Improving the forecast demand process for Emergency Medical Services: The case study of INEM

Nika Shahidian

Department of Engineering and Management, Instituto Superior Técnico, Universidade de Lisboa

December 2021

Abstract

Emergency Medical Services (EMS) are a vital component of pre-hospital medical care. These services are paramount to save lives as they focus on providing quality care and minimizing response times. The Portuguese EMS system, SIEM, is responsible for providing prompt and adequate medical care for mainland Portugal. Complex planning decisions must be made, which require accurate demand estimates to serve as inputs. This work aims to identify forecasting models exploring spatial-temporal datasets and contextual data to provide reliable information for decision-makers to support their decisions about resource allocation, and to meet the volume of calls answered at dispatch centres. The time-varying Gaussian mixture model is recognized in the literature as the most suitable option for operational forecasting. Together with other Machine Learning models, these predictive models are developed, optimized, and validated. Validation with SIEM's data shows that the Gradient Boosting model achieves, on average, 1.14% higher accuracy than the Gaussian mixture model. Furthermore, having surpassed all other models, boosting algorithms prove to be the most promising for similar applications.

Keywords: Emergency medical services; Operational planning; EMS demand forecasting; Ambulance service demand; Gaussian mixture models; Machine Learning.

1. Introduction

Emergency medical services (EMS) are complex systems that are designed to provide medical assistance to patients with serious injuries or illnesses, and are a vital component of pre-hospital medical care. Such systems play a key role in preserving lives, as patients benefit from these services from the moment they make a call to 112 until the moment they receive pre-hospital medical care and are transported to the hospital where they will receive the appropriate treatment according to their needs (Bélanger et al. 2019).

The Portuguese Integrated Emergency Medical System, *Sistema Integrado de Emergência Médica* (SIEM), is managed by the National Emergency Medical Institute, *Instituto Nacional de Emergência Médica* (INEM). Their primary goal is to provide adequate and prompt medical care to patients in mainland Portugal.

INEM is responsible for allocating scarce resources, which implies planning decisions that affect different planning levels. These decisions are, however, supported by intuition or simple averages at INEM due to the lack of decision support tools based on the state-of-the-art (Santos et al. 2019). Nevertheless, these decisions have an impact on the quality of the care that is provided, justifying the use of models addressed in the state-of-the-art to support decision-making.

These decision support tools usually require as input variables the volume of call arrivals and transportation needs. Based on that, improvement in the forecasting process is a basis to ensure efficient planning, and hence contribute to reducing response times.

The main goal of this work is to apply forecasting techniques to improve the planning of physical and human resources at INEM. The developed forecasting models should provide useful information for INEM to improve the management of its resources. The accuracy of these models contributes to matching demand and supply so that resources can be efficiently allocated both in time and location, and consequently, response times can be reduced. Secondary goals include characterizing the system and analysing historical data to provide managerial insights to INEM. Once the importance of accurate and reliable predictive models is recognized, the techniques that have been already applied in the literature are reviewed to identify potential gaps and improvement opportunities. In this context, forecasting models are developed and validated with the Portuguese case study, while comparing the performance of different techniques.

The remainder of this paper is organized as follows. Section 2 presents a literature review, including both explored and unexplored methods for the problem of EMS forecasting. Section 3 describes the problem and introduces the case study. Section 4 introduces the historical data, and presents an exploratory analysis. Section 5 presents the methodology of the work. Section 6 describes the elaboration process of the forecasting tool and shows the final results, including the limitations of the work. Section 7 proposes a decision support framework for future applications. Finally, Section 8 concludes the paper.

2. Literature Review

Accurate forecasts in fine time and space granularity provide information for EMS managers to make supported decisions at the operational level, and to assist with critical time-dependent decisions (Chen et al. 2016). In order to obtain these forecasts, data collected from EMS systems are used, which usually consist of a timestamp, occurrence location, priority level, vehicle(s) dispatched, etc. (Aringhieri et al. 2017). Patterns are identified based on this data since demand volumes vary throughout months, days of the week, and hours of the day. Even so, several attributes have been explored to explain the demand fluctuations in forecasting models, such as weather conditions, special events, and celebrations.

2.1 Explored Forecasting Methods

The literature presents three types of forecasting models to explain EMS demand, regression models, time series, and spatial-temporal models, and each model is explored with a variety of forecasting techniques (Steins et al. 2019).

2.1.1 Regression Models

Regression models are the first models addressed to forecast EMS demand, however, these models suffer from multicollinearity and difficulty in selecting relevant predictors (Steins et al. 2019). Despite this, they are still explored nowadays due to their simplicity and ease of application.

Aldrich et al. (1971) use thirty-two independent variables to apply a least squares regression forecasting model, while assuming a linear relationship between the dependent and the independent variables. The study indicates that aged people and single men generate more calls than the rest of the population. On the other hand, Siler (1975) adopts a nonlinear relationship and considers four socioeconomic variables to construct a multiple regression model.

The effect of population aging on pre-hospital EMS demand is considered in the regression model developed by McConnel & Wilson (1998). Svenson (2000) also identifies a dependency of EMS use rates with age, by adopting a Poisson multiple regression analysis. Wong & Lai (2010) use a multiple regression model to examine the weather effects on daily demand, by assessing the dependency of selected variables against weather factors. The impact of age on EMS demand rates is also identified by Lowthian et al. (2011) in a study to measure the impact of population growth and ageing in emergency ambulance services.

Recent regression models have aimed to incorporate the geographical area of EMS calls in addition to factors such as time, which is the case of Cramer et al. (2012). Recently, Steins et al. (2019) used a Zero Inflated Poisson regression model considering time as an independent factor, as well as socioeconomic and geographic factors.

2.1.2 Time Series Models

Since the 1980s, time series models such as autoregressive integrated moving average (ARIMA) and Holt-Winters methods have been explored to forecast call volumes and ambulance demand (Vile et al. 2016). Both methods are successful in overcoming many issues in regression techniques such as multicollinearity, autocorrelation, and the difficulty of selecting covariates (Vile et al. 2016).

Baker & Fitzpatrick (1986) adopt Winters' exponential smoothing model and use a multistep approach to determine the optimal parameters, while Channouf et al. (2007) recognize that EMS demand is influenced by when people work, commute, sleep, and celebrate. For daily volumes, they develop an autoregressive model and a doubly-season ARIMA model. Their results show that the autoregressive model's performance is superior and that the ARIMA model performs poorly when forecasting more than one week into the future. For hourly call volume rates, they consider a multinomial distribution conditional on the total daily call volume, and compare it to a time series model fit on data at the hourly level. Their results show that the conditional distribution approach generally worked better.

Contrasting the models proposed by Channouf et al. (2007) based on Gaussian linear time series, Matteson et al. (2011) assume that the hourly EMS call-arrival volume has a Poisson distribution. They combine an integer-valued time series models with a dynamic latent factor structure.

