American Journal of Theoretical and Applied Statistics
2025; 14(1): 12-29
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20251401.12
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

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

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

[1]Department of Statistics and Mathematics, Faculty of Science and Engineering, Bindura University of Science Education, Bindura, Zimbabwe
[2]School of Mathematics, Statistics and Computer Science, University of Kwa Zulu-Natal, Pietermaritzburg, South Africa

Email address:

tichaonamapuwei@yahoo.com (Tichaona Wilbert Mapuwei), bodhlyerao@ukzn.ac.za (Oliver Bodhlyera),
mwambih@ukzn.ac.za (Henry Mwambi)
*Corresponding author

To cite this article:

Tichaona Wilbert Mapuwei, Oliver Bodhlyera, Henry Mwambi. (2025). Impact of Varying Response Time on Ambulance Deployment Plans in Heterogeneous Regions Using Multiple Performance Indicators. *American Journal of Theoretical and Applied Statistics, 14*(1), 12-29. https://doi.org/10.11648/j.ajtas.20251401.12

Received: 10 September 2024; **Accepted:** 13 December 2024; **Published:** 14 January 2025

Abstract: The paper conducts an assessment of the impact of varying response time distributions on ambulance deployment plans by integrating 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. Increasing the fleet size influences the average response time below a certain threshold value across all the heterogeneous regions. However, when fleet size is increased beyond this threshold value, no significant changes occur in the performance indicators. Fleet size varied inversely to ambulance utilisation levels. As fleet size is gradually increased, utilisation levels also gradually decreased. Due care must be taken to avoid under-utilisation of ambulances during deployment. Under utilisation culminates to human and material equipment idleness and yet the resources available are scarce and should be deployed where needed most. 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

1. Introduction

According to [1], the provision of best possible service to the public remains as a global challenge for emergency medical services (EMS). Reference [2] defines EMS as public safety systems established to coordinate the provision of pre-hospital care to patients under medical emergency conditions. An effective EMS that endeavours to minimise or eradicate loss of human lives is an essential component of any health care system [3]. Due to the complexity in EMS, it has created problems for decision makers at strategic, tactical and operational levels in trying to provide equitable, effective and efficient service to the public. The major strategic problem is the location of ambulance stations and ambulances

while tactical problems involve the sizing of the fleet of ambulances to the ambulance stations. Operational challenges evolve around issues on deciding which ambulances should be dispatched and the relocation of ambulances to ensure optimal preparedness in a specific region.

Due to the ever evolving complexities of EMS, it has led to the development of several models around the interacting decision making pillars of location, relocation and dispatch of ambulance resources. According to [4], these models have been broadly classified into single coverage models, multiple coverage models, probabilistic and stochastic models, stochastic and robust location-allocation models, fuzzy models, and human based models. In their analysis they highlighted that despite derived benefits from considering patient survivability in location models, response time thresholds and coverage are still dominant and imperative measures used in evaluating EMS performance. Emphasis has always been placed on the response time and how it influences the allocation of ambulances to fixed bases. However, there is always an interacting influence between the response time and service time on the overall performance of the EMS when rendering service to patients in need of emergency medical services which has to be considered simultaneously. Thus, components of EMS can be viewed as interconnected queuing systems that interact and influence one another. Reference [5] generalised an EMS system as a server-to-customer system where servers (ambulances) travel to render assistance to customers (patients) in need of emergency services. Hence, there is always need to balance the service desired to be offered to the population against the infrastructure, equipment, human resources and financial investment to attain this desired level of service in the near or distant future.

To improve service delivery, there is growing need to holistically consider the influence of response time simultaneously with service time in the optimal allocation of ambulances to base stations for future demand. This can be achieved by adopting simulation modeling techniques that allows for designing of models of a real system and conducting experiments with these models for purposes of understanding the behavior of the system and or evaluating various strategies for the operation of the system. Several authors such as [1] have acknowledged that the probability that an ambulance is available at a station depends on the inter-arrival of calls, number of ambulances allocated to the station and the ambulance service time. Reference [6] insinuated that, even though EMS are designed specifically for a local context, many share common design elements. However, they further highlighted that the use of several variables coupled with the random nature of demand renders deterministic methods of analysis unattractive. Simulation modelling techniques allows one to acquire information in the operation of a system without necessarily disturbing it. Operation policies can be developed to improve the system performance by observing different system scenarios through animation and visualisation.

When applying different analytical methods in operations research such as network analysis, linear programming, heuristics, game theory, queuing theory and simulation, each has its own merits and demerits. Reference [7] concluded that despite the hypercube model remaining as a powerful modelling approach, it requires several assumptions with regards to the way ambulances are dispatched and creating a huge threat in convincing decision makers to adopt the predictions due to these model complexities. This has been a common feature for most analytical methods in the discipline of operations research when applied to EMS systems. Reference [8] insinuated that the need to assess the impact of changes before actual implementation coupled with the pressure for better services and low availability of resources in EMS has created a huge opportunity in increasing simulation and modelling in health-care. Reference [9] indicated that comparing different scenarios allows us to identify in an objective manner, different changes that can lead to enhanced performance before committing resources which are usually scarce. Reference [5] developed probabilistic models that used the queuing theory and considered ambulances as servers in a queuing system and were considered to be sometimes unavailable. However, there has been a wide range of studies and research aimed at integrating different techniques [2, 6, 10, 11] in order to improve the robustness of the analysis and results of the models developed. According to [12], simulation plays an important role in many problems of our daily life. Simulation has been used over the years in various disciplines all over the world such as production and planning, port logistics, mining, software integration, construction, and energy among many others [13-22].

Static ambulance deployment model entails having a fixed number of ambulances to known fixed stations. Arrival rate is considered as the number of calls received at a station per unit time. This has often been referred to as the demand of a geographical area in literature. Response time is defined by [23] as the time taken to reach a patient after an emergency call is received. Alternatively, response time can be defined as the time that elapses from the moment a call is received by the EMS call center until an ambulance arrives at the scene of the patient. Response time in EMS is critical as it might determine the difference between life and death of a patient. Service time is defined as the duration an ambulance is occupied with a call. Ambulance resource utilisation level used as a performance measure is defined as the total workload time divided by the total operating time. The throughput ratio, represents the total number of emergency ambulance calls that are completely served divided by the total number of emergency calls that enters the emergency response system, 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.

Several researchers have made similar attempts in conducting numerical experiments to specific areas of EMS across the world in recent years. Reference [24] focused on the allocation of ambulance vehicles to a set of ambulance stations with known locations and alluded that the action

to reduce response time due to delays (pre-trip delay and queuing delay) 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 appropriate ambulance vehicle and 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 this is often attributed to system congestion. References [24] and [25] agree that this aspect is prevalent especially when limited EMS vehicles are used to serve extensive emergency calls across large geographical zones. It often affects service reliability, defined as the availability of EMS service when it is desired. The prevalence of an unreliable EMS, in real life would lead to individuals seeking other alternatives which might be more expensive or leave everything to fate in an emergency situation. Travel time, a critical contributor to the overall response time, is considered as the time that elapses from the moment the ambulance crew begin their journey until they reach the patient requiring emergency services.

