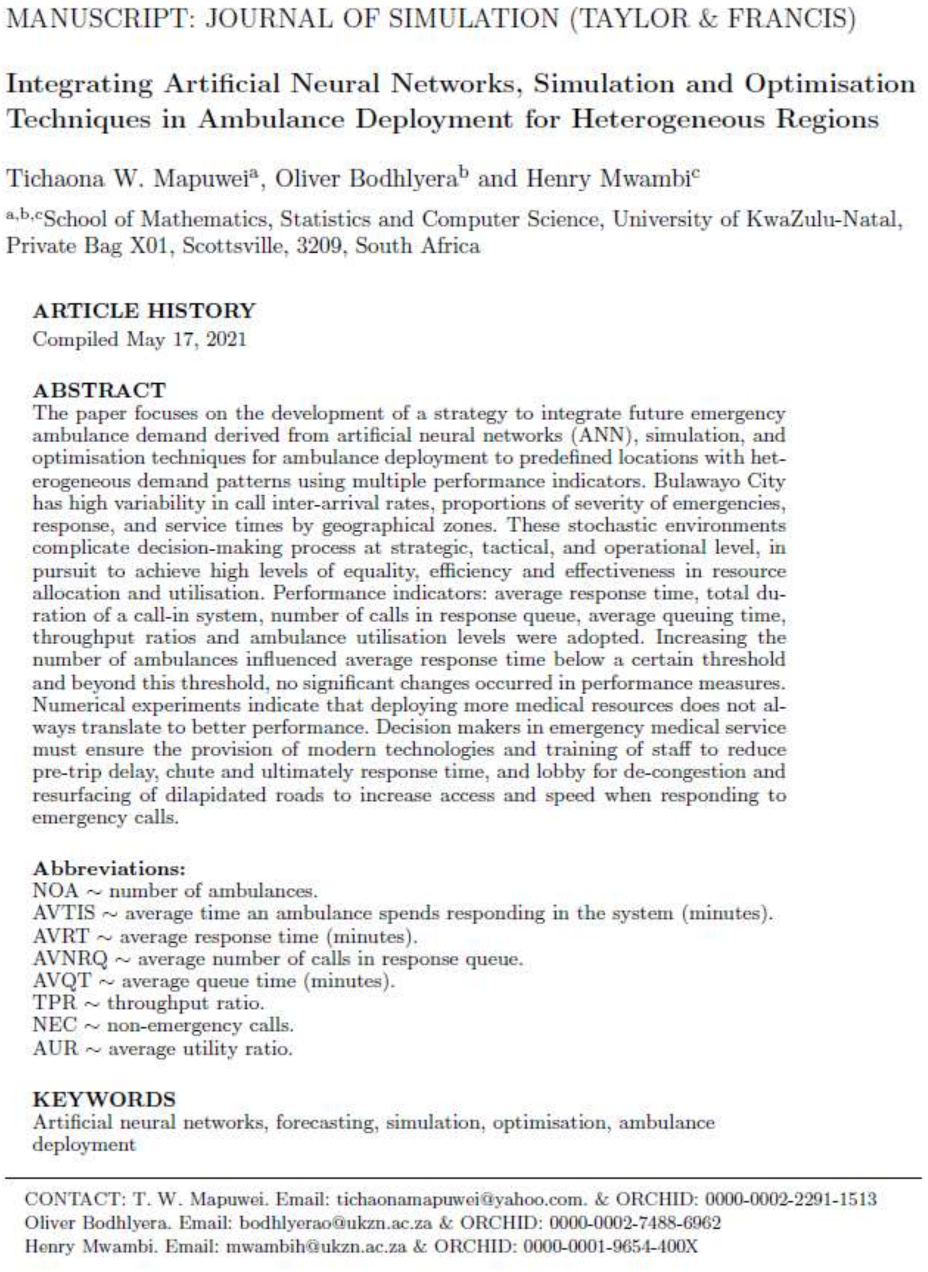
Journal of Simulation Taylor & Francis Taylor & Francis Group

Integrating Artificial Neural Networks, Simulation and Optimisation Techniques in Ambulance Deployment for Heterogeneous Regions