The non-parametric technique for time series analysis known as the Singular Spectrum Analysis (SSA) has been of growing interest due to its flexibility, as it is not dependent on parametric assumptions like linearity, stationarity, and normality (Al-Azzani et al. 2020). Vile et al. (2012) show that SSA produces superior long-term forecasts and comparable short-term forecasts to well-established methods.

Similar to Channouf et al. (2007), Ho Ting Wong & Lai (2014) also use ARIMA models to forecast daily demand. They show that by integrating weather factors such as temperature, the accuracy of daily EMS demand forecasts can be improved. ARIMA models have been modelled to incorporate seasonality, giving rise to SARIMA models, which have been explored by Gijo & Balakrishna (2016).

Recently, Ho Ting Wong & Lin (2020) focus on understanding the effects of weather to help EMS management, like Ho Ting Wong & Lai (2014) and Wong & Lai (2010) had previously done. They aggregate records in time series data according to patients' characteristics and then regress on meteorological data through multivariate forward regression. Also recently, Al-Azzani et al. (2020) compare the performance of four forecasting approaches, ARIMA, Holt-Winters, multiple regression, and SSA, on a selection of planning horizons (weekly, monthly, and 3-monthly). Their results show that ARIMA provides the most accurate forecasts for weekly and monthly predictions, and that long-term demand is best predicted by the SSA method.

2.1.3 Spatial-temporal Models

Contrary to time series and regression models, spatial-temporal models are capable of exploring both time and location. In addition to staffing and fleet size management, spatial-temporal demand estimates are critical to decisions such as the selection of station locations and for dynamic deployment planning (Zhou et al. 2015).

Setzler et al. (2009) use Artificial Neural Networks (ANN) to predict call volumes at fine spatial and temporal granularity. The model considers four temporal attributes: hour of the day, day of the week, month, and season. Their results show that ANN outperform the local adopted practice (moving average model) at low spatial granularity with marginal gains. However, significantly lower computational effort is associated to the three models presented in Zhou (2016): a time-varying Gaussian Mixture Model (GMM), a spatial-temporal Kernel Density Estimation, and a Kernel Warping method. They assume an independent non-homogeneous Poisson process, and consider spatial and temporal patterns such as location-specific seasonality, and daily and weekly seasonality.

Both the time-varying GMM applied by Zhou et al. (2015) and the Kernel Density Estimation model applied by Zhou & Matteson (2015) show higher statistical predictive accuracy than the current industry practice, with a comparable computational expense. While the Kernel Density Estimation model is easy to interpret and use by non-experts, the proposed GMM proves to be a more accurate method for fine time and location scales.

The time-varying GMM presented by Zhou et al. (2015) is further compared with the Kernel Warping method applied in Zhou & Matteson (2016) for Melbourne data. The complexity of this model lies in overcoming sparsity through smoothing, while capturing complex spatial-temporal patterns that require fine-resolution modelling. The results show that the Kernel Warping approach is slightly more adequate for highly complex spatial domains, although the accuracy improvement is not considered sufficient to justify the increased complexity of this model.

The Bayesian approach, applied by Nicoletta et al. (2017), allows the combination of available data with prior information, and then have those results be used as prior information once new data is available. The model and the parameters considered are, however, not sufficient to capture the complex spatial-temporal dynamics inherent to EMS demand. The methods addressed in Zhou (2016) represent a more accurate model of the data by overcoming data sparsity and representing complex spatial and temporal patterns through priors and weights.

2.2 Unexplored Forecasting Methods

ML algorithms have been used in the literature for diverse applications (Erickson et al. 2017; Yildiz et al. 2017), yet have not been directly applied to the problem of predicting EMS demand.

2.2.1 Machine Learning Models

ML algorithms are selected according to four learning approaches: supervised learning, unsupervised learning, semi-unsupervised learning, and reinforcement learning (Hafeez et al. 2021). The goal of supervised learning is to build a concise model capable of making predictions about future instances. To do so, it requires a training set with both inputs and outputs for each observation. Within this approach, problems can be modelled as regression or classification. Continuous values are predicted in regression problems, while a label or class is predicted in classification problems.

The following ML algorithms are mainly used in the context of supervised learning, for both regression and classification problems. Naïve Bayes and K-Nearest-Neighbour (KNN) have an explicit underlying probability model that gives a numerical probability as an output. When using these algorithms for classification problems, the model considers the probability of an instance belonging to each class. The Naïve Bayes algorithm assumes that each attribute is independent, while the KNN algorithm assigns the class or value of an unknown instance based on its nearest neighbours (Clark & Niblett 1989; Cover & Hart 1967). On the other hand, the Support Vector Machine (SVM) algorithm focuses on finding the optimal boundary between the training data (Boser et al. 1992).

The Random Forest algorithm is an ensemble method that builds multiple decision trees to diversify the use of the training data and build a generalized model. The logic behind the ensemble method is to build several models and combine those that perform best. The most popular are boosting and bagging (bootstrap aggregating). Boosting combines multiple weak learners into a single strong learner by sequentially training predictors, while bagging trains the models in parallel on different random subsets of the training dataset. Once the training stage is complete, the prediction for a new instance is made by aggregating the predictions of all the learners in the ensemble. The Random Forest algorithm trains models via bagging method, randomly selecting from two attributes to continue each decision tree (Ho 1995). The Extremely Randomized Trees (Extra Trees) algorithm is similar to Random Forest, only it adds additional randomization to each decision tree (Geurts et al. 2006). On the other hand, the Gradient Boosting algorithm is a boosting method that minimizes the loss function calculated for each observation, and creates decision trees to predict the errors (Breiman 1997). Adaptive Boosting (AdaBoost) is similar to Gradient Boosting except instead of working with full decision trees, it uses decision stumps which only have one node and two leaves (Freund & Schapire 1997). Finally, the Bootstrap Aggregating (Bagging) algorithm builds decision trees using different subsets from a dataset (Breiman 1996).

2.2.2 Performance Improvement

Training and validation are common processes in ML algorithms. The first consists of identifying patterns between the inputs and outputs from historical data. The second aims to ensure that the developed model is generalizable by using a test set to compare predictions with the real values. The validation process can be performed using cross-validation, which increases the reliability of the measured efficiency since it is based on resampling procedures (Raschka 2018).

Two problems can arise in the training process, known as overfitting and underfitting. Overfitting occurs when the model is not generalizable, i.e., it performs well on the training dataset but not on other sets of data. Contrarily, underfitting occurs when the model is not able to identify the patterns and underlying structure of the data (Géron 2019). For this purpose, learning curves are used to identify whether the training model is overfitting or underfitting the data.

