Article

A Data Mining Approach for Health Transport Demand

Jorge Blanco Prieto [1,*] , Marina Ferreras González [1] , Steven Van Vaerenbergh [2] and
Oscar Jesús Cosido Cobos [3]

1   UPintelligence S.L., 33011 Oviedo, Spain
2   Department of Mathematics, Statistics and Computation, University of Cantabria, 39005 Santander, Spain
3   Area of Computer Science and Artificial Intelligence, Department of Computer Science, University of Oviedo, 33007 Oviedo, Spain
*   Correspondence: jorge.blanco@upintelligence.es

Abstract: Efficient planning and management of health transport services are crucial for improving accessibility and enhancing the quality of healthcare. This study focuses on the choice of determinant variables in the prediction of health transport demand using data mining and analysis techniques. Specifically, health transport services data from Asturias, spanning a seven-year period, are analyzed with the aim of developing accurate predictive models. The problem at hand requires the handling of large volumes of data and multiple predictor variables, leading to challenges in computational cost and interpretation of the results. Therefore, data mining techniques are applied to identify the most relevant variables in the design of predictive models. This approach allows for reducing the computational cost without sacrificing prediction accuracy. The findings of this study underscore that the selection of significant variables is essential for optimizing medical transport resources and improving the planning of emergency services. With the most relevant variables identified, a balance between prediction accuracy and computational efficiency is achieved. As a result, improved service management is observed to lead to increased accessibility to health services and better resource planning.

Keywords: data mining; ambulance response performance; variable importance measure; ambulance demand prediction; exploratory data analysis

1. Introduction

E-health, a term coined to describe the integration of digital technologies into healthcare, represents a pivotal transformation in the delivery and management of health services [1]. It encompasses a broad spectrum of applications, from electronic health records (EHRs) to telemedicine and mobile health (mHealth) solutions [2]. The core objective of e-health is to leverage information and communication technologies to improve healthcare access, efficiency, and quality while empowering patients to take charge of their well-being through digital tools and platforms.

Within the realm of healthcare services, the concept of e-health plays a fundamental role in optimizing various facets, including the provision of healthcare transportation services. By harnessing digital advancements, such as real-time tracking systems, telehealth consultations, and data-driven logistics, e-health can significantly enhance the efficiency and effectiveness of healthcare transportation [3]. Whether it involves ambulance services, non-emergency patient transfers, or medical courier deliveries, the incorporation of e-health technologies aims to streamline operations, reduce response times, and ensure prompt and tailored care for patients during transit.

It is clear that in the healthcare system, medical transport plays a fundamental role in ensuring both care and the efficient planning and management of services. While its role in emergency situations stands out, it is equally necessary in the transfer of patients requiring specialized care. Achieving a correct operation is possible thanks to different

factors, including an adequate infrastructure of personnel and resources, as well as effective coordination between the agents involved.

Determining the precise number of personnel and resources required can be complicated, leading to occasional coordination lapses and two undesirable scenarios. On the one hand, under-allocating services can saturate the system, jeopardizing guaranteed healthcare and degrading care quality. On the other hand, if resources are overestimated, the system will incur unnecessary expenditures that will have an impact on the budgets allocated to other types of services. As a result, countries such as the United States have proposed solutions to address this cost overrun [4].

The analyzed transport system, recently updated and rich in data, positions ambulance services as ideal candidates for predictive model applications. Reference [5] indicates that time series analysis offers potent short-term forecasts of future ambulance service needs. Interestingly, simple models in this context may outperform complex and costly ones, emphasizing the importance of focusing on predicting service volume rather than solely relying on average run times and patient acuities.

Another critical area within healthcare, especially in organ transplantation, faces challenges in organ allocation. As per a different study [6], designing an optimal and efficient organ allocation approach is crucial to balance organ supply and demand, preventing the loss of patients awaiting suitable organs. Recent literature points to a gap in considering simultaneous medical and logistical factors in organ allocation strategies.

In response to these concerns, a number of studies [7–9] are looking at new models that can lead to better planning in the medical transport fleet. The first step to obtaining these accurate models is to know which are the determining variables in this prediction, which will allow simpler models to be much more powerful. Data mining techniques make it possible to recognize and analyze these characteristics in order to provide an answer to the questions posed.

Hence, integrating data mining techniques into ambulance service planning, alongside insights from organ allocation studies, could offer a holistic framework for optimizing resource allocation, predicting service needs, and enhancing the overall efficiency of healthcare transport systems. The primary aim of this research lies in leveraging statistical and data mining methodologies to identify and discern the most pivotal variables crucial for accurately predicting this demand.

This paper is divided into several sections. It commences with an analysis of existing studies to date on predicting healthcare transportation demand, distinguishing between the characterization of variables, the utilization of time series techniques, fleet management, and spatio-temporal prediction. Following this, a data mining section is presented, elucidating the underlying mathematical theory behind the applied techniques. This is followed by an examination of available data, along with their contextualization, and the exploration of external data sources that could be cross-referenced and correlated with the existing dataset. The results obtained from these correlations and the decisions derived from them are provided. Finally, the outcomes of implementing the techniques introduced in the data mining section are presented, accompanied by an interpretation of these results, culminating in a conclusion drawn from the conducted research.

2. Related Work

2.1. Health Transport Demand Prediction

Over the past few years, several research studies have been conducted with the aim of developing predictive models that can accurately estimate the demand for medical assistance in the ambulance setting [10,11]. These studies have used different methodological approaches and predictor variables to identify the factors that influence demand and provide a solid basis for the development of more accurate and reliable models.

One of the key issues addressed in the scientific literature is the consideration of temporal patterns of demand. Studies [12,13] have investigated the variability of demand throughout the day, week, or year, identifying specific demand patterns at different time

points. These analyses reveal the existence of peak demand at certain times of the day or days of the week, as well as seasonal variations that may be influenced by external factors.

In addition to temporal patterns, geographic variables have been explored as another influential factor in the demand for ambulance care [14,15]. Studies have analyzed the influence of geographic location, highlighting that areas with a higher population density or with specific geographic characteristics, such as mountainous or rural areas, may have a higher demand for ambulance services. Furthermore, geographic variability in ambulance use is large and is associated with variations in the health status and socioeconomic situation of patients [16].

The relationship between demographic variables and the demand for ambulance services has also been investigated [17]. Previous studies have examined the impact of demographic variables, such as population, age, and gender, on the demand for ambulance services. These demographic factors may be associated with certain types of medical emergencies, which implies that their consideration is essential to develop accurate predictive models.

In terms of the methodologies used, the scientific literature has covered a wide spectrum of approaches. They have ranged from traditional statistical models, such as linear regression [18] or time series [19], to more advanced techniques, such as neural networks [20], support vector machines, and machine learning algorithms [21]. These approaches have proven effective in predicting care demand, with promising results in terms of accuracy and generalizability.

2.2. Predicting Ambulance Demand for Care Using Time-Series Techniques

In [22], the importance of using time series prediction techniques for adequate health planning is highlighted. The cited study employs the Holt-Winters exponential smoothing model, a time series prediction technique, to detect seasonality patterns and demand evolution, allowing quality predictions in the short term. This method has advantages such as straightforward interpretation and implementation, as well as high reliability, being surpassed only by procedures that require a more complex and detailed comparative analysis. Therefore, this approach is recommended for routine use. Several adjustment measures are evaluated, such as RMSE, MAE, or MAPE (the definitions of these metrics are provided in Section 5), achieving short-term predictions with a MAPE of 5.9% at one week and 10.4% at three weeks.