Journal:	*Journal of Simulation*
Manuscript ID	TJSM-2021-OP-0135
Manuscript Type:	Original Paper
Date Submitted by the Author:	17-May-2021
Complete List of Authors:	Mapuwei, Tichaona; Bindura University of Science Education, STATISTICS AND MATHEMATICS Bodhlyera, Oliver; University of KwaZulu-Natal College of Agriculture Engineering and Science Mwambi, Henry; University of KwaZulu-Natal College of Agriculture Engineering and Science
Keywords:	Artificial neural networks, Forecasting, Simulation, Optimisation, Ambulance Deployment
Abstract:	The paper focuses on the development of a strategy to integrate future emergency ambulance demand derived from artificial neural networks (ANN), simulation, and optimisation techniques for ambulance deployment to predefined locations with heterogeneous demand patterns using multiple performance indicators. Bulawayo City used as a case study has high variability in call inter-arrival rates, proportions of severity of emergencies, response, and service times by geographical zones. These stochastic environments complicate decision-making process at strategic, tactical, and operational level, in pursuit to achieve high levels of equality, efficiency and effectiveness in resource allocation and utilisation. Performance indicators: average response time, total duration of a call-in system, number of calls in response queue, average queuing time, throughput ratios and ambulance utilisation levels were adopted. Increasing the number of ambulances influenced average response time below a certain threshold and beyond this threshold, no significant changes occurred in performance measures. Numerical experiments indicate that deploying more medical resources does not always translate to better performance. Decision makers in emergency medical service must ensure the provision of modern technologies and training of staff to reduce pre-trip delay, chute and ultimately response time, and lobby for de-congestion and resurfacing of dilapidated roads to increase access and speed when responding to emergency calls.

URL: https://mc.manuscriptcentral.com/ors-jos

Figure A.3: Research Paper Number 2: Journal Details



MANUSCRIPT: JOURNAL OF SIMULATION (TAYLOR & FRANCIS)

Integrating Artificial Neural Networks, Simulation and Optimisation Techniques in Ambulance Deployment for Heterogeneous Regions

Tichaona W. Mapuwei[a], Oliver Bodhlyera[b] and Henry Mwambi[c]

[a,b,c]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Private Bag X01, Scottsville, 3209, South Africa

ARTICLE HISTORY
Compiled May 17, 2021

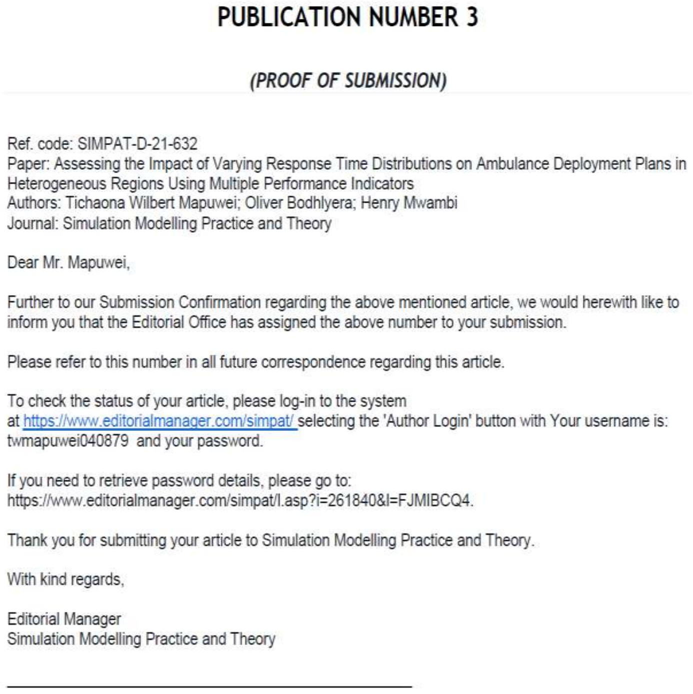
ABSTRACT
The paper focuses on the development of a strategy to integrate future emergency ambulance demand derived from artificial neural networks (ANN), simulation, and optimisation techniques for ambulance deployment to predefined locations with heterogeneous demand patterns using multiple performance indicators. Bulawayo City has high variability in call inter-arrival rates, proportions of severity of emergencies, response, and service times by geographical zones. These stochastic environments complicate decision-making process at strategic, tactical, and operational level, in pursuit to achieve high levels of equality, efficiency and effectiveness in resource allocation and utilisation. Performance indicators: average response time, total duration of a call-in system, number of calls in response queue, average queuing time, throughput ratios and ambulance utilisation levels were adopted. Increasing the number of ambulances influenced average response time below a certain threshold and beyond this threshold, no significant changes occurred in performance measures. Numerical experiments indicate that deploying more medical resources does not always translate to better performance. Decision makers in emergency medical service must ensure the provision of modern technologies and training of staff to reduce pre-trip delay, chute and ultimately response time, and lobby for de-congestion and resurfacing of dilapidated roads to increase access and speed when responding to emergency calls.

