American Journal of Applied Mathematics and Statistics, 2022, Vol. 10, No. 3, 80-94
Available online at http://pubs.sciepub.com/ajams/10/3/3
Published by Science and Education Publishing
DOI:10.12691/ajams-10-3-3

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

Tichaona Wilbert Mapuwei[1,*], Oliver Bodhlyera[2], Henry Mwambi[2]

[1]Department of Statistics and Mathematics, Bindura University of Science Education, Bindura, Zimbabwe
[2]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa
*Corresponding author: tichaonamapuwei@yahoo.com

Received October 10, 2022; Revised November 15, 2022; Accepted November 24, 2022

Abstract The paper focuses on the development 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. The metropolitan city of Bulawayo was used as a case study with high variability in call inter-arrival rates, response times, service times, and proportions of severity of emergencies by geographical zones covered by sub-stations. These stochastic environments complicate the 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. This paper proposes an integrated simulation optimisation methodology that integrates 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. Increasing the number of ambulances influences the average response time below a certain threshold. Beyond this threshold, no significant changes occur in the performance measures. As the fleet size is increased, the ambulance utilisation levels decreased, hence there is always need to balance resource allocation and capacity utilisation to avoid idleness of essential equipment and human resources. Numerical experiments conducted to align the response time to international standards resulted in reduction in number of ambulances required for optimal deployment. For medical resources such as ambulances, deploying more resources do not always translate to better performance, hence there is need to simultaneously consider multiple performance measures. 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 as it will go a long way in reducing the pre-trip delay time, chute time and ultimately the response time. 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.

Keywords: *forecasting, artificial neural networks, simulation, optimisation, ambulance deployment*

Cite This Article: Tichaona Wilbert Mapuwei, Oliver Bodhlyera, and Henry Mwambi, "Integrating Artificial Neural Networks, Simulation and Optimisation Techniques in Ambulance Deployment for Heterogeneous Regions under Stochastic Environment." *American Journal of Applied Mathematics and Statistics*, vol. 10, no. 3 (2022): 80-94. doi: 10.12691/ajams-10-3-3.

1. Introduction

Reference [1], in a review paper highlighted that there is 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 systems overall performance and effectiveness of the different pre-hospital interventions. An 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 ability for an EMS to provide timely response is affected by the fleet size and locations of the ambulances. Even though there is no global standard response time (RT), rapid response is the main goal of most EMS systems [2]. To achieve such ambitious goals, there is need to focus on EMS aspects of strategic, tactical and operational problems affecting the ambulance service value chain [3]. According to [4] the rising costs of medical equipment, increasing call volumes, worsening traffic conditions in urban areas make emergency medical service control centres face increasing pressure to meet performance targets. Such endeavours have been hampered by the uneven distribution of the population in the city, distribution of health centres, medical response

vehicles, technical staff and the ambulance service stations resulting in failing to meet the performance measures such as the response time. [5] 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.

Reference [6] purported that artificial neural networks are receiving a huge amount of interest in areas of applications such as forecasting, pattern recognition, classification and clustering. According to [7], short-term forecasting remains an integral component in public ambulance emergency preparedness. A neural network is defined as a non-linear statistical model represented by a network diagram and are good at modelling any complex function where the relationship between variables is unknown [8]. According to [9], an artificial neural network (ANN) can also be defined as an information processing system that has been developed for generalisation of mathematical models of human neural biology. [7] emphasised that 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. [10] highlighted that ANN have been successfully applied and proven to be useful in time series modelling where the future values of a variable is determined using its past values.

Over the years, the complexity in which EMS evolve has led to the development of different models relating to the interaction between location, relocation and dispatch decisions. These models have been broadly classified as single coverage models, multiple coverage models, probabilistic and stochastic models, Stochastic and robust location-allocation models, fuzzy models, and human outcome based models [11]. Single coverage models whose objective was to minimise the number of vehicles required whilst ensuring that all zones were covered, formed the backbone of EMS research. According to [11], multiple coverage models were focused on the stochastic or randomness of the calls for vehicles and their availability whilst increasing the chances or likelihood of a demand zone covered. Probabilistic and stochastic models often referred to as expected covering location models, relied on the calculation of expected values of variables that are often characterised by uncertainty in their occurrence. The stochastic and robust location-allocation models endeavoured to account for the randomness of the call arrivals. According to [11], unlike the other discussed demand coverage maximization models, location-allocation inspired models aim to minimize costs under demand satisfaction constraints. [12] 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. [13] 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 turnaround time in solving them and this has reduced their attractiveness for adoption. According to [5], the paradigm of fuzzy models is mostly applicable to deal with the uncertainty in the number of emergency calls when the stochastic framework or the probabilistic paradigm cannot be used. The fuzzy paradigm allows the use of qualitative data as well as expert-based knowledge by characterising them as linguistic terms. With the advent of human outcome-based paradigm, 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. Reference [14] 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. Reference [11] 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.

Reference [5] indicated that equity is one of the most challenges in the health-care 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. In general practice, EMS providers are deemed to be providing equitable services if they favour disadvantaged groups. Reference [11] 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 stakeholders; patients' perspective and the service providers perspective. The patient is mainly concerned about fairness in the context of patients outcomes and patients waiting time whilst stakeholder perspective is mainly focused on issues around fairness in the distribution of workload which directly affects the retention of skilled personnel and the levels of attraction of new employees to the organization. Reference [5], 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. Reference [15] 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. Reference [4] 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 the evaluation of operational performances of deployment plans using a detailed simulation model.

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 behaviour of the system and or evaluating various strategies for the operation of the system. According to [16], 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.

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. Reference [17] 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, [18] 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 Horizonte in Brazil and the UK system. [19] carried out a simulation study to improve the performance of the EMS 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. [19] 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.