The study conducted by [24] indicates that reducing the travel times usually requires adding ambulance stations or hospitals which is costly as the municipality is overwhelmingly under-resourced both in terms of equipment, human and financial resources. 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 come from. Here, the expectation is that reducing the response time plays a huge impact in improving service delivery, survival rates and patient satisfaction. Reference [26] alluded that the ability to provide timely response is hugely affected by fleet size and the locations of the ambulances. Reference [25] concurred that the average response time, which is immensely affected by the distribution of EMS vehicle stations and allocation of EMS vehicles to the incident areas, remains as a key measure to assess the efficiency and effectiveness of emergency responses.

Reference [24] argues that models that do not account for the uncertainty in the four components (call delay, chute, queuing and travel time) might overestimate or underestimate the number of ambulances required to provide a desired specific service level. The objective of optimisation for simulation is for searching the best set of inputs that optimises the performance of the system under predefined constraints. The methodology also considered the use of artificial neural networks, which are receiving a huge amount of interest in areas of forecasting because of their flexibility and remarkable features [27, 28]. Reference [29] defined an artificial neural network (ANN) as a nonlinear statistical model, which can be a classification or two-staged regression model usually represented as a network diagram and is good at modelling most complex functions where the relationships between variables is unknown. Alternatively, [30] defined an ANN as an information processing system developed for the generalisation of mathematical models of human neural biology. According to [31], 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. Among these areas of application include health, geophysics, geomechanics, stock markets, chemical engineering, electrical engineering, global logistics, construction engineering, financial business support, and insurance [31-42]. In the integrated approach adopted in this study, ANN were used to predict short-term annual demand for future planning using historical data. The methodology employed, removed several assumptions which are highly necessary in other operations research analytic models. The integrated research strategy developed can be used to manage other service systems with a set of service entities (such as taxis or service vehicles) that need to be located and allocated such that customer entities (such as passengers or customers) can be reached equitably, efficiently and effectively as proposed by [43]. This implies that the strategy can be replicated to other situations with minor adjustments.

2. Materials and Methods

2.1. Model Input Data

Historical public emergency ambulance demand data for BEMS from January 2010 to December 2018 was used for developing forecasting and simulation models. Numerical experiments were then conducted to assess the impact of varying response time distributions on the optimal ambulance deployments plans through the use of sensitivity analysis and optimisation techniques. A summary of the methodology is presented in Figure 1.

2.2. Artificial Neural Network Forecasting

A feed-forward neural network (FFNN) was trained using R-package by adopting the ¡§neuralnet¡¨ function, a network training function that updates weights and bias values during training. With the FFNN, information flows strictly from the input layer to the output layer without recurrent or backward connections as presented in Figure 2. Each layer has neurons and there is no connection between neurons that are in the same layer. The input vector is represented by Y_j denoted by $Y_j = \{y_1, y_2, y_3\}$; $W_{jk}(j = 1, 2, 3; k = 1, 2)$ 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. $\Theta_k(k = 1, 2)$ is the bias value of the hidden layer nodes and Θ is the bias value of the output layer.

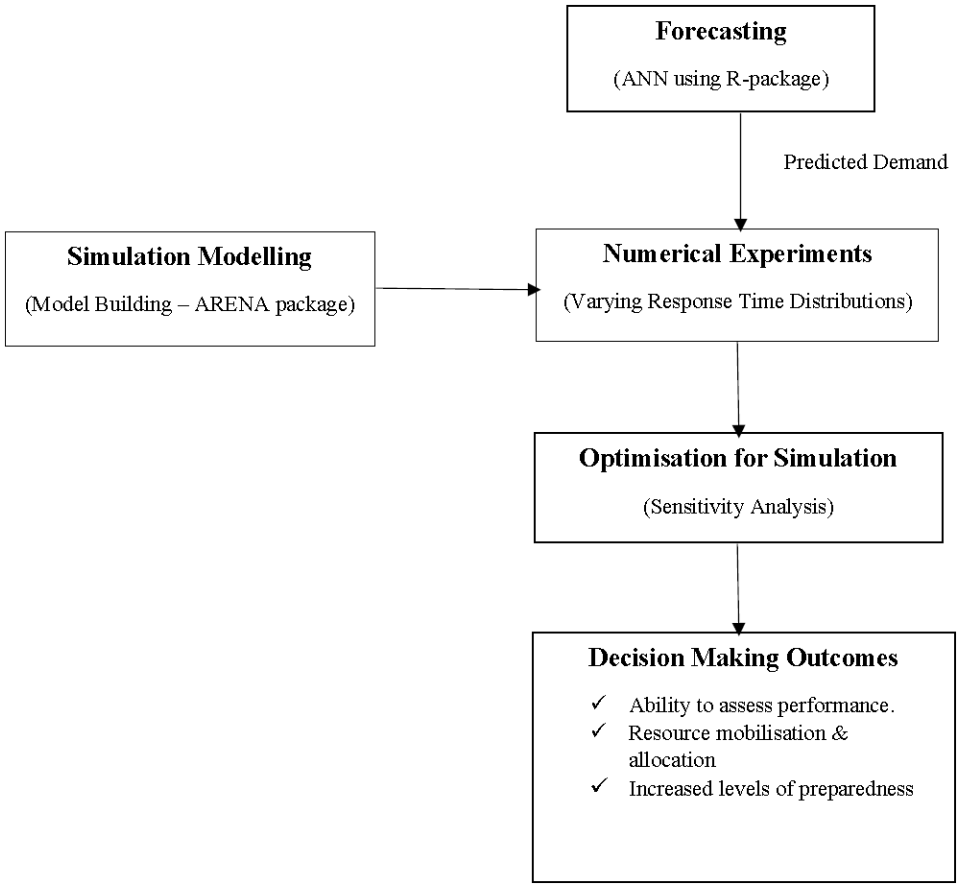

Figure 1. *An Integrated Simulation Modelling Approach.*

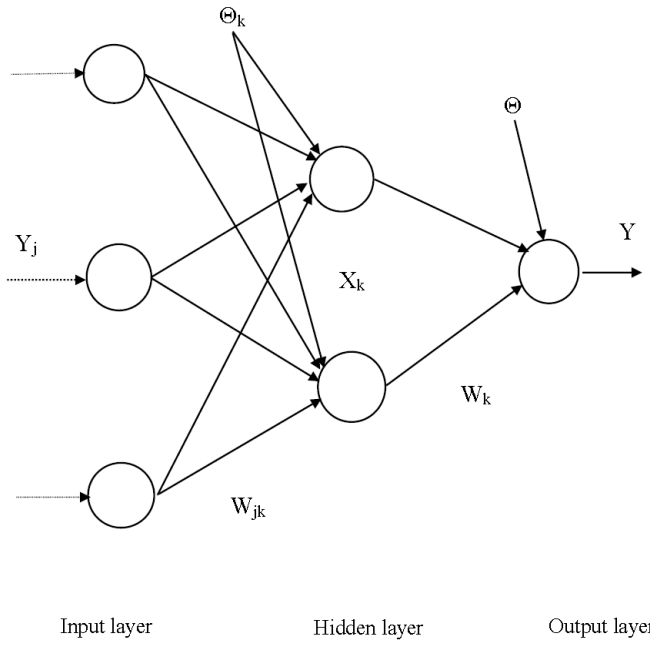

Figure 2. *Feed-forward neural network for time series forecasting.*

To facilitate the development and validation of the FFNN, the data splitting approach was adopted. Data from January 2010 to December 2017 was allocated for model building and 2018 was assigned for model cross-validation. Thus 96 observations were assigned for model building while 12 observations for model cross-validation. Data scaling using the minimum-maximum criteria to an interval (0,1) to prevent saturation in the hidden nodes and to increase the convergence rate during training of the neural network. Data was split into training and testing sets of seventy-two (72) and twenty-four (24) observations which translates to 75% and 25% respectively. The selection of the number of inputs in the model was based on trial and error as proposed by [6]. The general architecture of the FFNN can be generalised by equation 1.

