



EVA: Emergency Vehicle Allocation

DANIEL CACHEIRA DIAS

Outubro de 2023



EVA: Emergency Vehicle Allocation

Demand Forecasting and Spatial Allocation models for emergency medicine

Daniel Cacheira Dias

Student nº: 1170790

**Dissertation to obtain the master's degree in Artificial Intelligence
Engineering**

**Orientador: Doutor Carlos Fernando da Silva Ramos, Professor Coordenador Principal
do Instituto Superior de Engenharia do Instituto Politécnico do Porto**

Júri:

Presidente:

Doutor Luiz Felipe Rocha de Faria, Professor Coordenador do Instituto Superior de Engenharia
do Instituto Politécnico do Porto

Vogais:

Doutor Carlos Fernando da Silva Ramos, Professor Coordenador Principal do Instituto Superior
de Engenharia do Instituto Politécnico do Porto

Doutora Maria Goreti Carvalho Marreiros, Professora Coordenadora com Agregação do
Instituto Superior de Engenharia do Instituto Politécnico do Porto

Porto, September 2023

Resumo

A medicina de emergência desempenha um papel fundamental no desenvolvimento da Sociedade, onde o objetivo é prestar assistência médica no menor tempo possível. Consequentemente, os sistemas que apoiam as operações de emergência precisam de ser robustos, eficientes e eficazes na gestão dos recursos limitados. Para isso, são analisados dados históricos no intuito de encontrar padrões em ocorrências passadas que possam ajudar a prever o volume futuro de chamadas. Esta é uma tarefa demorada e muito complexa que poderia ser resolvida com o uso de soluções de *Machine Learning*, que têm funcionado adequadamente no contexto da previsão de séries temporais. Só depois de conhecida a demanda futura poderá ser feita a otimização da distribuição dos recursos disponíveis, com o objetivo de suportar zonas de elevada densidade populacional. O presente trabalho tem como objetivo propor um sistema integrado capaz de apoiar a tomada de decisão em operações de emergência num ambiente de tempo real, atribuindo um conjunto de unidades disponíveis dentro de uma área de serviço com base em previsões volume de chamadas a cada hora. A arquitetura de sistema sugerida emprega uma abordagem de micros serviços juntamente com comunicações baseadas em eventos para permitir interações em tempo real entre os componentes. Esta dissertação centra-se nos componentes de previsão do volume de chamadas e otimização da atribuição. Foram usados modelos de séries temporais tradicionais e *Deep Learning* para modelar dados históricos de chamadas de emergência de *Virginal Beach* entre os anos de 2010 e 2018, combinadas com informações relacionadas ao clima. As soluções de *Deep Learning* ofereceram melhores métricas de erro, com *WaveNet* a ter um valor MAE de 0,04. No que diz respeito à otimização da localização dos veículos de emergência, a solução proposta baseia-se num problema de Programação Linear para minimizar o número de veículos em cada estação, com um mecanismo de vizinho, denominado EVALP-NM, para adicionar unidades adicionais às estações próximas de uma zona de alta densidade de chamadas. Esta solução foi comparada com um algoritmo genético que teve um desempenho significativamente pior em termos de tempo de execução e resultados. O desempenho do EVALP-NM foi testado em simulações com configurações diferentes, como número de zonas, estações e ambulâncias.

Palavras-chave: Medicina de Emergência, Previsão de série temporal, Atribuição espacial

Abstract

Emergency medicine plays a critical role in the development of a community, where the goal is to provide medical assistance in the shortest possible time. Consequently, the systems that support emergency operations need to be robust, efficient, and effective when managing the limited resources at their disposal. To achieve this, operators analyse historical data in search of patterns present in past occurrences that could help predict future call volume. This is a time-consuming and very complex task that could be solved by the usage of machine learning solutions, which have been performed appropriately in the context of time series forecasting. Only after the future demands are known, the optimization of the distribution of available assets can be done, for the purpose of supporting high-density zones. The current works aim to propose an integrated system capable of supporting decision-making emergency operations in a real-time environment by allocating a set of available units within a service area based on hourly call volume predictions. The suggested system architecture employs a microservices approach along with event-based communications to enable real-time interactions between every component. This dissertation focuses on call volume forecasting and optimizing allocation components. A combination of traditional time series and deep learning models was used to model historical data from Virginal Beach emergency calls between the years 2010 and 2018, combined with several other features such as weather-related information. Deep learning solutions offered better error metrics, with WaveNet having an MAE value of 0.04. Regarding optimizing emergency vehicle location, the proposed solution is based on a Linear Programming problem to minimize the number of vehicles in each station, with a neighbour mechanism, entitled EVALP-NM, to add a buffer to stations near a high-density zone. This solution was also compared against a Genetic Algorithm that performed significantly worse in terms of execution time and outcomes. The performance of EVALP-NM was tested against simulations with different settings like the number of zones, stations, and ambulances.

Keywords: Emergency Medicine, Time Series Forecasting, Spatial Allocation

Acknowledgements

I would like to show my appreciation towards Polytechnic of Porto - School of Engineering (ISEP) for the continuous innovation of the curriculum, in order to make it possible to learn high demand topics, such as Artificial Intelligence, with the “hands-on” mindset I have become familiarized since my bachelor’s degree.

I would like to thank my family for their support during this journey, for cheering my achievements and smoothing my setbacks, and the constant encouragement to go above and beyond.

To my friends, to the ones that are and to the ones that were.

And to you, my Ana. For bringing me peace, joy, and a lot of laughter.

Index

1 Introduction	18
1.1 Contextualization	18
1.2 Research Questions and Objectives	19
1.3 Structure	20
2 State of the Art	21
2.1 Emergency Medical Services.....	21
2.1.1 Emergency Medical Services planning strategies	21
2.1.2 Challenges to the Emergency Management Services.....	22
2.2 Time Series Forecasting.....	23
2.2.1 Exponential Smoothing	23
2.2.2 Autoregressive Integrated Moving Averages.....	24
2.2.3 Prophet	25
2.2.4 Transformers	26
2.2.5 Evaluation Metrics.....	27
2.3 Spatial Density Grouping	28
2.3.1 K-Means Algorithm	28
2.3.2 Density-based Clustering.....	28
2.3.3 Evaluation Metrics.....	30
2.4 Maximum Coverage Problem	31
2.4.1 Coverage.....	32
2.4.2 Linear Programming.....	32
2.4.3 Genetic Algorithms.....	33
3 Literature Review	34
3.1 Forecasting demand	34
3.1.1 Regression Models	35
3.1.2 Traditional Time Series.....	35
3.1.3 Hybrid solutions	38
3.2 Ambulance deployment problem	38
3.2.1 Linear Programming.....	38
3.2.2 Genetic Algorithms.....	39
3.2.3 Tabu Search	39
3.3 Privacy, security, and ethical considerations.....	40
4 Event-based vehicle allocation system for future emergency demands	41
4.1 Event-Driven Architecture	41
4.1.1 Publisher/Subscriber Pattern.....	42
4.1.2 Microservices.....	44
4.2 Solution Design	45
4.2.1 Visualization Component.....	46

4.2.2 Control Centre Component	46
4.2.3 Forecast Demand Component	48
4.3 Density Demand Prediction	49
4.4 Optimization of ambulance allocation.....	50
4.4.1 Emergency vehicle allocation LP with a neighbour mechanism.....	50
4.4.2 Genetic Algorithm for Reducing RT.....	52
5 EVA - A Virginia Beach Case Study.....	55
5.1 Historical Data Source.....	55
5.1.1 Feature Engineering	57
5.1.2 Data Split and Tunning	58
5.1.3 Frameworks and Libraries.....	58
5.2 Models Performance	58
6 Conclusions.....	65
6.1 Main Conclusion	65
6.2 Future Work.....	66

List of Figures

Figure 1 – Non-stationary data (left) vs. stationary data (right) (R.J. Hyndman et. al, 2021).....	24
Figure 2 - The Transformer - model architecture (Vaswani et al., 2017).....	26
Figure 3 - Multi-head attention consists of several attention layers running in parallel (Vaswani et al., 2017).....	27
Figure 4 - Sample databases from Martin et al. (1996)	29
Figure 5 - Density-reachability and Density-connection concepts in DBSCAN.	29
Figure 6 - Clustering sequence of a DBSCAN algorithm.....	30
Figure 7 - Efficient Coverage Area.	32
Figure 8 - The general structure of GA by (Kuok et. al, 2011).....	33
Figure 9 – Comparison of MAE error rates from TSA and ML for St. Mary's Hospital. Source: Vollmer et al. (2021)	36
Figure 10 – Comparison of MAE error rates from TSA and ML for Charing Cross Hospital. Source: Vollmer et al. (2021)	37
Figure 11 – Comparison of RMSE errors of different time series models (UTSF vs MTSF). Source: Source: Vollmer et al. (2021)	37
Figure 12 - Event hierarchy by (Dunkel et al., 2011).....	42
Figure 13 - Publisher/Subscriber Pattern communication model	43
Figure 14 - Apache Kafka message protocol diagram.....	44
Figure 15 - Microservices architecture	45
Figure 16 - EVA's architecture.....	46
Figure 17 - Virginia Beach city council districts.....	47
Figure 18 - Domain diagram representing the State.	47
Figure 19 - MLOps Pipeline to produce forecasting models.....	48
Figure 20 - Demand prediction with Hot-Spot location clustering	49
Figure 21 - Sum of ambulances minimizer with a neighbour mechanism structure	51
Figure 22 - Visual representation of a solution (individual) in the proposed genetic algorithm for ambulance deployment	52
Figure 23 - Visual representation of the partition of each offspring for the next population in the proposed genetic algorithm for ambulance deployment.....	54
Figure 24 – Seasonal decomposition using moving averages between 2010-2018 in Virginia Beach city.....	56
Figure 25 - Autocorrelation analysis for the Virginia Beach dataset.	57
Figure 26 - Elbow Method to find the K value for the K-Means model.	60
Figure 27 – K-Means visual result of the clusters when K takes the value of 47	61
Figure 28 - Density clusters of emergency calls using different time windows.....	62
Figure 29 – Simulation 1 visual representation of the optimal solution ($A' = 7$) using EVALP-NM.	63
Figure 30 – Simulation 2 visual representation of optimal solution ($A' = 34$) using EVALP-NM.	64

List of Tables

Table 1 - Schema metadata of the Virginia Beach dataset.....	55
Table 2 - Call volume prediction models performance using MSE and MAPE evaluation metrics.	59
Table 3 – Spatial Clustering performance between K-Means and DBSCAN with hyperparameter optimization.	61
Table 4 - Simulation configurations and performance results for optimizing ambulance allocation.	62

Acronyms and Symbols

API	Application Processing Interface
ARIMA	Autoregressive Integrated Moving Average
CFS	Calls for Service
CHI	Cakinski-Harabasz Index
CPU	Central Processing Unit
DBC	Density-Based Clustering
DBI	Davies-Bouldin Index
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
EA	Emergency Administrators
EDA	Event-Driven Architecture
EMS	Emergency Medical Services
ES	Exponential Smoothing
EVA	Emergency Vehicle Allocation
EVALP	Emergency Vehicle Allocation Linear Programming
EVALP-NM	Emergency Vehicle Allocation Linear Programming with Neighbour Mechanism
GA	Genetic Algorithms
GRU	Gated Recurrent Unit
HoP	Horizon of Prediction
HSC	Hot-Spot Clustering
HW	Holt-Winters
IoT	Internet of Things
LP	Linear Programming
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error

MAPE	Mean Absolute Percentage Error
MCP	Maximum Coverage Problem
MLOps	Machine Learning Model Operationalization Management
MQQT	Message Queuing Telemetry Transport
MSE	Mean Squared Error
MTSF	Multivariate Time Series Forecasting
NLP	Natural Language Processing
O	Objectives
PNN	Poisson Neural Network
REC	Residual Error Correction
REST	Representational State Transfer
RM	Regression Models
RNN	Recursive Neuro Networks
RQ	Research Questions
RSME	Root Mean Squared Error
RT	Response Time
SA	Service Area
SAL	Stationary Ambulance Location
SARIMA	Seasonal Autoregressive Integrated Moving Average
SDG	Spatial Density Grouping
SI	Silhouette Index
SOAP	Simple Object Access Protocol
SSE	Server-Sent Events
TS	Tabu Search
TSF	Time Series Forecasting
TTS	Traditional Time Series

UTSF	Univariate Time Series Forecasting
VAR	Vector Autoregression

1 Introduction

This chapter presents some initial considerations of the work described in this dissertation such as the context, motivation, and relevance of the work, general objectives, and the document structure.

1.1 Contextualization

Over the years, Urbanization has been noticeably accelerating, bringing new challenges to how society should manage their cities. According to the United Nations, 55% of the world's population lives in urban areas, which should increase to 68% by the end of 2050 (2022). Cities have the potential to provide a better quality of life for their citizens, and they are the backbone of the economic development of many countries. At the same time, cities struggle with traffic congestion, poverty, and pollution at dangerous levels, and have the record of being the epicentres of massive epidemics with consequences for the well-being of the population (Kuddus et al., 2020; Louf et al., 2015). Because of inflation, food prices are also rising (Artuc et al., 2022), and together with the abundance of fast-food choices currently available (Wu et al., 2017), it will make it more challenging, particularly for low-income households, to have a healthy diet (Conklin et al., 2019). This has been pointed out to be one of the reasons behind the exponential increase in obesity cases which is known to cause several problems such as diabetes, cardiovascular diseases, cancer, high blood pressure, and strokes (World Health Organization. Regional Office for Europe, 2022). Also, transportation can suffer a noticeable impact since more people will be traveling within the city, which can increase traffic-related accidents (Louf et al., 2015), which cost most countries 3% of their gross domestic product. As a result, health service systems have a vital role in shaping the future of the population living in the cities.

Emergency medical services (EMS) are critical infrastructures as they manage every aspect of pre-hospital care, making them the first respondents when an emergency happens. However, the intensification of the number of calls driven by urbanization will pressure the resources

already in place, which can lead to a decrease in service quality (Kuddus et al., 2020). More than ever, all the systems involved in coordinating emergency operations must be efficient, quick to help, and able to adapt in case the parameters change. For this reason, city governments must be alert and provide some indicators to evaluate patient outcomes and the overall quality of the EMS. One of the most impactful is the response time (RT), which is the period between the moment the medical services register a new emergency occurrence and when the first respondents arrived at the scene to provide care.

Many pieces of research suggest that RT is a determinant factor for increasing survival rates in patients that experienced out-of-hospital cardiac arrests (OHCA), traffic accidents, strokes, and overall trauma accidents (Bürger et al., 2018; Byrne et al., 2019; Heemskerk et al., 2021; Holmén et al., 2020). Also, on a socioeconomic level, lower-income areas are known to be neglected in terms of care delivery. A US national study concluded that low-income regions had a 10% higher RT than high-income areas, which can explain the income-mortality gap also noticed between these locations (Friedson, 2018; Hsia et al., 2018). Aside from the differences in investment, profit-driven ambulance management companies are pointed out to be one of the reasons behind these discrepancies for delivering, deciding to cover areas where patients can pay better (Friedson, 2018). Other aspects can also be considered in patients' outcomes like bystanders' intervention, age, medical history, etc (Bürger et al., 2018; Byrne et al., 2019; Heemskerk et al., 2021; Holmén et al., 2020). Still, RT is the aspect that EMS can directly influence over it, and policymakers have begun to set standards to ensure low levels of mortality or complications related to RT (Alharbi et al., 2022).

Introducing modern planning approaches into the EMS is pivotal for increasing efficiency and quality of service. To reach the goal of seamless coordination it is important to have accurate forecasting systems, that can handle the uncertainty of the environment in which the EMS operates (Jaklič et al., 2015; Reuter-Oppermann et al., 2017). However, existing methodologies seem to be inadequate to manage resources, since it uses averages of a few data points that produce a lot of spikes resulting in under-resourcing and over-resourcing (Matteson, McLean, Woodard, & Henderson, 2011). On the other hand, intelligent dispatching strategies must also work alongside the forecasting systems to be able to maintain a suitable overall coverage. In many cities, when an ambulance becomes available (after finishing a service) it must return to a pre-defined base, instead of reallocating to another station that requires more operatives (Reuter-Oppermann et al., 2017).

1.2 Research Questions and Objectives

Even though the main goal of this master's thesis is to propose an intelligent EMS capable of supporting decision-making operations in all phases of pre-hospital emergency care, is possible to define a collection of research questions (RQ) that will guide this work, as follows:

- **RQ1** – What is the current state-of-the-art of EMS, especially in the context of forecasting emergency demands and allocating emergency vehicles?

- **RQ2** – Can traditional systems architectures support real-time decision-making platforms empowered by AI models in emergency medicine?
- **RQ3** – Do traditional forecasting models perform well for short-term predicting of emergency demands?
- **RQ4** – Does spatial allocation reduce RT in emergency resource management?