A series of steps are followed to ensure that a generalizable model with the best accuracy is obtained. The dataset is prepared by removing attributes that provide redundant information to the model in order to reduce the input dimensionality. Hence, this increases the training and prediction speed, and the models become more practical since fewer inputs are required. Attribute selection methods for supervised learning can be separated into four groups: filter methods, wrapper methods, embedded methods, and hybrid methods (Jain & Singh 2018; Venkatesh & Anuradha 2019). Filter methods select attributes regardless of the learning algorithm by exploring statistical measures, while wrapper methods use a heuristic approach to consider possible subsets of attributes, and evaluate the model's performance with each one. In embedded methods, the attribute selection process is incorporated in the training of the learning algorithm, and hybrid methods combine several approaches to take advantage of the benefits of different methods (Guerra-Manzanares et al. 2019; Venkatesh & Anuradha 2019).

Once a relevant dataset is obtained, the model's hyperparameter values are fine tuned to check for accuracy improvement possibilities. The three most popular hyperparameter tuning methods are Grid Search, Random Search (RS), and Genetic Algorithm (GA) (Liashchynskyi & Liashchynskyi 2019). Grid Search tests every possible combination of values in a pre-defined set, while RS is more efficient by testing random combinations (Bergstra & Bengio 2012). On the other hand, GA is an evolutionary search algorithm that sequentially selects, combines, and varies hyperparameters in a manner that simulates the process of natural selection (Liashchynskyi & Liashchynskyi 2019).

3. Case Study

EMS systems, such as INEM, are complex structures built to provide medical assistance as fast as possible to patients with serious injuries or illnesses. The benefit provided is a unique and vital component of the health care system, as it provides primary care and serves as a bridge to hospital care.

3.1 EMS Planning

To minimize response times and ensure fair service and efficient use of resources, EMS require thorough planning. Planning is, however, complex and challenging due to the variability in terms of volume, location, and priority of calls (Ingolfsson 2013).

There are three different planning levels defined, each referring to different planning horizons: strategic, tactical, and operational. On the strategic level decisions are made for several years, on the tactical level decisions refer to periods of one month to one year, and on the operation level decisions are made on a daily basis or in real time (Reuter-Oppermann et al. 2017). At each planning level different EMS planning decisions are made, such as the following (Bélanger et al. 2019):

- Strategic level: location of ambulance stations, fleet dimensioning, staff hiring;
- Tactical level: location of ambulances' standby sites, staff scheduling and crew pairing, fleet management strategies;
- Operational level: ambulance location and relocation, ambulance dispatching, assignment of calls to resources.

These planning decisions require accurate inputs, such as correct demand estimations, to ensure that there are sufficient resources available at the right time and place.

Forecasting Levels

Planning problems on all three levels require forecasting as an input. Forecasting can also be divided into the same three levels depending on which decision level it supports (Reuter-Oppermann et al. 2017).

Operational forecasts must be accurate as they are the input for critical operational planning decisions that may have a direct impact on the survivability of the victim if not optimized. Failure in accurately estimating demand can result in inefficient resource allocation, and preventable time inefficiencies.

Short-term forecasts provide real time decision support for dynamic ambulance deployment and hourly operational deployment plans, while large time period forecasting is useful for strategic planning and budgeting (Reuter-Oppermann et al. 2017).

3.2 SIEM

SIEM serves as an extension of the Emergency Departments of National Health Service hospitals in mainland Portugal. It is responsible for the intervention process from the moment a call is placed until the patient is transferred to an appropriate health unit. SIEM's intervention process consists of six stages (INEM 2013):

- Detection: emergency situation detected by civilian and 112 calls;
- Alert: screening, triage, priority level assignment and vehicle dispatch by Pre-hospital Emergency Technician (TEPH);
- Pre-aid: guidance and assistance to a caller to perform first-aid basic care if necessary;
- Initial aid in the accident's location: after vehicle arrival, stabilization of the victim and initial treatment;
- Transport and care during transit: transportation to the appropriate health unit and in-transit treatment;
- Transfer and treatment in health unit: transfer of the victim to receiving health unit to finalize treatment.

Resources

To achieve the objectives of the organization, INEM has financial, technological, logistic, and human resources (INEM 2018). The human resources of INEM are mainly Pre-hospital Emergency Technicians (TEPHs), nurses, and doctors. The logistic resources include all emergency medical vehicles at INEM's disposal. In 2018, the INEM fleet was made up of 658 emergency medical vehicles distributed throughout mainland Portugal, and an additional 62 seasonal reinforcement vehicles (INEM 2018).

There are a total of ten types of vehicles available to INEM. The most commonly dispatched vehicles owned and managed by INEM are Medical Emergency Ambulances (AEM), Vehicles of Medical Emergency and Reanimation (VMER), and Immediate Life Support Vehicles (SIV). Each one of these vehicles serves a different purpose:

- AEM vehicles are basic life support emergency vehicles that require a crew of two TEPH;
- VMER vehicles are advanced life support vehicles with advanced medical equipment, that require a doctor and a nurse;
- SIV vehicles are differentiated ambulances with immediate life support equipment, that require a nurse and a TEPH.

Call Triage

Urgent Patient Dispatching Centres (CODUs) ensure, for mainland Portugal, daily and continuous emergency medical call reception, forwarded through the European Emergency Number 112. The calls are answered by TEPHs, who are supported by a team of medical doctors and psychologists. They evaluate, through a system of triage algorithms and in the shortest possible time, the received aid requests to determine the necessary and adequate resources for each case (INEM 2018).

There has been a significant increase in emergency calls placed throughout the years. In 2018, 1,393,594 emergency calls were answered by CODU, representing an increase of 16% since 2013, and 1.9% from 2017. This rise in demand has been justified by the ageing population and increase in chronic diseases (INEM 2018).

Once a call is answered by CODU, a TEPH begins the triage process to determine the severity of the incident and the appropriate vehicles required to dispatch. Since 2012, this triage process has been done through the use of the Telephonic Triage and Counselling System (TETRICOSY), which allows for standardization of procedures and high efficiency. The software, developed by INEM, assigns a priority level for the incident according to the information provided by the call operator (INEM 2018). Although there is a total of nine priority levels (P1-P9), an emphasis is placed on those with the most frequency:

- Priority 1 (P1): critical life-threatening incidents, originating the dispatch of several advanced life support emergency medical vehicles;
- Priority 3 (P3): urgent situation and dispatching of basic life support emergency medical vehicles;
- Priority 5 (P5): non-urgent situation where the triage results in no vehicle dispatching and the call is transferred to the appropriate health support line;
- Other priorities: other situations that require differentiated assistance.

Throughout the years, the percentage of P3 calls has been increasing significantly, and they hold a large majority of the total number of emergency calls (around 70-75%). The second most common medical emergencies are of P1 priority (around 10-15%), followed by P5 calls (around 7-12%) (Santos et al. 2019). The remaining priorities represent less than 13% of the total volume of calls.

3.3 Problem Definition

In order to tackle the problem of estimating call volumes and emergency vehicles demand, addressed in the literature as Forecasting of EMS Demand and Ambulance Service Demand, this work aims to present a forecasting model validated with real data shared by INEM. Both types of forecasting are directly related to INEM's operational planning problems. Currently, the demand forecasts are based on averages from the historical ratio of calls. An updated decision support tool based on state-of-the-art methodologies will certainly contribute to more effective and efficient planning.