$$I - (H_1, H_2, H_3, ..., H_n) - O \tag{1}$$

where I represents the number of input nodes, H_k number of neurons in hidden layer k, and O the number of neurons in the output layer. An example is an ANN with four (3) input nodes, one hidden layer with two (2) neurons and one (1) output neuron can be represented as 3-(2)-1 respectively. Supervised training with resilient backpropagation was implemented using 2017 demand calls as target values in the training algorithm. Training rate factors of 0.5 and 1.2 were adopted as the minimum and maximum values. Default values for momentum were set, with a threshold value set at 0.01 for 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 varied systematically to obtain the best and accurate model based on the mean absolute error (MAE) and residual mean square error (RMSE) as performance measures. According to [28], RMSE and MAE are both measures of accuracy and the degree of of spread of data points. The MAE is a measurement of how close forecasts are to the actual data points; the average of the absolute errors [35]. The MAE apportions equal weighting to differences from actual values whilst the RMSE gives a huge penalty in these differences and is more suitable in identifying outliers. The formulations of MSE, RMSE and MAE are given by:

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

$$RMSE = \sqrt{\frac{1}{N}\Sigma_{t=1}^{N}(y_t - \hat{y_t})^2} \tag{3}$$

$$MAE = \frac{1}{N}\Sigma_{t=1}^{N}|(y_t - \hat{y_t})| \tag{4}$$

where y_t is the actual observation for the period t, $\hat{y_t}$ the forecast for the same period, and N the length of the test set. The mean square error (MSE) and mean absolute error (MAE) were used as 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.

2.3. Simulation Modelling

Bulawayo Emergency Medical Services (BEMS) adopted the regionalised response strategy where EMS teams are assigned to serve a pre-specified geographical area or region. It is assumed that if the assigned EMS team(s) is busy, the closest team must perform the task. An important benefit derived from this strategy is to minimise travel times due to the reduced size of the geographic zone that the EMS teams need to travel between call locations. Bulawayo City, for purposes of emergency response is demarcated into two broad regions, the eastern and western regions. The eastern region covers the low density suburbs characterised by low population densities and the the western region covers the high density suburbs characterised by high population densities. Both the Eastern and Western regions are further split into two sub-regions to which an ambulance station is assigned. There are four sub-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.

2.3.1. Description of Components of BEMS Response System

BEMS is operating using the Anglo-American response strategy where the EMS is separated from the medical system as it offers only paramedic care. BEMS uses different kinds of vehicles but fitted with the same equipment features to respond to emergency calls. Ambulance dispatch is performed by a dispatcher upon receiving calls requiring EMS. BEMS is inclined to call-initiated dispatch decision making strategy where the dispatcher is required to select one of the idle ambulance vehicles to be dispatched after the arrival of an emergency call. BEMS employs the first in first out (FIFO) dispatch strategy with priority given to road traffic accidents in the case where waiting calls are in the response system. 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 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 fueling of vehicle. Emergencies are broadly categorised in three (3) categories with an assigned unique code for tracking, rescue team deployment and reporting purposes. These are summarised in Table 1. 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 1. *Categories and Codes of Emergency Response*

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 study assumed a static ambulance deployment model which endeavors to allocate a fixed number of ambulances to a set of known fixed sub-stations to ensure that the best medical outcomes for patients and workers are met. A logical presentation of the BEMS multi-location dispatch model is presented in Figure 3.