In [17], a modified clustering genetic algorithm is applied to compare optimal ambulance locations, predict future ambulance locations, and determine the required number of vehicles. The study predicts variations in care demand, in this case for emergencies, by reassigning the location of ambulances to other nearby ones. This reduces the average response time by 57 s. The importance of the age variable in considering the number of services is highlighted, using demographic predictions to infer future cases of long-term emergency health services.

A similar approach is explored in [23], wherein neural networks incorporating predefined trajectories are used to predict the location of future demand for ambulance services. Leveraging these forecasts obtained, ambulances are relocated before actual emergencies occur.

These studies demonstrate the value of time series techniques, genetic algorithms, and neural networks in forecasting ambulance demand for care, allowing for efficient resource allocation and improved response times in emergency situations.

2.3. Fleet Management

Ambulance fleet management has been the subject of most of the studies reviewed in the literature, which have focused mainly on vehicle location and relocation, often sidelining the predictions regarding the number of services. A review of modeling approaches used in ambulance fleet management was conducted in [24]. This examination considered factors such as objectives, coverage and location constraints, number of ambulances, and geographic region. Among the most commonly used techniques are branch and bound,

branch and cut, heuristic methods, tabu search, the ant colony algorithm, and genetic algorithms. In addition, Bayesian approaches are noted to have been proposed for predicting the number of emergency calls in each area.

The Bayesian approach has proven to be an important feature in health management applications as it allows the combining of available data with prior information in a sound theoretical framework. In this way, subsequent inference can be used as prior information when new data become available, as reported in [25]. For example, a Bayesian model has been used to estimate the distribution of ambulance travel times in different road segments of a city [26], as well as to predict the demand of patients attended by the home care service [27] and emergency calls [28].

2.4. Spatio-Temporal Prediction

Another strand of research has centered on both temporal and spatio-temporal prediction of medical transport demand. To forecast the spatio-temporal demand for ambulances in Toronto, Ref. [29] considered weekly seasonality, daily seasonality, and short-term serial dependence during some specific hours. Notably, they addressed the seasonality of the area without considering the exact routes. Prior to applying machine learning (ML) algorithms in their work, a prediction was formed using an averaging formula over a spatial region of 1 km² for a duration of one hour. Previous research has used different methods to predict aggregate ambulance demand as a temporal process, including autoregressive moving averages [30], factor models considering hourly and daily seasonality [31], spectral analysis [19], and grid-based neural networks in discrete time and space [32]. In the case of [29], a discrete-time and continuous-space Gaussian model was adopted to predict emergency call volumes.

More recent research has used decision models, such as a hybrid decision tree using a naive Bayes classifier, to predict ambulance offload delay [33]. In this case, the predictor variables used in the decision tree include the day of the week, time of day, call volume, free ambulance rate, and total number of ambulances. Figure 1 shows the significance of these variables obtained after applying the hybrid decision tree with a naive Bayes classifier [33].
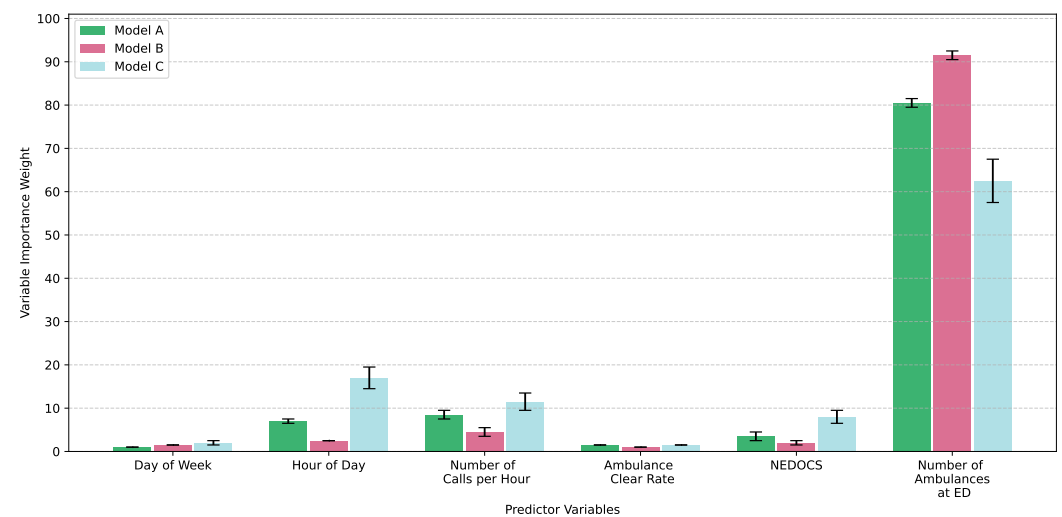

Figure 1. Importance of variables obtained after applying a hybrid decision tree with a naive Bayes classifier to three different models. It is observed that, in all models, the most important variable is the number of ambulances at ED, followed by the hour of the day or the number of calls per hour, and the values provided by the National Emergency Department Overcrowding Scale (NEDOCS). Less important are the day of the week and the ambulance clear rate. Figure redrawn from [33].

However, this is not the first time that ML techniques have been used to estimate ambulance demand. In [18], these techniques were employed to quantify the characteristics that influence demand. Among these characteristics, age plays an important role, as a

region is likely to experience higher demand due to a larger elderly population. For this purpose, actual patient data sets and demographic data from the past 10 years were used. Other relevant factors include the day of the week and month, as demand tends to have a periodic pattern, as well as short- and long-term historical demands in each region (e.g., outbreaks, sporting events). The variables considered in this study were classified into spatial (region ID), temporal (day of the week, day of the month, day of the year), demographic (number of people over 50 years old in a specific year), short-term historical demands (7 variables corresponding to the demands of the last 7 days in that region), and long-term historical aggregate demand (total demand for the last 30 days, last 7 days, week to sample date, month to sample date).

In addition, socioeconomic variables have been added in this research to analyze whether the demand for ambulances is related to the socioeconomic characteristics of the inhabitants of a region. For this purpose, methods such as regional moving average, linear regression, support vector regression (SVR), multilayer perceptron (MLP), radial basis function neural network (RBFN), and light gradient boosting machine (LightGBM) were used. After comparing the different methods, it was concluded that the best solution was LightGBM (regression tree). The most important characteristics were the ID of the region in which the demand was predicted, the demand of the previous 30 days, and the demand of the previous 7 days. The number of people over 50 years of age in the region was also considered important. Table 1 shows a comparative table of the accuracies of the different models used in ambulance demand prediction [18].

Table 1. Comparative table of accuracies for the different models used in ambulance demand prediction. Table from [18].

Method	WAPE (%)	MAE	MSE
Regional Moving Average	25.8	2.20	11.2
Linear Regression	**24.5**	**2.09**	**10.1**
MLP	24.6	2.10	**10.1**
RBFN	25.1	2.14	10.8
SVR	25.2	2.15	11.2
LightGBM	24.5	2.09	10.2

Bold indicates the best results for each column. WAPE: weighted percentage error; MAE: mean absolute error; MSE: mean squared error; MLP: multiplayer perceptron; RBFN: radial basis function network; SVR: support vector regression; LightGBM: light gradient boosting machine.