4. Exploratory Data Analysis

Historical data from 2017 to 2018 shared by INEM, together with a large number of attributes, are presented in two datasets. The data is aggregated in discrete time and spatial intervals. The first dataset contains hourly volumes of calls of priorities P1 and P3 answered per each of the eighteen districts in mainland Portugal. The second dataset has the number of dispatches of vehicles SIV, VMER, and AEM per 8-hour shift from each of the twenty bases in the municipality of Lisbon. Other than the differences in time intervals and spatial origin of demand, the remaining attributes are equal for both datasets.

4.1 Attributes

The sixty-five attributes available in the datasets are grouped into nine categories: *weather*, *special-event*, *resident population*, *age*, *gender*, *employment*, *seasonal patterns*, *accident/crime*, and *occurrence type*. While most of the attributes are obtained from sources such as the National Institute of Statistics, *occurrence type* contains historical information registered by INEM in the form of thirty-seven types of medical occurrences. The majority of these occurrences suffered a small growth in 2018.

The relationship between attributes is measured through a coefficient of correlation, which indicates the strength of the statistical association between two attributes. Pearson's correlation coefficient is selected to evaluate linear relationships due to its wide application in the literature (Liu et al. 2020; Rastegari et al. 2019). Although with some exceptions, high correlation is found between attributes within the same category. Between categories, little to no correlation is observed with categories *weather*, *special-event*, and *employment*. A positive correlation is found between most other categories, except for *age* that presents negative correlation with the others.

Out of the thirty-seven attributes of category *occurrence type*, ten of them can be considered as rare since they have low daily occurrence rates, resulting in a lack of correlation with other attributes due to their sporadic behaviour.

4.2 Target Variables

There are a total of five target variables for the problem at hand: P1 and P3 call volumes available in the call dataset; and SIV, VMER, and AEM vehicle dispatches from the vehicle dataset.

P1 calls typically result in the dispatch of two out of the three vehicles under analysis, while P3 calls usually only result in the dispatch of one life support emergency medical vehicle, such as AEM or SIV.

Calls of priority P3 are significantly more frequent than P1, P3 having on average 5.24x higher demand, and this gap increased in 2018. Nonetheless, both priority calls had a significant demand growth from 2017 to 2018, with 3.9% and 5.43% increases for P1 and P3 calls, respectively. The highest demand volumes are observed in the most populated districts, Lisbon, Porto, and Setúbal. On the other hand, Faro, Beja, and Portalegre present the greatest volumes per resident population.

AEM vehicles are significantly more issued than VMER and SIV vehicles. In 2017, the proportion of dispatches was the following: 4.77% SIV, 16.54% VMER, and 78.69% AEM. The next year, SIV and VMER vehicles represented a slightly bigger portion of dispatches, leaving AEM vehicles with 77.29% of dispatches. Although it is not significant, this could indicate a trend towards a more balanced distribution.

Dispatches of vehicles SIV, VMER, and AEM in the municipality of Lisbon are spatially aggregated according to the base from which the vehicle left. From 2017 to 2018, demand only increased for SIV vehicles at 5.72%, whilst VMER demand decreased by 0.59%

and AEM decreased by 7.07%. In the two years under analysis, the majority of SIV dispatches originated from two bases: b0 (77.79%) and b4 (15.58%). Similarly, the majority of dispatches involving VMER vehicles were concentrated from three bases: b9 (29.06%), b10 (35.68%), and b11 (34.09%). AEM dispatches are the most varied, although ten out of twenty bases issued less than 1.10% of all AEM vehicles. The vast number of zero dispatches per shift is noteworthy, with no dispatches of SIV, VMER, and AEM vehicles in 94.07%, 85.63%, and 61.45% of observations, respectively.

Both call and vehicle demand are dependent on the time of day, day of week, month, and season. Regarding the time of day, a distinct demand pattern is identified with the lowest volumes during the night, peak in the morning, and slow decrease throughout the day. Standard deviation is also lower during the night. Although not as pronounced, weekdays have higher demand comparatively to weekends. Also, Monday is the day of the week with the highest demand and Sunday typically has the lowest. Regarding yearly patterns, the beginning and end of the year are the periods with the highest demand, and demand decreases sequentially throughout the middle of the year. There is a slight increase in call volumes in August, although it is not replicated in vehicle demand. The standard deviation of August is also significantly higher than the remaining months, which is a probable consequence of the intense heat that frequently impacts Portugal during this month. These yearly patterns may also be related to the effects of the season. Higher call and vehicle demand volumes are observed during autumn and winter, summer being the season with the lowest demand and winter with the highest.

5. Methodology

The workflow identified for the development of predictive models based on ML algorithms is used as the computational methodology of this work (Al-Janabi et al. 2017; Stetco et al. 2019; Were et al. 2015). The data was collected, pre-processed, and treated for outliers prior to this work, so these procedures are not performed.

Following the initial exploratory data analysis, presented in the previous section, the prediction approach (regression or classification) is defined to determine the need for label designation and to select the appropriate evaluation metrics. For each ML algorithm that is applied, the subset of attributes is selected through a suitable method, and the model is trained and tested via cross-validation using default hyperparameters. After all the results are obtained, the models with the best performance are selected and subjected to hyperparameter tuning procedures to maximize accuracy. The final models are compared with one another through appropriate measures, conclusions about their performance are obtained, and a final model capable of producing accurate predictions is chosen.

6. Experimental Results

With the goal of elaborating a forecasting tool directed towards operational planning, the GMM and eight other ML algorithms are selected and tested to identify the best performing model. These eight ML algorithms are the following: Naïve Bayes, KNN, SVM, Random Forest, Extra Trees, Gradient Boosting, AdaBoost, and Bagging. The developed model is expected to be an improvement on the existing prediction methods currently in use by INEM. However, this improvement is not measurable because the planning decisions are currently guided by intuition.

The experiments are conducted on a laptop with a 2.90GHz Intel Core i7-7500U processor and 8.00GB of RAM, with two cores, running on Windows 10.

6.1 Data Preparation

A separate model is developed for each target demand volume that is predicted, i.e., for each type of vehicle (SIV, VMER, and AEM) and each priority level call (P1 and P3). The highest available granularity level is used since the forecasts are directed towards short-term operational planning. This means that call demand is predicted on an hourly level and vehicle demand for 8-hour shifts. The twenty bases in Lisbon represent the spatial location of the vehicle demand and call demand is grouped by the eighteen districts in mainland Portugal.

The two datasets are subjected to a normalization procedure to obtain uniform data due to the extreme differences in the range of values of the attributes. The homogenization of the data facilitates the training process by ensuring a lower computational cost and increasing the ability of the model to rapidly converge. The normalization process is done via a min-max scaling method, rescaling the dataset so the values of each attribute are in the same [0,1] range.