Abbreviations:
NOA ∼ number of ambulances.
AVTIS ∼ average time an ambulance spends responding in the system (minutes).
AVRT ∼ average response time (minutes).
AVNRQ ∼ average number of calls in response queue.
AVQT ∼ average queue time (minutes).
TPR ∼ throughput ratio.
NEC ∼ non-emergency calls.
AUR ∼ average utility ratio.

KEYWORDS
Artificial neural networks, forecasting, simulation, optimisation, ambulance deployment

CONTACT: T. W. Mapuwei. Email: tichaonamapuwei@yahoo.com. & ORCHID: 0000-0002-2291-1513
Oliver Bodhlyera. Email: bodhlyerao@ukzn.ac.za & ORCHID: 0000-0002-7488-6962
Henry Mwambi. Email: mwambih@ukzn.ac.za & ORCHID: 0000-0001-9654-400X

Figure A.4: Research Paper Number 2: Abstract of Research Paper



PUBLICATION NUMBER 3

(PROOF OF SUBMISSION)

Ref. code: SIMPAT-D-21-632
Paper: Assessing the Impact of Varying Response Time Distributions on Ambulance Deployment Plans in Heterogeneous Regions Using Multiple Performance Indicators
Authors: Tichaona Wilbert Mapuwei; Oliver Bodhlyera; Henry Mwambi
Journal: Simulation Modelling Practice and Theory

Dear Mr. Mapuwei,

Further to our Submission Confirmation regarding the above mentioned article, we would herewith like to inform you that the Editorial Office has assigned the above number to your submission.

Please refer to this number in all future correspondence regarding this article.

To check the status of your article, please log-in to the system at https://www.editorialmanager.com/simpat/ selecting the 'Author Login' button with Your username is: twmapuwei040879 and your password.

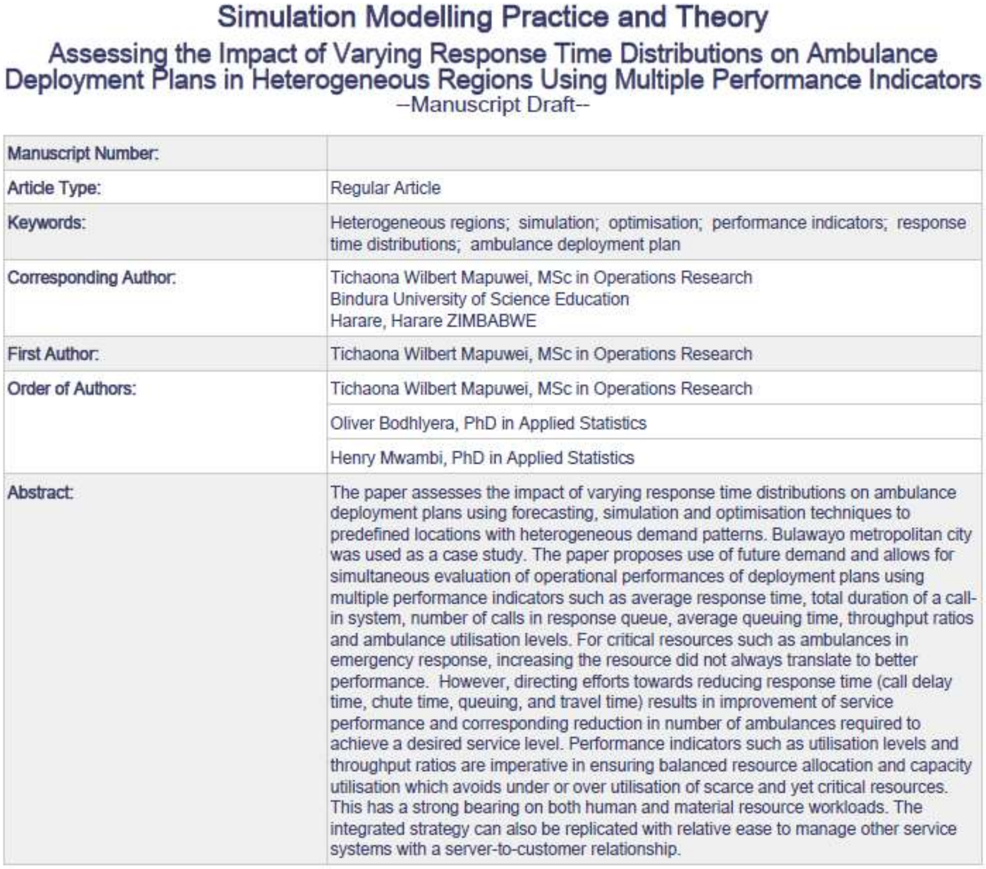
If you need to retrieve password details, please go to:
https://www.editorialmanager.com/simpat/l.asp?i=261840&l=FJMIBCQ4.