3. Data Mining

Data mining [34] refers to the process of extracting interesting new knowledge from large data sets. Importantly, before building data models for the prediction of the demand for healthcare transportation, various preprocessing techniques are used to determine the key variables to be used in constructing them. In this study, descriptive mining techniques are initially applied in order to explore the data in depth and determine the key variables in the prediction of the demand for healthcare transportation.

3.1. Association Rules

Association rules are data mining techniques used to discover frequent patterns between variables or elements in large data sets, as in our case. These associations are used a posteriori in decision-making, behavior prediction, or product recommendation.

According to [34], an association rule is defined as an implication of the form $A \implies B$, where $A \subset I$, $B \subset I$, $A \neq \varnothing$, $B \neq \varnothing$, and $A \cap B = \varnothing$. This rule is satisfied on the set of total transactions, D, with a support s, which represents the percentage of transactions in D containing $A \cup B$ and is computed as the probability of $A \cup B$, $P(A \cup B)$. In addition, it has a confidence c on the set of total transactions, D, where c is the percentage of transactions in D that contain A and also contain B, and is calculated as the conditional probability $P(B|A)$.

Another important measure for determining the importance of an association rule in a data set is the *lift* measure. This measure is based on the idea that if $P(A \cup B) = P(A)P(B)$, then A and B are independent. Therefore, the calculation of the *lift* measure, defined as $P(A \cup B)/(P(A)P(B))$, allows the dependence between A and B to be determined. Values close to 1 indicate that the occurrence of A is not related to B, while larger or smaller values indicate positive or negative dependence, respectively.

In this study, the a priori algorithm, proposed by R. Agrawal and R. Srikant in 1994, is used to obtain association patterns in the data set. This algorithm is based on prior knowledge of frequent transactions and uses an iterative approach for its implementation.

3.2. Dickey–Fuller Test and Autocorrelation Functions

Time series are composed of three elements: the trend, which represents smooth changes over the medium to long term; the seasonal component, which shows periodicity over time; and the random component, which follows no discernible pattern [35]. Identifying the presence of a seasonal component in a time series can improve the quality of predictions and allow the application of more robust techniques. For this purpose, there are statistical tests and tests to verify the existence of seasonal components in time series, and the Dickey–Fuller test is one of the most widely used.

The Dickey–Fuller test is a statistical test developed by David Dickey and Wayne Fuller in 1979. This test evaluates the presence of unit roots in a time series. A unit root indicates a stochastic trend, i.e., the absence of stationarity [36]. Under the null hypothesis (H0), the time series has a unit root and, therefore, is not stationary [37]. Under the alternative hypothesis (H1), the time series does not have a unit root and, therefore, is stationary.

Once it is determined that a time series is not stationary, it is important to know if it has seasonal patterns, i.e., periodicity. For this purpose, there are tools such as the autocorrelation function (ACF), which allows us to identify dependence patterns and temporal structures in the data. The ACF does not serve as a similarity index to measure how much similarity exists between the behavior of a time series at present and other dates [38]. Instead, it aids in identifying the presence of a linear relationship between the observations of the time series at various lags. However, the ACF can be affected by intermediate values, so the partial autocorrelation function (PACF) is used, which measures the correlation between the time series and its own values at one particular lag when the effect of all the other lags is removed.

3.3. Gradient Boosting

There are many diverse existing approaches to evaluate the importance of predictor variables in prediction algorithms. Some of the most commonly used methods include, as seen in previous research, the calculation of correlation coefficients and the use of gradient-boosting models.

One of the techniques used in data mining for decision-making is decision trees. Decision trees are a type of machine learning algorithm used for both classification and regression tasks. They construct a tree-like model where each internal node represents a decision based on input features, and each leaf node represents the final prediction. Decision trees are easy to interpret and visualize, making them valuable for understanding the decision-making process in the model. In their construction, the most relevant variables tend to appear in the first splits. Measures such as the Gini index, which assesses the purity of a data split based on the distribution of classes, are used to identify relevance. Smaller values of the Gini index after a split indicate a higher relevance of the variable used for the split.

However, the use of decision trees can be highly dependent on the seed and initial node, which reduces their reliability. For this reason, ensemble techniques such as random forest or gradient boosting are used. In gradient boosting, trees are built sequentially, with each tree correcting the errors of the previous one, allowing the model to fit the data more closely [39]. The final prediction is made by summing the weighted predictions of all the

trees. On the other hand, in random forest, trees are built independently in parallel, each trained on a random subset of the data. During prediction, the predictions of all the trees are combined through voting (in classification) or averaging (in regression) to obtain the final prediction. Both methods are highly effective and widely used in the field of machine learning due to their ability to handle complex data and provide accurate and reliable results. Ensemble techniques tend to generalize well and provide robust performance across various datasets due to the diversity of trees. They are widely used in machine learning for their ability to handle high-dimensional data and deliver reliable results.

In this case, the gradient boosting algorithm is used, where the importance of the variables is evaluated according to three metrics:

- The weight metric considers the frequency with which a variable is used to split the nodes in the individual trees.
- The gain metric takes into account the improvement of the loss function obtained by performing a split in the data using a particular variable.
- The cover metric evaluates the proportion of samples affected after splitting the data set based on a variable.

In the specific case considered, the extreme gradient boosting (XGB) algorithm is used because it allows working with qualitative variables in Python without requiring additional transformations [40]. Most algorithms require the variables to be quantitative, and the most common method for this involves transforming categorical variables into dummy variables. This involves creating as many variables as there are possible values for each qualitative variable, which can generate a large number of variables. Once the most important variables have been identified and it has been verified that reducing the number of predictor variables does not significantly affect accuracy, it is possible to use other methods without incurring such a high computational cost.

4. Materials and Methods

4.1. Analysis of Available Data

The available database is a structured historical database given by means of an extended star or snow-like model. It contains data from 2016 to 2022, with a total of two million services carried out in the Principality of Asturias, Spain (Table 2). The variables provided by the entity include:

- Date and time of service: In one column, date in the format "d/m/Y" and, in another column, time in the format "HH:MM" at which the service starts.
- Classification as urgent or scheduled: A column indicating YES when corresponding to urgent services and NO when corresponding to scheduled services.
- Origin of the call: Which entity the call is made from.
- Origin and destination of the service provided at different levels: the health area, council, type of premises/building, and address. There are 8 columns corresponding to the origins (Starting) and destinations (Arrival) at the four levels indicated above. (For example, starting area, arrival area, starting council, arrival council, etc.)
- Stretcher requirement: A column with YES if there is a need for a stretcher and NO if not needed.
- Stretcher bearer requirement: A column with YES if there is a need for a stretcher bearer and NO if not needed.
- Nurse requirement: A column with YES if there is a need for a nurse and NO if not needed.
- Companion requirement: A column with YES if there is a need for a companion and NO if not needed.

Table 2. Number of services available for analysis. The table shows the available data, disaggregated by health area (from 1 to 8), and whether the services are emergency (E) or scheduled (S). A total of 1,350,971 scheduled services and 605,251 emergency services were provided.