As a result of the conception of the above RQ, five objectives (O) were also defined, as follows:

- **O1** – Investigate the current state-of-the-art in demand forecasting and geographical allocation.
- **O2** – Identifying what is lacking in the current techniques for demand forecasting and geographical allocation.
- **O3** – Propose a new architecture to implement an EMS to support the use of forecasting and allocation models in real-time.
- **O4** – Testing the multiple forecasting models to identify the suitable model architecture for short-window forecasting.
- **O5** – Testing the proposed allocation of emergency vehicles optimization in several geographical environments.

1.3 Structure

This document is organized into multiple sections, described as follows. Chapter 2 describes the general *modus operandi* of an EMS and the state of the art regarding solutions that could be used to solve both problems. The literature review is presented on Chapter 3. Chapter 4 delineates the proposed architecture of the EVA system, with greater detail on the time series forecasting and spatial allocation components. On Chapter 5, the results of each solution in both contexts are presented using a selected dataset. Chapter 6 discusses this work's overall findings and proposes future work.

2 State of the Art

More than ever, patients expect Healthcare services to meet the highest standards of care with state-of-the-art medical equipment, and that begins with emergencies. Decisions made by emergency administrators (EA) can have an impact beyond the call they are operating. They are experienced professionals, but the human species has a limited capacity to process data. Therefore, advanced information technologies need to assist in the decision-making process to reach the best possible considering all the variables.

2.1 Emergency Medical Services

Aside from countries' regulations, EMS systems can be categorized into two types: the Anglo-American system and the Franco-German system which can be differentiated by the first having paramedics trained to administer any treatment necessary, while the latter having specialized doctors that are called only in severe cases (Reuter-Oppermann et al., 2017).

2.1.1 Emergency Medical Services planning strategies

The main goal of the EMS is to reach patients in the shortest time possible to assess, stabilize and, if necessary, transport them to hospitals for further treatment. Every day, EA need to make decisions in two phases: call assessment and vehicle dispatch. First, they need to determine the severity and urgency of the occurrence to know which type and how many vehicles it is needed on the scene. Only then, can decisions be made to better dispatch the ambulances. Consequently, planning strategies are essential for this complex system to respond to entire communities effectively. It is an ongoing task in an uncertain environment that requires constant monitoring of a collection of performance indicators that assert the various aspects of the system (Brotcorne et al., 2003).

According to Matinrad (2022), emergency planning can be divided into four phases: Mitigation, Preparedness, Response, and Recovery. In each stage, the necessary measures are accessed to minimize the outcomes of daily emergencies. Bélanger et al. (2019), detailed that decisions are made at three levels: Strategic, Tactical, and Operational. The first two have long-term and short-term implications, respectively, and are usually done during the Preparedness phase (e.g., location of ambulance stations and personnel's shift schedule). On the other hand, the Operational level decisions are related to the Responsive stage which encompasses real-time actions carried out after an emergency occurrence (e.g., dispatching and relocation).

Most of the recent work regarding planning strategies in the EMS has focused on solving static allocation problems that fall on the Strategic and Tactical levels. Although these solutions improve the quality of care and equity of service, demands can change during short periods, causing the current plan to not meet the standards. Therefore, a continuous assessment is necessary to be able to adapt to unexpected circumstances (Aringhieri et al., 2017; Bélanger et al., 2019).

2.1.2 Challenges to the Emergency Management Services

Over the years, much research has been done to mitigate the challenges in the EMS context, which can be categorized into four sections to be presented as follow:

- **Forecasting** – To better provide assistance and to prevent the waste of resources, EMS systems need to account for future demands to better plan their course of action in an effective way. At the moment, EMS uses basic approaches to forecast the volume of emergency presentations, using the averages of past days in relation to the current day (King et al., 2022).
- **Static Location** – Out-of-hospital resources need to be dispersed throughout the area it is destined to serve. Stationary base locations and ambulance base assignments are the main pains since they have longer consequences.
- **Dispatching** – EA have policies to evaluate the urgency at the emergency scene that helps dispatch the correct number and type of ambulances needed at the emergency scene. However, EA also need help managing the available resources to avoid over-dispatching without taking into consideration the void that it will produce.
- **Relocation** – Flexibility is required to always maintain adequate levels of coverage, meaning that even though ambulances are assigned to a station, if there is a sudden rise in demand anywhere else, these resources could rearrange.

Most of the research usually concentrates on only one of these problems so there is a lack of focus on the integration aspect in the literature. Forecasting demands are pivotal for effective allocation and dispatch models since they are usually trying to maximize coverage, minimize response times, etc., whose values result from assessing historical data and trying to infer future needs. Therefore, to have a complete support system at every step of the EMS response actions, it needs to have state-of-art forecasting models to analyse the necessary demands to serve as input for the allocation, dispatching, and relocation problems.

2.2 Time Series Forecasting

Time Series Forecasting (TSF) is an analytical technique that specializes in the prediction of sequential future occurrences of events based on patterns and trends identified in historical data. TSF techniques involve using statistical, machine learning and deep learning-like models to capture cause-effect relationships in temporal data and extrapolate the same patterns into the future. Not all temporal data can be modelled using TSF methods.

TSF can be classified into Univariate (UTSF) and Multivariate (MTSF) models depending on the number of variables it is trying to predict. To use the TSF approach, the data must have some characteristics (Brockwell et al., 2002):

- Trend: Normal behaviour of the variable value in the time series. Normally refers to an increase or decrease over time.
- Seasonality: Patterns can be observed within a timeframe (e.g., daily, monthly, yearly, etc.). This can be due to external factors that only happen at certain times like holidays and weather conditions based on the season.
- Residual: This encompasses the rest of the data that outliers from the above characteristics (e.g., holidays).

2.2.1 Exponential Smoothing

Holt (2004) and Winters (1960) suggested Exponential Smoothing (ES) to produce accurate estimations of future values based on a weighted average of past occurrences. “Weighted” in this context refers to giving a level of confidence to the past data, whose value diminishes as the data is older. This is because recent data have more correlation with future prediction by helping identify the current trends, than older data. Holt-Winters method involves one forecast equation and three smoothing equations, as follow:

$$\hat{y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)} \quad (1)$$

$$l_t = \alpha(y_t - s_{t-m}) + (1 - \alpha)(l_{t-1} + b_{t-1}) \quad (2)$$

$$b_t = \beta^*(l_t - l_{t-1}) + (1 - \beta^*)b_{t-1} \quad (3)$$

$$s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1 - \gamma)s_{t-m}, \quad (4)$$

where l_t denotes an estimate of the level of the series at the time t , b_t denotes an estimate of the trend of the series at time t , α is a smoothing parameter for the level and β^* is the smoothing parameter for the trend. s_t is the seasonal equation and shows a weighted average between the current seasonal index and the seasonal index of the same previous m period.

2.2.2 Autoregressive Integrated Moving Averages

Like other regression models, an autoregression model forecasts the target variable using the linear combination of the past values of the variable (R.J. Hyndman et al., 2021).

$$y_t = c + \phi_\rho y_{t-\rho} + \varepsilon_t, \quad (5)$$

where ε_t is noise and $\phi_\rho y_{t-\rho}$ are lagged values of y_t of order ρ .

Autoregressive models work with stationary data only and can become very complex when $\rho > 3$. Therefore, rather than using past values of the target variable, the moving average model uses past forecast errors for the regression model.

$$y_t = c + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \dots + \phi_q \varepsilon_{t-q}, \quad (6)$$

where ε_t is noise and $\phi_q \varepsilon_{t-q}$ are lagged values of the noise of order q . The combination of these two approaches can result in a non-seasonal autoregressive integrated moving average model (ARIMA).

Stationarity is a key concept in time series prediction since is used to assess if it is possible to find patterns in the data or if it has a high percentage of uncertainty and randomness. Stationary data has a constant mean and variance over time which makes it easy to model and produce accurate forecasts. On the other hand, non-stationary time series have mean and variance values changing because the data show fluctuations over time (R.J. Hyndman et al., 2021). Figure 1 represents the Google stock price over 200 days in which it can be seen trends and levels (non-stationary) and the annual total of lynx trapped in the McKenzie River district of northwest Canada (stationary).

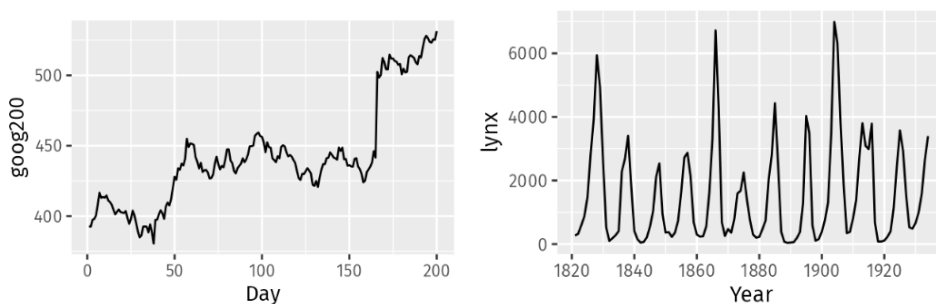


Figure 1 – Non-stationary data (left) vs. stationary data (right) (R.J. Hyndman et. al, 2021).

This property of the data is important for models, including ARIMA, since to make reliable predictions, data need to be stationary. If a time series is non-stationary, a technique called differencing can be applied by subtracting the current value from the previous one resulting in a series of differences between consecutive data points. Differencing helps stabilize the mean and variance of the series, making it more amenable to modelling and prediction.

Even though ARIMA models are widely used, these models are more suggestible to large trend errors, especially in long-term predictions (e.g., when patterns have major gaps between each occurrence) or are very sensible to outliers such as sudden rises because of holidays.

2.2.3 Prophet

Prophet is an innovative forecasting model developed Taylor et al. (2018) for Facebook and revolves around the idea that a time series can be analysed in a way that allows one to identify the different components contributing to a specific behaviour. Prophet combines three main factors such as trend, seasonality, and holidays, as follow:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t, \quad (7)$$

where $g(t)$ denotes the non-periodic changes, $s(t)$ the periodic changes and $h(t)$ represents the holidays which are irregular occurrences with a random duration period typically short.

The main component of the Prophet model is the trend term, $g(t)$ which tries to infer how data will behave in the future based on how it behaved in the past. Currently, two trend models are available to use: a linear and a nonlinear saturation growth model. Growth forecasting is based on analysing the evolution of predicted variable and how is expected to continue to evolve. For variables that does not experience saturation growth the trend component can be defined as follow:

$$g(t) = (k + a(t)^T \delta)t + (m + a(t)^T \gamma), \quad (8)$$

where k is a constant for growth rate, m is an offset parameter. $a(t)^T \delta$ and $a(t)^T \gamma$ are adjustments made to each constant.

On the other hand, if a saturation growth is present the data is usually modelled based of a logistic function, as follow:

$$g(t) = \frac{C(t)}{1 + \exp(-(k + a(t)^T \delta)(t - (m + a(t)^T \gamma)))}, \quad (9)$$

where $C(t)$ is the carrying capacity at time t . In the EMS context, calls are expected to increase based on the growth of the population within the city.

Seasonality is captured using periodic functions to analyse the effects on a time series that repeats over several periods, as

$$s(t) = X(t)\beta \quad (10)$$

where $X(t)$ is a standard Fourier series to model arbitrary smooth seasonal effects and β constant to provide smoothing. Emergency calls fluctuate during each day, but it can be seen a slight increase after office hours for short-term seasonality.

The last component is related to the know events that disrupted the normal flow of the time series and since there is not a specific pattern of occurrence it cannot be modelled. Instead, Prophet has the option of including these events as a list when modelling the data.

2.2.4 Transformers

The Transformers model was introduced by Vaswani et al., 2017 as a groundbreaking encoder-decoder architecture (Figure 2) to process sequential data, having the most significance in the field of natural language processing (NLP).

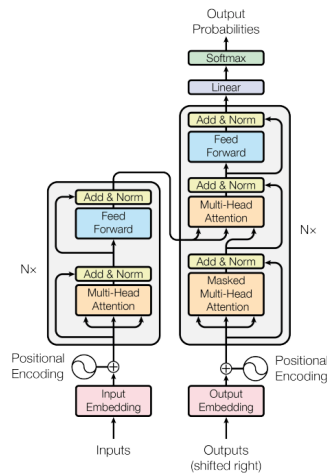


Figure 2 - The Transformer - model architecture (Vaswani et al., 2017).

The encoder maps the input sequence, whose representation enables the decoder to generate the output sequence by applying autoregressive mechanisms. An encoder and decoder layer have each two sub-layers: the multi-head attention and a feed-forward network. The decoder has an additional multi-head attention layer for the output of the encoder (Vaswani et al., 2017). The attention mechanism is what captures the meaningful relationships between different elements of sequential data. Each input is associated to a learnable matrix $m(V, K, Q)$, that is project h times and applied a *softmax* function (Scaled Dot-Product Attention) (Ahmed et al., 2023). The outputs are then concatenated to produce the final solution as it can be seen in Figure 2 - The Transformer - model architecture (Vaswani et al., 2017).Figure 2.

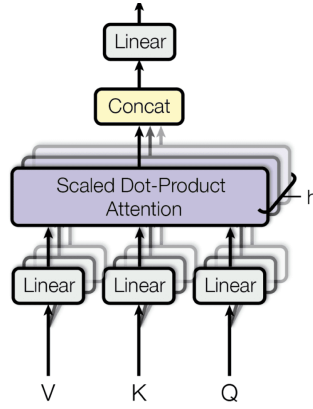


Figure 3 - Multi-head attention consists of several attention layers running in parallel (Vaswani et al., 2017)

Like previous mentions of autoregression in Section 2.2.2, Transformers models use previous outputs as the inputs for each step which is suitable for scenarios where forecasts are influenced by their own past predictions. Also, these models offer a way to tackle some of the problems of other deep learning architectures such as recursive neural networks (RNN), long short-term memory (LSTM) and gated recurrent unit (GRU) face, by having mechanisms such as self-attention and positional encoding that address the problems with overfitting predictions (Ahmed et al., 2023).

2.2.5 Evaluation Metrics

- Mean Absolute Error (MAE)

It is a simple metric for evaluating the performance of a forecasting model that calculates the average of the differences between the prediction and the real values, given by:

$$MAE = \frac{1}{n} * \sum_{i=0}^n |real_i - predicted_i| \quad (11)$$

, where n is the number of data points in the dataset, $real_i$ represents the actual value at the data point i , $predicted_i$ represents the predicted value at the data point i .

- Mean Absolute Percentage Error (MAPE)

This metric is the percentage equivalent to the MAE value, given by:

$$MAPE = \frac{1}{n} * \sum_{i=0}^n |real_i - predicted_i| * 100 \quad (12)$$

- Mean Squared Error (MSE)

MSE is very commonly used when using forecasting models, because of the extra impact it has when the models have bigger errors between what was predicted and the real value. It can be represented as:

$$MSE = \frac{1}{n} * \sum_{i=0}^n (real_i - predicted_i)^2 \quad (13)$$

2.3 Spatial Density Grouping

Spatial Density Grouping (SDG) is a useful technique for finding different patterns between geographical data and assessing how they can be grouped together. Clustering is the basis of SDG and as an unsupervised learning method, it does not rely on labelled data for training.

2.3.1 K-Means Algorithm

K-Means is a partitional clustering algorithm which means it aims to group data into a pre-specified number of clusters (K) so that the sum of the squared distances between data points and the cluster's centroid is minimized (Praveen et al., 2016).

The algorithm starts by randomly choosing K centroids following the objective function and assigning the data points more closely to each centroid, separating the plane into several regions, equivalent to a Voronoi diagram. Each iteration optimizes the locations of the centroids and regroups the data points from one cluster to another, halting when there is no more movement between the current clustering layout and the previous.

Although K-Means algorithm offers a multitude of opportunities such as simplicity in implementing, speed and scalability when dealing with many features and large datasets. However, it also has some limitations such as having to define the K value in advance, being dependant on the initialization of the centroids leading to multiple solutions depending on initial centroid placement, and assumptions relating the clusters being equally spherical and having similar densities.

2.3.2 Density-based Clustering

Density-based Clustering (DBC) was firstly presented by Martin et al. (1996) with the name of Density-Based Spatial Clustering of Applications with Noise (DBSCAN), addressing the limitations of the K-Means algorithm by introducing the concept of density when generating the clusters. This property improves the definition of each cluster by evaluating the concentration of data in certain spaces, which also helps spot the noise in the dataset since it is usually less dense. This type of clustering is useful for samples like the ones shown in Figure 4.