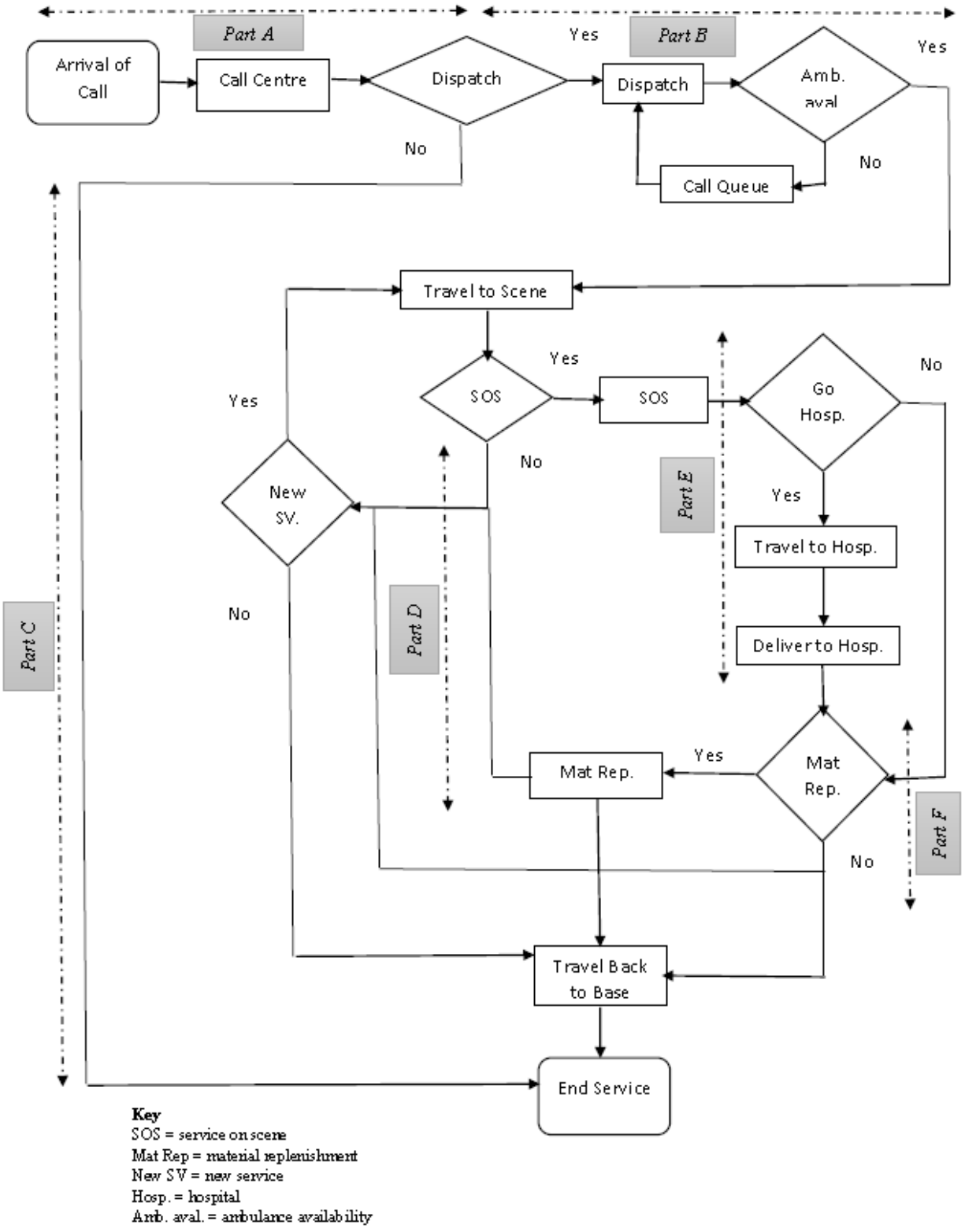

Figure 3. *Bulawayo Emergency Medical Service Multi-location Dispatch Model.*

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 were a decision is made whether to replenish the medical resources or not. It also includes aspects of vehicle service or refueling.

2.3.2. Assumptions of the Simulation Model

The simulation model will incorporate the randomness in call arrivals, travel times and service time. The model will assume the following:

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

2.3.3. Mathematical Representation of the Simulation Model

The optimisation problem according to [44] can be formulated as follows:

$$Min_{\theta \varepsilon \Theta} J(\theta) = E\left[C(\theta, w)\right] \tag{5}$$

subject to:

$$R(\theta, w) \leq RT \tag{6}$$

$$N_i(\theta, w) \leq NLS \sim \forall i \tag{7}$$

where:

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

As suggested by [6], to solve the problem, an optimiser and a simulator must work together as shown in Figure 4. The simulator evaluates the performance of each candidate solution, while the optimiser seeks to identify the candidate solutions.

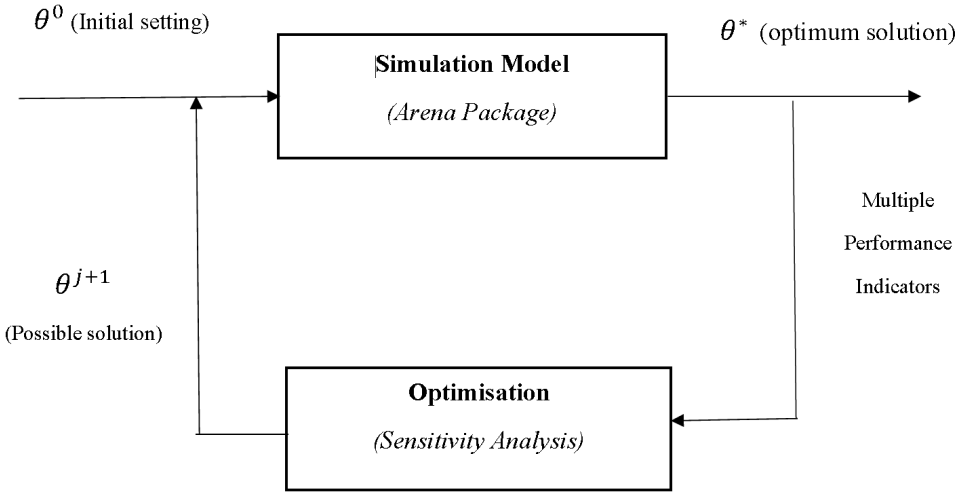

Figure 4. *Schematic Diagram for Optimisation for Simulation.*

2.3.4. Simulation Model Performance Indicators

The performance indicators considered for developing optimum simulation models and sensitivity analysis are: the average duration of a call in the system, average response time, average response queue time, average number of calls in response queue, throughput ratio 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: Optimum static ambulance deployment to predicted ANN demand, maintaining the RTD and Assessment of the influence of varying the response time to meet international standards by adopting uniform distributions given by $U(10, 15)$ and $U(5, 10)$.

2.3.5. Estimation 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) for each of the heterogeneous sub-station. The NSOS emanate from FAG and FAGI and usually results in less time required by the responding crew team since no pre-hospital care is not provided. However, these occur in different proportions in the heterogeneous geographical zones of service and were computed separately. The selection of the best distribution is based primarily on the square error (s.e) 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.

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) were 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 also based on the 2018 annual historical data.

3. Results and Discussion

Results of forecasting from ANN, simulation and optimisation techniques for ambulance deployment are presented and discussed in this section.

3.1. 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 are presented in Table 2.

Table 2. *Simulation Model Distributions of the 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 center. 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 3. 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 nature of the four 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 3. 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.

Table 3. *Summary of Expected Simulation Model Proportions by Sub-station.*

Description	Proportion	Famona	Northend	Nketa	Nkulumane
Call	EC	0.92	0.92	0.92	0.92
Filter	NEC	0.08	0.08	0.08	0.08
	Total	1	1	1	1
Call	CAT A	0.69	0.58	0.56	0.62
Category	CAT B	0.26	0.34	0.37	0.36
Classification	CAT C	0.05	0.08	0.07	0.02
	Total	1	1	1	1
Nature	SOS	0.84	0.84	0.93	0.94
of Service	NSOS	0.16	0.16	0.07	0.06
	Total	1	1	1	1
Expected Monthly Call Proportion	EMCP (%)	18.1%	17.6%	47.1%	17.2%

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.

3.2. Artificial Neural Network Public Ambulance Demand Forecasts

An architecture of the FFNN given by $(7-(4)-1)$, with seven input nodes, one hidden layer (4 neurons) and one output neuron was identified as the best appropriate model having the lowest MAE of 94.0 and RMSE of 137.19 as summarised in Table 4. The $*$ in the table represents the minimum value of the performance measure across all models.

Table 4. *Feed forward neural network model selection.*

Model	Structure	Testing set (MSE)	Testing set (MAE)	Validation (RMSE)	Validation (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*

Paired sample t-test at 5% level of significance was applied to validate any significant differences between actual values and predicted values of the selected FFNN model using Minitab statistical package. The calculated p-value for FFNN was $0.493(> 0.05)$ (Table 5) and we conclude that there is no significant difference between the actual values and predicted public emergency ambulance demand. The selected neural network model (7-(4)-1) was used to forecast demand for 2019 and results are presented in Table 6. Demand is expected to be high in January, March, September and December whilst lower demand is projected for April, June and July 2019.