A classification prediction approach is selected for this application, and four metrics are selected to evaluate the models: Receiver Operating Characteristics (ROC) Area Under the Curve (AUC), accuracy, Precision, and Recall.

This approach requires the transformation of the continuous outputs of the target variables into classes. These classes are defined through unsupervised learning, via the K-means clustering algorithm. An optimal number of two clusters is obtained for all target variables, resulting in binary classification problems. It is important to note that the obtained classes demonstrate imbalance in terms of the number of observations in each class. For all the target variables, the number of instances in cluster 0 is significantly higher (>75%) than cluster 1.

6.2 GMM Application

Two methods are explored for the selection of the most relevant subset of attributes, a filter and a wrapper method. The selected filter method, Pearson’s correlation coefficient, is used for a correlation analysis (CA). The attributes with a correlation higher than 0.4 or lower than -0.4 with the target variable are selected. After this, attributes with correlation higher than 0.8 or lower than -0.8 with each other are removed, remaining only the one with the highest correlation with the target variable (Guerra-Manzanares et al. 2019; Pallonetto et al. 2019; Rastegari et al. 2019). GA is selected as the wrapper and it is applied over 5 generations with a crossover probability of 0.5, mutation of 0.2, and with a population size of 15. The algorithm converges within the 5 generations since all the individuals in the final population present the same solution.

The hyperparameters of the GMM are tuned via two methods, RS and GA, both with the goal of maximizing accuracy. For RS, the parameters are sampled uniformly although not all parameter values are tried out, but rather 10 parameter settings are sampled from the specified listed values. The GA is run through 5 generations with a population size of 15, crossover probability of 0.9 and mutation of 0.03.

Stratified cross-validation with 5 folds is used to preserve the percentage of samples in each class. Four GMM models are obtained at this stage: two are built on the CA dataset, having the hyperparameters of one been tuned with RS and the other with GA; the other two models are constructed on the GA dataset, with the same distinction in tuning processes. Regarding attribute selection, CA overall provides better results than GA. Additionally, the CA method is significantly faster computationally, further supporting the choice to use this practice. Different preferences are shown in regard to hyperparameter tuning methods depending on the evaluation metrics, with AUC and accuracy recommending RS and GA, respectively.

6.3 ML Application

Similar to the GMM application, both CA and GA are used as attribute selection methods in the ML algorithms. The CA results in exactly the same selection of attributes since it is independent of the learning algorithm. This is not the case for the GA method, which must be repeated for each of the eight algorithms.

For this application, the hyperparameter tuning process is only conducted after the evaluation of the ML models with default hyperparameters to identify the ones that best model the data. Preliminary results of accuracy and AUC obtained from a stratified 5-fold cross-validation identify algorithms Gradient Boosting and AdaBoost as the ones that result in the best performing models. The third best performing algorithm is SVM, however, the lengthy computational time of this algorithm does not justify its further exploration for operation planning.

An analysis of the AUC and accuracy metrics shows that Gradient Boosting performs best with the attributes selected through GA, and AdaBoost performs best with the CA dataset. The main difference between these two datasets is the number of attributes that are selected for each. For this data, the CA method selects significantly fewer attributes, 44.8 on average, while the GA method selects on average 83.3 attributes. Therefore, in this case, Gradient Boosting performs best with additional attributes while AdaBoost prefers fewer attributes.

Identically to the GMM application, hyperparameter tuning is performed both via RS and GA. Performance improvements for both models are more significant with RS.

6.4 Results Analysis

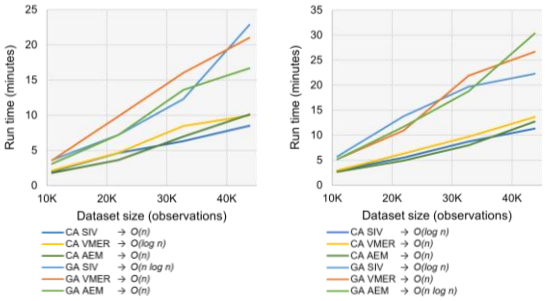
The models are studied through analyses of learning curves, ROC curves, Precision-Recall curves, and confusion matrices, to obtain a deeper understanding of the performance of each one.

Learning curves measure the quality of the model's fit to the data and provide a view of its generalization ability. In general, the models show adequate behaviour, reaching a point of stability with a small gap between the training and the validation curves. There are two situations where the learning curves demonstrate ill-fitting models, however, they are corrected through small adjustments of hyperparameter values to fix underfitting and overfitting problems.

ROC curves give a visual understanding of each model's capability to distinguish class 1 from class 0, as they plot the true positive rate (Recall) versus the false positive rate for all possible cut-off values. The same logic is applied for the Precision-Recall curves, although they provide a view directed towards understanding each model's ability to predict the minority class (class 1). Overall, both ROC curves and Precision-Recall curves are significantly worse for the two GMM models when compared to the two ML boosting models, which is especially true for P1 and P3 models.

Computational Complexity

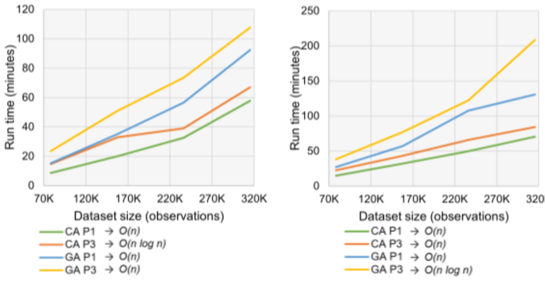
Although previous studies in the literature show that ML algorithms typically follow a O(log n) curve (Hafeez et al. 2021), this analysis aims to understand the complexity of the models when combined with additional particularities such as attribute selection. The analysis is performed by evaluating the time required to train models Gradient Boosting and AdaBoost with RS tuning performed for 10 iterations, and with a 5-fold stratified cross-validation. A performance comparison is shown between using attribute selection methods CA and GA. Figure 1 shows the computational complexity for vehicle models.



(a) Emergency vehicles – Gradient Boosting (b) Emergency vehicles – AdaBoost

Figure 1 – Computational complexity of emergency vehicle models.

While the vehicle dataset consists of 43,800 observations, the call dataset is significantly larger, with 315,360 observations. Despite the differences in data sizes, the computational complexity of the call models is similar, as is shown in Figure 2.



(a) Emergency calls – Gradient Boosting (b) Emergency calls – AdaBoost

Figure 2 – Computational complexity of emergency call models.

The models with GA attributes have longer run times than CA due to the additional cost that this selection process represents. Although not all curves represent the expected O(log n), the results do not show high model complexity. Compared to CA, the curves of GA models are further from the expected, which suggests that the algorithm follows a different complexity behaviour. Overall run time is superior for AdaBoost, which represents a significant advantage for Gradient Boosting models in comparison.