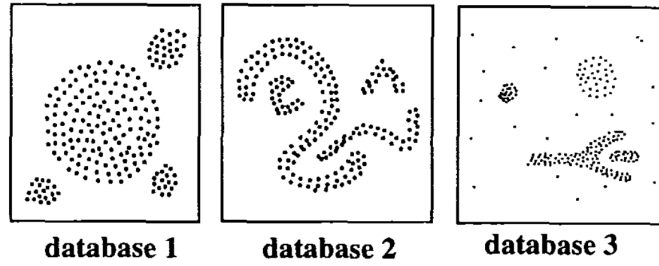


Figure 4 - Sample databases from Martin et al. (1996)

The algorithm evaluates if a point is related to another if the concept of reachability and connectivity are present. Let p and q be two points in space, p is density-reachable from q if they can be connected by a set of intermediate point that are density-connected. On other hand, p and q are density-connected if there is an intermediate point o that can connect them within their area. Density-reachability and density-connectivity are represented in Figure 5.

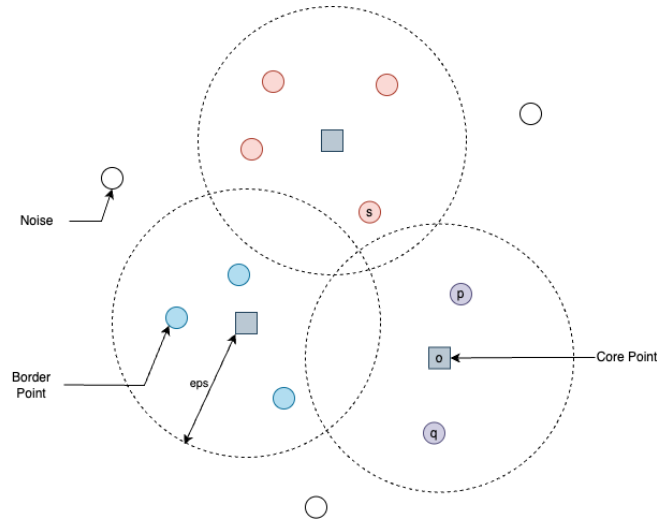


Figure 5 - Density-reachability and Density-connection concepts in DBSCAN.

As shown in Figure 5, p and q are density-connected by and o , therefore belong to the same cluster, while s is outside of the radius projected by o , making s belong to another cluster. This radius value is one of DBSCAN parameters, often represented as ϵ , alongside a threshold value of the number of data points necessary for a cluster to be recognized as dense, $minPts$. The clustering process of DBSCAN models is represented in Figure 6.

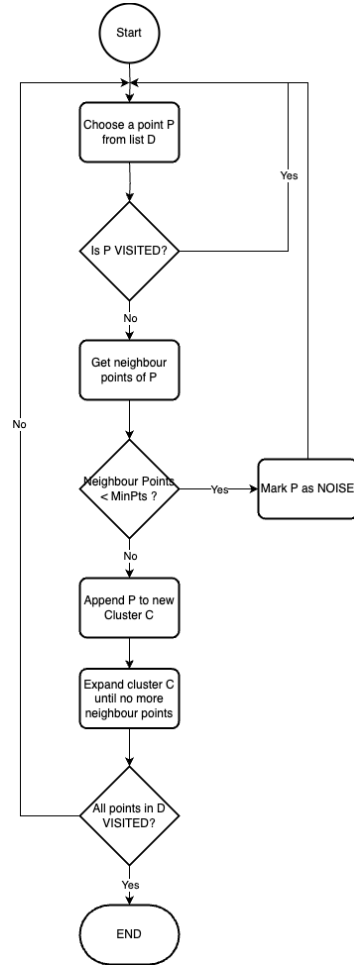


Figure 6 - Clustering sequence of a DBSCAN algorithm.

The algorithm iterates over a set of points and tries to produce a cluster by recursively getting the neighbours points within the radius ε . The ambiguity of this parameter difficulties the tuning of an DBSCAN model.

2.3.3 Evaluation Metrics

In this section some performance metrics are presented that can be used to evaluate clustering models.

- Silhouette Index (Shutaywi et al., 2021)

Silhouette Index (SI) measures how each data point in a cluster is similar to other members of the same cluster compared to the rest of the clusters. It helps evaluate the separation of clusters and can be represented as:

$$SI(x_i) = \frac{b(x_i) - a(x_i)}{\max\{b(x_i), a(x_i)\}} \quad (14)$$

where, $a(x_i)$ is the average distance between the data point to all the members of the same cluster and $b(x_i)$ is the average distance from the data point to all the other outside its cluster.

$SI(x_i)$ can take the value -1 to 1 by the first indicates that the data point was assign to the wrong cluster and the latter being that the data was well clustered.

- Cakinski-Harabasz Index (X. Wang et, 2019)

Cakinski-Harabasz Index (CHI), also known as the Variance Ratio Criterion, measure assesses the quality of the cluster, K , by comparing the variance inter-cluster, $B(K)$, to the variance of intra-cluster $W(K)$, as:

$$CHI(K) = \frac{B(K)(N-K)}{W(K)(K-1)}, \quad (15)$$

The higher the value for CHI the more in common each data point in a cluster have with the member of the same cluster.

- Davies-Bouldin Index (Ros, Riad, & Guillaume, 2023)

Davies-Bouldin Index (DBI) evaluates the degree of similarity between a cluster and other clusters, and can be represented as,

$$DBI(k) = \frac{1}{k} \sum_{i=1}^k \max_{i \neq j} \left\{ \frac{S_i + S_j}{d(\bar{x}_i, \bar{x}_j)} \right\} \quad (16)$$

where, S_i is the average distance between the cluster's centroid and all the members. This measure helps determine the optimal number of clusters, in case it is needed to be predefined before training. The lower the DBI shows that each cluster is distinct and well separated.

2.4 Maximum Coverage Problem

Maximum Coverage Problem (MCP) is a type of optimization problem for the purpose of finding an ideal strategy when selecting a subset of elements from a larger group to maximize the coverage of a certain property. MCP practical applications can be seen in many domains such as:

- Logistics – Choosing the location of a new facility to ensure the expansion of services to places where it could positively impact the business (Shariff et al., 2012).
- Wireless networks – In under-provisioned networks the placement of limited devices (e.g., IoT devices) needs to guarantee optimal coverage of the targets while in over-provisioned networks have also the problem of minimizing the number of devices (Alibeiki et al., 2021).
- Advertising and Marketing – “One-fit-for-all” campaigns are not as effective as target advertising when trying to maximize the impact of a campaign and, therefore, reach the most relevant audience (Zhao et al., 2019).

Fundamentally, an MCP can be formulated as:

$$\text{maximize } \sum_{e_j \in E} y_j \quad (17)$$

$$\text{subject to } \sum x_i \leq k \quad (18)$$

$$\sum_{e_j \in S_i} x_i \geq y_j \quad (19)$$

$$y_i \in \{0,1\} \quad (20)$$

$$x_i \in \{0,1\} \quad (21)$$

where, y_j is the coverage of each element to be maximized, x_i is a set of elements that will cover the target, k is a constant to limit the number of sets to select.

2.4.1 Coverage

For the development of MCP models, defining the notion of coverage within the requirements of the problem is vital to achieving high-quality solutions. Chen et al. (2023) defines coverage as a property that specifies how well a certain target can rely on the availability of a service. This translates into coverage being represented as a space/area surrounding the element providing the service as represented in Figure 7.

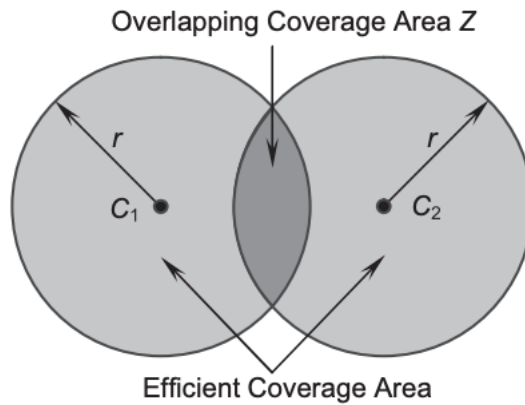


Figure 7 - Efficient Coverage Area.

2.4.2 Linear Programming

Linear Programming (LP) is an optimization method that aims to use mathematical functions to achieve the best solution for a given problem, and it is used in a wide range of fields. At its core, the LP algorithm can be defined as:

find x

maximize/minimize $F(x)$

subject to C

where x is the set of variables to be determined, $F(x)$ is the objective function to optimize and C is the set of constraints for the solution space of the function.

The process of developing an LP solution involves converting real-world problems into a mathematical representation, which can be a hard endeavour for problems that are far too complex to be represented as an equation or in cases where there are many solutions.

2.4.3 Genetic Algorithms

Initially proposed by Holland J. (1992, Genetic Algorithm (GA) is a metaheuristic approach that offers optimal solutions when there is a lot of uncertainty, by using adapted principles of natural selection such as mutation, crossover, and selection. A GA utilizes a set of potential solutions to a problem, denominated population, with the primary goal of improving these solutions over successive iterations. In each step, each solution is evaluated on how well it performs with respect to the problem's objective, using a fitness function, whose values will impact the selection of the solutions that will proceed to produce the solutions for the next iteration. To refine the selected solutions, crossover and mutation mechanisms are implemented to help explore the solution space, introduce diversity, and prevent suboptimal solutions. Poor quality solutions are disregarded, and the process restarts until some termination conditions are met (e.g., a good enough solution is found, no improvement between iterations, execution time conditions, etc.). The general structure of a GA is represented in Figure 8.

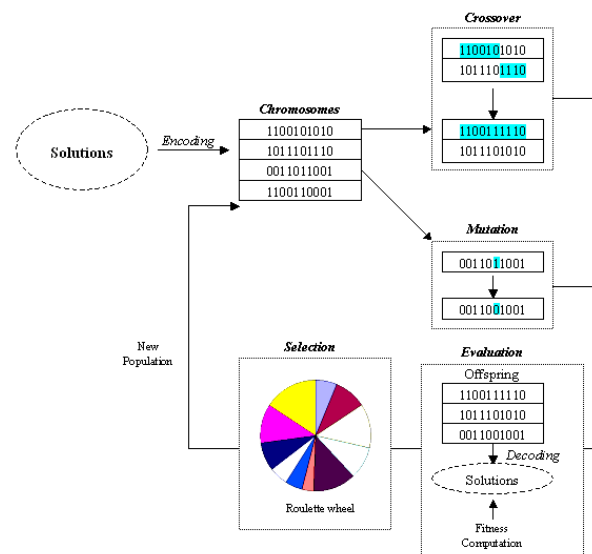


Figure 8 - The general structure of GA by (Kuok et. al, 2011)

To develop a GA for any context several parameters and settings need to be configured that influence the behaviour and performance of the model, such as population size, selection method, crossover and mutation rates and termination criteria.

3 Literature Review

An emergency can happen at any time and place, making uncertainty an important aspect to consider in EMS planning. To achieve high levels of efficiency in such an ambiguous environment, EMS needs to be able to predict the demands of service based on historical data. Only after assessing what is necessary based on the current parameters can a plan be formulated or adjusted.

3.1 Forecasting demand

According to the literature, forecasting models are built with the assumption that future outcomes can be predicted based on the circumstances that lead to their occurrence in the past. Future predictions have been attempted to be done across centuries, but in the last 50 years forecasting competitions (or M Competitions) have become a trend among the research community. The last was the M5 Competition was focused on retail forecasting demand that supports supermarket companies in their planning strategies (e.g., stocks management, and promotions...) (Makridakis et al., 2022). These events have been used to enhance old forecasting models' accuracy and propose new approaches for this problem in several fields (e.g., energy, tourism, and retail) (Rob J. Hyndman, 2020). Consequently, these competitions opened the door for the same foundations and models to be used in other domains such as in the EMS.

According to recent systematic reviews and comparative studies (Etu et al., 2022; Sudarshan, Brabrand, Range, & Wiil, 2021; Whitt & Zhang, 2019) forecasting demand solutions can be categorized into Regression models (RM), Traditional Time Series (TTS) and hybrid solutions.

3.1.1 Regression Models

Regression models provide a function that describes a correlation between one or multiple independent variables and a target, dependable, variable. Many types of regression models are used in prediction analysis. To use the correct regression approach to model the data, researchers commonly do a first comprehensive analysis to evaluate the quality of the data with one of the most important aspects being the data's distribution.

Steins et al. (2019) studied the use of a Poisson regression approach to predict real-time demand at the emergency department. It provided a collection of explanatory factors that were added to the data that are known to correlate with changes in emergency demand (e.g., population age groups, road length, number of points of interest). Even though the Poisson regression model performed better than the current implementation in the EMS, it was not clear that changing to the new model would be beneficial. However, the Poisson regression model gave relevant information about which and at what length variables influence the demands.

King et al., (2022) present a forecast pipeline to aggregate multiple predictions for effective bed management attending to current and future needs. Unlike the work presented before, this pipeline computes current emergency department patients' probability of admission accumulated with the predictive volume of patients yet to arrive, within a prediction window. The patient's probability of admission was generated using a binary logistic loss function with an extreme gradient boosting classifier model. For the patient's forecast, the author used the Poisson regression model with temporal predictors such as prediction time, seasonality, and day of the week.

3.1.2 Traditional Time Series

Many authors use an ARIMA model to predict emergency demands over various time series. Boyle et al. (2012) develop a software solution that could be able to predict hospital admissions to assist in bed management. It was discovered that the models performed worse - have a higher MAPE – on shorter time intervals due to the smaller sample size (Monthly: 2%; Daily: 11%, 4-hourly: 38%; Hourly: 50%). The methods perform poorly for real-time assessment of the demand and were no evidence of impact analysis after its implementation. Wong & Lai (2014) studied whether climate forecasting data (e.g., temperature, humidity) could improve the ARIMA model's 7-day prediction capability. The metric selected for evaluating this model was the root mean squared error (RMSE) which is a better metric than MAPE to understand the model's performance. It was found that including average temperature improved the RMSE of the models by decreasing the metric's value by 8.8% of the 1-day forecast (RMSE: 53 vs. 58).

Vollmer et al. (2021) propose a new approach that compares RM and TTS models to be used in online platforms that provide demand predictions of one, three, and seven days into the future. Data sources were used from eight years of daily presentations at the St. Mary's and Charing Cross Hospitals in London to train and validate the models. For both hospitals, TTS

performed worse than RM, regarding MAPE and MAE (Figure 9;

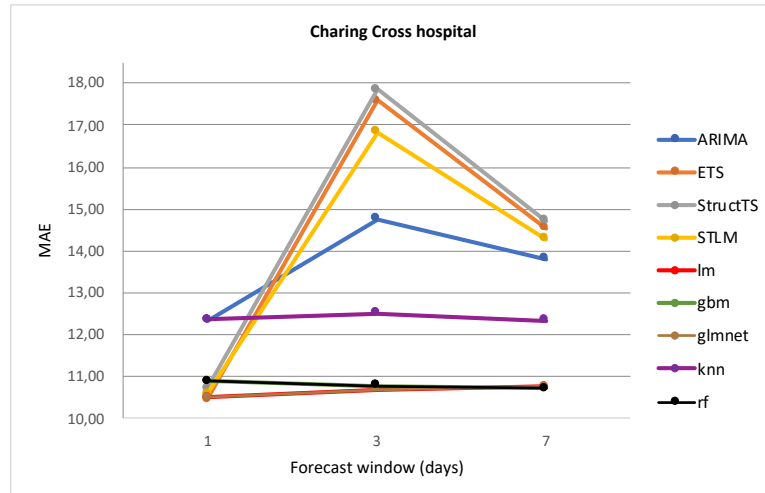


Figure 10). MAE values for RM get higher on longer prediction intervals, but the same cannot be verified in TTS. These models had a higher value in the three-prediction period than in the seven-prediction period. Aside from the modelling development, there was presented a comprehensive analysis of the data resulted in the identification of weekly and monthly trends that can be used as predictors (e.g., demand increases during flu seasons and Mondays have the highest demand...).