Table 5. *Paired sample t-test for 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 6. *Expected Daily Ambulance Demand Per Station from ANN Forecasts.*

	Expected Values	Jan.	Feb.	March	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Y2019	ANN Forecasts	1622	1494	1713	1368	1482	1318	1391	1526	1572	1541	1532	1638
	Days (2019)	31	28	31	30	31	30	31	31	30	31	30	31
Famona	Monthly Calls	294	270	310	248	268	239	252	276	285	279	277	296
	Daily Calls (N)	9	10	10	8	9	8	8	9	10	9	9	10
Northend	Monthly Calls	285	263	301	241	261	232	245	269	277	271	269	289
	Daily Calls (N)	9	9	10	8	8	8	8	9	9	9	9	9
Nketa	Monthly Calls	764	704	807	644	698	620	655	719	740	726	722	771
	Daily Calls (N)	25	25	26	21	23	21	21	23	25	23	24	25
Nkulumane	Monthly Calls	279	257	295	235	255	227	239	262	270	265	264	282
	Daily Calls (N)	9	9	10	8	8	8	8	8	9	9	9	9

3.3. Simulation Model Building for the Heterogeneous Sub-stations

In developing the simulation model, number of ambulances were gradually increased from one (1) to the allocated fleet size at each of the heterogeneous sub-stations, while changes in performance indicators were simultaneously being observed and recorded. Current unoptimised ambulance deployment plan has Famona one (1), Northend one (1), Nketa three (3) and Nkulumane one (1) allocated ambulance(s) respectively. A summary of results for the simulation models is presented 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
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

Models developed are a resemblance of the current prevailing EMS process for Bulawayo city. Average response times are relatively high when compared to internationally recommended standards of 5 to 10 minutes. Average queuing times and number of ambulances queuing remain undesirably high posing a threat to human based outcomes of safety and satisfaction. An introspect into Nketa Station simulation model reveals important insights into the influence of varying fleet size on key performance indicators such as average time of a call in system, average response time, average number of calls in response queue and the corresponding average time in queue, throughput ratio (%) and average ambulance utility (%) as presented in Figure 5.

It was observed that as the number of ambulances increases, there is corresponding positive changes in performance indicators. The average time of a call in system, average response time, average number of calls in response queue, and average time of a call in queue decrease as the number of ambulance fleet size increases. Throughput ratios, increases with increase in allocated number of ambulances as expected. However, the ambulance utilisation levels decreases with increase of ambulance fleet size. For an efficient and effective EMS, it is a general expectation that no call should queue for service with response time reduced to internationally recognised levels of 5 to 10 minutes in urban areas. Hence, there is need to determine the optimum ambulance deployment models that minimise the number of ambulances needed to provide a specific and desired service level. The next section seeks to improve the EMS model performances by adopting optimisation for simulation through the use of sensitivity analysis whilst integrating future demand forecasts.

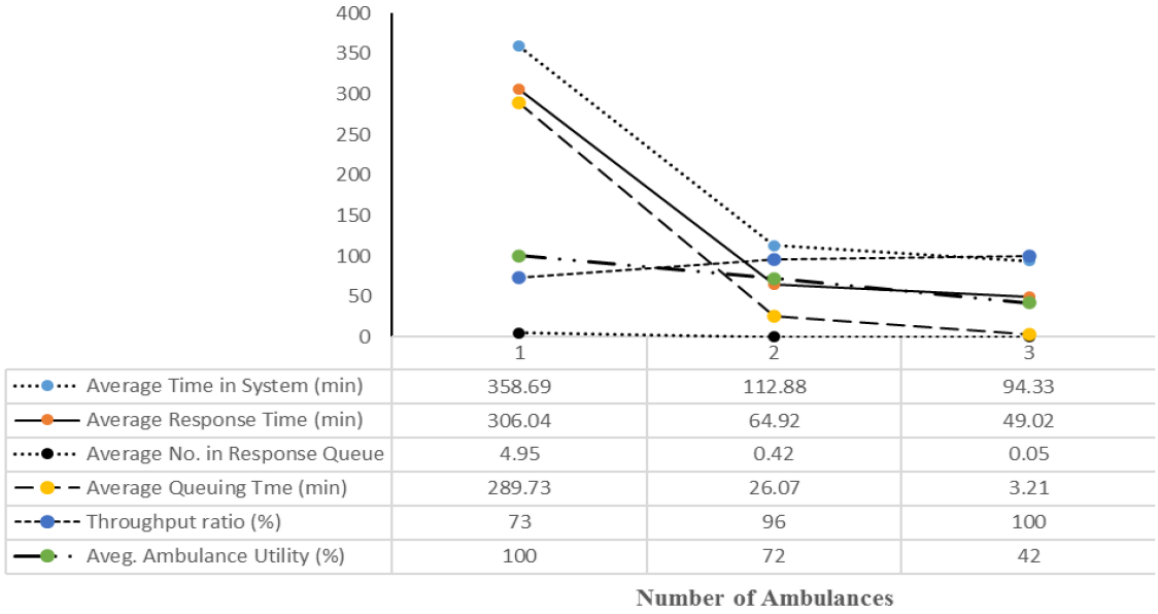

Figure 5. *Influence of varying fleet size on performance indicators.*

3.4. Sensitivity Analysis

Sensitivity analysis was conducted to determine optimal static ambulance deployment plans by varying the response time against the predicted ANN values. Firstly, optimal deployment plans using the initial response time distributions were developed. Later, uniform distributions $U(10;15)$ and $U(5;10)$ were adopted to represent response times that will vary between 10 and 15 minutes and 5 to 10 minutes respectively. 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 not changed.

3.4.1. Optimum Deployment Plan for BEMS Using Initial Response Distribution Functions

The ANN forecasts in Table 6 indicate that for 2019, across the different 12 months would assume values of 8, 9 and 10 as expected daily calls for Famona, Northend and Nkulumane sub-stations. Nketa sub-station assumed values of 21, 23, 24, 25 and 26 respectively. Summaries of the processes in determining the optimum development plans by integrating forecasts, simulation and optimisation techniques are presented in Table 8 and Table 9. Ambulance fleet sizes were incremented from one (1) whilst monitoring performance indicators.

Table 8. *Optimum Deployment Plans for Famona, Northend and Nkulumane Stations: Initial RTD.*

Station/RTD	Calls (N)	NOA	AVTIS (min.)	AVRT (min.)	AVNRQ	AVQT (min.)	TPR	AUR	NSOS (min.)	SOS (min.)
		1	86.15	38.80	0.03	5.32	8/8	0.45	16.44	65.9
	N=8	*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
		1	86.15	40.51	0.04	6.58	8/9	0.48	16.44	65.9
Famona	N=9	*2*	*73.06*	*32.29*	*0.0*	*0.0*	*8/9*	*0.23*	*15.62*	*64.37*
2+GAMM(22;1.48)		3	73.06	32.29	0.0	0.0	8/9	0.15	15.62	64.37
		1	86.15	40.51	0.04	6.58	8/9	0.48	16.44	65.9
	N=10	*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
		1	93.87	47.63	0.09	17.91	7/7	0.37	31.82	52.0
	N=8	*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
		1	93.79	46.67	0.09	15.67	8/8	0.43	31.82	52.22
Northend	N=9	*2*	*85.65*	*49.26*	*0.0*	*0.0*	*6/6*	*0.18*	*27.78*	*53.63*
2+GAMM(23.9;1.36)		3	85.65	49.26	0.0	0.0	6/6	0.13	27.78	53.63