Thank you for submitting your article to Simulation Modelling Practice and Theory.

With kind regards,

Editorial Manager
Simulation Modelling Practice and Theory

Figure A.5: Research Paper Number 3: Submission Communication Letter



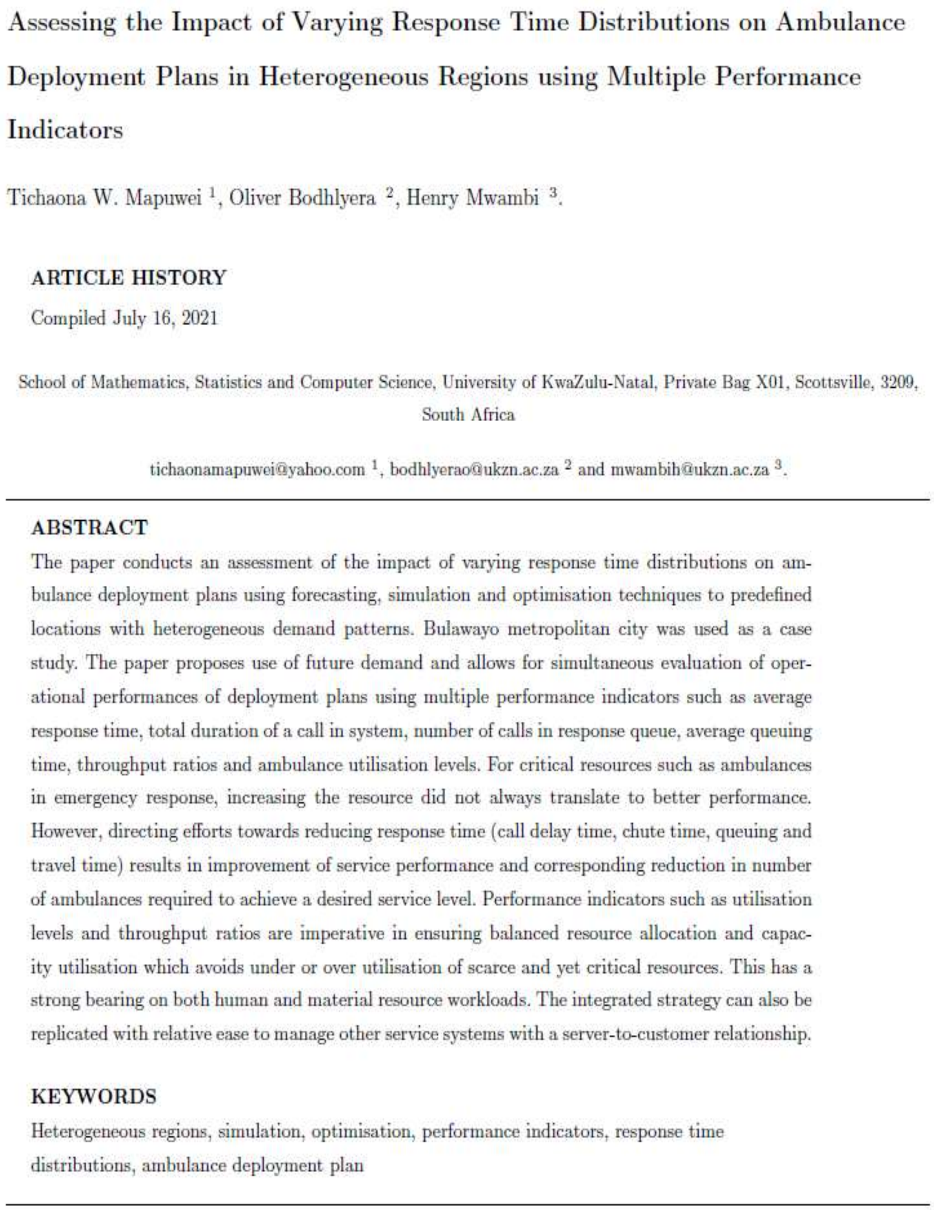
Simulation Modelling Practice and Theory

Assessing the Impact of Varying Response Time Distributions on Ambulance Deployment Plans in Heterogeneous Regions Using Multiple Performance Indicators