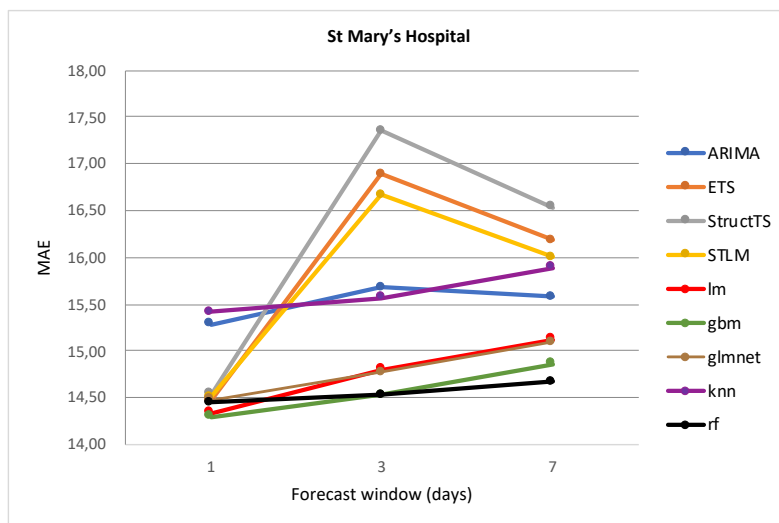


Figure 9 – Comparison of MAE error rates from TSA and ML for St. Mary's Hospital. Source: Vollmer et al. (2021)

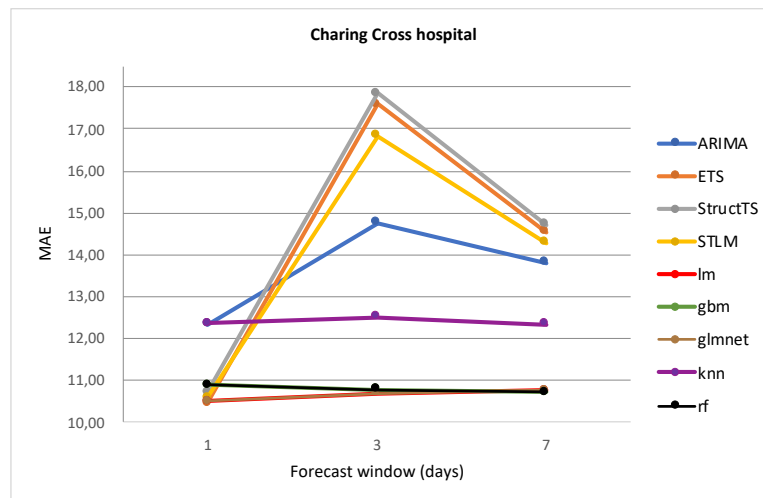


Figure 10 – Comparison of MAE error rates from TSA and ML for Charing Cross Hospital.
Source: Vollmer et al. (2021)

Etu et al., (2022) compared the performance between UTSF and MTSF models in predicting emergency patient arrivals during the COVID-19 pandemic over different time series (1, 7, 14, 21, and 30 days). RMSE was used to assess performance for both UTSF and MTSF models, which can be seen in Figure 11. On UTSF models Holt-Winters (HW) was the one that performed better in short time series (1 - 7 days), seasonal autoregressive integrated moving average (SARIMA) on the 14 days horizon, and Univariate LSTM had substantially decent results on the longer time series (21 and 30 days). On MTSF models Multivariate LSTM had the lowest RMSE value in all time horizons.

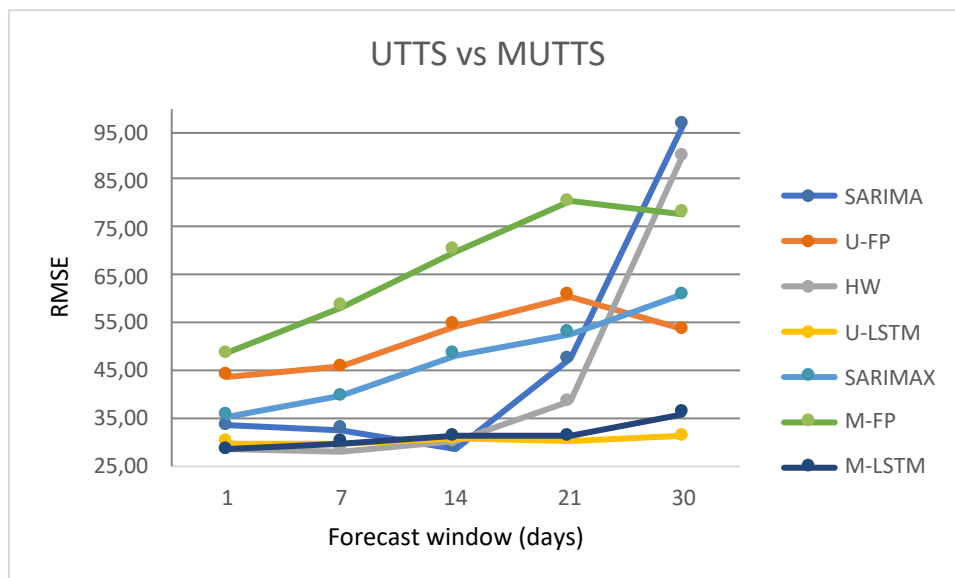


Figure 11 – Comparison of RMSE errors of different time series models (UTSF vs MTSF).
Source: Source: Vollmer et al. (2021)

3.1.3 Hybrid solutions

Huang et al. (2019) propose a Poisson Neural Network (PNN) which is a combination of the Poisson regression model with a neural network. A GA was used to train the PNN as an optimization mechanism that is known to improve the neural network parameters (weights and thresholds). Also, to further improve the accuracy of the PNN it included a Residual Error Correction (REC) model that tries to predict the residual error to sum to the PNN predicted result. A version of each individual model used in the combined model was used to compare and evaluate the PNN in terms of MAPE and the RMSE values, which showed that PNN was superior (MAPE: 7.44%; RMSE: 6.46%).

3.2 Ambulance deployment problem

Ambulance location and relocation problems have been the focus of a lot of research work, particularly since it is one of the biggest challenges in the EMS context (Aringhieri et al., 2017). First studies develop static models based on deterministic and probabilistic static approaches (Aringhieri et al., 2017; Brotcorne et al., 2003) that were very useful for the strategic level but lacked the flexibility for real-time changes during operation.

3.2.1 Linear Programming

Andersson et al. (2004) and Andersson et al. (2016) propose a new approach for assessing coverage of allocation plans that encompasses the current and future demand needs using the preparedness measure. Some of the recent work in ambulance deployment uses the preparedness equation as the objective function in their implementations (Carvalho et al., 2020; T. C. van Barneveld et al., 2016; T. van Barneveld et al., 2018).

Naoum-Sawaya et al. (2013) developed a two-stage stochastic optimization model that aims firstly to minimize ambulance relocations within a time horizon while maintaining an acceptable service level. The second stage is related to the dispatching mechanism that assigns vehicles to emergency calls, excluding those that do not meet response time requirements. This work introduces the use of thresholds to assess the need for relocations, making the system flexible but at the same time reliable.

(Chen et al., 2023) propose a mixed integer programming model to select a set of new facilities to open that will maximize the coverage for the customer's demands and minimize the cost of opening a new facility by having a budget. This paper introduces the notion of cost, which is monetary in this case, but could be extended in the context of relocation constraints when moving ambulances dynamically.

3.2.2 Genetic Algorithms

Shariff et al. (2012) propose a GA-based model to address the capacity constraints of the limited ambulances and stations available. By assuming all demand nodes will not change, the objective function tries to maximize the number of ambulances in each node, which results in a hard dependency between the prediction of demand and the allocation. This could cause not having enough backup units to accommodate sudden rises in demand in lower-demand zones.

Zaheeruddin et al. (2021) developed a multi-component framework that focuses on providing a deployment plan that allocates a set of ambulances on stations based on the frequency of accidents registered in a specific area and the RT of each ambulance to the sites. This framework poses a great asset to analyse how impactful the number of ambulances is on the average RT but evaluates the current solution based on RT only, neglecting to provide a way to incorporate redeployment measures in case of variable changes as well as demonstrating the system performance in real-time conditions.

Kaur (2021) uses a GA to maximize the overall coverage by assigning each available ambulance to some coordinates independent from the stations. This can be achieved by computing the area that each ambulance is covering as a circle and making sure that in each iteration the GA produces fewer and fewer ambulances with zero crossing points with another ambulance. One important advantage of this approach is that the area coverage increases exponentially within a few iterations which is helpful to test this solution in a real-time environment. However, a downside was pointed out by the author related to the limited number of ambulances that this method supports.

Wang et al. (2022) proposed a serial-number-coded GA to improve the computation time by shortening the length of the chromosome (possible solution) and the quality of the convergence against a standard GA. The chromosomes are represented as an array and were encoded by having the serial number of the medical unit in each element and the demand points serial numbers as the index. In the crossover, it was used a multi-point approach that demonstrates very effective in exploring the search space and did not have the fittest random solutions covered together making it a hard dependency on the initialization of the population.

3.2.3 Tabu Search

Gendreau et al. (2001) suggest a tabu search (TS) model based on a parallel heuristic approach for real-time relocations. The authors also include a penalty constraint to avoid multiple relocations for the same ambulance and longer relocation time between locations. On the simulation data, the TS model was able to have a suitable functioning since 98% of urgent calls were covered within 7 minutes. However, TS failed to provide feasible redeployments on concurrent calls or even when the time between calls is too short which is a common occurrence in EMS.

3.3 Privacy, security, and ethical considerations

Data protection legislation enforces some policies for the storage and use of personal information. Care organizations have a commitment to the public in keeping their medical information secure and not sharing medical records without patients' informed consent. Artificial Intelligence tools often use sensitive data to make predictions, and recommendations, and support both patients and medical personnel. Aside from ensuring privacy and security, transparency in addressing the public should be focused on by the organizations to bring awareness about what their data is being used (Ngiam et al., 2019).

One way to ensure privacy is making sure only those with legitimate needs for using the data have access to it. Data repositories need to be secure and when sharing data with third parties, organizations need to handle only the necessary data points and, in case of sensitive data, perform anonymization and be active in the feedback to the public. In the case of the forecasting models, these need data on the emergency calls received in the past to make predictions of the future. Therefore, implementing this tool would require access to emergency records and for those to be updated sometimes daily.

Ethical concerns have been raised for years about the use of intelligent approaches in healthcare. An unbalanced and unrepresentative dataset could originate a model that has biases that could originate inequalities, unfairness, and discrimination when providing care. On forecasting demands, data should be analysed prior to modelling to ensure the quality of information and the importance of each data feature on the predictions. This will help to effectively evaluate the performance of the model. Also, predictions should offer "explainability" to be able to assert more confidence for the decision-makers. For the allocation problems, the system must give special attention to high call rate areas and those in remote locations, but also make sure to have enough resources for outliers that come with such an uncertain domain. Since variables can rapidly change, continuous assessment of the current allocation plan should be a must in ensuring a rapid diagnostic of coverage deficiencies and correcting them.

The author commits to performing ethical and responsible research practices and ensuring the integrity and credibility of the research results.

4 Event-based vehicle allocation system for future emergency demands

This chapter describes the proposed solution to integrate demand prediction of emergency incidents with vehicle allocation algorithms in an event-based solution. Initially, the foundations of an event-driven software architecture are introduced, with a focus on communication patterns. In the following sections, it is proposed an architecture solution based on events, that can support real-time responsiveness, with particular attention to the presentation of the forecasting modelling and vehicle allocation optimization.

4.1 Event-Driven Architecture

Event-driven architecture (EDA) is a modern software design concept used in many real-time decision-support systems across multiple domains such as streaming services, Internet of Things (IoT) applications and geospatial data infrastructures. Events in EDA are random occurrences that have a direct influence on the system state (Dunkel et al., 2011). These events can originate from a multitude of sources, and they can be categorised based on their importance for the decision process as seen in Figure 12.

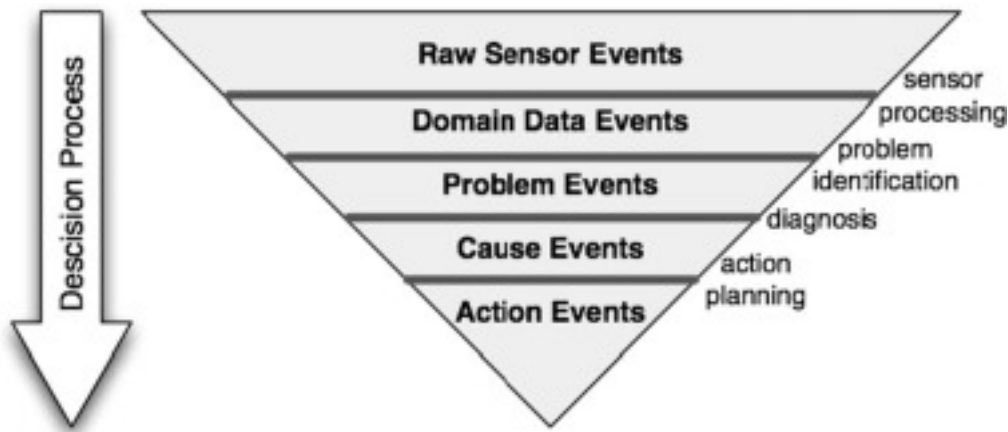


Figure 12 - Event hierarchy by (Dunkel et al., 2011).

EDA relies on asynchronous communication which is a characteristic not achievable by other architectures such as Representational State Transfer (REST) or Simple Object Access Protocol (SOAP), which are better suited for request-response interactions. Asynchronous means that events are independent of each other and do not require immediate responses. Other characteristics include:

- Lightweight data transfer – Often event payloads do not contain a lot of information, since the focus is more towards the fact that something happened.
- Decoupled infrastructure – services communicating via events have no dependencies between them, in fact, they are usually not aware that other components exist. This offers flexibility and security to continue adding new components with minimal interferences.
- Horizontal scalability – Due to the flexibility mentioned before, the workload can be managed by distributing more event processing units to handle an increase in events.

4.1.1 Publisher/Subscriber Pattern

As previously mentioned, EDAs follow an asynchronous communication model between every component that enables the consumer to target the types of events that are relevant to their functionality. On the other hand, the events need to be created by another service that acts as a producer of specific messages. This pattern is known as the publisher/subscriber pattern where the subscriber consumes events from a topic of interest, while the publisher produces events related to a topic (Rieke et al., 2018) The interaction between the publisher and the subscriber is achieved by a middleware-like module often addressed in the literature as the message broker, event channel or event server, with the main responsibility of maintaining the state of the topics and their subscriptions as shown in Figure 13.

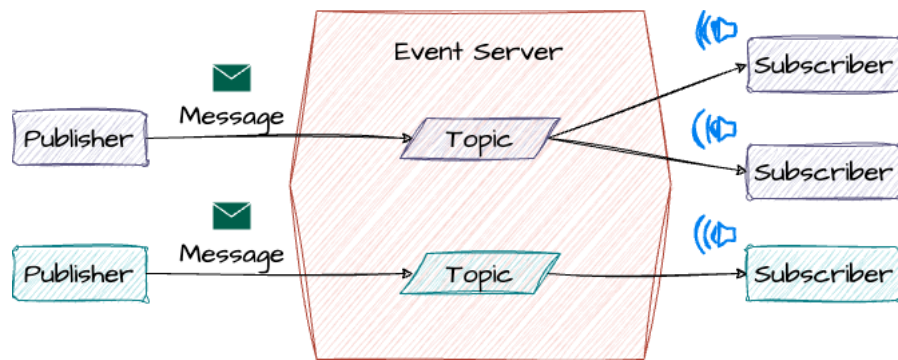


Figure 13 - Publisher/Subscriber Pattern communication model

There are two models for the subscriber to consume a topic: client-initiated (pull-based) or server-initiated (push-based). The difference between both approaches relies on whether it is necessary for the consumer to be connected to the topic or if the producer pushes an event to a consumer whenever the event occurs (Jacobsen, 2009). Several message protocols follow the client or server-initiated approaches, some as follows:

- WebSocket – is a bidirectional communication protocol to send data across a connection channel if the connection is alive. This protocol is used on real-time applications with a front-end component where data displayed is continuously being sent by the server (V. Wang et al., 2013).
- Message Queuing Telemetry Transport (MQQT) – is a TCP/IP-based message protocol with minimal message payloads, making it efficient in low-bandwidth networks. MQQT is the preferred protocol for IoT systems as they often have limitation of central processing unit (CPU) and battery life (V et al., 2022).
- Server-Sent Events (SSE) – This protocol implements a server-initiated approach in the sense that instead of clients continually pooling for new information, the producer pushes as soon as new data is available. The use cases for SSE are live notification displays, location tracking and financial services (Torre et al., 2019).
- Apache Kafka (<https://kafka.apache.org/>) – Apache Software Foundation solution for data streams comes to target concerns regarding data traffic, especially for complex and heavy messages (images, videos, live streams). Kafka is a distributed streaming platform in the sense that messages arriving at the broker (cluster) are separated and allocated into different partitions enabling parallel message transfer (Figure 14).