Model Limitations

The GA used for attribute selection converged within 5 generations since the individuals in the final population were identical. This same number of generations was used for the hyperparameter tuning processes, although convergence is not obtained for all of the models. The best course of action to avoid local optimal solutions is to increase the considerably low number of generations. This is limited by the computational time of the process, as well as the need to run multiple models for each algorithm explored in this work. RS proved to be a good alternative. The GA with 5 generations took on average 2.3x longer than the RS with 10 iterations. Adding this to the fact that superior results were obtained for both Gradient Boosting and AdaBoost models, the RS is recommended as an overall better hyperparameter tuning process.

Despite the overall high accuracy of the models, issues related to data imbalance present a great challenge. Recall is poor (lower than 0.5) for P1 models regardless of the algorithm, meaning that the majority of positives (class 1) are predicted incorrectly. Low Recall is especially bad considering the nature of the service, where incorrect low demand predictions can lead to unpreparedness and failure of EMS resources. Since Precision-Recall is a trade-off, the nature of the service may favour sacrificing Precision to obtain higher Recall values. However, Recall is bad regardless of the algorithm that is used, suggesting that this problem is related to the imbalance of the data. The fundamental structure of the data implies a lack of representation due to a large number of observations in the original data with 0 demand. This deficit is related to the set granularity levels which, if reduced, would help overcome data sparsity.

Final Remarks

The final results of the two boosting models, summarized in Table 1, show that Gradient Boosting presents higher values for most evaluation metrics. Furthermore, based on all of the analyses that were performed, Gradient Boosting has the overall best performance out of the evaluated models, including lower computational complexity. For this reason, the recommended model is Gradient Boosting trained with attributes selected via GA and with hyperparameters tuned through RS. Although the recommended GA Gradient Boosting model is appropriate for operational planning, CA models may be preferred if the main priority goal is computational cost minimization. Nonetheless, RS is identified as a better option for hyperparameter tuning when compared to GA in regard to both model performance and computational efficiency.

Table 1 – Results from stratified 5-fold cross-validation of the boosting models.

		Gradient Boosting	AdaBoost
SIV	AUC	**0.9923**	0.9846
	Accuracy	**0.9834**	0.9774
VMER	AUC	0.9910	**0.9913**
	Accuracy	**0.9673**	0.9672
AEM	AUC	**0.9737**	0.9569
	Accuracy	**0.9151**	0.8933
P1	AUC	**0.8607**	0.8595
	Accuracy	**0.8436**	0.8425
P3	AUC	**0.9837**	0.9827
	Accuracy	**0.9616**	0.9599

Overall, the boosting models explored in this work outperformed other ML models for this set of highly sparse data. This includes the GMM, which performed similarly to other ML models and was ultimately surpassed.

7. Decision Support Framework

Data analytics is commonly split into four sections: descriptive, diagnostic, predictive, and prescriptive analytics (Mustafee et al. 2018). The aim of descriptive analytics is to identify and summarize what happened using data visualization and key performance indicators, to assess performance and compare it against targets. Sequentially, diagnostic analytics follows the information obtained and intends to identify why something happened. Then, the predictive analytics level focuses on identifying what is likely to happen by developing estimates of outcomes based on planned inputs. While descriptive, diagnostic, and predictive analytics are information focused, prescriptive is decision focused. The aim is to obtain prescriptions of specific actions that lead to the desired outcome, with an emphasis on the concrete decision problem. A solution path is suggested by prescribing one or more courses of action and informing on the likely outcome of each one. The success of prescriptive analytics is mainly dependent on the assessment of the alternatives generated from the prediction phase and the impact they have on performance (Mustafee et al. 2018).

Considering prescriptive analytics in the context of EMS planning, each of the identified decision problems requires a concrete plan to determine the most suitable actions.

Predictive Analytics

Predictors such as time interval, date, and location are common throughout any model. Other attributes can be added, although they should be carefully selected as they could add useless information and increase computational time. Available attributes should be submitted to an attribute selection process to determine the most relevant ones. This study explores the contrasts between CA and GA, and the conclusions obtained from this comparison can be extended to other methods. Similar to what has been observed in the literature, wrapper methods typically provide better results while the computational burden of filter methods is insignificant with small accuracy deterioration (Venkatesh & Anuradha 2019).

Different parameter values should be explored to achieve the most accurate model, and RS proved to be a superior method to GA in this work. The selection of the number of iterations for a RS is dependent on the number of hyperparameters and values to test, as well as the available time to search for an efficient combination. If run time is not an important factor, for instance in strategic planning, RS with a large number of iterations should be run. While RS is appropriate for large search spaces due to the trade-off between run time and quality of solution, Grid Search can be used for small search spaces since it explores all possible combinations.

Regarding forecasting methods, time series models have mostly surpassed regression models due to their superior performance and ease of use. The literature has explored ARIMA and Holt-Winters methods to predict EMS demand, as well as a non-parametric technique, SSA. While ARIMA models provide accurate forecasts for tactical level planning, SSA has been successful in producing accurate long-term forecasts (Al-Azzani et al. 2020; Vile et al. 2012). Also, seasonality has been recently incorporated in ARIMA models, making SARIMA models an improved option for tactical forecasting (Gijo & Balakrishna 2016).

Spatial-temporal models are improvements on time series since they can model both time and location. ANN achieves high accuracy, but it is most appropriate for tactical and strategic level forecasts due to the associated computational cost and large volumes of data required (Setzler et al. 2009). The results of Zhou (2016) show that Kernel Warping is highly accurate, closely followed by GMM. The complexity of the Kernel Warping is its main limitation, while GMM is easily applicable with only a slight decrease in accuracy. ML models present a promising opportunity since they have achieved great results in other fields. From the experiments made in this work, boosting models demonstrate great performance, and are recommended for similar applications.

Prescriptive Analytics

After relevant descriptive and diagnostic analyses, predictive analytics should be applied following the presented guidelines. These predictions are fed to prescriptive analytics tools such as simulation, resulting in a clear solution path. This process represents the framework that should be followed to solve a decision problem, and is represented in Figure 2. The input data is used at all stages of this framework and the ultimate results, actionable recommendations, are used to assist the original decision problem.

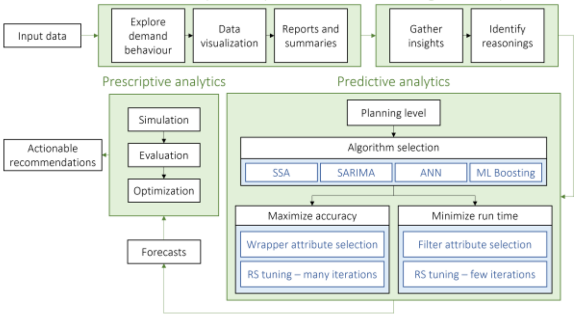


Figure 2 – Guideline for descriptive, diagnostic, predictive, and prescriptive analytics.

The value of the predictions obtained in the predictive analytics stage derives from its utility in decision-making. Depending on the decision problem, forecasts referent to different time horizons are constructed and obtained in different run times.