--Manuscript Draft--

Manuscript Number:	
Article Type:	Regular Article
Keywords:	Heterogeneous regions; simulation; optimisation; performance indicators; response time distributions; ambulance deployment plan
Corresponding Author:	Tichaona Wilbert Mapuwei, MSc in Operations Research Bindura University of Science Education Harare, Harare ZIMBABWE
First Author:	Tichaona Wilbert Mapuwei, MSc in Operations Research
Order of Authors:	Tichaona Wilbert Mapuwei, MSc in Operations Research
	Oliver Bodhlyera, PhD in Applied Statistics
	Henry Mwambi, PhD in Applied Statistics
Abstract:	The paper assesses the impact of varying response time distributions on ambulance deployment plans using forecasting, simulation and optimisation techniques to predefined locations with heterogeneous demand patterns. Bulawayo metropolitan city was used as a case study. The paper proposes use of future demand and allows for simultaneous evaluation of operational performances of deployment plans using multiple performance indicators such as average response time, total duration of a call-in system, number of calls in response queue, average queuing time, throughput ratios and ambulance utilisation levels. For critical resources such as ambulances in emergency response, increasing the resource did not always translate to better performance. However, directing efforts towards reducing response time (call delay time, chute time, queuing, and travel time) results in improvement of service performance and corresponding reduction in number of ambulances required to achieve a desired service level. Performance indicators such as utilisation levels and throughput ratios are imperative in ensuring balanced resource allocation and capacity utilisation which avoids under or over utilisation of scarce and yet critical resources. This has a strong bearing on both human and material resource workloads. The integrated strategy can also be replicated with relative ease to manage other service systems with a server-to-customer relationship.

Figure A.6: Research Paper Number 3: Journal Details



Assessing the Impact of Varying Response Time Distributions on Ambulance Deployment Plans in Heterogeneous Regions using Multiple Performance Indicators

Tichaona W. Mapuwei [1], Oliver Bodhlyera [2], Henry Mwambi [3].

ARTICLE HISTORY

Compiled July 16, 2021

School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Private Bag X01, Scottsville, 3209, South Africa

tichaonamapuwei@yahoo.com [1], bodhlyerao@ukzn.ac.za [2] and mwambih@ukzn.ac.za [3].

ABSTRACT

The paper conducts an assessment of the impact of varying response time distributions on ambulance deployment plans using forecasting, simulation and optimisation techniques to predefined locations with heterogeneous demand patterns. Bulawayo metropolitan city was used as a case study. The paper proposes use of future demand and allows for simultaneous evaluation of operational performances of deployment plans using multiple performance indicators such as average response time, total duration of a call in system, number of calls in response queue, average queuing time, throughput ratios and ambulance utilisation levels. For critical resources such as ambulances in emergency response, increasing the resource did not always translate to better performance. However, directing efforts towards reducing response time (call delay time, chute time, queuing and travel time) results in improvement of service performance and corresponding reduction in number of ambulances required to achieve a desired service level. Performance indicators such as utilisation levels and throughput ratios are imperative in ensuring balanced resource allocation and capacity utilisation which avoids under or over utilisation of scarce and yet critical resources. This has a strong bearing on both human and material resource workloads. The integrated strategy can also be replicated with relative ease to manage other service systems with a server-to-customer relationship.

KEYWORDS

Heterogeneous regions, simulation, optimisation, performance indicators, response time distributions, ambulance deployment plan

Figure A.7: Research Paper Number 3: Abstract of Research Paper

Bibliography

Aboueljinane, L., Sahin, E., Jemai, Z., and Marty, J. (2014). A simulation study to improve the performance of an emergency medical service: application to the french val-de-marne department. *Simulation modelling practice and theory*, 47:46–59.

Adebiyi, A. A., Adewumi, A. O., and Ayo, C. K. (2014). Comparison of arima and artificial neural networks models for stock price prediction. *Journal of Applied Mathematics*, 2014.

Aish, A. M., Zaqoot, H. A., and Abdeljawad, S. M. (2015). Artificial neural network approach for predicting reverse osmosis desalination plants performance in the gaza strip. *Desalination*, 367:240–247.

Alpaslan, F., Egrioglu, E., Aladag, C., and Tiring, E. (2012). An statistical research on feed forward neural networks for forecasting time series. *American Journal of Intelligent Systems*, 2(3):21–25.

Anderson, J. A. (1995). *An introduction to neural networks.* MIT press.

Anderson, O. D. (1977). The box-jenkins approach to time series analysis. *RAIRO - Operations Research - Recherche Opérationnelle*, 11(1):3–29.

Aras, S. and Kocakoç, İ. D. (2016). A new model selection strategy in time series forecasting with artificial neural networks: Ihts. *Neurocomputing*, 174:974–987.

Aringhieri, R., Bruni, M. E., Khodaparasti, S., and van Essen, J. T. (2017). Emergency medical services and beyond: Addressing new challenges through a wide literature review. *Computers & Operations Research*, 78:349–368.