Health	Area	1	2	3	4	5	6	7	8
2016	E	3525	1659	12,406	23,130	21,981	4756	6271	6238
	S	10,707	12,347	27,195	48,274	38,947	9770	14,524	20,411
2017	E	3281	1699	13,045	23,939	23,016	4682	6194	6397
	S	10,050	10,557	30,008	52,057	39,305	14,113	15,465	22,136
2018	E	3385	1709	13,341	24,750	24,515	5072	6327	6615
	S	10,908	11,751	29,396	53,521	41,307	12,454	17,016	23,120
2019	E	3675	1632	13,644	26,765	25,830	4907	6147	6607
	S	1106	10,911	28,931	56,794	43,615	12,025	15,896	17,862
2020	E	3556	1669	13,627	26,545	26,114	4402	6277	6520
	S	8312	9140	24,456	53,850	35,991	10,132	14,931	16,762
2021	E	3835	1732	13,909	27,529	25,559	5136	6894	6244
	S	10,665	11,546	31,271	60,386	42,087	11,963	17,847	17,706
2022	E	3846	1782	13,750	26,182	25,484	5137	6222	6160
	S	10,850	10,589	29,667	63,872	43,832	11,921	15,357	17,404

The Principality of Asturias is an autonomous community of Spain located on the north coast of the Iberian Peninsula, bordered by the Cantabrian Sea on the north. It has a population of approximately 1 million inhabitants, according to data available as of September 2023. Asturias has a territorial extension of about 10,604 square kilometers and is organized by zones, divided administratively into 78 councils that are responsible for the management of resources and services at a municipal level. It is divided into eight health areas (Figure 2), which in turn contain 68 basic zones and 16 special health zones. The health areas are responsible for managing both primary care services and hospitals and other health centers in the corresponding region. The number of councils in each health area varies according to population density, administrative organization, geographical distribution, and primary care needs.
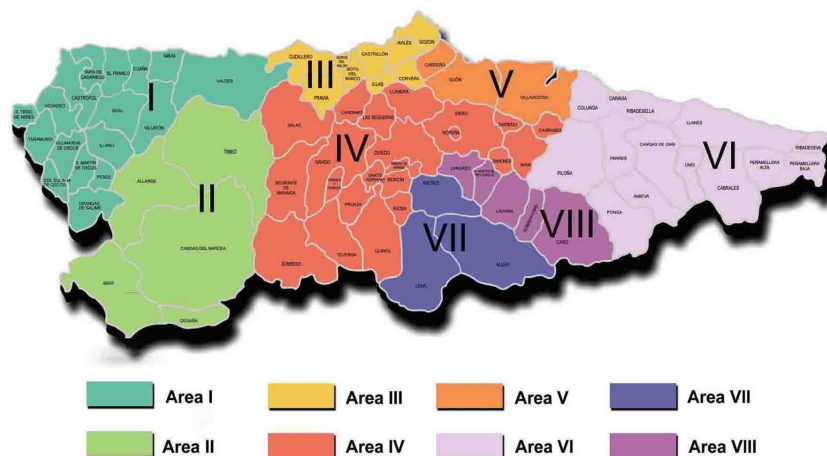

Figure 2. Division of health areas of Asturias. The map shows the 78 councils into which the Principality of Asturias is divided and, colored in different colors, the different regions that belong to each of the eight health areas. Image extracted from https://tematico8.asturias.es/repositorio/sanidad-ambiental/articulos/articulo_1372503041940.html, accessed on 12 December 2023.

We started by performing an exploratory analysis of the data, in which we checked the temporal evolution of the number of daily services for each year and council, differentiating between whether these services correspond to emergency or scheduled.

In carrying out this exploratory analysis, the peaks were identified manually for each of the 78 councils and years of available data. Then, a search was carried out in the media and virtual newspaper libraries on the events and/or occurrences that took place in that space and time.

The identification of events or occurrences associated with the peaks observed in the time evolution (Figure 3) gives an idea of the external databases that are related to the future peaks to be predicted. On the other hand, the identification of the reason why a peak appears is necessary for subsequent steps to identify seasonality in the data when ignoring punctual events.
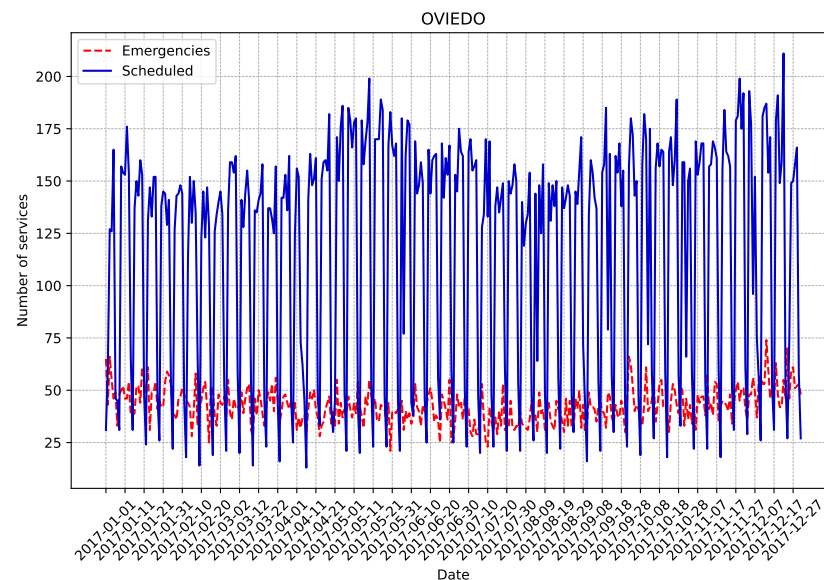

Figure 3. Example of the temporal evolution of the number of services during 2017 in the council of Oviedo. In red, the number of scheduled daily services is shown, and in blue, the number of emergency services. In the scheduled services, the difference between weekdays and weekends can be seen. A slight increase in the number of emergency services during the winter months is observed.

The identification of these peaks has led to different conclusions. On the one hand, emergency services are usually related to traffic or occupational accidents. Asturias is a mining province, so accidents in mines stand out, accompanied by assaults, suicide attempts, and drowning. Natural disasters, such as fires, storms, floods, flash floods, and landslides, usually require emergency services as well. On the other hand, parties and festivals, although they usually have scheduled services, sometimes coincide with peaks in emergencies. Other events that have scheduled ambulance services are sporting events, such as rallies, cycling, trail events, popular races, or soccer tournaments.

Other conclusions obtained in the exploratory analysis are the fluctuations from March 2020 to the end of 2021 due to COVID-19. The number of scheduled services linked both to sporting events and to rehabilitations and transfers decreased, increasing the number of transfers due to infection and/or symptoms of the virus.

4.2. External Databases Related to the Demand for Care

4.2.1. National and Local Holidays

Regardless of the size of the council analyzed, local and national holidays, as well as weekends, are associated with a decrease in the number of scheduled services. However, there is no significant variation in the number of emergency services, except for holidays associated with a celebration.

4.2.2. Sporting Events

Sporting events accompanied by a large influx of population take place in more densely populated municipalities, which is associated, as will be shown later, with a greater number of daily services and on holidays or weekends, so that an increase in the number of services due to a sporting event is hidden within this decrease. During the analysis of the data, it was found that there is a direct relationship between events held in less populated municipalities, such as rallies or trail races. However, it is not possible to draw clear conclusions about the events held in more populated localities. The sports events analyzed are:

- Rallies: The annual calendar of the FAPA (Federación de Automovilismo del Principado de Asturias) is divided into different categories. The data provided by these calendars from 2017 to 2022 is used. The contrast of the database with the available data concludes that on the days when there is a rally race, there is a pronounced peak, being higher in the councils with a lower population. In other categories, such as Rallysprint or historic rallies, the correlations are less appreciable, and in others, such as autocross or mountain and slalom, there are no peaks.
- Races: Races that congregate a larger number of people, such as popular races, take place in more populated areas, so no correlation with peaks is observed. However, mountain races such as trail races correspond to significant peaks in less populated areas.
- Soccer: An analysis is made of the matches of both third division and higher categories that have taken place during the years for which data are available. As in previous cases, soccer matches take place in the most populated areas and coincide with holidays or weekends, so there is no direct correlation between the matches and the number of services.