Several researchers have made similar attempts in conducting numerical experiments to specific areas across the world involving emergency medical services. [20] 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 pre-trip and queuing delays are far more 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 ambulances are available either busy attending to other calls. 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. They argued that reducing the response time by 20 seconds, is actually 20 seconds saved and it does not matter which component of response time these savings come from. The expectation is that reducing the response time has a huge bearing in improving service delivery, survival rates and patient satisfaction. [3] presented an 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 were able to demonstrate that as the response time decreases, there is 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 increase in demand for EMS provision.

Reference [4] 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. An important contribution of this paper is to demonstrate on how to integrate forecasting, simulation and optimisation techniques for ambulance deployment in a heterogeneous region under multiple performance measures. We conduct simulation modelling and experiments in which multiple performance measures (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) and different objectives (minimising average response time, minimising average queuing time and maximizing throughput ratios) are simultaneously conducted.

2. Materials and methods

In this section, sources of data, univariate time series analysis model using artificial network, simulation and optimisation techniques are discussed in detail.

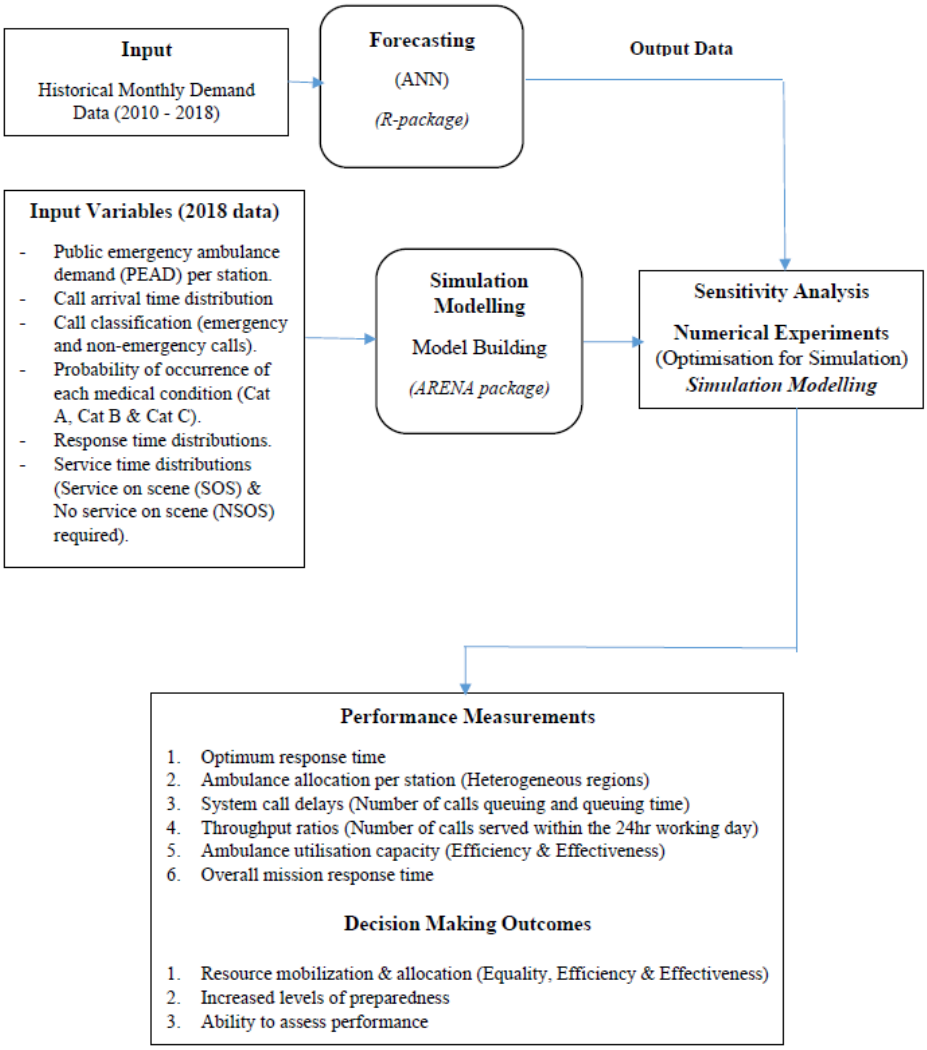

Figure 1. An Integrated Strategy for Ambulance Deployment

2.1. Model Input Data

Historical data of public emergency ambulance demand for the Bulawayo Emergency Medical Services (BEMS) January 2010 to December 2018, were retrieved from the archives for purpose of developing forecasting and simulation models. The flowchart of the methodology is presented in Figure 1.

2.2. Feed-Forward Neural Networks

The title should be formatted in an hourglass style; the first line longer than the second, the second line shorter than the third. Use numerical superscript callouts as shown in this template to link authors with their affiliations. Corresponding author should be denoted with an asterisk as shown. Email address is compulsory for the corresponding author.

The feed-forward neural network (FFNN) 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 layer to the output layer without recurrent or backward connections. Each layer consists of neurons and there is no connection between neurons that are in the same layer. The data splitting approach was adopted in order to develop and validate the feed-forward neural network. 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. This translates to 96 observations for model building and 12 observations for model cross-validation.

Data was scaled done using the minimum-maximum criteria to an interval (0,1) to prevent saturation in the hidden nodes and to assist t in the optimisation of the convergence rate during training of the neural network. Data was is to be split into training and testing sets of seventy-two (72) and twenty-four (24) observations respectively. This translates to 75% for training set and 25% for training set. The selection of the number of inputs in the model was based on trial and error as proposed by [10]. The general architecture of the FFNN can be generalised by equation 1.

$$I-(H_1,H_2,H_3,...,H_n)-O \quad (1)$$

where I represent the number of input nodes, H_n number of neurons in hidden layer n, and O the number of neurons in the output layer. An example is an ANN with seven (4) input nodes, one hidden layer with three (2) neurons and one (1) output neuron can be represented as 4-(2)-1 respectively. A three layer 4-(2)-1 FFNN architecture is shown in Figure 2.