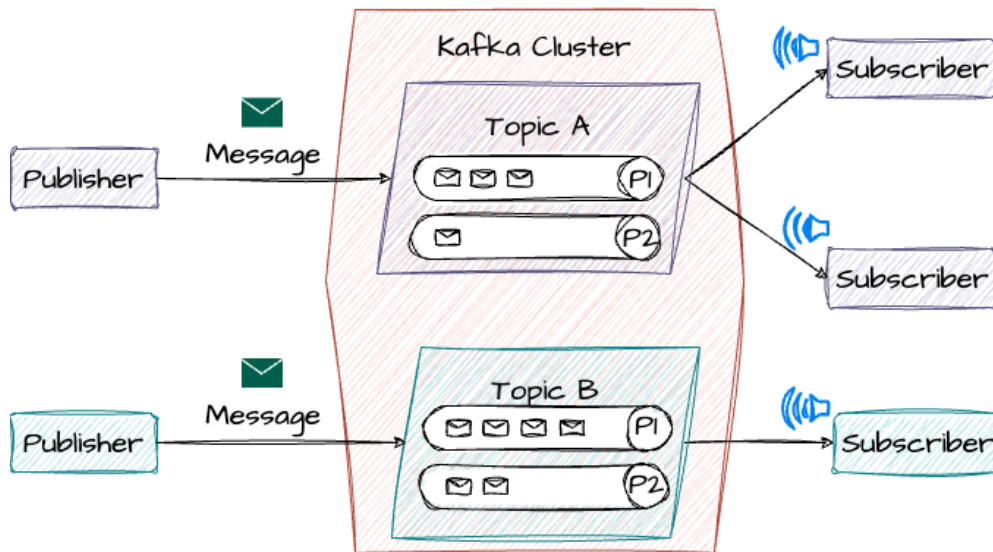


Figure 14 - Apache Kafka message protocol diagram.

4.1.2 Microservices

EDA enables the separation of a complex application into independent modules integrated with each other using communication models such as the ones presented in Section 4.1.1. This is the concept behind the microservices technology which has gained popularity in recent years, due to their flexible, scalable, and agile way of designing applications. Each microservice can be developed, deployed, and maintained without affecting other microservices' functionality, and the communications between them are done using well-defined application processing interfaces (API). The main concern when dealing with microservices architectures is that services need to be specialized in a segment of the whole system, and, if necessary, broken down into smaller services in case complexity increases (Newman, 2015). Figure 15 presents a visual representation of a microservice architecture, the interactions between each other and other external components. Common services usually are segregated into containers to also incentivize the low decoupling strategy, having at least one service communicating with external components such as storage or clients.

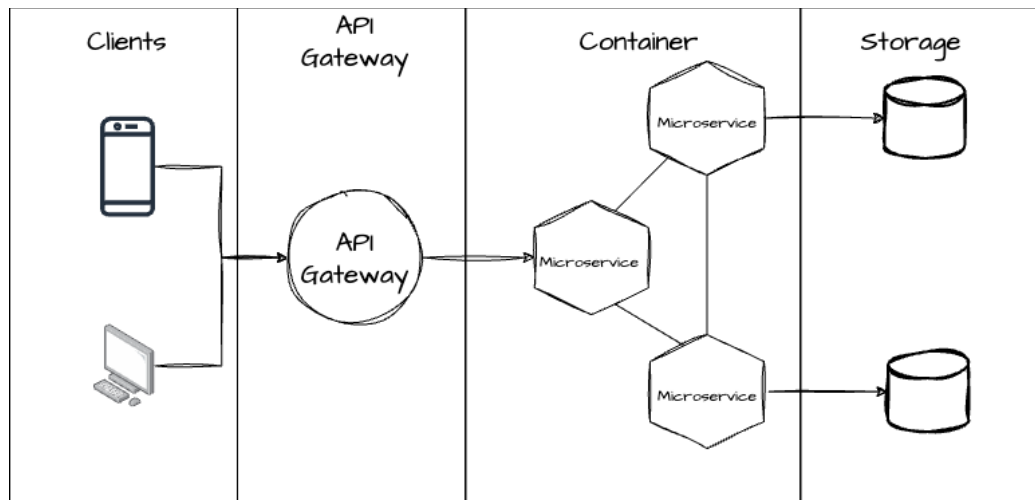


Figure 15 - Microservices architecture

4.2 Solution Design

The architectural design is fundamental for its proper functioning of a system, especially one that relies heavily on real-time data for enabling decision-making processes. As an event-based critical computing system, EMS need to keep up to date on new ways to improve performance, robustness, and security. Traditional architectural designs fail meet the requirements to develop end-to-end solutions mainly because of scalability issues, where microservices architectures does not seem to be affected.

The current work proposes the boilerplate architecture presented for emergency vehicle allocation (EVA) in Figure 16 to develop decoupled EMS. This architecture is based on a microservices approach, and each interaction between different services is done through an event-like, low-latency communication channel. Each individual service works separately without being dependent on other components, which improves the maintenance of the system in an agile development environment. The goal of this system is to support a visual platform with the information needed to build dashboards, maps, and other types of graphical components to help users understand the status of operations and how can they be improved.

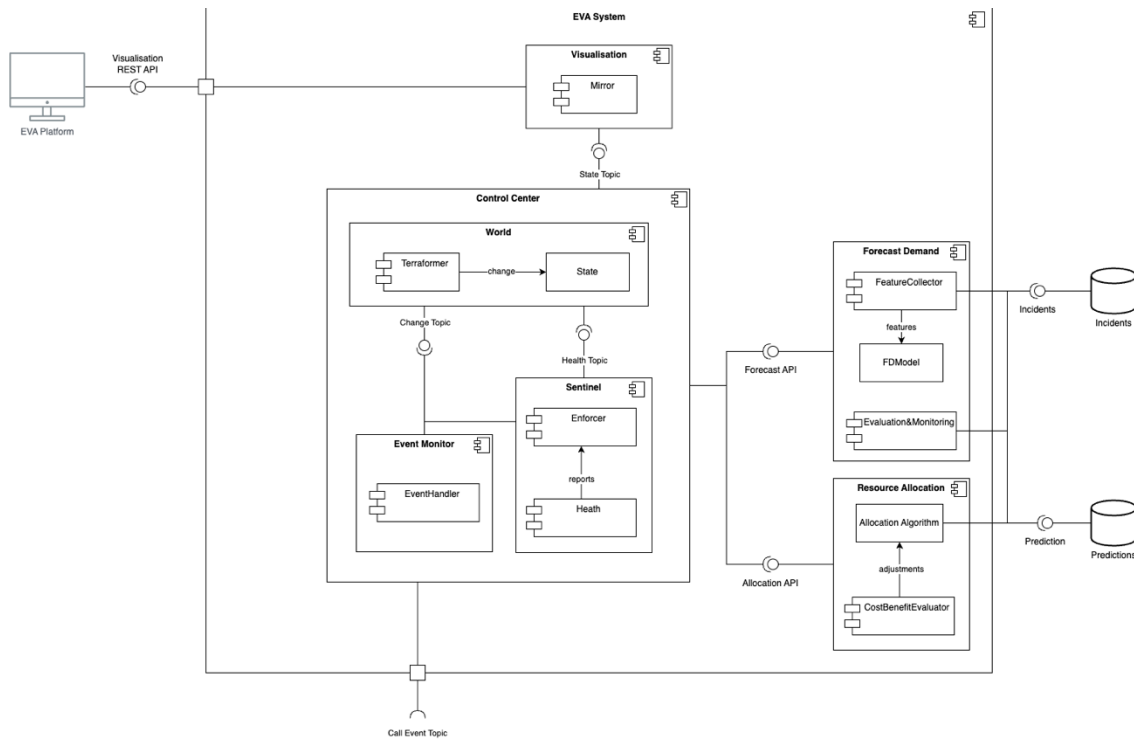


Figure 16 - EVA's architecture.

4.2.1 Visualization Component

Visualization tools such as dashboards have been the preferred way of simplifying large amounts of data into visual digestive information. These tools empower the user to make informed decisions based on the analysis at hand, especially for real-time systems that need to always have the most updated version of the data available.

In such complex architecture, accurate data representation holds an important role in the usability of the proposed solution. The visualization component has the responsibility to provide a simplified view of the ongoing conditions of the allocation layout of each ambulance like location and status, and each zone's coverage evaluation. This component is the gateway between the EVA's platform, which is the user interface, and the business logic behind what is presented.

4.2.2 Control Centre Component

Logistics operations often need a centralized facility to manage the entirety of the operations. The same principle was applied in the proposed architecture, represented by the Control Centre component functioning as the supervisor of the current state and decision-maker of the necessary actions to be taken.

For the sake of representing the current layout of operations and their relationships, the solution proposes the State, an entity serving as a single point of truth. Countries have multiple EMS that serve in a specific service area (e.g., city), which can be further divided into zones based on census information (Figure 17) to make it easy to serve the population.

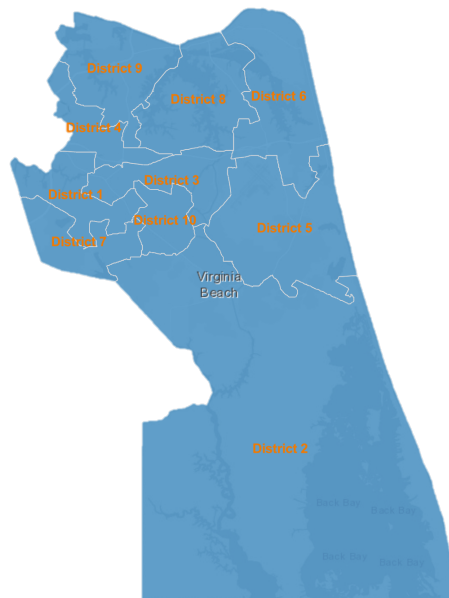


Figure 17 - Virginia Beach city council districts.

Each zone may differ in demand for emergency services, so the zone's individual coverage needs to be considered during the allocation of emergency vehicles. To ensure adequate coverage, stations are constructed to which a set of ambulances can be assigned, especially in rural areas. Station locations are long-term decisions that restrict future improvements due to the costs of building the facility and adding a new facility may not be the most viable decision. On the other hand, the number of ambulances in each station can be adjusted during shift management. This relationships can be each entity is shown in the diagram of Figure 18.

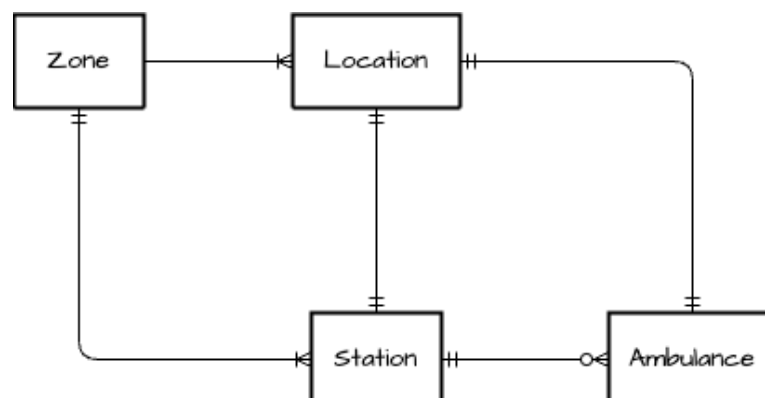


Figure 18 - Domain diagram representing the State.

The Control Centre need to hold the real state of the operations and be the only one that can apply changes to it, through the means of the *Terraformer* module. As the name suggested, *Terraformer* is the only module with the business logic to deal with updates on the State, to avoid discrepancies entering to have redundancy on update a single object. However, the *Terraformer* alone cannot make any changes without having an event that causes it to enforce some action.

The Sentinel component is the protection layer between the external services and the State. It focuses on two areas: load balancing for *Terraformer* and inspecting the quality of the State. The first is managing the external requests for changes from outside of the Control Centre. The second. Is checking the health of the state, meaning if there is an optimal solution currently in the state or it has been degradation in terms of quality.

4.2.3 Forecast Demand Component

As mentioned before, effective planning for the distribution of emergency vehicles encompasses anticipating the demands of the population, which essentially means the number of calls the EMS will receive. The Forecast Demand component's primary goal is to offer a state-of-the-art forecasting model that results from training several types of machine learning and deep learning architectures with the historical data of past emergency calls. To achieve this, it is necessary to implement several Machine Learning Model Operationalization Management (MLOps) Pipelines like the one presented in Figure 19.

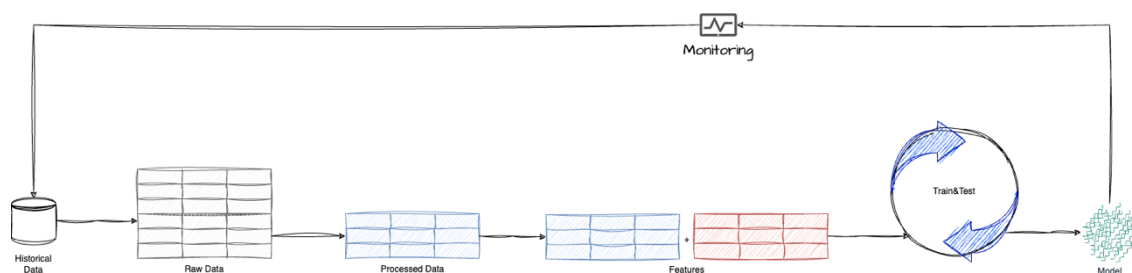


Figure 19 - MLOps Pipeline to produce forecasting models.

Each pipeline begins with fetching the necessary historical data from a source, followed by applying data processing techniques to evaluate data quality and clean the dataset. Featuring engineering is then used to enrich with additional information, that together with the original data could become features if they capture relevant patterns and relationships that can lead to more accurate predictions. Next, the features will be used to model against a target variable and then test the accuracy for new predictions. The monitoring stage is pivotal in evaluating the degradation of the model's accuracy over time, by comparing the prediction made with the real values and computing a score. This score will determine if the model needs to be retrained, which could happen if old patterns are deprecated.

The choice of which is the right forecasting model to use in a production environment benefits from this Pipeline approach. The literature review does not provide a clear answer on what is

the best architecture for volume/frequency forecasting instead, it offers a variety of approaches that perform better under certain scenarios (e.g., short vs. long window forecasting). This means that to choose the best, a sort of competition should be encouraged among the known model's architecture.

4.3 Density Demand Prediction

As mentioned before, predicting the future volume of emergency incidents is crucial in any EMS. Besides this, having a model capable of selecting areas where it has been reported a bigger density of calls can improve the preparedness of the emergency services when managing their resources.

Temporal prediction alight with geographical density clustering is a powerful combination for a single model to achieve. Therefore, this paper proposes a hybrid solution represented in Figure 20.

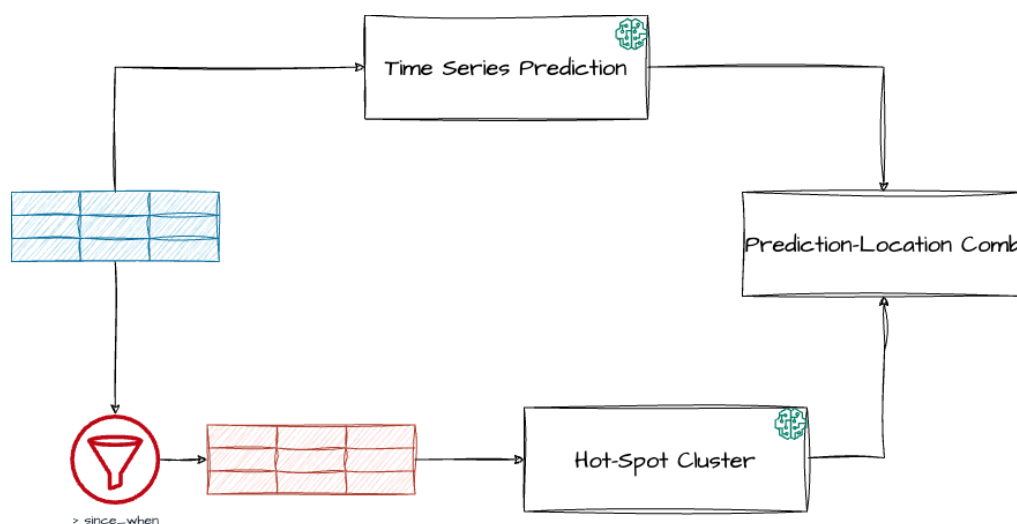


Figure 20 - Demand prediction with Hot-Spot location clustering

There are three key parts of the hybrid model:

- Time Series Prediction

It is a Time Series model focused on finding patterns that could help predict the volume of calls for a specific time window (e.g., hourly).

- Hot-Spot Clustering

Hot-Spot Clustering model (HSC) is dedicated to spatial prediction of the volume of calls to spot high density areas where the demand for emergency medicine was greater. The model is based on the DBSCAN algorithm presented on Section 2.3.2, tailored to focus on the clusters with the higher density of data points. It is used the same data for training as for the time series prediction, with a restriction on how further back the records would go, as a configurable

parameter (*since_when*) according to the domain knowledge. This decision was made under the assumption that location patterns for emergency calls can change with time, so it should take into consideration the amount historical data feed to the model.

- Prediction-Location Combination

Both above models focus on a specific part of the problem, consequently their outputs need to be merged representing the future prediction for each individual hot-spot zone HS_{iw} , as follow:

$$HS_{iw} = D_i * prediction_w \quad (22)$$

where, D_i is the density score of the cluster i , resulting from the HSC model and $prediction_w$ the output from the time series model which is the value of the volume of calls for the next window w .