4.2.3. Demographic and Socioeconomic Data

An analysis of the literature indicates the existence of a correlation between the demand for emergency ambulance services and demographic or socioeconomic variables. In this case, the following variables are considered:

- Population: Number of people residing in the health area. It represents the number of people living in relation to the geographic extension of the area and is typically expressed as the number of inhabitants per unit area, such as people per square kilometer.
- Youth index: A measure that indicates the ratio between the number of people under 20 years of age and the number of people over 60 years of age.
- Overall dependency ratio: Demographic indicator that establishes the ratio between the number of dependent persons, who are under 16 years of age or over 64 years of age, and the number of persons of working age, who are between 15 and 64 years of age.
- Aging rate: Demographic measure that establishes the ratio between the number of persons over 64 years of age and the number of persons under 15 years of age.
- Labor force: Total number of persons who are of working age and are employed or actively seeking employment. This category includes persons who are employed in paid work, as well as those who are unemployed but actively seeking work.

As can be seen in Figure 4, the number of services per inhabitant in each health area, in emergency and scheduled services, is plotted, showing that it varies according to the area. As for the correlation between the number of services and the population, we obtain a correlation of 0.9956 for emergency services, while for scheduled services, the value is 0.9836, so the relationship in both cases is almost linear.
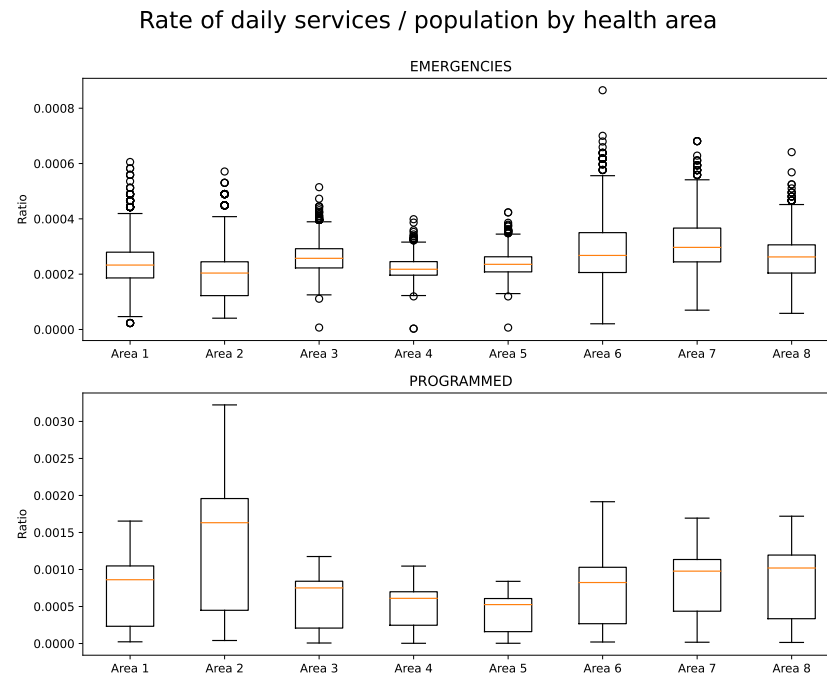
Rate of daily services / population by health area

Figure 4. Rate of number of services as a function of population for the different health areas.

Youth index and aging rate:

A significant direct correlation was observed between the youth index and the number of ED services (r = 0.9059) and scheduled services (r = 0.9091). Similarly, when analyzing the correlation with the aging rate, values of −0.8513 for ED and −0.8437 were found for scheduled. These results suggest that an increase in the proportion of young people in the population is related to an increase in both ED and scheduled services.

Overall dependency ratio:

For overall dependency, a negative correlation of −0.7302 was identified for ED services, and a positive correlation of 0.7615 was identified for scheduled services. Therefore, it is concluded that an increase in the dependency ratio is associated with a decrease in the number of ED services but an increase in scheduled services.

Active population:

Finally, when considering the size of the working population, a highly positive correlation of 0.9934 was found for ED services and 0.9802 for scheduled services. These results indicate that as the active population increases, there is a greater demand for both emergency and scheduled services.

4.3. Analysis of Internal Variables

We initially proceeded to visualize and analyze the data provided. A determining factor in distinguishing between emergency and scheduled services is the origin from which they are requested. A total of 1598 different origins were identified in the dataset, so they were grouped into categories considered to be the most relevant:

- Hospital,
- Health center,
- Home,
- Residence,
- Public road,
- Daycare center,
- Outpatient clinic,
- Other.

The initial analysis shows that some services that are continuous over time, such as transfers, may belong to both emergency and scheduled services. This analysis is performed at a council level to identify possible patterns of interest.

It is found that, in most of the councils, the origin of the most frequent service calls corresponds to homes, exceeding 50% of the total number of calls. In relation to scheduled services, in municipalities with hospitals, most of the services are scheduled from these facilities. In the smaller municipalities, the scheduling of services from daycare centers, residences, and homes stands out.

The next step in the analysis consisted of calculating the mean and standard deviation of daily services for each council, distinguishing between emergencies and scheduled services. Since there are differences in population between councils, and this may affect the variability of the data, the standard deviation is not considered to be a meaningful measure. Instead, the coefficient of variation, which is defined as the standard deviation divided by the mean, is used.

It is concluded that the variation in the number of emergency services is inversely proportional to the size of the council's population. Furthermore, this relationship follows an exponential function, i.e., the variation decreases exponentially as the population increases. However, in the case of scheduled services, the councils with the least variation are those with the reference hospitals for each of the established health areas.

From this point on, taking into account that the objective is to predict at the health area level, this spatial scope will be considered in the analysis of the data.

4.3.1. Analysis of Transfers between Health Areas

The analysis begins with the visualization of transfers between health areas using heat maps. It can be seen that the number of transfers within the same area is considerably higher compared to those between different areas. In order to highlight the latter, a logarithmic scale is used in the visualization (Figure 5).

A greater number of transfers is observed in Area IV, especially with regard to programmed services, which is understandable given that this area houses the province's central hospital. On the other hand, it is observed that, in the case of emergency services, transfers between areas hardly occur, except in those health areas that include large cities (III, IV, and V).

Considering the previous conclusions, where it was indicated that most of the programmed transfers originated in hospitals, an analysis of the destinations of these transfers was carried out. However, it was found that the classification of destinations in the recorded data is indeterminate in most cases, which prevents concrete conclusions from being drawn from this analysis.