Station/RTD	Calls (N)	NOA	AVTIS (min.)	AVRT (min.)	AVNRQ	AVQT (min.)	TPR	AUR	NSOS (min.)	SOS (min.)
		1	103.01	54.85	0.14	22.61	9/9	0.50	31.82	52.84
	N=10	*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
		1	97.81	43.21	0.02	3.67	7/7	0.46	50.86	56.06
	N=8	*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
		1	97.81	43.21	0.02	3.67	7/7	0.46	50.86	56.09
Nkulumane	N=9	*2*	*83.04*	*43.95*	*0.0*	*0.0*	*6/7*	*0.20*	*34.5*	*53.54*
0.999+GAMM(21.8;1.62)		3	83.04	43.95	0.0	0.0	6/7	0.13	34.5	53.54
		1	97.81	43.21	0.02	3.67	7/7	0.46	50.86	56.09
	N=10	*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

Table 9. *Optimum Deployment Plan for Nketa Station: RTD ∼ -0.001+ERLA(18.5;2).*

Calls (N)	NOA	AVTIS (min.)	AVRT (min.)	AVNRQ	AVQT (min.)	TPR	AUR	NSOS (min.)	SOS (min.)
	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
N=21	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
	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
N=23	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
	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
N=24	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
	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
N=25	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
	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
N=26	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

Key insights are derived from Nketa Station with a wide variation in expected number of calls and ambulance fleet size required for deployment in achieving optimal solution. A summary of key indicators were tracked as fleet sizes were incremented to optimal levels as presented in Figure 6. As the fleet size is increased there is 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, 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 increase in fleet size. The optimum deployment plans still have high average response time which are far above the recommended international standards, hence the need to adjust accordingly through the numerical experiments.

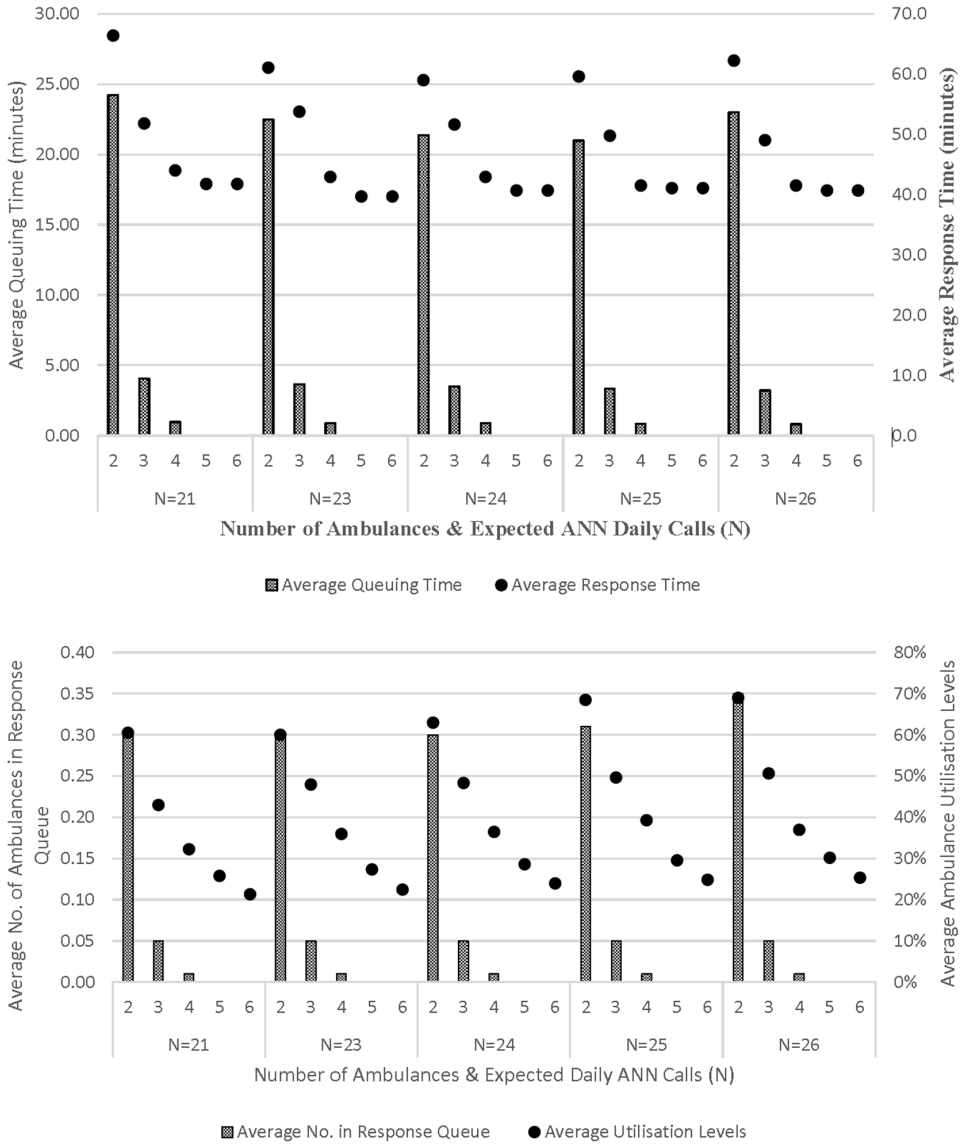

Figure 6. *Implications of varying fleet sizes on performance indicators.*

3.4.2. Comparison of Optimum Deployment Plans by Varying Response Time Distributions

A comparison of the different performance changes due to the influence of the changes in response time distributions on the optimal deployment plan for all the stations was conducted. The same optimisation process of gradually increasing the fleet sizes while observing the multiple performance measures was performed. Summaries are presented in Table 10, Table 11, Table 12 and Table 13.

Table 10. *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.)
	2+GAMM(22;1.48)	2	76.06	33.07	0	0	0	8/8	0.21	15.62	64.37
N=8	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*
	2+GAMM(22;1.48)	2	73.06	32.29	0	0	0	8/9	0.23	15.62	64.37
N=9	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*
	2+GAMM(22;1.48)	2	73.06	32.29	0	0	0	8/10	0.24	15.62	64.37
N=10	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 11. *Comparison of Optimum Deployment Plans For Northend Station.*

ANN Forecasts	Response Time Distribution	Opt. NOA	AVTIS (min.)	AVRT (min.)	AVNRQ	AVQT (min.)	NEC	TPR	AUR (min.)	NSOS (min.)	SOS
	2+GAMM(23.9;1.36)	2	94.34	49.26	0	0	2	6/6	0.20	34.47	55.68
N=8	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*
	2+GAMM(23.9;1.36)	2	85.65	49.26	0	0	3	6/6	0.18	27.78	53.63
N=9	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*
	2+GAMM(23.9;1.36)	2	78.39	44.40	0	0	3	7/7	0.19	26.84	53.63
N=10	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*

Table 12. *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.)
	-0.001+ERLA(18.5;2)	5	92.21	41.75	0	0	1	20/20	0.26	11.84	52.49
N=21	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*
	-0.001+ERLA(18.5;2)	5	89.23	39.71	0	0	1	22/22	0.27	11.84	51.32
N=23	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*
	-0.001+ERLA(18.5;2)	5	89.73	40.7	0	0	1	23/23	0.29	19.56	53.45
N=24	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*
	-0.001+ERLA(18.5;2)	5	88.97	41.07	0	0	1	24/24	0.30	20.14	53.45
N=25	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*
	-0.001+ERLA(18.5;2)	5	87.43	40.68	0	0	1	25/25	0.30	15.78	52.65
N=26	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*