4.4 Optimization of ambulance allocation

The main goal of optimizing allocation is for it to be efficient and use the minimum of resources without compromising the quality of service. It is known that for every emergency call at least an ambulance is deployed to assist on scene. As protocol dictates, the chosen ambulance is always the one closer (in time) to the local that originated the incident.

To allocate ambulances in a service area (SA) it is required some à priori knowledge about variables that can impact the chosen location, such as:

- The number of ambulances – this number will impact the allocation plan since one of the goals of this optimization is deploying only the essential number of vehicles.
- Demand points – this is the output of the forecasting models in and the focus of the objective function for the optimization models.
- Stationary ambulance location (SAL) – this usually represents the stations where ambulances can park and wait for a new dispatch order, but it could also be parking spaces or other facilities that provide support to available ambulances (e.g., fire stations).

This section provides a comprehensive understanding of the developed algorithms supporting ambulance distribution.

4.4.1 Emergency vehicle allocation LP with a neighbour mechanism

Emergency Vehicle Allocation LP with a neighbour mechanism (EVALP-NM) was developed to distribute the ambulances for each station with the assumption that stations already provide suitable coverage of the entire SA. It is constituted by two LP problems as represented in Figure 21.



Figure 21 - Sum of ambulances minimizer with a neighbour mechanism structure

The first problem allocates the available ambulances by the stations, giving the remaining number of available ambulances to be allocated to stations to serve as support to adjacent stations, in the second problem.

- Ambulance allocation with prediction

Let A_s , the decision variable of this problem, be the number of ambulances located at the station s , which can assume the value of $1 \dots N_s$, with N_s being the total number of stations in a service area. The objective function can be represented as:

$$\text{MINIMIZE } \sum_{i=1}^{N_s} A_s \quad (23)$$

subject to,

$$\sum_{i=1}^{N_s} A_s < N_A \quad (24)$$

$$A_s \geq 1 \quad (25)$$

$$A_s \geq \text{prediction}_z / N_{sz}, \quad (26)$$

where prediction_z is the value of the number of calls forecasted for the zone z and N_{sz} is the number of stations in zone z .

- Neighbour mechanism

The thought behind the neighbour mechanism was that stations near stations from other zones could have a buffer of available units. When predicting call frequency for a zone, there is no indication of the distribution of calls in that zone, meaning that the calls made near a border should increase both zones' demands. Following the first allocation algorithm, a set of ambulances that were not allocated to any station, N'_A , can be used to allocate to stations near the border with zones that have a higher prediction of calls. The objective function for this mechanism can be represented as:

$$\text{MINIMIZE } \sum_{i=1}^{N_s} A_s \quad (27)$$

subject to,

$$\sum_{i=1}^{N_s} A_s < N'_A \quad (28)$$

$$A_s \geq \text{neighbour_zones}_s \quad (29)$$

where, neighbour_zones_s is the number of zones where the borders are within a station coverage area, as follows:

$$\sum_{i=1}^{N_s} Z_s \quad (30)$$

subject to,

$$d_{sz} \geq r_{\text{coverage}} \quad (31)$$

$$\text{prediction}_z \geq \alpha * \text{prediction}_{z'} \quad (32)$$

where, d_{sz} is the distance from station s to zone's border z , r_{coverage} is the radius of the coverage area provided by the stations, $\text{prediction}_{z'}$ is the volume of calls predicted for the neighbour zone and α is a constant value to spot neighbour zones with impactful differences in predictions.

4.4.2 Genetic Algorithm for Reducing RT

A GA is a suitable approach for testing multiple solutions when there is a lot of uncertainty, which is the case for the ambulance deployment problem. The main objective is minimizing the RT of the ambulances to the demand points while maintaining good overall coverage of an SA.

- Initial population

GA requires generating an initial population with, in most cases, a random set of possible solutions. Each individual is a vector with integers whose values represent the number of ambulances located at each SAL. A graphic representation of a solution is portrayed in Figure 22.

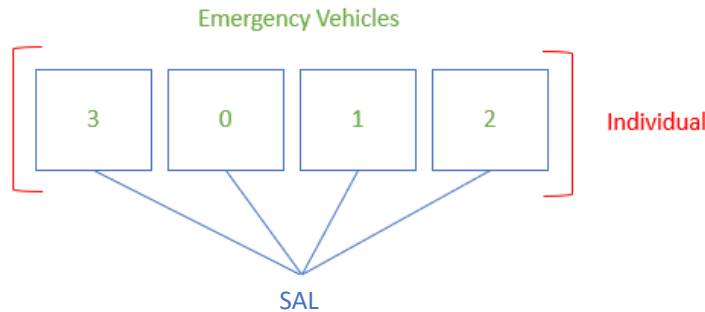


Figure 22 - Visual representation of a solution (individual) in the proposed genetic algorithm for ambulance deployment

In the end, the fittest solution will be decoded to identify the coordinates for the SAL by using the index of each position of the vector as an identifier. Each individual component of the GA can be described, as follows:

- Fitness function

The objective function is used on the selection phase of the GA in each iteration and computes the sum of the RT of the proposed solution using a geolocation service. It is important to mention that it was assumed that each ambulance has the same speed for each RT calculation. Therefore, the fitness function can be defined as:

$$\min \left\{ \sum_{i=0}^n RT_i \right\} \quad (33)$$

$$RT_i = \text{distance}(SAL_i, DP_i) * v, \quad (34)$$

where RT_i is the RT of each index i given by the distance between the SAL_i and the DP_i times the velocity v which is a constant.

- Selection

The selection mechanism applied to the ambulance deployment problem is based on a simple version of the Roulette Wheel Selection which assigns a probability to each individual in the population proportional to their fitness value, so they have a better chance to be chosen to perform crossover.

- Crossover and Mutation

Regarding crossover, this step is done by using a single-point crossover technique on the results of the selection, by enforcing two crossovers: i) between the best individuals of the selection and ii) between the best individuals and the worst ones. This was done because the worst individuals could improve the already admissible solutions and improve them further. Concerning mutation, it is performed randomly by incrementing/decrementing 1 value to a random position of the vector. However, some constraints must be considered since the sum of the genotype of the mutated individual cannot be above the limit of the number of emergency vehicles or lower than 0.

The final step of the iteration is generating the next population based on the best individuals to keep, the offspring's individuals resulting from crossover and mutation steps, and some randomly generated individuals to achieve the population size and to impose randomness in the process. Figure 23 is a visual representation of the next generation population based on the probabilities of the offspring to keep.

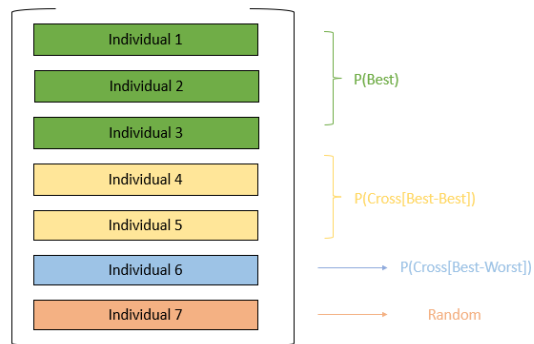


Figure 23 - Visual representation of the partition of each offspring for the next population in the proposed genetic algorithm for ambulance deployment

5 EVA - A Virginia Beach Case Study

In this chapter, the proposed solution will be applied to the Virginia Beach city case study. The reason behind the selection of this case study was the availability and quality of the provided by the City of Virginia Beach.

5.1 Historical Data Source

Forecasting models need a substantial amount of data to make more accurate inferences based on historical data. To improve transparency and support the scientific communities, EMS departments began to make available their data related to the calls for service (CFS). In this work, it was used a dataset from Virginia Beach, a city in the state of Virginia in the United States of America, with the CFS from 2010 until 2018. This data is from a public repository for data published by the City of Virginia Beach (<https://gis.data.vbgov.com/>). The data has no explicit personal data that can be used to identify an individual, nor is relevant to the current work. Table 1 has some metadata insights about the dataset, divided into 12 columns, summing to 326609 rows of raw data.

Table 1 - Schema metadata of the Virginia Beach dataset.

Column Name	Type	Format
EMS Call Number	Integer	N/A
Block Address	String	N/A
City	String	Virginia Beach
State	String	VA
Call Priority	Integer	1-5
Rescue Squad Number	String	N/A
Call Date and Time	Date Time	YYYY-MM-DD hh:mm:ss
Entry Date and Time	Date Time	YYYY-MM-DD hh:mm:ss
Dispatch Date and Time	Date Time	YYYY-MM-DD hh:mm:ss

En route Date and Time	Date Time	YYYY-MM-DD hh:mm:ss
On Scene Date and Time	Date Time	YYYY-MM-DD hh:mm:ss
Close Date and Time	Date Time	YYYY-MM-DD hh:mm:ss

EMS Call Number, City, State, Rescue Squad Number, and all the *DateTime* columns except *Call Date and Time* are irrelevant and were dropped from the dataset and the analysis to reduce noise. *Call Date and Time* columns were split into the date and the time to improve analysis, data visualization, and storage.

As previously mentioned, forecast prediction is only possible if the data follows some pattern (trend), and if that pattern is observed within a timeframe (seasonality), which was detected in the Virginia Beach dataset using the *Call Date and Time* (Figure 24).

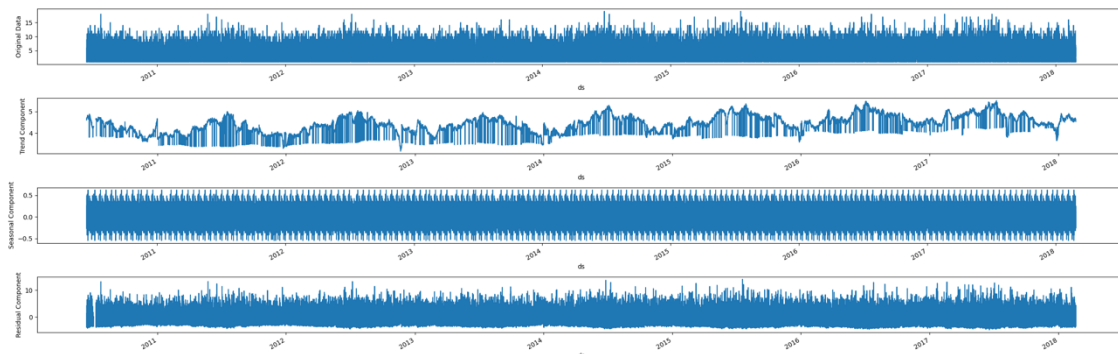


Figure 24 – Seasonal decomposition using moving averages between 2010-2018 in Virginia Beach city.

However, in such a larger dataset is advantageous to use a more advanced technique to evaluate and understand the relationships between the time series data with itself over time. For the current work it was used Autocorrelation Analysis which shows a weak correlation between the values of the time series at different time instances, suggesting that past observations lack the predictive power to forecast future occurrences. This makes it difficult for models that heavily rely on seasonality and trend to be used in this context, such ARIMA and ES. The results of this study can be seen in the graphic present in Figure 25, where the values are relatively close to 0.

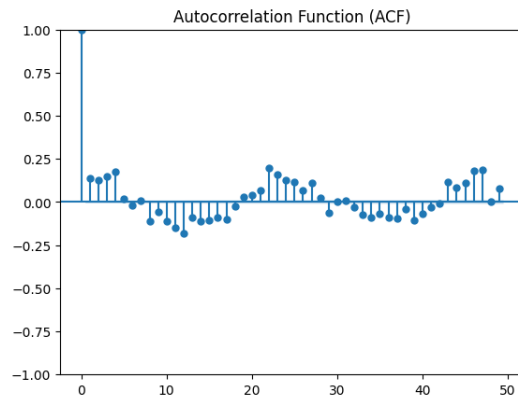


Figure 25 - Autocorrelation analysis for the Virginia Beach dataset.

Although it could be used the plot above to visualize the distribution of data over time, a deep analysis using boxplots was done in Appendix A.1. It detected some outliers that were not removed because, in the current context, models should be sensible to outliers, like Christmas.

5.1.1 Feature Engineering

The original dataset lacks additional information that could be used to capture complex relationships and enhance the model's accuracy and reliability, such as:

- Geographical metadata – the original dataset column *Block Address* was used to retrieve coordinates using a geolocation API, alongside the service area a specific call belongs to. These features can help the model recognize localized trends and disparities in emergency call volumes.
- Weather insights – weather-related conditions such as temperature, humidity, precipitation, snow depth, wind speed and visibility can help understand how weather conditions impact emergency call patterns.
- Temporal parameters – day of the week and holidays can help the model to identify very specific variations.
- Historical call volume – Embedding the prior context before an increase/decrease in call volume in each hour, the model can gain the ability to anticipate future demands by analysing short-term trends. It was added two columns with the number of calls in the previous hour and the number of calls in the previous day.

These new features could make the model more susceptible to subtle variations leading to predictions close to real-world values. With newly added features is important to have a statistical test to determine the usefulness of the feature. The Augmented Dickey-Fuller unit root test assesses the presence of a unit root, which indicates non-stationarity in the dataset. ADF tests were used on the features mentioned above and selected the variables that exhibit consistent patterns.

5.1.2 Data Split and Tunning

In TSF models training and testing are usually dependent on the amount of time series into the future the model will try to predict, also known as the horizon of the prediction (HoP). The training set in each training iteration will be all the data less the rows corresponding to the value of the horizon.

Hyperparameters tuning efforts were performed to find optimal values for some of the hyperparameters of the models such as the use of the cross-validation approach. The goal is to split the data as mentioned, generate iteratively random values for the hyperparameters, fit the data using the generated values, and compare each performance metric resulting from the iterations. Additionally, for the clustering models, it was used the Grid Search Cross Validation from *sklearn* for tuning, by using different combinations of the specified hyperparameters, calculating the performance under a specific metric and choosing the best combination.

5.1.3 Frameworks and Libraries

Several libraries were used to help the development of the models, as follows:

- *StatsForecast* (<https://nixtla.github.io/statsforecast/>) – Offers a pool of univariate time series forecasting models optimized for high performance and scalability.
- *Scikit-learn* (<https://scikit-learn.org/>) – It provides many efficient tools for predictive data analysis integrated with data processing, visualization, and evaluation metrics.
- *Tensorflow* (<https://www.tensorflow.org/>) – Is an open-source platform for building and deploying a multitude of machine and deep learning models.
- *Openrouteservice* (<https://openrouteservice.org/>) – Provides geographic services using free data directly from OpenStreetMap. It was used to compute travel times between the locations and the demand points.
- *Pulp* (<https://coin-or.github.io/pulp/>) – LP modeller to solve optimization problems. It was used in the allocation models.

5.2 Models Performance

As presented in the architectural design of the proposed solution, a collection of models is used to enable the function of EVA's system. For each individual category, it was studied several models, based on what is present in the literature, and their performance when modelling the historical data – for the volume prediction and clustering – or assessing their robustness – for the allocation of ambulances.

For call volume forecasting it was used some traditional time series forecasting models such as ARIMA and ES, more specifically the HW, which are univariate. This was done to stablish a baseline for the impact of additional features when predicting calls. Both models have the

season_length and *horizon* have hyperparameters and since the focus is on hourly data, it was set to 24 and 1, respectively.

For regression, it was chosen the Vector Autoregression (VAR) model, which supports multivariate forecasting. The hyperparameter *lag_order* represents the amount of data needed prior to the data point in order to predict its value and is essential to understanding the seasonality and trends. It can be determined using domain knowledge about the features and how they impact the target variable. It was used the value using the Autoregression Analysis shown previously in Section 5.1.1, which was 23.

Because of the limited autocorrelation, other models capable of capturing complex patterns in the data like GRU, WaveNet, Prophet and Transformer. For all these architectures, hyperparameters were tuned to find the best configuration.

All the models were trained and validated using the complete dataset, except for the Transformer model due to hardware constraints. For this situation, the dataset was divided into smaller ones (yearly) and each sub-dataset was used to train a transformer model. The results from each model alongside relevant configurations are presented in Table 2.

Table 2 - Call volume prediction models performance using MSE and MAPE evaluation metrics.

Model	Hyperparameters	Evaluation Metrics	
		MSE	MAPE (%)
ARIMA	season_length = 24 horizon = 1	6.27	-
ES (HW)	season_length = 24 horizon = 1	9.12	-
VAR	lag_order = 22	7.62	80.85
GRU	First GRU Layer: - units = 64 - input_shape = number_of_features - return_sequences = True Second GRU Layer: - units = 32 - input_shape = from the 1st layer - return_sequences = False Dropout: 20% Output Layer: - units = 1 Optimizer: Adam	7.26	85.15
WaveNet	Input Layer: units = number_of_feature	0.04	66.48

Prophet	It takes a set of <i>regressors</i> which are the features.	5.22	60.36
Transformer	input_dim = number_of_features output_dim = 1 num_heads = 4 num_layers = 2 optimizer = Adam	1.26*	72.14*

As expected, traditional time series models, ARIMA, ES and VAR, performed poorly compared to the deep learning alternatives. WaveNet was the one that had the lowest MAE value, and therefore the model chosen to implement for the forecast component in EVA system.