4.3.2. Relationship between the Number of Daily Emergencies and the Number of Scheduled Services

As a result of the conclusions obtained, the question arises as to the existence of a possible relationship between the number of daily emergencies and the number of scheduled services. To address this question, a correlation analysis was carried out both in general and for each health area, obtaining values that were always positive, although not very significant. The correlation coefficients are higher in the health areas that include more populated cities. Consequently, there is a weak direct correlation between the number of emergency services and scheduled services. However, the results obtained are not sufficiently conclusive to draw reliable conclusions.
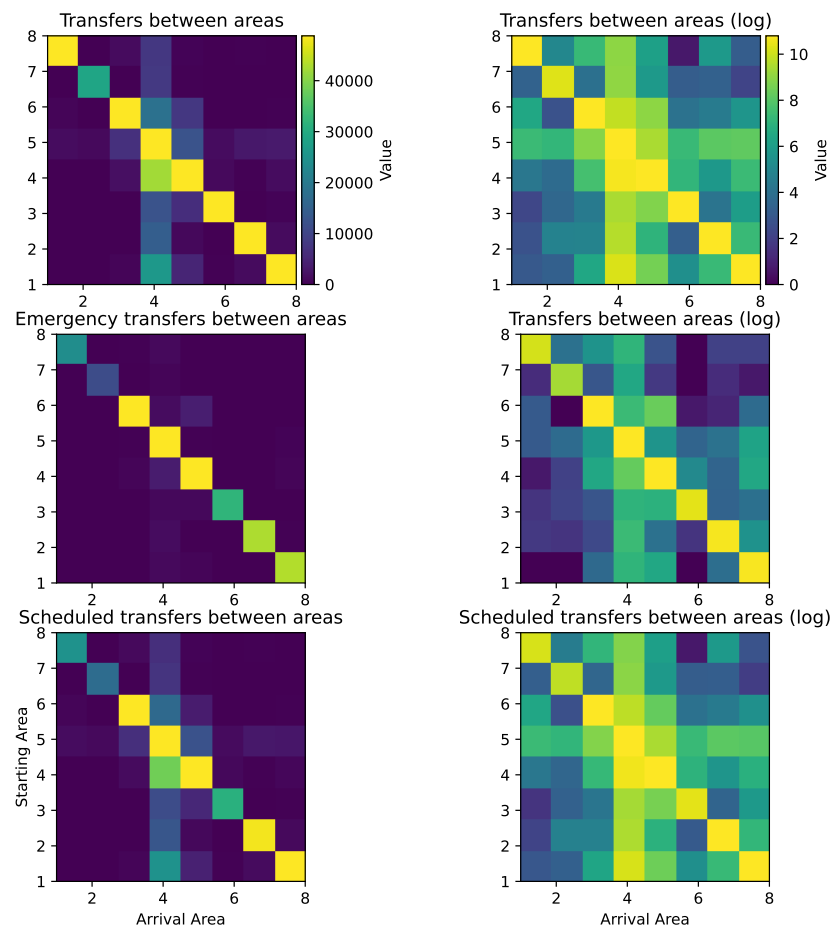
Figure 5. Transfers between health areas. The *y* axis represents the health area of departure, and the *x* axis, the health area of arrival.

4.3.3. New Time Variables

Other variables that may be determinant in the target prediction are the time variables. However, these variables are very precise and do not provide much information by themselves. Nevertheless, from them, it is possible to derive new variables that continue to provide relevant information. According to the literature analyzed and as will be verified later, variables such as the day of the week, the day of the month, and the day of the year have been identified as important variables in this context. Therefore, an intermediate processing of the data will be performed to obtain the following variables:

- From the time of service variable, a new variable corresponding to 6 time slots will be created:

  - From 00:00 to 04:00;
  - From 04:00 to 08:00;
  - From 08:00 to 12:00;
  - From 12:00 to 16:00;
  - From 16:00 to 20:00;
  - From 20:00 to 00:00.

- From the date of service variable, new variables will be created that may be more significant, such as:

  - Day of the week;
  - Day of the month;
  - Week of the year;
  - Month of the year;

– Year.

With the new variables, we proceeded to visualize the data and obtain new conclusions. Some of them can be seen in Figures 6 and 7. It can be concluded that, in the case of emergency services, the distribution over the days of the week is more or less uniform, although a slight increase is observed on Mondays and a decrease on weekends. However, in the case of scheduled services, there is a marked difference between Saturdays and Sundays compared to the rest of the days.

Frequency distribution by day of the week
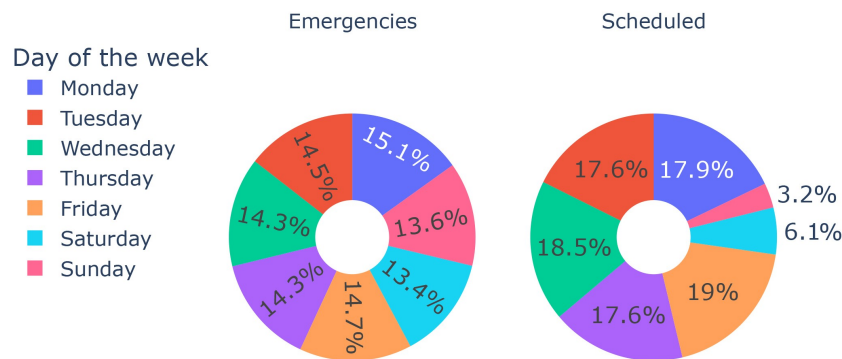
Figure 6. Distribution of frequencies in the number of services according to the day of the week.

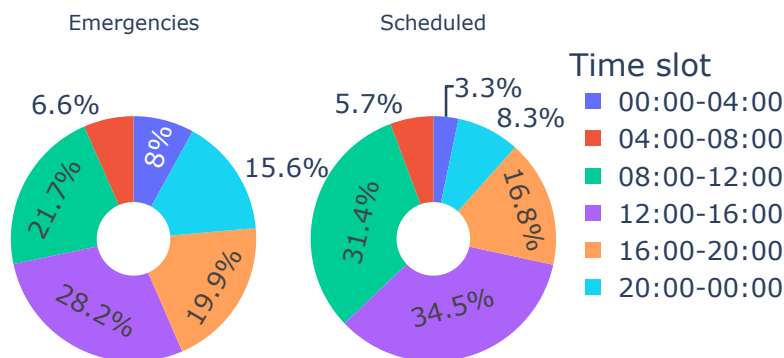## Frequency distribution by time slot

Figure 7. Distribution of frequencies in the number of services according to time slot.

In terms of time slots, for both emergency and scheduled services, a higher concentration of services is observed between 08:00 and 16:00, followed by the afternoon hours. For scheduled services, there is a notable decrease after 20:00, while in the case of emergency services, this decrease does not occur until 00:00.

5. Results

After the exploratory analysis of the data, the data mining techniques explained in Section 3 were applied. For this, a transactional database is required. Before applying the algorithm, it is necessary to prepare the database by eliminating single or irrelevant variables. In this case, the following considerations are taken into account:

- The service time is divided into 6 time slots.
- The date variable is eliminated, but new variables related to the day of the week, day of the month, week of the year, month of the year, and year are created.

Taking into account the amount of data available, a support value of 0.1 was selected, avoiding discarding transactions with a lower frequency that could be relevant. In addition,

a confidence value of 0.5 was set to ensure that the rules generated have an acceptable level of accuracy.

With the selection of these parameters, different combinations of variables such as no escort, stretcher bearer, nurse and/or stretcher were identified, but their presence in the association rules is due to the fact that in more than 84% of the transactions, none of these services are required. Therefore, it is considered that these variables do not provide significant information due to the imbalance in the data, and it is suggested that they should not be taken into account in future analyses.

It can be seen that the origin "Home" is the main one in the scheduled services, something already mentioned above. In addition, there is an evident association between the area and the referral hospital in the area. Another origin, denoted as "collective support", is always related to scheduled services and is required from home.

Regarding discharges, it is concluded that they are almost always scheduled, also obtaining an association between "Emergencies" and 112 (European Union emergency assistance telephone number) and SAMU (Emergency Medical Care Service—a specialized system that is part of "112" and is specifically dedicated to emergency medical care) calls. Scheduled services tend to be concentrated mainly between 08:00 and 14:00 hours, and a greater number of these are performed from home on Mondays, Wednesdays, and Fridays.