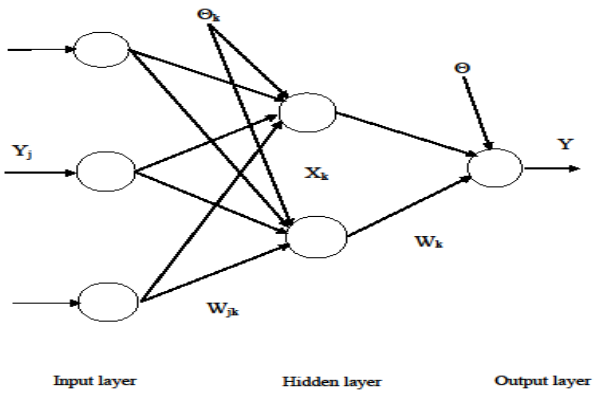

Figure 2. An Integrated Strategy for Ambulance Deployment

The input vector is represented by Y_j denoted by $Y_j = \{y_1 , y_2 , y_3 , y_4 \}$; W_{jk} (j = 1, 2, 3, 4; k = 1, 2) 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) is the vector of k neurons in the hidden layer; W_k (k = 1, 2) 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. Θ_k (k = 1, 2) is the bias value of the hidden layer nodes and Θ is the bias value of the output layer.

Supervised training with resilient backpropagation was adopted with 2017 demand calls as target values in the training algorithm. Training rate factors of 0.5 and 1.2 were implemented as the minimum and maximum values respectively. The momentum was set to assume default values whilst a threshold value was set at o.01 for the training data. The logistic function was implemented as activation function in the hidden layer. A single output neuron with a linear activation function was assumed. Number of hidden layers and neurons were systematically varied to obtain accurate models and the best model is based on mean absolute error (MAE) and means square error (MSE) as performance measures.

2.3. Simulation Modelling

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. The Bulawayo City, for the purposes of emergency response is demarcated into two broad regions, the eastern and western regions respectively. The eastern region covers basically 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 study will consider the geographical distribution of emergency calls in reference to the four stations: Famona, Northend, Nketa and Nkulumane. The study will determine the inter-arrival rates distributions for each station using the historical data in ARENA.

2.3.1. Call features, Types of Emergency and Ambulance Requirements

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. 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. The EMS crew is expected to provide service at the scene, deliver a patient to a medical institution, perform hand-over and take over procedure at medical centre, restocking and fuelling of vehicle.

The types of emergencies were broadly categorised in three (3) distinct categories with assigned unique codes for tracking, rescue team deployment and reporting purposes. These are summarised in Table 1. For simulation modelling purposes, the model will adopt the codes Cat A, Cat B and Cat C for distinguishing the different emergency response categories.

Table 1. Emergency Response Categories and Codes

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

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 assist in decision making in determining 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 Figure 3.

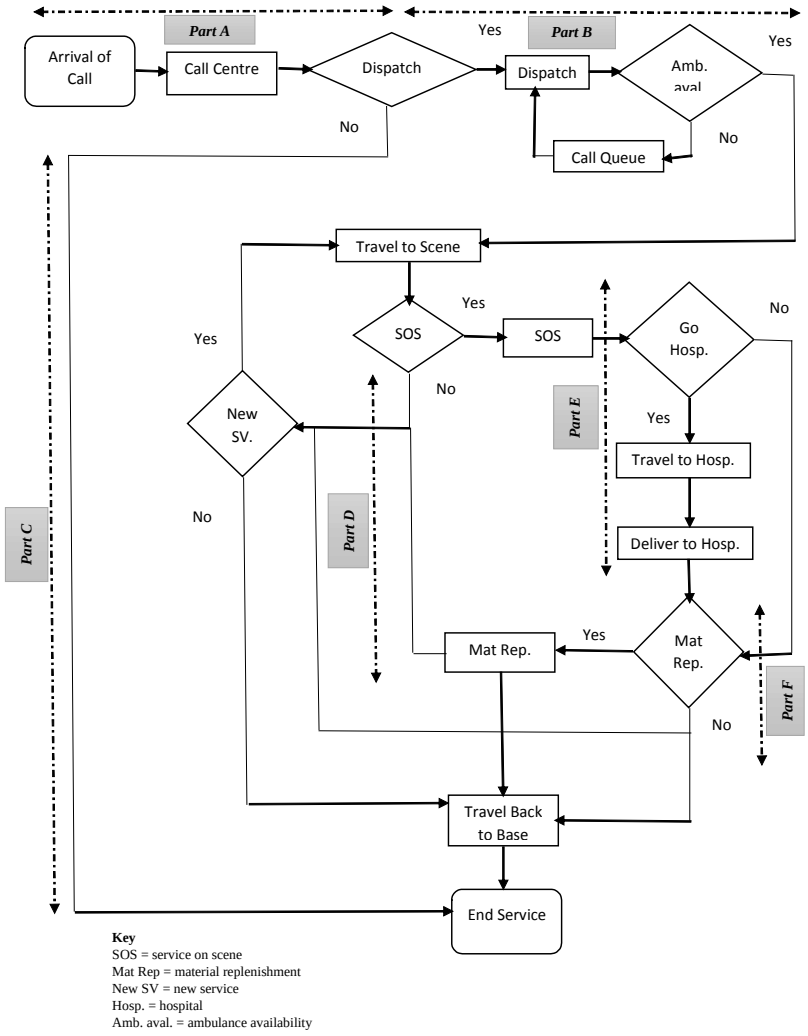

Figure 3. BEMS Multi-location Dispatch Model

Part A of the dispatch model represents 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 were service on the scene is not required. Normally these are calls recorded as false and malicious alarm (FAM) or false alarm good intention (FAGI). Part E represents a case were service on the scene is required and 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 refuelling.