Azadivar, F. (1999). Simulation optimization methodologies. In *Proceedings of the 31st conference on Winter simulation: Simulation—a bridge to the future-Volume 1*, pages 93–100.

Başar, A., Çatay, B., and Ünlüyurt, T. (2012). A taxonomy for emergency service station location problem. *Optimization letters*, 6(6):1147–1160.

Batta, R., Dolan, J. M., and Krishnamurthy, N. N. (1989). The maximal expected covering location problem: Revisited. *Transportation Science*, 23(4):277–287.

Beaudry, A., Laporte, G., Melo, T., and Nickel, S. (2010). Dynamic transportation of patients in hospitals. *OR spectrum*, 32(1):77–107.

Bélanger, V., Ruiz, A., and Soriano, P. (2019). Recent optimization models and trends in location, relocation, and dispatching of emergency medical vehicles. *European Journal of Operational Research*, 272(1):1–23.

Belayneh, A., Adamowski, J., Khalil, B., and Ozga-Zielinski, B. (2014). Long-term spi drought forecasting in the awash river basin in ethiopia using wavelet neural network and wavelet support vector regression models. *Journal of Hydrology*, 508:418–429.

Boujemaa, R., Jebali, A., Hammami, S., Ruiz, A., and Bouchriha, H. (2018). A stochastic approach for designing two-tiered emergency medical service systems. *Flexible Services and Manufacturing Journal*, 30(1-2):123–152.

Bowersox, D. J., Calantone, R. J., and Rodrigues, A. M. (2003). Estimation of global logistics expenditures using neural networks. *Journal of Business Logistics*, 24(2):21–36.

Channouf, N., L'Ecuyer, P., Ingolfsson, A., and Avramidis, A. N. (2007). The application of forecasting techniques to modeling emergency medical system calls in calgary, alberta. *Health care management science*, 10(1):25–45.

Chatfield, C. and Weigend, A. S. (1994). Time series prediction: Forecasting the future and understanding the past: Neil a. gershenfeld and andreas s. weigend, 1994,'the future of time series', in: As weigend and na gershenfeld, eds.,(addison-wesley, reading, ma), 1-70. *International Journal of Forecasting*, 10(1):161–163.

Cheng, B. and Titterington, D. M. (1994). Neural networks: A review from a statistical perspective. *Statistical science*, pages 2–30.

Chronicle (2018). Ten cholera cases in bulawayo. https://www.chronicle.co.zw/Accessed 6 June 2019.

Church, R. and ReVelle, C. (1974). The maximal covering location problem. In *Papers of the regional science association*, volume 32, pages 101–118. Springer-Verlag.

Ciaburro, G. and Venkateswaran, B. (2017). *Neural Networks with R: Smart models using CNN, RNN, deep learning, and artificial intelligence principles.* Packt Publishing Ltd.

Daskin, M. S. (1983). A maximum expected covering location model: formulation, properties and heuristic solution. *Transportation science*, 17(1):48–70.

Daskin, M. S. and Stern, E. H. (1981). A hierarchical objective set covering model for emergency medical service vehicle deployment. *Transportation Science*, 15(2):137–152.

Doerner, K. F., Gutjahr, W. J., Hartl, R. F., Karall, M., and Reimann, M. (2005). Heuristic solution of an extended double-coverage ambulance location problem for austria. *Central European Journal of Operations Research*, 13(4):325–340.

Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

Eaton, D. J., U, H. M. L. S., Lantigua, R. R., and Morgan, J. (1986). Determining ambulance deployment in santo domingo, dominican republic. *Journal of the Operational Research Society*, 37(2):113–126.

El Sayed, M. J. (2012). Measuring quality in emergency medical services: a review of clinical performance indicators. *Emergency medicine international*, 2012.

Eldabi, T. and Young, T. (2007). Towards a framework for healthcare simulation. In *2007 Winter Simulation Conference*, pages 1454–1460. IEEE.

Erkut, E., Ingolfsson, A., and Erdoğan, G. (2008). Ambulance location for maximum survival. *Naval Research Logistics (NRL)*, 55(1):42–58.

Fu, M. C. (2002). Optimization for simulation: Theory vs. practice. *INFORMS Journal on Computing*, 14(3):192–215.

Galvão, R. D., Chiyoshi, F. Y., and Morabito, R. (2005). Towards unified formulations and extensions of two classical probabilistic location models. *Computers Operations Research*, 32(1):15 – 33.

Gendreau, M., Laporte, G., and Semet, F. (2001). A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel computing*, 27(12):1641–1653.