Regarding spatial clustering of the emergency calls, it was selected K-Means and DBSCAN models. The Elbow Method was used to discover the optimal K value to be used as an hyperparameter for K-Means, which was 6 as shown in Figure 26.

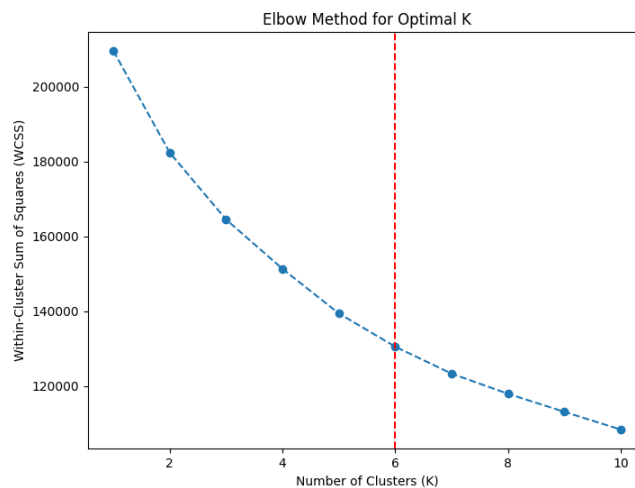


Figure 26 - Elbow Method to find the K value for the K-Means model.

As for DBSCAN, finding the best configuration for the hyperparameters ε (eps) and *minPts* is more complex in the current context, due to the unclear meaning of the ε value. Initially, was used the GridSearchCV to determine the optimal values combination for the model. After, these parameters could be tuned further to have clusters that are dense but not large since the main goal for these models is to find zones in a city where there is a greater concentration of demand for emergency medicine. Table 3 shows the performance of both models, alongside their configuration. In this table, was achieved satisfactory metric values, except when we tried to increase the number of clusters on the K-Means model.

Table 3 – Spatial Clustering performance between K-Means and DBSCAN with hyperparameter optimization.

Model	Hyper-Parameters	Evaluation Metrics			
		Clusters	Silhouette Score	Cakinski-Harabasz Score	Devies-Bouldin Score
K-Means	K = 5	5	0.0289	550.2886	5.7615
	K = 47	47	-0.0587	1925.9542	11.0123
DBSCAN	eps = 0.001 minPts = 5	333	0.2494	41.9474	1.7869
	eps = 0.0045 minPts = 22	47	0.2609	439.5925	1.9747

The Silhouette Score for K=47 is negative which indicates that some data points are assigned to the wrong clusters, due to many factors such as overlapping of clusters, which can be seen in Figure 27. On the other hand, when we tuned the DBSCAN to reduce the number of clusters, we see an improvement in the scores.

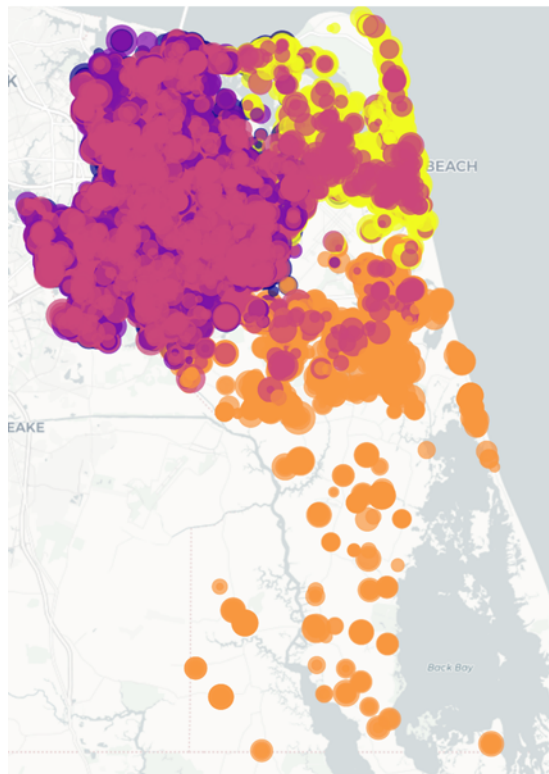


Figure 27 – K-Means visual result of the clusters when K takes the value of 47

Using all the data available since 2010, would allow it to capture the overall distribution of calls, which may not align with the current situation. This means that the density of the clusters can change over time, which is not visible if the models use the entire dataset. Figure 28 represents the differences between clusters modelled by DBSCAN with data from 2010 and 2017. Although

some clusters can be found in both years, others shifted, their density changed or even disappear.

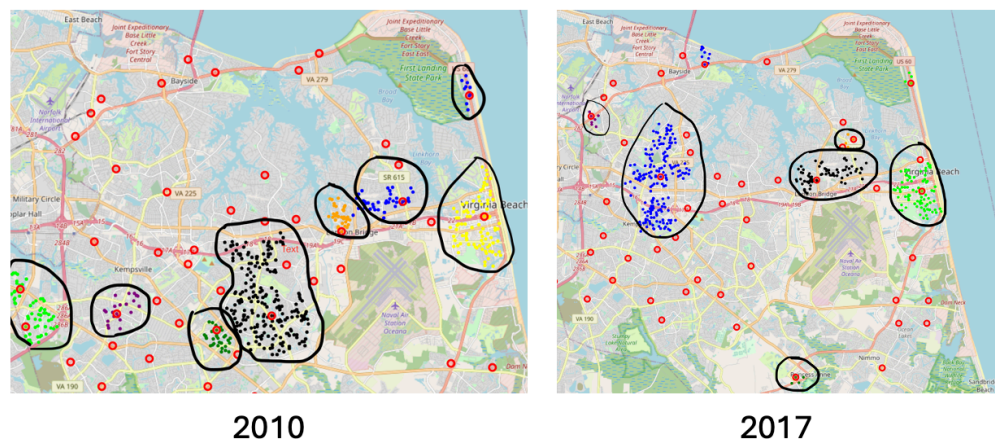


Figure 28 - Density clusters of emergency calls using different time windows.

Finally, the allocation models were developed with an extra goal, besides optimizing allocation, which was they had to be fast to be able to support the decision-making. Alongside the evaluation of the solutions provided by them, the time that takes to reach them was also assessed. The optimization models selected were the EVALP, EVALP-NM and a GA described in Section 2.4.

In order to evaluate these models, some simulations were created, which essentially are various combinations of the input variables. The simulation configurations and performance metrics are presented in *Simulation 1* is relatively simple with 3 zones, and 5 stations, and tries to allocate 10, 7, and 3 ambulances. *Simulation 2* increases the number of zones to 11, stations to 20, and tries to allocate 40, 34 and 20 ambulances.

Table 4. Let Z be the number of zones with their volume forecast, S the number of stations (each assigned to their respective zone) and A the number of ambulances as the input variables; A' the number of ambulances allocated by the optimization model and T the execution time of the model, as the performance metrics. It was decided to maintain the number of Z and S in the same simulation but change A in order to assess if the model does perform well in minimizing when there is a higher number than what is needed, and otherwise when the number is not enough for optimal solution. *Simulation 1* is relatively simple with 3 zones, and 5 stations, and tries to allocate 10, 7, and 3 ambulances. *Simulation 2* increases the number of zones to 11, stations to 20, and tries to allocate 40, 34 and 20 ambulances.

Table 4 - Simulation configurations and performance results for optimizing ambulance allocation.

Model	Simulation 1					Simulation 2				
	Input Variables			Performance		Input Variables			Performance	
	Z	S	A	A'	T	Z	S	A	A'	T

EVALP	3	5	10	5	0.072s	11	20	40	29	0.061s		
			7	5	0.068s			34	29	0.056s		
			3	3	0.056s			20	20	0.035s		
EVALP-NM			10	5 + 2	0.096s			40	29 + 5	0.085s		
7			5 + 2	0.053s	34			29 + 5	0.053s			
3			3 + 0	0.034s	20			20 + 0	0.050s			
GA			10	6	71.583s			40	40	7.482m		
7			5	4.594s	34			34	6.293m			
3			-	Inf	20			20	6.044m			

The LP-based optimization models prevailed against the GA model, which did not offer appealing execution times, which is a huge disadvantage, especially in a context where time is crucial for their usability. Also, in terms of solutions, it also failed to achieve optimal solutions since it often used all the available ambulances in the allocation. Regarding the two LP solutions, EVALP-NM has a slightly higher execution time, which is not seen as a downside because of the gain the neighbour mechanism provides to the quality of the solutions. As it can be seen, when for the cases of allocating 3 and 20 ambulances both models cannot find the optimal solutions, however, EVALP-NM prioritizes the allocation of the available units before adding extra to support the neighbours.

The visualization of the optimal solution modelled by EVALP-NM for *Simulation 1* is presented in Figure 29 where can be seen the effects of the NM on zones 2 and 3. They both have a predicted call volume of 1 call but for stations 2 and 4 it was added an additional unit to support Zone 1 which has the double predicted volume of calls.

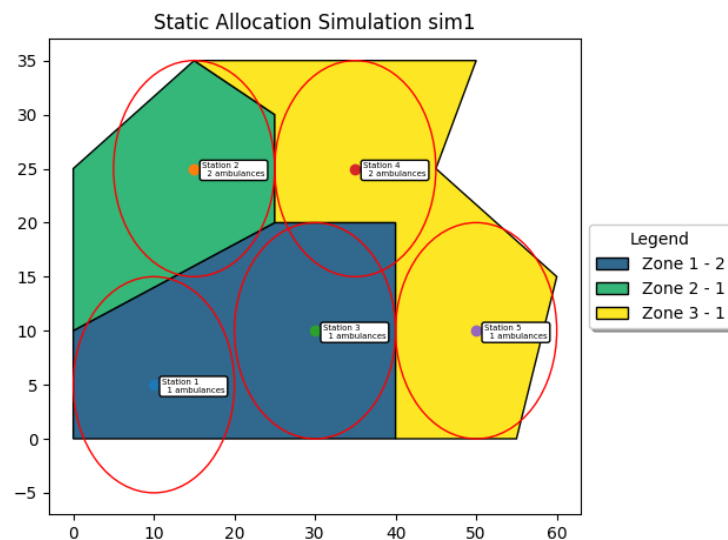


Figure 29 – Simulation 1 visual representation of the optimal solution ($A' = 7$) using EVALP-NM.

The same was done for *Simulation 2*, which outputs the visual representation shown in Figure 30, where NM can also be viewed, this time in zone 11 with stations 5 and 18 having an

additional unit being added. It is relevant to note that even though Zone 8 has a higher value, there is no station with a cover zone intercepting this zone.

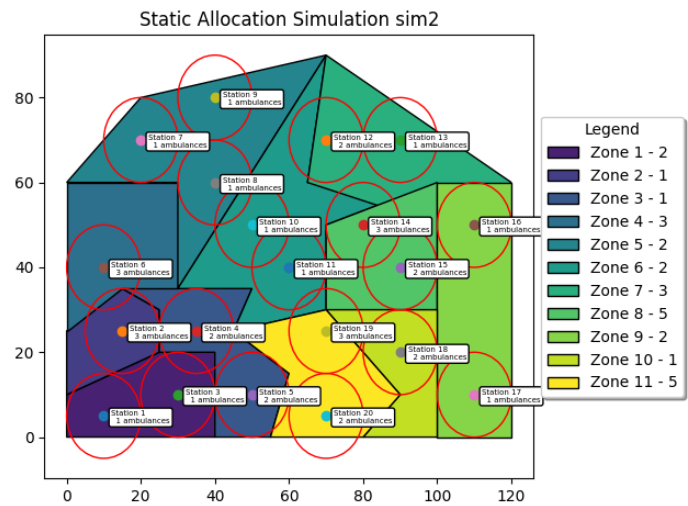


Figure 30 – Simulation 2 visual representation of optimal solution ($A' = 34$) using EVALP-NM.

6 Conclusions

This chapter presents the main conclusion drawn from this work, circling back to the objective and research questions, and identifying future work to improve the proposed solution.

6.1 Main Conclusion

As cities continue their demographic expansion, inefficient resource management can reduce the quality of life of the citizens, especially in the medical domain. Pre-hospital care solutions need to have demand information to make accurate and reliable decisions to impact ambulance distributions to reduce RT and prevent service overloading or underloading. To achieve this, EMS systems need to be embedded with state-of-the-art models with the modelling power to analyse large data sources of information, correlate many features and perform at a reasonable speed. On the other hand, optimizing ambulance's locations to serve the found needs, is an asset in maintaining the coverage of hot-spot areas while keeping enough redundancy for unpredictable scenarios.

The objectives introduced in Chapter 1 were accomplished during the course of this dissertation. O1 was addressed in Chapter 2 with the presentation of multiple architectures for forecasting demands and special allocation of vehicles. O2 was tackled in Chapter 3, by presenting a review of the literature for both topics and assessing the improvements needed on the presented solutions. O3 was answered in Chapter 4, by proposing a base model for an architecture that could sustain the use of forecasting and allocation models focused on the important aspect of this system: real-time. O3 and O4 were achieved in Chapter 5, with the presentation of the performance results of using these models in a real-case scenario.

With the completion of the defined objectives, it was possible to formulate the answers to the research questions also presented in Chapter 1, as follows:

- **RQ1** – What is the current state-of-the-art of EMS, especially in the context of forecasting emergency demands and allocating emergency vehicles?
 - Forecasting tools have been widely available for contexts related to business intelligence, but the same principles can be applied in emergency care since both try to model historical trends that could explain some future phenomenon. However, in emergency care can be difficult to consider models that depend on very rigid and defined trends, like TSF models, which leads to the search for a more complex solution that also analyses sequential data such as deep learning techniques.
 - Optimizing ambulance location has a hard dependency on knowing the demands so that the model can allocate emergency vehicles towards a goal. LP-based solutions seem to be the most effective and time-efficient.
- **RQ2** – Can traditional systems architectures support real-time decision-making platforms empowered by AI models in emergency medicine?
 - No, traditional system architectures based on REST or SOAP lack some important features required for a robust real-time system, such as low latency, lightweight communication channels, and scalability with multiple connections. An alternative approach lies in microservices architecture with event-based communication. These types of architectures have already been used for complex IoT systems.
- **RQ3** – Do traditional forecasting models perform well for short-term predicting of emergency demands?
 - No, TTS models require data to have strict, well-defined patterns that could explain past occurrences in order to use the same knowledge to predict future occurrences. In the Virginia Beach case, this cannot be observed in the datasets, reasons why it impacted the performance results of these models. However, it was also shown that deep learning models can model these complex patterns, even in short-term prediction windows.
- **RQ4** – Does spatial allocation reduce RT in emergency resource management?
 - While the optimization metric was not the RT itself, the fact that the proposed allocation models have a neighbour mechanism to increase the area of coverage by adding some redundancy, can put ambulances near the scene of the emergency, therefore reducing the time needed to get there. On the other hand, it is a minimization problem, which leaves some available units to be distributed if edge cases are noticed.

6.2 Future Work

Although the objectives of this master thesis were accomplished, the system's complexity opens the door for future expansions into the concepts that build this solution. One critical area of improvement is the implementation and demonstration of the architecture in a practical setting, where external factors (such as the arrival of emergency calls) could be simulated as

closely as in a real environment. Every component should work seemly together, but independently at the same time, bringing to the user the full picture of what is happening and what decisions can be made to improve resource management.

On a more focused revision, emergency demand forecasting models should turn the focus to deep-learning architectures like transformers, which have been getting impressive results in modelling sequential data. On the side of the allocation optimization, the model should be improved to consider allocating a single ambulance which has become available into an already defined distribution.