Table 13. *Comparison of Optimum Deployment Plans For Nkulumane Station.*

ANN (N)	Response Time Distribution	Opt. NOA	AVTIS (min.)	AVRT (min.)	AVNRQ	AVQT (min.)	NEC	TPR	AUR	NSOS (min.)	SOS (min.)
	0.999+GAMM(21.8;1.62)	2	83.04	43.95	0	0	0	6/7	0.20	34.5	53.54
N=8	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*
	0.999+GAMM(21.8;1.62))	2	83.04	43.95	0	0	0	6/7	0.20	34.50	53.54
N=9	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*
	0.999+GAMM(21.8;1.62)	2	83.04	43.95	0	0	0	6/7	0.20	34.50	53.54
N=10	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*

3.4.3. Optimal Fleet Sizes for ANN Forecasts and Response Time Distributions

A summary of the optimum deployment plans when integrating ANN forecast and the proposed response time distributions are summarised in Table 14.

Table 14. *Optimal Fleet Sizes for ANN Forecasts and The Response Time Distributions.*

Station	RTD	Jan.	Feb.	March	April	May	June	July	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		1	2	2	1	1	1	1	1	2	1	1	2
Northend	U(10;15)	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		1	2	2	1	1	1	1	1	2	1	1	2
Northend	U(5;10)	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

A summary for all the four deployment plans is presented in Figure 7. The prevailing deployment plan has six (6) ambulances available across the twelve (12) months. If optimisation is applied by integrating the ANN demand forecasts without varying the response time distributions, eleven (11) ambulances are required. However, when $U(10, 15)$ and $U(5, 10)$ response time distributions are applied, ambulance deployment plans vary between seven (7) and eight (8) ambulances as aligned to demand monthly patterns. Implications are that reducing the response time distribution from a range of 10 to 15 minutes to a range of 5 to 10 minutes did not change the deployment plans across the whole year.

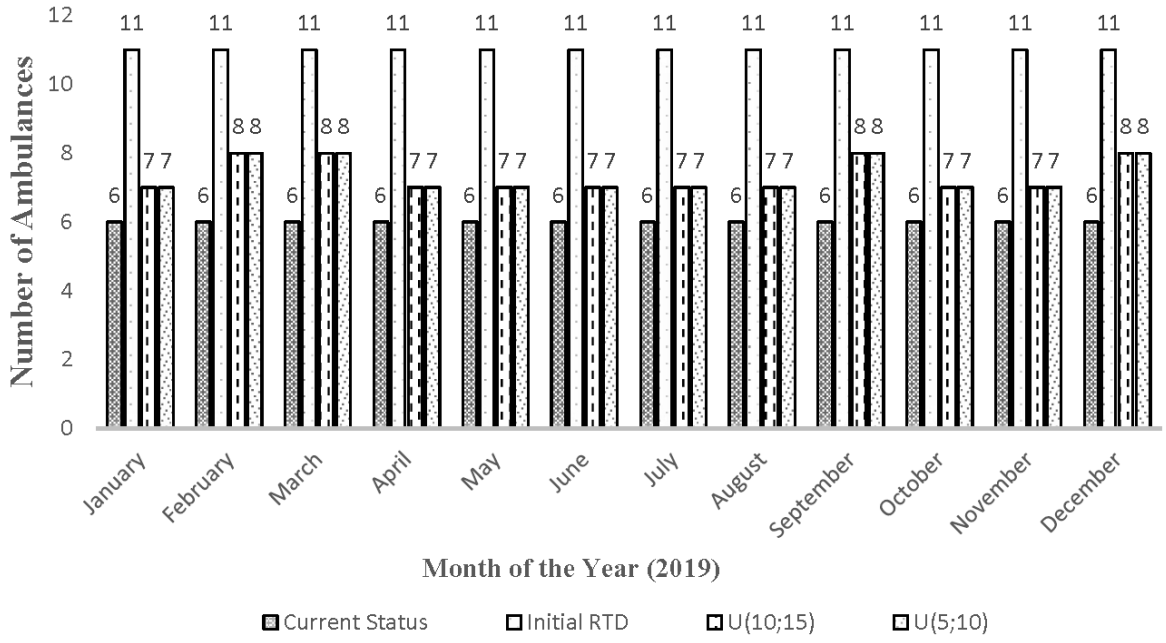

Figure 7. *A Comparative of Deployment Plans By Varying Response Time Distributions.*

4. Conclusion

The paper integrated forecasting, simulation and optimisation techniques to assess the impact of varying response time distributions on ambulance deployment in heterogeneous regions using multiple performance indicators. Future ambulance demand were predicted using an ANN model with a 7-(4)-1 architecture. Simulation model input parameters were developed to capture the stochastic nature of demand (inter-arrival rates of calls), response time, service time, occurrences of emergency calls and their levels of severity for the four heterogeneous demand zones in Bulawayo Metropolitan City.

Simulation models developed resembled the current EMS being provisioned by BEMS with six (6) operational ambulances. The general simulation models developed indicated that average response times are well above 15 minutes, and characterised by 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. Hence, there was need to determine the optimum ambulance deployment plans that minimised the average response time, number of calls in queue and queuing time, and adjusting for fleet size to provide a specific service level using optimisation for simulation whilst monitoring other performance indicators such as utilisation levels and throughput ratios.

It was observed that increasing the fleet size influences the average response time below a certain threshold value across all the heterogeneous regions. When fleet size is increased beyond this threshold value, no significant changes would occur in the performance indicators. Fleet size varied inversely to ambulance utilisation levels. As fleet size is gradually increased, utilisation levels also gradually decreased. Due care must be taken to avoid under-utilisation of ambulances during deployment. Under utilisation culminates to human and material equipment idleness and yet the resources available are scarce and should be deployed where needed most. For all the other performance indicators, increase in fleet size resulted in the decrease of the values as expected.

Standardising the response time between 10 to 15 minutes by adopting a uniform distribution ($U(10, 15)$) resulted in significant decrease in the number of ambulances required for deployment from eleven (11) down to a maximum of eight (8) ambulances. The decrease in ambulances deployed together with the response time distributions improved 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 number of calls in queue and queuing time to zero. The ambulance utilisation levels and the throughput ratios remained relatively high. Thus, reducing the response time culminated in the reduction of the number of ambulances required to achieve optimum deployments. However, further reduction of response time to between 5 and 10 minutes by adopting a uniform distribution ($U(5, 10)$) did not affect the deployment plans as compared to a scenario when a $U(10, 15)$ distribution was adopted. Despite a decrease in average response time across all the heterogeneous regions, the total number of ambulances required to obtain optimum deployment plans did not change.