2.3.2. Data Manipulation and Analysis

The simulation model to be developed will incorporate the randomness in call arrivals, travel times and service time. The model will assume that: (1) The arrival rate of calls may vary and is time dependent, (2) Calls are related with socio-economic conditions of the population, (3) Calls are serviced as per first come first serve (FIFO), (4) An ambulance could only serve one call at a time, (5) Ambulances have the same capacity in terms of size and equipment and that each ambulance team is made up of the driver and an attendant, (6) Ambulances are to be allocated randomly, (7) Response time is the time between the receipt of a call and to when the ambulance team arrive at the scene, (8) 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 vehicle is ready to depart for station and available to perform another task, and (9) Total duration in the system is the time from when a call is received up to when the ambulance is ready to depart for station ready to perform another task.

2.3.3. Estimation of Statistical Distributions of Simulation Model Parameters

The call inter-arrival time, response time, and service time distributions were generated in ARENA simulation package using the Input Analyser module on the 2018 historical data. The service time distributions were separated for cases were service needs to be rendered on scene (SOS) and cases were service on scene is not required (NSOS). The NSOS emanate from FAG and FAGI and in such cases results in less time required by the responding crew team. However, these occur in different proportions in the heterogenous regions of service and were computed separately. The selection of the best distribution is based primarily on the square error (she) and test for goodness of fit, which was performed using

non-parametric tests (Chi-square and the Kolmogorov-Smirnov tests), both embedded in the ARENA Input Analyser.

2.3.4. Estimation of Proportions of Model Parameters

The monthly and daily occurrences of demand per station will be 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 or category (Cat A, Cat B and Cat C) shall be computed in Excel based on the 2018 annual historical data.

2.3.5. Simulation Model Development and Performance Measures

The simulation model will be developed using 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, average number of calls in response queue, throughput ratios and capacity utilization levels of ambulances. Sensitivity and numerical experiments were conducted to achieve an in-depth analysis of the simulation models developed. Sensitivity analysis and numerical experiments entails changing model parameters and subsequently observing how these changes affect the general model performance and the deployment plan. The research will explore the following scenarios as part of simulation model development, sensitivity analysis and numerical experiments: (1) Optimum static ambulance deployment maintaining response time distributions (RTD), (2) Optimum static ambulance deployment to predicted ANN demand, maintaining the RTD and (3) Assess the influence of standardising the response time to international standards on the optimal ambulance deployment plan by adopting a uniform distribution given by U (10, 15).

3. Results and Discussion

Results on integrating forecasting, simulation and optimisation techniques for ambulance deployment are presented and discussed in this section.

3.1. Neural Network Ambulance Demand Forecasts

Several models of different architectures were systematically selected starting with two hidden units 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 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 2. 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.

Table 2. 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 minimum value of the performance measure across all models.

The selected neural network model was used to forecast the public emergency ambulance demand for 2019 and results are summarised in Table 3 and Figure 4. Important quantitative measures such as weekly and daily forecasts can be derived from such forecasts and can be fully utilized for strategic planning purposes. Demand is expected to be high in January, March, September and December whilst lower demand is projected for April, June and July 2019.

Table 3. 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

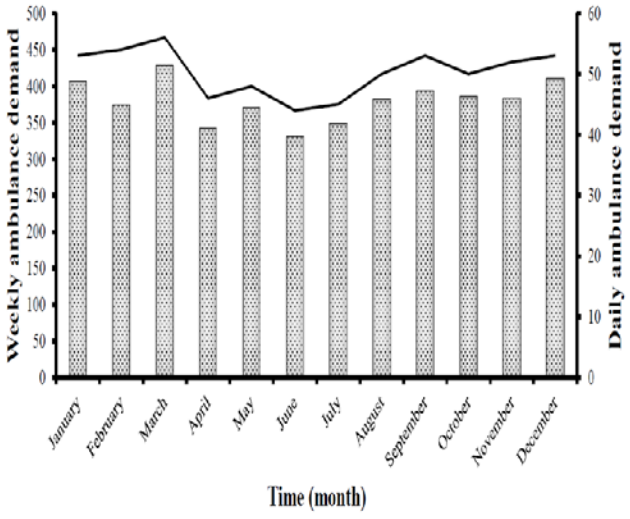

Figure 4. Projected monthly, weekly and daily ambulance demand calls for 2019

Table 4. Simulation Model Distributions of The Sub-stations

Famona	Inter-arrival time 0.999+WEIB (180;1.17)	Response Time 2+GAMM(22;1.48)
	Service on scene delay -0.001+164*BETA(2.7;6.47)	No-service on scene -0.5+72*BETA(0.606;1.2)
Northend	Inter-arrival time 3+GAMM(143;1.33)	Response Time 2+GAMM(23.9;1.36)
	Service on scene delay N(51.6;23.7)	No-service on scene -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
	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
	N(53;19.8)	-0.5+63*BETA(0.484;0.79)

3.2. 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 ARENA simulation package using the 2018 historical data. A summary of the results is presented in Table 4.

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 and presented in Table 5.

Table 5. Summary of Model Input Parameters

Item	Parameter	Frequency	Proportion	% Proportion
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 as: Cat A, Cat B or Cat C, together with their corresponding probability of occurrences. As these vary from one sub-station to another due to the heterogeneous regions they render service, computations were done separately for each sub-station. The service on scene delay (SOS) and no service on scene delay (NSOS) proportions of occurrence were also computed and the statistics are summarised in Table 6. The no service required on scene (NSOS) emergencies emanate from the FAM and FAGI where the general service time is smaller as compared to cases where service on the scene (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).

Table 6. 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	0.38	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 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 FAGI cases. This however, justifies the need for equitable deployment of ambulance resources to meet the heterogeneous needs of the populace.

3.3. Simulation Model Building for all Sub-stations

In developing the simulation model, the number of ambulances were incremented from one (1) to the stipulated number of allocated ambulances to each sub-station whilst changes in performance measures 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. According to the official reports from the department of fire brigade, Famona was allocated one (1), Northend one (1), Nketa three (3) and Nkulumane one (1) ambulance(s) respectively. A summary of the simulation models is given in Table 7.