References

- Ahmed, S., Nielsen, I. E., Tripathi, A., Siddiqui, S., Ramachandran, R. P., & Rasool, G. (2023). Transformers in Time-Series Analysis: A Tutorial. *Circuits, Systems, and Signal Processing*. Retrieved from <https://doi.org/10.1007/s00034-023-02454-8>
- Alharbi, R. J., Lewis, V., & Miller, C. (2022). International Perspectives of Prehospital and Hospital Trauma Services: A Literature Review. *Trauma Care*, 2(3), 445–462. Retrieved from <https://doi.org/10.3390/traumacare2030037>
- Alibeiki, A., Motameni, H., & Mohamadi, H. (2021). A new genetic-based approach for solving coverage problem in directional sensor networks. *Journal of Parallel and Distributed Computing*, 154, 16–26. Retrieved from <https://doi.org/10.1016/j.jpdc.2021.03.006>
- Andersson, T., Petersson, S., & Varbrand, P. (2004). Calculating the preparedness for an efficient ambulance health care. In *Proceedings. The 7th International IEEE Conference on Intelligent Transportation Systems (IEEE Cat. No.04TH8749)* (pp. 785–790). IEEE. Retrieved from <https://doi.org/10.1109/ITSC.2004.1399002>
- Andersson, T., & Värbrand, P. (2016). Decision Support Tools for Ambulance Dispatch and Relocation. In *Operational Research for Emergency Planning in Healthcare: Volume 1* (pp. 36–51). London: Palgrave Macmillan UK. Retrieved from https://doi.org/10.1057/9781137535696_3
- Aringhieri, R., Bruni, M. E., Khodaparasti, S., & van Essen, J. T. (2017). Emergency medical services and beyond: Addressing new challenges through a wide literature review. *Computers & Operations Research*, 78, 349–368. Retrieved from <https://doi.org/10.1016/j.cor.2016.09.016>
- Artuc, E., Falcone, G., Port, G., & Rijkers, B. (2022). War-induced food price inflation imperils the poor. In *Global Economic Consequences of the War in Ukraine Sanctions, Supply Chains and Sustainability* (pp. 155–162). London: CEPR PRESS.
- Bélanger, V., Ruiz, A., & Soriano, P. (2019). Recent optimization models and trends in location, relocation, and dispatching of emergency medical vehicles. *European Journal of Operational Research*, 272(1), 1–23. Retrieved from <https://doi.org/10.1016/j.ejor.2018.02.055>
- Boyle, J., Jessup, M., Crilly, J., Green, D., Lind, J., Wallis, M., ... Fitzgerald, G. (2012). Predicting emergency department admissions. *Emergency Medicine Journal*, 29(5), 358–365. Retrieved from <https://doi.org/10.1136/emj.2010.103531>
- Brockwell, P. J., & Davis, R. A. (Eds.). (2002). Introduction. In *Introduction to Time Series and Forecasting* (pp. 1–44). New York, NY: Springer New York. Retrieved from https://doi.org/10.1007/0-387-21657-X_1

- Brotcorne, L., Laporte, G., & Semet, F. (2003). Ambulance location and relocation models. *European Journal of Operational Research*, 147(3), 451–463. Retrieved 16 January 2023 from [https://doi.org/10.1016/S0377-2217\(02\)00364-8](https://doi.org/10.1016/S0377-2217(02)00364-8)
- Bürger, A., Wnent, J., Bohn, A., Jantzen, T., Brenner, S., Lefering, R., ... Fischer, M. (2018). The Effect of Ambulance Response Time on Survival Following Out-of-Hospital Cardiac Arrest. *Deutsches Ärzteblatt International*. Retrieved from <https://doi.org/10.3238/arztebl.2018.0541>
- Byrne, J. P., Mann, N. C., Dai, M., Mason, S. A., Karanicolas, P., Rizoli, S., & Nathens, A. B. (2019). Association Between Emergency Medical Service Response Time and Motor Vehicle Crash Mortality in the United States. *JAMA Surgery*, 154(4), 286. Retrieved from <https://doi.org/10.1001/jamasurg.2018.5097>
- Carvalho, A. S., Captivo, M. E., & Marques, I. (2020). Integrating the ambulance dispatching and relocation problems to maximize system's preparedness. *European Journal of Operational Research*, 283(3), 1064–1080. Retrieved from <https://doi.org/10.1016/j.ejor.2019.11.056>
- Chen, L., Chen, S.-J., Chen, W.-K., Dai, Y.-H., Quan, T., & Chen, J. (2023). Efficient presolving methods for solving maximal covering and partial set covering location problems. *European Journal of Operational Research*, 311(1), 73–87. Retrieved from <https://doi.org/10.1016/j.ejor.2023.04.044>
- Conklin, A. I., Daoud, A., Shimkhada, R., & Ponce, N. A. (2019). The impact of rising food prices on obesity in women: a longitudinal analysis of 31 low-income and middle-income countries from 2000 to 2014. *International Journal of Obesity*, 43(4), 774–781. Retrieved from <https://doi.org/10.1038/s41366-018-0178-y>
- Dunkel, J., Fernández, A., Ortiz, R., & Ossowski, S. (2011). Event-driven architecture for decision support in traffic management systems. *Expert Systems with Applications*, 38(6), 6530–6539. Retrieved from <https://doi.org/10.1016/j.eswa.2010.11.087>
- Etu, E.-E., Monplaisir, L., Masoud, S., Arslanturk, S., Emakhu, J., Tenebe, I., ... Krupp, S. (2022). A Comparison of Univariate and Multivariate Forecasting Models Predicting Emergency Department Patient Arrivals during the COVID-19 Pandemic. *Healthcare*, 10(6), 1120. Retrieved from <https://doi.org/10.3390/healthcare10061120>
- Friedson, A. I. (2018). Income and Ambulance Response Time Inequality. *JAMA Network Open*, 1(7). Retrieved from <https://doi.org/10.1001/jamanetworkopen.2018.5201>
- Gendreau, M., Laporte, G., & Semet, F. (2001). A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Computing*, 27(12), 1641–1653. Retrieved from [https://doi.org/10.1016/S0167-8191\(01\)00103-X](https://doi.org/10.1016/S0167-8191(01)00103-X)
- Heemskerk, J. L., Domingo, R. A., Tawk, R. G., Vivas-Buitrago, T. G., Huang, J. F., Rogers, A., ... Freeman, W. D. (2021). Time Is Brain: Prehospital Emergency Medical Services Response

- Times for Suspected Stroke and Effects of Prehospital Interventions. *Mayo Clinic Proceedings*, 96(6), 1446–1457. Retrieved from <https://doi.org/10.1016/j.mayocp.2020.08.050>
- Holland, J. H. (1992). Genetic Algorithms. *Scientific American*, 267, 66–73.
- Holmén, J., Herlitz, J., Ricksten, S., Strömsöe, A., Hagberg, E., Axelsson, C., & Rawshani, A. (2020). Shortening Ambulance Response Time Increases Survival in Out-of-Hospital Cardiac Arrest. *Journal of the American Heart Association*, 9(21). Retrieved from <https://doi.org/10.1161/JAHA.120.017048>
- Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1), 5–10. Retrieved 5 February 2023 from <https://doi.org/10.1016/J.IJFORECAST.2003.09.015>
- Hsia, R. Y., Huang, D., Mann, N. C., Colwell, C., Mercer, M. P., Dai, M., & Niedzwiecki, M. J. (2018). A US National Study of the Association Between Income and Ambulance Response Time in Cardiac Arrest. *JAMA Network Open*, 1(7), e185202. Retrieved from <https://doi.org/10.1001/jamanetworkopen.2018.5202>
- Huang, H., Jiang, M., Ding, Z., & Zhou, M. (2019). Forecasting Emergency Calls With a Poisson Neural Network-Based Assemble Model. *IEEE Access*, 7, 18061–18069. Retrieved from <https://doi.org/10.1109/ACCESS.2019.2896887>
- Hyndman, R.J., & Athanasopoulos, G. (2021). *Forecasting: principles and practice* (3rd edition). Melbourne, Australia: OTexts.
- Hyndman, Rob J. (2020). A brief history of forecasting competitions. *International Journal of Forecasting*, 36(1), 7–14. Retrieved from <https://doi.org/10.1016/j.ijforecast.2019.03.015>
- Jacobsen, H.-A. (2009). Channel-Based Publish/Subscribe. In *Encyclopedia of Database Systems* (pp. 321–323). Boston, MA: Springer US. Retrieved from https://doi.org/10.1007/978-0-387-39940-9_1207
- Jaklič, T. K., & Kovač, J. (2015). The Impact of Demographic Changes on the Organization of Emergency Medical Services: the Case of Slovenia. *Organizacija*, 48(4), 247–258. Retrieved from <https://doi.org/10.1515/orga-2015-0021>
- Kaur, B. (2021). Maximum Geographical Coverage Deployment of Ambulances using Genetic Algorithm. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* (pp. 1493–1499). IEEE. Retrieved from <https://doi.org/10.1109/ICICV50876.2021.9388381>
- King, Z., Farrington, J., Utley, M., Kung, E., Elkhodair, S., Harris, S., ... Crowe, S. (2022). Machine learning for real-time aggregated prediction of hospital admission for emergency

- patients. *Npj Digital Medicine*, 5(1), 104. Retrieved from <https://doi.org/10.1038/s41746-022-00649-y>
- Kuddus, M. A., Tynan, E., & McBryde, E. (2020). Urbanization: a problem for the rich and the poor? *Public Health Reviews*, 41(1), 1. Retrieved from <https://doi.org/10.1186/s40985-019-0116-0>
- Kuok, K. K., Harun, S., & Chiu, P.-C. (2011). AUTO-CALIBRATION OF DAILY AND HOURLY TANK MODEL'S PARAMETERS USING GENETIC ALGORITHM. *Malaysian Journal of Civil Engineering*, 12–28.
- Louf, R., & Barthelemy, M. (2015). How congestion shapes cities: from mobility patterns to scaling. *Scientific Reports*, 4(1), 5561. Retrieved from <https://doi.org/10.1038/srep05561>
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4), 1346–1364. Retrieved from <https://doi.org/10.1016/j.ijforecast.2021.11.013>
- Martin, E., Hans-Peter, K., Jörg, S., & Xiaowei, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–231.
- Matinrad, N. (2022). *Models for Dispatch of Volunteers in Daily Emergency Response*. Retrieved from <https://doi.org/10.3384/9789179291853>
- Matteson, D. S., McLean, M. W., Woodard, D. B., & Henderson, S. G. (2011). Forecasting emergency medical service call arrival rates. *The Annals of Applied Statistics*, 5(2B). Retrieved from <https://doi.org/10.1214/10-AOAS442>
- Naoum-Sawaya, J., & Elhedhli, S. (2013). A stochastic optimization model for real-time ambulance redeployment. *Computers & Operations Research*, 40(8), 1972–1978. Retrieved 23 January 2023 from <https://doi.org/10.1016/J.COR.2013.02.006>
- Newman, S. (2015). *Building Microservices* (O'Reilly Media, Vol. 52). CA, USA.
- Ngiam, K. Y., & Khor, I. W. (2019). Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, 20(5), e262–e273. Retrieved from [https://doi.org/10.1016/S1470-2045\(19\)30149-4](https://doi.org/10.1016/S1470-2045(19)30149-4)
- Praveen, P., Babu, Ch. J., & Rama, B. (2016). Big data environment for geospatial data analysis. In *2016 International Conference on Communication and Electronics Systems (ICCES)* (pp. 1–6). IEEE. Retrieved from <https://doi.org/10.1109/CESYS.2016.7889816>
- Reuter-Oppermann, M., van den Berg, P. L., & Vile, J. L. (2017). Logistics for Emergency Medical Service systems. *Health Systems*, 6(3), 187–208. Retrieved from <https://doi.org/10.1057/s41306-017-0023-x>

- Rieke, M., Bigagli, L., Herle, S., Jirka, S., Kotsev, A., Liebig, T., ... Stasch, C. (2018). Geospatial IoT—The Need for Event-Driven Architectures in Contemporary Spatial Data Infrastructures. *ISPRS International Journal of Geo-Information*, 7(10), 385. Retrieved from <https://doi.org/10.3390/ijgi7100385>
- Ros, F., Riad, R., & Guillaume, S. (2023). PDBI: A partitioning Davies-Bouldin index for clustering evaluation. *Neurocomputing*, 528, 178–199. Retrieved from <https://doi.org/10.1016/j.neucom.2023.01.043>
- Shariff, S. S. R., Moin, N. H., & Omar, M. (2012). Location allocation modeling for healthcare facility planning in Malaysia. *Computers & Industrial Engineering*, 62(4), 1000–1010. Retrieved from <https://doi.org/10.1016/j.cie.2011.12.026>
- Shutaywi, M., & Kachouie, N. N. (2021). Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. *Entropy*, 23(6), 759. Retrieved from <https://doi.org/10.3390/e23060759>
- Steins, K., Matinrad, N., & Granberg, T. (2019). Forecasting the Demand for Emergency Medical Services. Retrieved from <https://doi.org/10.24251/HICSS.2019.225>
- Sudarshan, V. K., Brabrand, M., Range, T. M., & Wiil, U. K. (2021). Performance evaluation of Emergency Department patient arrivals forecasting models by including meteorological and calendar information: A comparative study. *Computers in Biology and Medicine*, 135, 104541. Retrieved from <https://doi.org/10.1016/j.combiomed.2021.104541>
- Taylor, S. J., & Letham, B. (2018). Forecasting at Scale. *The American Statistician*, 72(1), 37–45. Retrieved from <https://doi.org/10.1080/00031305.2017.1380080>
- Torre, L. de la, Chacon, J., Chaos, D., Dormido, S., & Sánchez, J. (2019). Using Server-Sent Events for Event-Based Control in Networked Control Systems. *IFAC-PapersOnLine*, 52(9), 260–265. Retrieved from <https://doi.org/10.1016/j.ifacol.2019.08.218>
- United Nations Human Settlements Programme. (2022). *Envisaging the Future of Cities*. Retrieved 3 December 2022 from https://unhabitat.org/sites/default/files/2022/06/wcr_2022.pdf
- V, S., A, V., & Pattar, S. (2022). MQTT based Secure Transport Layer Communication for Mutual Authentication in IoT Network. *Global Transitions Proceedings*, 3(1), 60–66. Retrieved from <https://doi.org/10.1016/j.gltp.2022.04.015>
- van Barneveld, T. C., Bhulai, S., & van der Mei, R. D. (2016). The effect of ambulance relocations on the performance of ambulance service providers. *European Journal of Operational Research*, 252(1), 257–269. Retrieved from <https://doi.org/10.1016/j.ejor.2015.12.022>
- van Barneveld, T., Jagtenberg, C., Bhulai, S., & van der Mei, R. (2018). Real-time ambulance relocation: Assessing real-time redeployment strategies for ambulance relocation. *Socio-*

- Economic Planning Sciences*, 62, 129–142. Retrieved from <https://doi.org/10.1016/j.seps.2017.11.001>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention Is All You Need.
- Vollmer, M. A. C., Glampson, B., Mellan, T., Mishra, S., Mercuri, L., Costello, C., ... Bhatt, S. (2021). A unified machine learning approach to time series forecasting applied to demand at emergency departments. *BMC Emergency Medicine*, 21(1), 9. Retrieved from <https://doi.org/10.1186/s12873-020-00395-y>
- Wang, S., Wu, Y. J., & Li, R. (2022). An Improved Genetic Algorithm for Location Allocation Problem with Grey Theory in Public Health Emergencies. *International Journal of Environmental Research and Public Health*, 19(15), 9752. Retrieved from <https://doi.org/10.3390/ijerph19159752>
- Wang, V., Salim, F., & Moskovits, P. (2013). The WebSocket Protocol. In *The Definitive Guide to HTML5 WebSocket* (pp. 33–60). Berkeley, CA: Apress. Retrieved from https://doi.org/10.1007/978-1-4302-4741-8_3
- Wang, X., & Xu, Y. (2019). An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. *IOP Conference Series: Materials Science and Engineering*, 569(5), 052024. Retrieved from <https://doi.org/10.1088/1757-899X/569/5/052024>
- Whitt, W., & Zhang, X. (2019). Forecasting arrivals and occupancy levels in an emergency department. *Operations Research for Health Care*, 21, 1–18. Retrieved 22 January 2023 from <https://doi.org/10.1016/J.ORHC.2019.01.002>
- Winters, P. R. (1960). Forecasting Sales by Exponentially Weighted Moving Averages. *Management Science*, 6(3), 324–342. Retrieved from <https://doi.org/10.1287/mnsc.6.3.324>
- Wong, H.-T., & Lai, P.-C. (2014). Weather factors in the short-term forecasting of daily ambulance calls. *International Journal of Biometeorology*, 58(5), 669–678. Retrieved from <https://doi.org/10.1007/s00484-013-0647-x>
- World Health Organization. Regional Office for Europe. (2022). *WHO European Regional Obesity : Report 2022*.
- Wu, Y., Xue, H., Wang, H., Su, C., Du, S., & Wang, Y. (2017). The impact of urbanization on the community food environment in China. *Asia Pacific Journal of Clinical Nutrition*, 26(3), 504–513. Retrieved from <https://doi.org/10.6133/apjcn.032016.09>
- Zaheeruddin, Gupta, H., & Mehrotra, D. (2021). Optimizing Ambulance Deployment using Genetic Algorithm. In *2021 9th International Conference on Reliability, Infocom*

Technologies and Optimization (Trends and Future Directions) (ICRITO) (pp. 1–5). IEEE.
Retrieved from <https://doi.org/10.1109/ICRITO51393.2021.9596352>

Zhao, Z., Anand, R., & Wang, M. (2019). Maximum Relevance and Minimum Redundancy Feature Selection Methods for a Marketing Machine Learning Platform. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 442–452). IEEE. Retrieved from <https://doi.org/10.1109/DSAA.2019.00059>

Appendices

Appendix A Data Analysis

Appendix A.1 Boxplots representing the monthly distribution of the Virginia Beach dataset between 2010 and 2018.