8. Conclusions and Future Work

EMS systems are complex structures vital to preserving human lives and delivering fast and effective health care to the population. Demand for EMS has been increasing largely due to population ageing and growth, resulting in a need to evaluate and study these systems. Reliable forecasting tools capable of supporting planning decisions are required to optimize resource allocation and improve effectiveness. Short-term demand forecasts are a vital input for operational planning, and accurate estimates obtained in low computational times are needed for detailed planning on a daily, hourly, and real time basis. Currently, INEM uses simple averaging techniques to obtain future demand provisions, which does not meet with the forecasting techniques addressed in the state-of-the-art.

For the problem of operational planning, an extensive search of models applied in the literature for EMS demand forecasting allowed the identification of the time-varying GMM as the most promising option. Further research showed that common ML algorithms, frequently applied for predictive modelling in a variety of other problems, had not yet been directly applied to predict EMS demand. Recognizing this, a wide selection of ML models is chosen to train and validate on real data shared by INEM, and allow a comparison with the selected GMM. The models are explored to incorporate additional attributes in order to explain demand fluctuations. In addition to training and testing predictive models on INEM's data, optimization methods such as attribute selection and hyperparameter tuning processes are explored.

An in-depth analysis of INEM's historical data from 2017-2018 identifies demand patterns similar to those already recognized in the literature. These include relationships between demand and times when people sleep and go to work, as well as the month of the year likely related to vacation periods. The provided datasets are used to train and validate the models and experiment multiple improvement opportunities. The attribute selection processes identify both CA and GA as appropriate methods, the former having significantly lower computational time. Alternative hyperparameter tuning procedures are investigated and the highest improvements are obtained with the RS method in low computational times. The results obtained from a stratified 5-fold cross-validation recognize Gradient Boosting as the best model out of those that were explored in this work, closely followed by AdaBoost. The GMM model achieved results similar to those of other ML models, although it was surpassed by both of the explored boosting models.

The limitations of this work include the limited utility of the predictions due to the use of unsupervised learning for class definition. The exploration of other methods is recommended, validated by the decision-maker, in order to produce more relevant and overall better predictions.

Future work should attempt to further explore boosting ML algorithms considering the problem as regression. Although not incorporated in this work, location-specific and temporal seasonality represent an interesting addition for future applications of ML.

References

Al-Azzani, M. A. K., Davari, S., & England, T. J. (2020). An empirical investigation of forecasting methods for ambulance calls - a case study. *Health Systems*, *00*(00), 1–18. https://doi.org/10.1080/20476965.2020.1783190

Al-Janabi, M., De Quincey, E., & Andras, P. (2017). Using supervised machine learning algorithms to detect suspicious URLs in online social networks. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017*, 1104–1111. https://doi.org/10.1145/3110025.3116201

Aldrich, C. A., Hisserich, J. C., & Lave, L. B. (1971). An analysis of the demand for emergency ambulance service in an urban area. *American Journal of Public Health*, *61*(11), 2158–2161. https://doi.org/10.2105/AJPH.61.11.2158

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 and Operations Research*, *78*(July 2015), 349–368. https://doi.org/10.1016/j.cor.2016.09.016

Baker, J. R., & Fitzpatrick, K. E. (1986). Determination of an optimal forecast model for ambulance demand using goal programming. *Journal of the Operational Research Society*, *37*(11), 1047–1059. https://doi.org/10.1057/jors.1986.182

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. https://doi.org/10.1016/j.ejor.2018.02.055

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*, 281–305.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140. https://doi.org/10.3390/risks8030083

Breiman, L. (1997). *ARCING THE EDGE Leo Breiman Technical Report 486 , Statistics Department University of California, Berkeley CA. 94720*. *4*, 1–14. https://pdfs.semanticscholar.org/65b7/b1a0d61fd012f10cfce642d4aa4dec9a5829.pdf

Channouf, N., L'Ecuyer, P., Ingolfsson, A., & Avramidis, A. N. (2007). The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta. *Health Care Management Science*, *10*(1), 25–45. https://doi.org/10.1007/s10729-006-9006-3

Chen, A. Y., Lu, T. Y., Ma, M. H. M., & Sun, W. Z. (2016). Demand Forecast Using Data Analytics for the Preallocation of Ambulances. *IEEE Journal of Biomedical and Health Informatics*, *20*(4), 1178–1187. https://doi.org/10.1109/JBHI.2015.2443799

Clark, P., & Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning*, *3*(4), 261–283. https://doi.org/10.1007/bf00116835

Cover, T. M., & Hart, P. E. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Cramer, D., Brown, A. A., & Hu, G. (2012). Predicting 911 calls using spatial analysis. *Studies in Computational Intelligence*, *377*(January 2011), 15–26. https://doi.org/10.1007/978-3-642-23202-2-2

Erickson, B. J., Korfiatis, P., Akkus, Z., & Kline, T. L. (2017). Machine learning for medical imaging. *Radiographics*, *37*(2), 505–515. https://doi.org/10.1148/rg.2017160130

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn , Keras & TensorFlow*.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Gijo, E. V., & Balakrishna, N. (2016). SARIMA models for forecasting call volume in emergency services. *International Journal of Business Excellence*, *10*(4), 545–561. https://doi.org/10.1504/IJBEX.2016.079252

Guerra-Manzanares, A., Bahsi, H., & Nomm, S. (2019). Hybrid feature selection models for machine learning based botnet detection in IoT networks. *Proceedings - 2019 International Conference on Cyberworlds, CW 2019*, 324–327. https://doi.org/10.1109/CW.2019.00059

Hafeez, M. A., Rashid, M., Tariq, H., Abideen, Z. U., Alotaibi, S. S., & Sinky, M. H. (2021). Performance improvement of decision tree: A robust classifier using tabu search algorithm. *Applied Sciences (Switzerland)*, *11*(15). https://doi.org/10.3390/app11156728

Ho, T. K. (1995). Random Decision Forests Tin Kam Ho Perceptron training. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 278–282.

INEM. (2013). *Sistema Integrado de Emergência Médica*. https://www.inem.pt/wp-content/uploads/2017/06/Sistema-Integrado-de-Emergência-Médica.pdf

INEM. (2018). Relatório Anual. In *Relatório anual de atividades e contas*. https://www.inem.pt/wp-content/uploads/2019/07/Relatório-Anual-Atividades-e-Contas-de-2018-2.pdf

Ingolfsson, A. (2013). EMS planning and management. *International Series in Operations Research and Management Science*, *190*, 105–128. https://doi.org/10.1007/978-1-4614-6507-2_6

Jain, D., & Singh, V. (2018). Feature selection and classification systems for chronic disease prediction: A review. *Egyptian Informatics Journal*, *19*(3), 179–189. https://doi.org/10.1016/j.eij.2018.03.002

Liashchynskyi, P., & Liashchynskyi, P. (2019). Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS. *ArXiv*, *2017*, 1–11.