We continued with the Dickey–Fuller test on the objective variable, the number of services, differentiating by health area and whether they are urgent or scheduled services. A significance level of $\alpha = 0.05$ was used. The results indicate that most of the health areas show non-stationarity in scheduled services, except for IV (Oviedo, the capital) and VIII. As for emergency services, non-stationarity cannot be affirmed in most cases, and those that are not stationary have p-values higher than those of the scheduled services.

For those cases in which the Dickey–Fuller test has rejected the null hypothesis and the non-stationarity of the series has been determined, the ACF and PACF were calculated, taking into account up to a difference of 400 days to detect annual, monthly, and weekly stationarity, etc.

In the scheduled services, a weekly seasonality is observed (as shown by the significant autocorrelation at lag 7, 14, and 28 in Figure 8). This is logical since scheduled services tend to decrease during weekends, which generates a weekly periodicity.
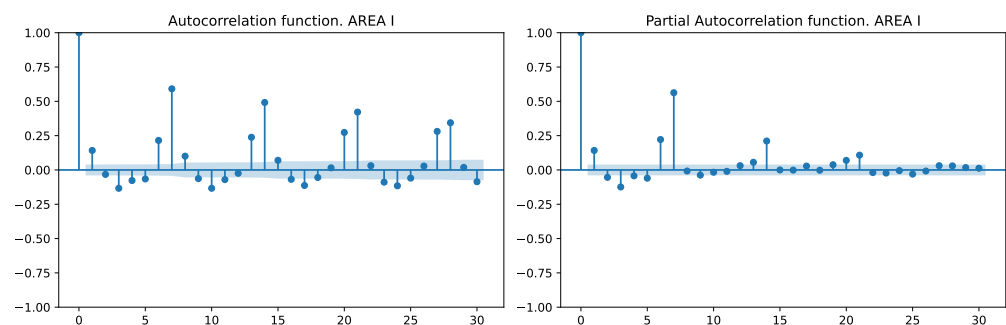

Figure 8. Autocorrelation and partial autocorrelation function for scheduled services in Area I.

In the health areas where the Dickey–Fuller test identified seasonal patterns, observations in emergency services indicate prominent periods occurring every 1 or 8 days (Figure 9). This pattern is attributed to the presence of both weekly seasonality and a short-term component at lag 1, with the occurrence at lag 8 arising from the interaction of these factors.
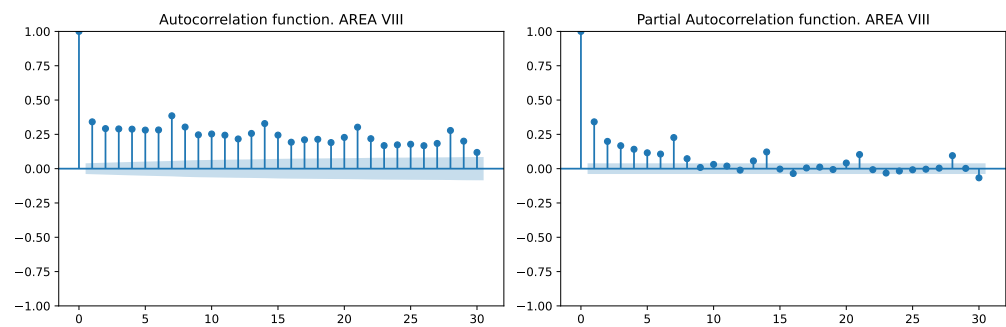
Figure 9. Autocorrelation and partial autocorrelation function for the emergency services of Area VIII.

In this case, only variables that can be useful for prediction are considered, so variables such as destination health areas and the need for stretcher-bearers are not taken into account. The variables taken into account are divided into two groups: on the one hand, those obtained directly from the table provided as the starting health area, and on the other hand, those obtained after processing the time of service and date of service variables: day of the week, week of the year, the distinction between emergency and scheduled, day of the month, day of the year, time slot and month of the year.

To compare the effect of taking the stated variables of greatest importance and to be able to make future comparisons, we first started with a predictive model. For this, we used, as a training data set, those services performed between 2016 and 2021 and, as a test sample, the first six months of 2022 since the following ones had incomplete data.

The objective is to evaluate how the use of different predictor variables affects the accuracy of a model according to their correlation with the target variable. As mentioned in Section 3, the XGB algorithm is used because it allows working with qualitative variables in Python without requiring additional transformations. After the algorithm was fitted with the available data, depending on the indicated metric (weight, gain, or cover), the provided variables were directly displayed and sorted according to their importance.

Taking into account any of the three metrics, the four most important variables, although the order changes depending on the metric, are the same:

- Starting health area,
- Time zone,
- Distinction between urgent and scheduled,
- Day of the week.

It was decided to use the gain metric, which indicates how much the loss function is reduced when dividing a data set according to a specific characteristic.

In situations such as this, where there is a large number of data and variables (even if some have been discarded), determining the importance of predictor variables and observing how models using fewer variables can be equally accurate can significantly reduce the computational cost. The complete order of the variables used is as follows:

1. Time slot,
2. Starting health area,
3. Day of the week,
4. Distinction between emergency and scheduled,
5. Month of the year,
6. Day of the year,
7. Day of the month,
8. Week of the year.

In the first instance, the model only considers the time slot, and subsequently, the other variables are added in new models until reaching one that includes all of them. For each model, four metrics are calculated:

- **Mean squared error (MSE):** It is a measure of the average squared difference between the actual and predicted values in a regression model. It quantifies the overall model performance, with lower MSE values indicating a better fit of the model to the data.
- **Root mean squared error (RMSE):** It is the square root of the MSE and represents the standard deviation of the residuals (prediction errors). RMSE is commonly used to interpret the error magnitude in the same units as the target variable.
- **Coefficient of determination (R-squared, R2):** It is a statistical metric that represents the proportion of the variance in the dependent variable (target) that is predictable from the independent variables (features) in a regression model. It ranges from 0 to 1, with higher values indicating a better fit of the model.
- **Mean absolute error (MAE):** It is a metric that measures the average absolute difference between the actual and predicted values in a regression model. Similarly to MSE, it is used to assess the model's accuracy, but it is less sensitive to outliers since it takes the absolute value of the errors.

Figure 10 shows that using only the four most important variables yields even better results than considering all variables. Halving the number of predictor variables implies a significant reduction in the computational cost associated with the predictions. Having fewer predictor variables to process reduces the complexity of the model and speeds up the execution time, saving both computational resources and CPU and memory time, which is especially useful when handling large data sets or when the model is to be used in real-time applications. Furthermore, with this reduction in variables, the possibility of over-fitting is reduced, and by focusing only on the most important variables, the interpretability of the model is improved.