Goutorbe, B., Lucazeau, F., and Bonneville, A. (2006). Using neural networks to predict thermal conductivity from geophysical well logs. *Geophysical Journal International*, 166(1):115–125.

Güler, İ. and Übeyli, E. D. (2005). An expert system for detection of electrocardiographic changes in patients with partial epilepsy using wavelet-based neural networks. *Expert Systems*, 22(2):62–71.

Günay, M. E. (2016). Forecasting annual gross electricity demand by artificial neural networks using predicted values of socio-economic indicators and climatic conditions: Case of turkey. *Energy Policy*, 90:92–101.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Henderson, S. and Mason, A. (2005). Ambulance service planning: Simulation and data visualisation. in: Brandeau m.l., sainfort f., pierskalla w.p. (eds) operations research and health care. *Springer*, 70:77 – 102.

Herliansyah, R. et al. (2017). Feed forward neural networks for forecasting indonesia exchange composite index. *GSTF Journal of Mathematics, Statistics & Operations Research*, 4(1).

Ingolfsson, A., Budge, S., and Erkut, E. (2008). Optimal ambulance location with random delays and travel times. *Health Care management science*, 11(3):262–274.

Jere, S., Kasense, B., and Bwalya, B. B. (2017). Univariate time-series analysis of second-hand car importation in zambia. *Open Journal of Statistics*, 7(4):718–730.

Kamenetzky, R. D., Shuman, L. J., and Wolfe, H. (1982). Estimating need and demand for prehospital care. *Operations Research*, 30(6):1148–1167.

Karami, A. (2010). Estimation of the critical clearing time using mlp and rbf neural networks. *European Transactions on Electrical Power*, 20(2):206–217.

Kheirkhah, A., Azadeh, A., Saberi, M., Azaron, A., and Shakouri, H. (2013). Improved estimation of electricity demand function by using of artificial neural network, principal component analysis and data envelopment analysis. *Computers & Industrial Engineering*, 64(1):425–441.

Kitapcı, O., Özekicioğlu, H., Kaynar, O., and Taştan, S. (2014). The effect of economic policies applied in turkey to the sale of automobiles: Multiple regression and neural network analysis. *Procedia - Social and Behavioral Sciences*, 148:653 – 661. 2nd International Conference on Strategic Innovative Marketing.

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

Lee, S. (2012). The role of centrality in ambulance dispatching. *Decision Support Systems*, 54(1):282–291.

Leknes, H., Aartun, E. S., Andersson, H., Christiansen, M., and Granberg, T. A. (2017). Strategic ambulance location for heterogeneous regions. *European Journal of Operational Research*, 260(1):122–133.

Mapuwei, T. W., Tsododo, M., Jakata, O., and Gondo, C. (2016). Demand modelling of alcoholic beverages in manicaland province using time series analysis. *International Journal of Innovative Science, Engineering Technology*, 3(10):333–340.

Matteson, D. S., McLean, M. W., Woodard, D. B., Henderson, S. G., et al. (2011). Forecasting emergency medical service call arrival rates. *The Annals of Applied Statistics*, 5(2B):1379–1406.

Mitrea, C., Lee, C., and Wu, Z. (2009). A comparison between neural networks and traditional forecasting methods: A case study. *International journal of engineering business management*, 1:11.

Montgomery, D. C. and Runger, G. C. (2014). *Applied statistics and probability for engineers*. John Wiley and Sons.

Neter, J., Wasserman, W., and Kutner, M. H. (1989). *Applied linear regression models*. Irwin Homewood, IL.

Noor, R. M., Ahmad, Z., Don, M. M., and Uzir, M. (2010). Modelling and control of different types of polymerization processes using neural networks technique: a review. *The Canadian Journal of Chemical Engineering*, 88(6):1065–1084.

Pinto, L., Silva, P., and Young, T. (2015). A generic method to develop simulation models for ambulance systems. *Simulation Modelling Practice and Theory*, 51:170–183.

Potts, D. (2000). Urban unemployment and migrants in africa: evidence from harare 1985–1994. *Development and Change*, 31(4):879–910.

Potts, D. (2006). 'restoring order'? operation murambatsvina and the urban crisis in zimbabwe. *Journal of Southern African Studies*, 32(2):273–291.

Prasad Y, J. and Bhagwat, S. S. (2002). Simple neural network models for prediction of physical properties of organic compounds. *Chemical Engineering & Technology: Industrial Chemistry-Plant Equipment-Process Engineering-Biotechnology*, 25(11):1041–1046.

Proietti, T. and Lütkepohl, H. (2013). Does the box–cox transformation help in forecasting macroeconomic time series? *International Journal of Forecasting*, 29(1):88–99.