Table 7. Simulation Model Performance Measures

Description	**Abbrev.**	**Famona**	**Northend**	**Nketa**			**Nkulumane**
Ambulance Numbers	**NOA**	**1**	**1**	**1**	**2**	**3**	**1**
Average time in system (min)	AVTIS	86.15	102.74	358.69	112.88	94.33	97.81
Average response time (min)	AVRT	40.51	58.7	306.04	64.92	49.02	43.21
Aveg. No in response queue	AVNRQ	0.04	0.19	4.95	0.42	0.05	0.02

Description	Abbrev.	Famona	Northend	Nketa			Nkulumane
Ambulance Numbers	**NOA**	**1**	**1**	**1**	**2**	**3**	**1**
Average queue time (min)	AVQT	6.58	24.99	289.73	26.07	3.21	3.67
Throughput ratio	TPR	8/9	11/11	16/22	22/23	24/24	7/7
Non-emergency calls	NEC	0	0	0	0	0	0
Amb. 1 utility		0.48	0.6	1.0	0.78	0.54	0.46
Amb. 2 utility					0.66	0.45	
Amb. 3 utility						0.53	
Average utility ratio	AUR	0.48	0.6	1.0	0.72	0.42	0.46

The models developed are adequately mimicking the prevailing EMS process for Bulawayo city. Results from Nketa station which was allocated more than one ambulance indicate that as the number of ambulance size increases, there is 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 response queue and the corresponding average time in queue also decrease as the number of ambulance size increases. The throughput ratio, increases with 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 recommended international standards of 10 to 15 minutes. 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 need to determine the optimum ambulance deployment models that minimise the number of ambulances needed to provide a specific service level. The next section seeks to address the issue by adopting optimisation for simulation through the use of sensitivity analysis.

3.4. Integrating ANN PEAD Forecasts in Ambulance Deployment

The objective is to integrate the ANN public emergency ambulance demand (PEAD) forecasts and optimisation for simulation through the use of sensitivity analysis, in determining the optimal deployment plans by varying the fleet sizes while observing the multiple performance measures.

3.4.1. 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. The proportions computed were as follows: Famona (18.1%), Northend (17.6%), Nketa (47.1%) and Nkulumane (17.2%) respectively. Theses proportions were then used to apportion the 2019 ANN forecasts as a build-up in the integration of forecasting and simulation concepts for future EMS preparedness. A summary of the expected monthly public ambulance demand calls per station are presented in Table 8.

Table 8. Expected Monthly Demand Per Station

Month	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

The expected average daily calls per station were further computed and results are summarised in Table 9. These calculated expected daily calls will be incorporated in the determination of optimal number of ambulances to be allocated in each station every month of 2019.

Table 9. Expected Daily Ambulance Demand Per Day

Month	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

3.4.2. Optimum Deployment Plan Strategy for BEMS

The ANN forecasts in Table 9 indicate that over the whole year of 2019, across the different 12 months would assume values of 8, 9 and 10 as expected daily number of calls for Famona, Northend and Nkulumane sub-stations. Nketa sub-station would assume values of 21, 23, 24, 25 and 26. The ambulance fleet sizes were incremented from one (1) whilst monitoring the performance measures. Summaries of the processes in determining the optimum development plans by integrating forecasting, simulation and optimisation techniques are presented as: Famona - Table 10, Northend - Table 11, Nketa - Table 12 and Nkulumane - Table 13.

It was observed that increasing the number of ambulances influences the average response time below a certain threshold. When fleet size is increased beyond this threshold, no significant changes occur in the performance measures. As the fleet size is increased, the ambulance utilisation levels decreased. Hence, there is need to balance resource allocation and capacity utilisation to avoid idleness of essential equipment and human resources. 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. The Optimum ambulance deployment plan are summarised in Table 14.

Table 10. Optimum Deployment Plan: Famona Station

Calls (N)	NOA	AVTIS (min)	AVRT (min)	AVNRQ	AVQT (min)	TPR	AUR (%)	NSOS	SOS
8	1	86.15	38.80	0.03	5.32	8/8	45	16.44	65.9
	2	76.06	33.07	0.0	0.0	8/8	21	15.62	64.37
	3	73.06	33.07	0.0	0.0	8/8	14	15.62	64.37
9	1	86.15	40.51	0.04	6.58	8/9	48	16.44	65.9
	2	73.06	32.29	0.0	0.0	8/9	23	15.62	64.37
	3	76.06	32.29	0.0	0.0	8/9	15	15.62	64.37
10	1	86.15	40.51	0.04	6.58	8/9	48	16.44	65.9
	2	73.06	32.29	0.0	0.0	8/9	24	15.62	64.37
	3	73.06	32.29	0.0	0.0	8/9	15	15.62	64.37

Table 11. Optimum Deployment Plan: Northend Station

Calls (N)	NOA	AVTIS (min)	AVRT (min)	AVNRQ	AVQT (min)	TPR	AUR (%)	NSOS	SOS
8	1	93.87	47.63	0.09	17.91	7/7	37	31.82	52.0
	2	94.34	49.26	0.0	0.0	6/6	20	34.47	55.68
	3	94.34	49.26	0.0	0.0	6/6	13	34.47	55.68
9	1	93.79	46.67	0.09	15.67	8/8	43	31.82	52.22
	2	85.65	49.26	0.0	0.0	6/6	18	27.78	53.63
	3	85.65	49.26	0.0	0.0	6/6	13	27.78	53.63
10	1	103.01	54.85	0.14	22.61	9/9	50	31.82	52.84
	2	78.89	44.40	0.0	0.0	7/7	19	26.84	53.63
	3	78.89	44.40	0.0	0.0	7/7	13	26.84	53.63