Liu, Y., Mu, Y., Chen, K., Li, Y., & Guo, J. (2020). Daily Activity Feature Selection in Smart Homes Based on Pearson Correlation Coefficient. *Neural Processing Letters*, *51*(2), 1771–1787. https://doi.org/10.1007/s11063-019-10185-8

Lowthian, J. A., Jolley, D. J., Curtis, A. J., Currell, A., Cameron, P. A., Stoelwinder, J. U., & McNeil, J. J. (2011). The challenges of population ageing: Accelerating demand for emergency ambulance services by older patients, 1995-2015. *Medical Journal of Australia*, *194*(11), 574–578. https://doi.org/10.5694/j.1326-5377.2011.tb03107.x

Matteson, D. S., McLean, M. W., Woodard, D. B., & Henderson, S. G. (2011). Forecasting emergency medical service call arrival rates. *Annals of Applied Statistics*, *5*(2 B), 1379–1406. https://doi.org/10.1214/10-AOAS442

McConnel, C. E., & Wilson, R. W. (1998). The demand for prehospital emergency services in an aging society. *Social Science and Medicine*, *46*(8), 1027–1031. https://doi.org/10.1016/S0277-9536(97)10029-6

Mustafee, N., Powell, J. H., & Harper, A. (2018). RH-RT: A data analytics framework for reducing wait time at emergency departments and centres for urgent care. *Angewandte Chemie International Edition, 6(11), 951–952.*, *2017*(July 2017).

Nicoletta, V., Lanzarone, E., Guglielmi, A., Bélanger, V., & Ruiz, A. (2017). Nicoletta, V., Lanzarone, E., Guglielmi, A., Belanger, V. (2017). A Bayesian Model for Describing and Predicting the Stochastic Demand of Emergency Calls.pdf. *Springer Proceedings in Mathematics & Statistics*, *194*(Bayesian Statistics in Action), 203–212. https://doi.org/https://doi.org/10.1007/978-3-319-54084-9_19

Pallonetto, F., De Rosa, M., Milano, F., & Finn, D. P. (2019). Demand response algorithms for smart-grid ready residential buildings using machine learning models. *Applied Energy*, *239*(October 2018), 1265–1282. https://doi.org/10.1016/j.apenergy.2019.02.020

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *ArXiv*.

Rastegari, E., Azizian, S., & Ali, H. (2019). Machine Learning and Similarity Network Approaches to Support Automatic Classification of Parkinson's Diseases Using Accelerometer-based Gait Analysis. *Proceedings of the 52nd Hawaii International Conference on System Sciences*, *6*, 4231–4242. https://doi.org/10.24251/hicss.2019.511

Reuter-Oppermann, M., van den Berg, P. L., & Vile, J. L. (2017). Logistics for Emergency Medical Service systems. *Health Systems*, *6*(3), 187–208. https://doi.org/10.1057/s41306-017-0023-x

Santos, G., Marques, I., & Barbosa-Póvoa, A. (2019). *Improving Emergency Medical Services Through Vehicle Location Optimization*. *December*.

Setzler, H., Saydam, C., & Park, S. (2009). EMS call volume predictions: A comparative study. *Computers and Operations Research*, *36*(6), 1843–1851. https://doi.org/10.1016/j.cor.2008.05.010

Siler, K. F. (1975). Predicting demand for publicly dispatched ambulances in a metropolitan area. *BMC Health Services Research*, *10*(3), 254–263.

Steins, K., Matinrad, N., & Grandberg, T. A. (2019). *Forecasting the Demand for Emergency Medical Services*. https://doi.org/10.24251/HICSS.2019.225

Stetco, A., Dinmohammadi, F., Zhao, X., Robu, V., Flynn, D., Barnes, M., Keane, J., & Nenadic, G. (2019). Machine learning methods for wind turbine condition monitoring: A review. *Renewable Energy*, *133*, 620–635. https://doi.org/10.1016/j.renene.2018.10.047

Svenson, J. E. (2000). Patterns of use of emergency medical transport: A population-based study. *American Journal of Emergency Medicine*, *18*(2), 130–134. https://doi.org/10.1016/S0735-6757(00)90002-0

Venkatesh, B., & Anuradha, J. (2019). A review of Feature Selection and its methods. *Cybernetics and Information Technologies*, *19*(1), 3–26. https://doi.org/10.2478/CAIT-2019-0001

Vile, J. L., Gillard, J. W., Harper, P. R., & Knight, V. A. (2012). Predicting ambulance demand using singular spectrum analysis. *Journal of the Operational Research Society*, *63*(11), 1556–1565. https://doi.org/10.1057/jors.2011.160

Vile, J. L., Gillard, J. W., Harper, P. R., & Knight, V. A. (2016). Time-dependent stochastic methods for managing and scheduling Emergency Medical Services. *Operations Research for Health Care*, *8*, 42–52. https://doi.org/10.1016/j.orhc.2015.07.002

Were, K., Bui, D. T., Dick, Ø. B., & Singh, B. R. (2015). A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecological Indicators*, *52*, 394–403. https://doi.org/10.1016/j.ecolind.2014.12.028

Wong, H. T., & Lai, P. C. (2010). Weather inference and daily demand for emergency ambulance services. *Emergency Medicine Journal*, *29*(1), 60–64. https://doi.org/10.1136/emj.2010.096701

Wong, Ho Ting, & Lai, P. C. (2014). Weather factors in the short-term forecasting of daily ambulance calls. *International Journal of Biometeorology*, *58*(5), 669–678. https://doi.org/10.1007/s00484-013-0647-x

Wong, Ho Ting, & Lin, J. J. (2020). The effects of weather on daily emergency ambulance service demand in Taipei: a comparison with Hong Kong. *Theoretical and Applied Climatology*, *141*(1–2), 321–330. https://doi.org/10.1007/s00704-020-03213-4

Yildiz, B., Bilbao, J. I., & Sproul, A. B. (2017). A review and analysis of regression and machine learning models on commercial building electricity load forecasting. *Renewable and Sustainable Energy Reviews*, *73*(March 2016), 1104–1122. https://doi.org/10.1016/j.rser.2017.02.023

Zhou, Z. (2016). *Predicting Ambulance Demand: Challenges and Methods*. 11–15. http://arxiv.org/abs/1606.05363

Zhou, Z., & Matteson, D. S. (2015). Predicting ambulance demand: A spatio-temporal kernel approach. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *2015-Augus*, 2297–2303. https://doi.org/10.1145/2783258.2788570

Zhou, Z., & Matteson, D. S. (2016). Predicting melbourne ambulance demand using kernel warping. *Annals of Applied Statistics*, *10*(4), 1977–1996. https://doi.org/10.1214/16-AOAS961

Zhou, Z., Matteson, D. S., Woodard, D. B., Henderson, S. G., & Micheas, A. C. (2015). A Spatio-Temporal Point Process Model for Ambulance Demand. *Journal of the American Statistical Association*, *110*(509), 6–15. https://doi.org/10.1080/01621459.2014.941466