It can be concluded that increasing the fleet size helps to improve service delivery to a certain level. For ambulances as a medical resource, the more resources deployed does not always translate to better performance. To complement this, efforts must be directed in reducing the response time, which only does not improve service delivery, but help reduce the number of ambulances to attain the desired level of service. Reducing the response time is easier, cheaper and feasible for management as this is directly linked to key components and contributors of response time such as pre-trip delays, chute time and queuing time which are more concentrated in the call centre. It is also equally imperative to simultaneously consider multiple performance indicators to complement the average response time. This goes a long way in balancing resource mobilisation, allocation and capacity utilisation. Digitisation of switch boards in the call center, continuous staff training coupled with provision of standard modern equipment to response teams will go a long way in reducing the response time. An important contribution of this paper is that the simplified methodology strategy can further be adopted to similar cases that involve a server-to-customer operation environment with relative ease with few adjustments.

Abbreviations

EMS	Emergency Medical Services
ANN	Artificial Neural Network
FFNN	Feed-forward Neural Network
MAE	Mean Absolute Error
RMSE	Residual Mean Square Error
MSE	Mean Square Error
BEMS	Bulawayo Emergency Medical Services
FIFO	First In First Out
FAM	False and Malicious
FAGI	False Alarm with Good Intent
SOS	Service on Scene
NSOS	No Service on Scene

Acknowledgments

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.

Conflicts of Interest

The authors declare that there is no conflict of interest.

References

[1] 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.

[2] 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.

[3] McCormack, R., and Coates, G. (2015). A simulation model to enable the optimisation of ambulance fleet allocation and base station location for increased patient survival. *European Journal of Operations Research*, 247(2015), 294-309. http://dx.doi.org/10.1016/j.ejor.2015.05.040

[4] 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.

[5] Galvao, R. D, and Morabito, R. (2008). Emergency service systems: the use of the hypercube queueing model in the soulution of probabilistic location problems. *International Transactions of Operations Research*, 15(2008), 525-549.

[6] 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.

[7] Handerson, S. and Mason, A. (2005). Ambulance planning: Simulation and data visuallisation in: Brandeau m. I., saintfort f., pierskalla w. p. (eds) operations research and

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

[9] 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.

[10] 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.

[11] 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.

[12] Reshid, T. M. (2020). Sampling distribution and simulation in R. *International Journal of Statistics and Mathematics*, 7(2), 154-163.

[13] Wang, T., Guinet, A., Belaidi, A., and Besombes, B. (2009). Modelling and simulation of emergency services with ARIS and Arena. Case study: the emergency department of Saint Joseph and Saint Luc Hospital. *Production Planning and Control*, 20(6), 484-495. https://doi.org/10.1080/09537280902938605

[14] Jeon, S. M., and Kim, G. (2016). A survey of simulation modeling techniques in production planning and control (PPC). *Production Planning and Control*, 27(5), 360-377.

[15] Dermici, E. (2003). Simulation modelling and analysis of a port investment. *Simulation*, 79(2), 94-105. https://doi.org/10.1177/0037549703254523

[16] Esmer, S., Yildiz, G., and Tuna, O. (2013). A new simulation modelling approach to continuous berth allocation. *International Journal of Logistics Research and Applications*, 16(5), 398-409. https://doi.org/10.1080/13675567.2013.813920

[17] Cordeau, J. F., Legato, P., Mazza, R. M., and Trunfio, F. (2015). Simulation-based optimization for housekeeping in a container transshipment terminal. *Computers and Operations Research*, 53(2015), 81-95. http://dx.doi.org/10.1016/j.cor.2014.08.001

[18] Cimpeanu, R., Devine, M. T., and O'Brien, C. (2017). A simulation model for the management and expansion of extended port terminal operations. *Transportation Research Part E*, 98(2017), 105-131.

[19] Ataeepou, N., and Baafi, E. Y. (1999). ARENA simulation model for truck-shovel operation in despatching and non-despatching modes. *International Journal of Surface Mining, Reclamation and Environment*, 13(1999), 125-129.

[20] Borodin, V., Bourtembourg, J., Hnaien, F., and Labadie, N. (2019). COTS software integration for simulation optimization coupling: case of ARENA and CPLEX products. *International Journal of Modelling and Simulation*, 39(3), 178-189. https://doi.org/10.1080/02286203.2018.1547814

[21] Abdelmegid, J., Legato, P., Hopfe, C., Rezgui, Y., and Sweet, T. (2015). A conceptual framework to support solar PV simulation using an open-BIM data exchange standard. *Automation in Construction*, 53(2015), 81-95. http://dx.doi.org/10.1016/j.cor.2014.08.001

[22] Gupta, A., Cemesova, A., Mazza, R. M., and Trunfio, R. (2014). Simulation-based optimization for housekeeping in a container transshipment terminal. *Computers and Operations Research*, 37(2014), 166-181. http://dx.doi.org/10.1016/j.autcon.2013.10.005

[23] Lee, S. (2012). The role of centrality in ambulance dispatching. *Decision Support Systems*, 54, 282-291.

[24] 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.

[25] Liu, Y., Li, Z., Liu, J., and Patel, H. (2016). A double standard model for allocating limited emergency medical service vehicle resources ensuring service reliability. *Transportation Research Part C*, 69(2016), 120-133. http://dx.doi.org/10.1016/j.trc.2016.05.023

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

[27] 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.

[28] Aras, S., and Kocakoc, I. D. (2016). A new model selection strategy in time series forecasting with artificial neural networks: IHTS. *Neurocomputing*, 174, 974-987.

[29] 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)

[30] 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.

[31] 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.

[32] Sut, A., and Senocak, 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.

[33] 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.

[34] Sadiq, T., Gharbi, R. B., and Juvkam-Wold, H. C. (2003). Use of neural networks for the prediction of frictional drag and transmission of axial load in horizontal wellbores. *International Journal for Numerical Methods in Geomechanics*, 27, 111-131.

[35] Herliansyah, R., and Jamilatuzzahro, J. (2017). Feed forward neural networks for forecasting Indonesia exchange composite index. *GSTF Journal of Mathematics, Statistics and Operations Research*, 4(1), 8-15.

[36] Mat Noor, R. A., Ahmed, Z., and Mat Don, M. (2010). Modelling and control of different types of polymerization process using neural networks. *The Canadian Journal of Chemical Engineering*, 88, 1065-1084.

[37] Prasad, J. Y., and Bhagwat, S. S. (2002). Simple neural network models for prediction of physical properties of organic compounds. *Chemical Engineering Technology*, 25(11), 1041-1046.

[38] Karami, A. (2010). Estimation of the critical clearing time using MLP and RBF neural networks. *European Transactions on Electrical Power*, 20, 206-217.

[39] Bowersox, D. J., and Calantone, R. J. (2003). Estimation of global logistics expenditures using neural networks. *Journal of Business Logistics*, 24(2), 21-3.

[40] Zichun, Y. C. (2012). The BP artificial neural network model on expressway construction phase risk. *Systems Engineering Procedia*, 4, 409-415.

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

[42] Shah, 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.

[43] Schmid, V., and Doerner K. F. (2010). Ambulance location and relocation with time-dependent travel times. *European Journal of Operations Research*. 207(2010), 1293-1303. http://dx.doi.org/10.1016/j.ejor.2010.06.033

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