Table 12. Optimum Deployment Plan: Nketa Station

Calls (N)	NOA	AVTIS (min)	AVRT (min)	AVNRQ	AVQT (min)	TPR	AUR (%)	NSOS	SOS
21	2	120.75	66.38	0.30	24.20	18/18	61	53.18	54.61
	3	101.16	51.82	0.05	4.05	19/19	43	0.0	49.34
	4	93.72	44.02	0.01	0.97	20/20	32	11.84	51.7
	5	92.21	41.75	0.0	0.0	20/20	26	11.84	52.49
	6	92.21	41.75	0.0	0.0	20/20	26	11.84	52.49
23	2	113.3	61.06	0.3	22.48	19/19	60	53.18	52.06
	3	102.59	53.74	0.05	3.66	21/21	48	0.0	48.84
	4	94.47	42.95	0.01	0.88	22/22	36	11.84	53.41
	5	89.23	39.71	0.0	0.0	22/22	27	11.84	51.32
	6	89.23	39.71	0.0	0.0	22/22	23	11.84	51.32
24	2	112.16	59	0.3	21.36	20/20	63	53.18	53.16
	3	98.97	51.62	0.05	3.5	22/22	48	23.42	49.75
	4	95.74	42.95	0.01	0.88	22/22	37	11.84	54.74
	5	89.73	40.7	0.0	0.0	23/23	29	19.56	53.45
	6	89.73	40.7	0.0	0.0	23/23	24	19.56	53.45
25	2	115.24	59.58	0.31	20.98	21/21	69	53.18	56.08
	3	96.87	49.79	0.05	3.35	23/23	50	23.42	49.33
	4	95.74	41.49	0.01	0.84	22/23	39	11.84	54.74
	5	88.97	41.07	0.0	0.0	24/24	30	20.14	53.45
	6	88.97	41.07	0.0	0.0	24/24	25	20.14	53.45
26	2	113.51	62.25	0.35	23.01	22/22	69	43.77	52.93
	3	94.33	49.02	0.05	3.21	24/24	51	17.09	49.33
	4	95.74	41.49	0.01	0.81	22/23	37	11.84	54.74
	5	87.43	40.68	0.0	0.0	25/25	30	15.78	52.65
	6	87.43	40.68	0.0	0.0	25/25	25	15.78	52.65

Table 13. Optimum Deployment Plan: Nkulumane Station

Calls (N)	NOA	AVTIS (min)	AVRT (min)	AVNRQ	AVQT (min)	TPR	AUR (%)	NSOS	SOS
8	1	97.81	43.21	0.02	3.67	7/7	46	50.86	56.06
	2	83.04	43.95	0.0	0.0	6/7	20	34.5	53.54
	3	83.04	43.95	0.0	0.0	6/7	13	34.5	53.54
9	1	97.81	43.21	0.02	3.67	7/7	46	50.86	56.09
	2	83.04	43.95	0.0	0.0	6/7	20	34.5	53.54
	3	83.04	43.95	0.0	0.0	6/7	13	34.5	53.54
10	1	97.81	43.21	0.02	3.67	7/7	46	50.86	56.09
	2	83.04	43.95	0.0	0.0	6/7	20	34.5	53.54
	3	83.04	43.95	0.0	0.0	6/7	13	34.5	53.54

Table 14. Computational Fleet Sizes on Expected Daily Calls (N) from ANN Forecasts

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

Table 15. Optimal Fleet Size for ANN Expected Daily Ambulance Demand Forecast

	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

A summary of the annual deployment plan by integrating the annual ANN expected daily demand forecasts is presented in Table 15. The results imply that an optimum deployment plan of eleven (11) ambulances is adequate to meet future demand as predicted by ANN.

3.5. Numerical Experiments

Numerical experiments were conducted to determine optimal static ambulance deployment plan by varying the response time to international standards against the predicted ANN values. A uniform distribution U (10; 15) was adopted to represent the response time and would allow the response time to vary between 10 and 15 minutes. Performance measures such as 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 were used to evaluate the models. The simulation model parameters such as the inter-arrival of calls and service time distributions were maintained.

3.5.1. Comparison of Optimum Deployment Plans: Famona Station

There was need therefore to compare the different performance changes due to the influence of the changes in response time distributions on the optimal deployment plan for Famona Station. A summary of statistics of the optimum deployments are presented in Table 16.

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 from two (2) to one (1) as the response time is set between 10 to 15 minutes using the uniform distribution U (10; 15) for all the cases. Therefore, reducing the response time impacts positively on the performance measures. 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 despite the fact that less 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.

3.5.2. Comparison of Optimum Deployment Plans: Northend Station

A comparison on the performance changes due to the influence of changes in response time distribution on optimal deployment plan for Northend Station was performed. Summary statistics are presented in Table 17.

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 average total duration time of call-in 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 ~ U (10; 15).

3.5.3. Comparison of Optimum Deployment Plans: Nketa Station

A comparison on the performance changes due to the influence of changes in response time distribution on optimal deployment plan for Nketa Station was performed. A summary of statistics is presented in Table 18 for all cases (N=21, 23, 24, 25 and 26) respectively. Results indicate that reducing the response time by adopting a uniform distribution U (10; 15) resulted in the decrease of the optimal number of ambulances required to achieve an optimal deployment plan. The adoption of U (10; 15) would result in a drop in the threshold fleet size from five (5) to three (3) ambulances for all cases (N=21, 23, 24, 25 and 26). Utilisation capacity levels increased with reduced response time regardless of the decrease in ambulance fleet sizes for all 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.