Qasim, A. W. (2013). United nations development programme (undp). human development report 2013. *Pakistan Development Review*, 52(1):95–96.

Raeesi, M., Mesgari, M., and Mahmoudi, P. (2014). Traffic time series forecasting by feed-forward neural network: a case study based on traffic data of monroe. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40(2):219.

Rather, A. M., Agarwal, A., and Sastry, V. (2015). Recurrent neural network and a hybrid model for prediction of stock returns. *Expert Systems with Applications*, 42(6):3234–3241.

Repede, J. F. and Bernardo, J. J. (1994). Developing and validating a decision support system for locating emergency medical vehicles in louisville, kentucky. *European Journal of Operational Research*, 75(3):567 – 581.

Restrepo, M., Henderson, S. G., and Topaloglu, H. (2009). Erlang loss models for the static deployment of ambulances. *Health care management science*, 12(1):67.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Sadiq, T., B. Gharbi, R., and C. Juvkam-Wold, H. (2003). Use of neural networks for the prediction of frictional drag and transmission of axial load in horizontal wellbores. *International journal for numerical and analytical methods in geomechanics*, 27(2):111–131.

Schilling, D., Elzinga, D. J., Cohon, J., Church, R., and ReVelle, C. (1979). The team/fleet models for simultaneous facility and equipment siting. *Transportation Science*, 13(2):163–175.

Setzler, H., Saydam, C., and Park, S. (2009). Ems call volume predictions: A comparative study. *Computers Operations Research*, 36(6):1843 – 1851.

Sha, D. (2008). A new neural networks based adaptive model predictive control for unknown multiple variable non-linear systems. *International Journal of Advanced Mechatronic Systems*, 1(2):146–155.

Shumway, R. H. and Stoffer, D. S. (2000). Time series analysis and its applications. *Studies In Informatics And Control*, 9(4):375–376.

Shumway, R. H. and Stoffer, D. S. (2017). *Time series analysis and its applications: with R examples*. Springer.

Silva, P. M. S. and Pinto, L. R. (2010). Emergency medical systems analysis by simulation and optimization. In *Proceedings of the 2010 winter simulation conference*, pages 2422–2432. IEEE.

Storbeck, J. E. (1982). Slack, natural slack, and location covering. *Socio-Economic Planning Sciences*, 16(3):99–105.

Su, Q., Luo, Q., and Huang, S. H. (2015). Cost-effective analyses for emergency medical services deployment: A case study in shanghai. *International Journal of Production Economics*, 163:112 – 123.

Süt, N. and Şenocak, M. (2007). Assessment of the performances of multilayer perceptron neural networks in comparison with recurrent neural networks and two statistical methods for diagnosing coronary artery disease. *Expert Systems*, 24(3):131–142.

Swalehe, M. and Aktas, S. G. (2016). Dynamic ambulance deployment to reduce ambulance response times using geographic information systems: A case study of odunpazari district of eskisehir province, turkey. *Procedia Environmental Sciences*, 36:199–206.

Tawodzera, G. (2011). Vulnerability in crisis: urban household food insecurity in epworth, harare, zimbabwe. *Food Security*, 3(4):503–520.

Tealab, A. (2018). Time series forecasting using artificial neural networks methodologies: A systematic review. *Future Computing and Informatics Journal*, 3(2):334 – 340.

Toregas, C., Swain, R., ReVelle, C., and Bergman, L. (1971). The location of emergency service facilities. *Operations research*, 19(6):1363–1373.

Tsai, C.-F. (2008). Financial decision support using neural networks and support vector machines. *Expert Systems*, 25(4):380–393.

Wanas, N., Auda, G., Kamel, M. S., and Karray, F. (1998). On the optimal number of hidden nodes in a neural network. In *Conference Proceedings. IEEE Canadian Conference on Electrical and Computer Engineering (Cat. No. 98TH8341)*, volume 2, pages 918–921. IEEE.

Wilson, G. T., Reale, M., and Haywood, J. (2015). *Models for dependent time series*. Chapman and Hall/CRC.

Zhang, Z.-H. and Li, K. (2015). A novel probabilistic formulation for locating and sizing emergency medical service stations. *Annals of Operations Research*, 229(1):813–835.

Zhen, L., Wang, K., Hu, H., and Chang, D. (2014). A simulation optimization framework for ambulance deployment and relocation problems. *Computers & Industrial Engineering*, 72:12–23.

Zichun, Y. et al. (2012). The bp artificial neural network model on expressway construction phase risk. *Systems Engineering Procedia*, 4:409–415.