Table 16. Comparison of Optimal Deployment Plans: Famona Station

ANN Forecasts	Response Time Distribution	Opt. NOA	AVTIS (min)	AVRT (min)	TPR	AUR (%)	NSOS (min)	SOS (min)
N=8	2+GAMM(22;1.48)	2	76.06	33.07	8/8	21	15.62	64.37
	U(10;15)	1	52.92	12.92	8/8	29	15.62	64.37
N=9	2+GAMM(22;1.48)	2	73.06	32.29	8/9	23	15.62	64.37
	U(10;15)	1	52.92	12.92	8/9	29	15.62	64.37
N=10	2+GAMM(22;1.48)	2	73.06	32.29	8/9	24	15.62	64.37
	U(10;15)	1	52.92	12.92	8/10	29	15.62	64.37

Table 17. Comparison of Optimum Deployment Plans: Northend Station

ANN Forecasts	Response Time Distribution	Opt. NOA	AVTIS (min)	AVRT (min)	TPR	AUR (%)	NSOS (min)	SOS (min)
N=8	2+GAMM(23.9;1.36)	2	94.34	49.26	6/6	20	34.47	55.68
	U(10;15)	2	56.61	11.83	6/6	12	34.47	55.68
N=9	2+GAMM(23.9;1.36)	2	85.65	49.26	6/6	18	27.78	53.63
	U(10;15)	2	58.04	11.63	7/7	14	34.47	55.36
N=10	2+GAMM(23.9;1.36)	2	78.39	44.40	7/7	19	26.84	53.63
	U(10;15)	2	58.90	11.63	8/8	17	34.74	54.95

Table 18. Comparison of Optimal Deployment Plans

ANN Forecasts	Response Time Distribution	Opt. NOA	AVTIS (min)	AVRT (min)	TPR	AUR (%)	NSOS (min)	SOS (min)
N=21	-0.001+ERLA(18.5;2)	5	92.21	41.75	20/20	26	11.84	52.49
	U(10;15)	3	65.49	12.77	19/19	29	28.41	57.34
N=23	-0.001+ERLA(18.5;2)	5	89.23	39.71	22/22	27	11.84	51.32
	U(10;15)	3	65.48	12.63	20/21	31	28.41	57.10
N=24	-0.001+ERLA(18.5;2)	5	89.73	40.7	23/23	29	19.56	53.45
	U(10;15)	3	64.29	12.57	21/22	32	28.41	55.53
N=25	-0.001+ERLA(18.5;2)	5	88.97	41.07	24/24	30	20.14	53.45
	U(10;15)	3	64.29	12.68	21/23	33	28.41	55.53
N=26	-0.001+ERLA(18.5;2)	5	87.43	40.68	25/25/	30	15.78	52.65
	U(10;15)	3	64.29	12.68	21/23	33	28.41	55.53

3.5.4. Comparison of Optimum Deployment Plans: Nkulumane Station

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. Summary statistics are presented in Table 19 for all cases (N=8, 9 and 10). For all the cases, the threshold of one (1) ambulance was achieved and any increase beyond this fleet size threshold would not positively influence changes in the performance measures. Ambulance utility levels (AUR) increased as the number of ambulances allocated decreased. Results indicate that reducing the response time between 10 and 15 minutes by adopting a U (10, 15) distribution resulted in decrease of the optimal number of ambulances required to achieve an optimal deployment plan, without having calls queuing for ambulance response services.

3.5.5. Optimal Fleet Sizes for ANN Forecasts and RTD ~ U (10; 15)

A summary of the optimum deployment plan when integrating ANN forecast and the proposed response time distribution: RTD ~ U (10; 15) is summarised in Table 20. Generally, the number of ambulances required are high in the Western suburbs as compared to the Eastern suburbs.

To determine the annual allocation of the ambulances across the stations and months, reference is made to Table 9 which represents the expected ANN daily calls forecasts per station across the months of the year. The integration of the expected daily forecasts from ANN and the optimisation for simulation modelling process resulted in an optimal deployment plan presented in Table 21.

Table 19. Comparison of Optimum Deployment Plans: Nkulumane Station

ANNForecasts	Response Time Distribution	Opt.NOA	AVTIS (min)	AVRT (min)	TPR	AUR (%)	NSOS (min)	SOS (min)
N=8	0.999+GAMM(21.8;1.62)	2	83.04	43.95	6/7	20	34.5	53.54
	U(10;15)	1	57.64	12.71	7/7	28	34.5	52.76
N=9	0.999+GAMM(21.8;1.62)	2	83.04	43.95	6/7	20	34.5	52.76
	U(10;15)	1	57.64	12.71	7/7	28	34.5	52.76
N=10	0.999+GAMM(21.8;1.62)	2	83.04	43.95	6/7	20	34.5	52.76
	U(10;15)	1	57.64	12.71	7/7	28	34.5	52.76

Table 20. Optimum Fleet Sizes on ANN Forecasts: RTD ~ U (10; 15)

Region	Station	Expected Daily Calls (N)							
		N=8	N=9	N=10	N=21	N=23	N=24	N=25	N=26
Eastern Suburbs	Famona	1	1	2	-	-	-	-	-
	Northend	2	2	2	-	-	-	-	-
Western Suburbs	Nketa	-	-	-	3	3	3	3	3
	Nkulumane	1	1	1	-	-	-	-	-

Table 21. Annual Optimal Fleet Size for ANN Forecasts: RTD ~ U (10; 15)

	Famona	Northend	Nketa	Nkulumane	Fleet Size
January	1	2	3	1	7
February	2	2	3	1	8
March	2	2	3	1	8
April	1	2	3	1	7
May	1	2	3	1	7
June	1	2	3	1	7
July	1	2	3	1	7
August	1	2	3	1	7
September	2	2	3	1	8
October	1	2	3	1	7
November	1	2	3	1	7
December	2	2	3	1	8

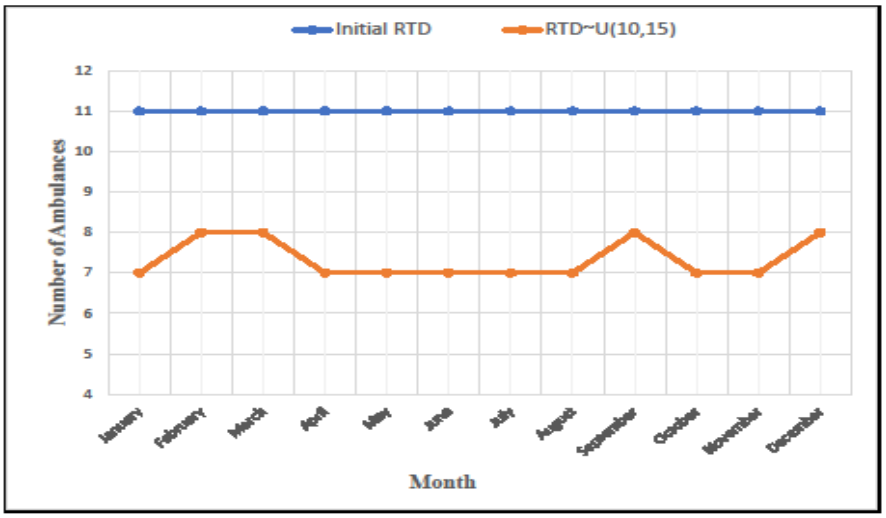

Figure 5. A Comparison of Optimum Deployment Plans

The deployment plan indicates that eight (8) ambulances are required in the months of February, March, September and December whilst seven (7) ambulances are required for the months of January, April, May, June, July, August, September, October and November. A comparison of the two ambulance deployment plans before and after adjusting for the response time distribution is shown in Figure 5.

3.5.6. Implications of Numerical Experiments

Standardising the response time between 10 to 15 minutes by adopting a uniform distribution (U (10, 15)) had a positive influence on the optimal deployment plan. It resulted in 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 decreases of average response time, average total duration of call in the system and reduced queuing time to zero. The ambulance utilisation levels and the throughput ratios remained relatively high. 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 and scarce resource in EMS. With reduced number of ambulances, brings about reduced requirement of human capital and reduced workloads as members would be able to be rotated more frequently as per international standards. One way is to focus on reducing the per-trip delays, chute delay time and the queuing time which is operationally possible. There is need therefore to manage closely activities within the call centre.

4. Conclusion

The paper developed a strategy of integrating forecasting, simulation and optimisation techniques for am- balance deployment in a heterogeneous region under multiple performance measures. 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. 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. The number of false alarm malicious (FAM) and false alarm good intent (FAGI) calls were prevalent in the eastern suburbs as compared to the western suburbs. 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, significantly high average queuing times and 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 sat- is faction with regards to service delivery. The general expectation is that no call should queue for service. Hence, there was 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 conducted by simultaneously minimising the average response time, average queuing time and maximizing throughput ratios. Increasing the number of ambulances influences the average response time below a certain threshold, beyond this threshold, the average response time stays at a certain level rather than decreasing gradually and no significant changes occur in other performance measures. Ambulance util- station inversely varied to increase in the fleet size. A total of eleven (11) ambulances are required to meet future demand. 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 10 to 15 minutes by adopting a uniform distribution given by U (10; 15) 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 (10; 15) 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. It is also imperative to simultaneously consider multiple performance indicators to complement the average response time. This goes a long way in balancing resource allocation and capacity utilisation to avoid 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 centre, training of the paramedics and provision of relevant modern equipment to the response teams. This also translates to reduced workloads on the response teams. 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. Training and provision of appropriate and modern equipment to the response teams will go a long way in reducing the pre-trip delay time, chute time and ultimately the response time. 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 in other operations research models.

Acknowledgements

The authors would like to thank the Bulawayo City Council and its personnel for their support in providing data and relevant information for the execution of this research.

References

[1] Sayed, M.J. (2012). Measuring quality in emergency medical services: a review of clinical performance indicators. Emergency Medicine International, 2012.

[2] Zaffar, M.A., Rajagopalan, H.K., Saydam, C., Mayorga, M., and Sharer, E. (2016). Coverage, survivability or response time: A comparative study of performance statistics in ambulance location models via simulation-optimisation. Operations Research for Health Care, 11, 1-12.

[3] Aartun, H.A., Andersson, E.S., Christiansen, H., Granberg, M., and Anderson, T. (2017). Strategic ambulance location for heterogeneous regions. European Journal of Operational Research, 260(1), 122-133.

[4] Zhen, L., Wang, K., Hongtao, H., and Daofang, C. (2014). A simulation optimization framework for ambulance deployment and relocation problems. Computers and Industrial Engineering, 72, 12-23.

[5] 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 and Operations Research, 78, 349-368.

[6] Kitapci, O., Ozekiciglu, H., Kaynar, O., and Tastan, S. (2014). The effect of economic policies applied in Turkey to the sale of automobiles: multiple regression and neural network analysis. Social and Behavioral Sciences, 148, 653-661.

[7] Rather, A.M., Agarwal, A., and Sastry, V.N. (2015). Recurrent neural network and a hybrid model for prediction of stock returns. Expert Systems with Applications, 42, 3234-3241.

[8] Hastie, T., Tibshirani, R., and Friedman, J. (2008). The Elements of statistical learning: Data mining, inference and prediction. (2nd ed.). New York: Springer Science + Business Media. (Chapter 11).

[9] Mitrea, C.A., Lee, C.K.M., and Wu, Z. (2009). A comparison between neural networks and traditional forecasting methods: A case study. International Journal of Engineering Business Management, 1(2), 19-24.

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

[11] Belanger, V., Ruiz, A., and Soriano, P. (2019). Recent optimization models and trends in location, relocation, and dispatching of emergency medical vehicles. European Journal of Operations Research, 272(1), 1-23.

[12] 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.

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

[14] Erkut, E., Ingolfsson, A., and Erdogan, G. (2008). Ambulance location for maximum survival. Naval Research Logistics (NRL), 55(1), 42-58.

[15] Handerson, S. and Mason, A. (2005). Ambulance planning: Simulation and data visualization in: Brandeau M.I., Saintfort F., Pierskalla w.p. (eds) Operations Research and Health Care: A Handbook of Methods and Applications; pages 77-102. Kluwer, Academic, Boston, 2004.

[16] Eldabi, T., and Young, T. (2007). Towards a framework for healthcare simulation. In 2007 Winter, Simulation Conference, pages 1454-1460. IEEE.

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

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

[19] 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.

[20] 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.