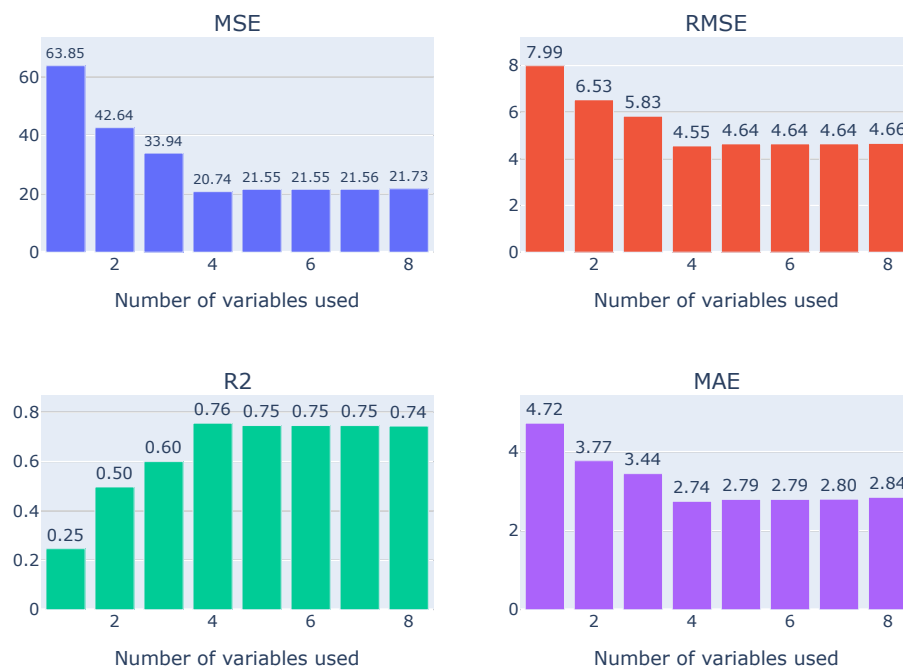Variation of metrics as a function of the number of variables used.

Figure 10. Variation in the metrics on the test set as a function of the number of variables used. The figure shows four graphs corresponding to the four metrics used: MSE, RMSE, R2, and MAE.

6. Conclusions

In health transport research, a broad spectrum of data analysis approaches and considerations have been explored. These range from defining clear objectives to selecting the appropriate methods and models for deriving insightful conclusions. Studies specifically targeting spatial and temporal prediction of the number of transport services emphasize

the importance of identifying relevant external variables and determining which are critical to the predictions.

Although it was not the main objective of the research, a previous analysis of the correlation and dependence with external variables, as well as the relationship between internal variables, has led to numerous conclusions that can be used not only in the definition of future models but also in current planning.

A study of external variables has been carried out, analyzing their correlation with the number of services. Unique events, such as sports or festive occasions, have been identified to elevate the demand for medical transport services in both emergency and scheduled contexts. This surge in demand is especially evident in less populated areas, where a singular event can amplify the average daily service count by as much as sevenfold. The combination of internal and external variables, along with future research involving other factors such as meteorological conditions, adds greater richness to the data and enables the derivation of new conclusions.

In addition, a strong correlation exists between demographic factors and health transport. In particular, a direct relationship has been found between the overall population and the active population segment with medical transport requirements across various health areas. Both these factors show a high correlation with the number of services, both emergency and scheduled. Interestingly, a significant correlation, close to 0.9, was observed between medical transport and the youth rate, indicating that areas with a higher proportion of young population tend to demand more medical transport services, a finding that might be counterintuitive to some. These results may be useful when determining fleet rates per population since, intuitively, one may think that older people require more health transport services, but this may be more oriented to scheduled services, while younger people require more emergency services.

During the examination of internal variables, significant insights emerged that aid in variable selection. The transfers that occur most frequently between different health areas were identified, along with the most common origins and destinations. Furthermore, specific time slots and weekdays were identified where transfers are more prevalent. Notable seasonality patterns emerged in scheduled services, which were validated through statistical measures such as the Dickey–Fuller test and the analysis of the ACF (autocorrelation function) and PACF (partial autocorrelation function).

The main objective of the study was to identify the determining variables in the prediction of the demand for medical transport. Based on an analysis and review of the literature, eight relevant variables were identified. However, only four of these proved essential for achieving comparable predictive results for service numbers. This reduction in the number of variables not only reduces the computational cost of the prediction models but also improves the interpretability of the results. The four most important variables in the prediction are time slot, health area of departure, day of the week, and the distinction between emergency and scheduled.

In this case, the amount of fleet (which was the most important factor in Figure 1) was not of interest since it is something that can be determined when the prediction based on it is known. This research has made it possible, on the one hand, to counter previous research with a new data set and, on the other hand, to determine that the gradient boosting algorithm yields similar results to previous ones. In addition, as a novelty, an analysis of the variation in the metrics has been carried out as the number of variables used increases.

Several promising lines of future research have emerged from this study. Firstly, exploring different predictive models and their respective hyperparameters is a logical next step. Using a comparative benchmarking approach, balancing computational expense with model interpretability, can help identify the best strategies. Additionally, exploring real-time data integration could further enhance prediction accuracy.

Author Contributions: Conceptualization, J.B.P., M.F.G., S.V.V. and O.J.C.C.; methodology, J.B.P., M.F.G. and S.V.V.; formal analysis, J.B.P. and M.F.G.; investigation, J.B.P., M.F.G. and O.J.C.C.; resources, J.B.P. and M.F.G.; writing—original draft preparation, J.B.P., M.F.G. and S.V.V.; writing—

review and editing, S.V.V. and O.J.C.C.; visualization, J.B.P., M.F.G. and S.V.V.; project administration, M.F.G. and S.V.V.; funding acquisition, J.B.P. and O.J.C.C. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: Authors Jorge Blanco Prieto and Marina Ferreras González were employed by the company UPintelligence S.L. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Eysenbach, G. What is e-health? *J. Med. Internet Res.* **2001**, *3*, e833. [CrossRef] [PubMed]
2. Kovačić, M.; Mutavdžija, M.; Buntak, K. e-Health Application, Implementation and Challenges: A Literature Review. *Bus. Syst. Res. Int. J. Soc. Adv. Innov. Res. Econ.* **2022**, *13*, 1–18. [CrossRef]
3. Rogers, H.; Madathil, K.C.; Agnisarman, S.; Narasimha, S.; Ashok, A.; Nair, A.; McElligott, J.T. A systematic review of the implementation challenges of telemedicine systems in ambulances. *Telemed. e-Health* **2017**, *23*, 707–717. [CrossRef] [PubMed]
4. Birmingham, L.E.; Arens, A.; Longinaker, N.; Kummet, C. Trends in ambulance transports and costs among Medicare beneficiaries, 2007–2018. *Am. J. Emerg. Med.* **2021**, *47*, 205–212. [CrossRef] [PubMed]
5. Tandberg, D.; Tibbetts, J.; Sklar, D.P. Time series forecasts of ambulance run volume. *Am. J. Emerg. Med.* **1998**, *16*, 232–237. [CrossRef]
6. Kargar, B.; Pishvaee, M.S.; Jahani, H.; Sheu, J.B. Organ transportation and allocation problem under medical uncertainty: A real case study of liver transplantation. *Transp. Res. Part E Logist. Transp. Rev.* **2020**, *134*, 101841. [CrossRef]
7. Zahiri, B.; Tavakkoli-Moghaddam, R.; Mohammadi, M.; Jula, P. Multi-objective design of an organ transplant network under uncertainty. *Transp. Res. Part E Logist. Transp. Rev.* **2014**, *72*, 101–124. [CrossRef]
8. Mantecchini, L.; Paganelli, F.; Peritore, D.; Trapani, S.; Morabito, V.; Oliveti, A.; Costa, A.N. Transport of human organs in Italy: Location, time, and performances. *Transplant. Proc.* **2017**, *49*, 622–628. [CrossRef]
9. Cacchiani, V.; Malandri, C.; Mantecchini, L.; Paganelli, F. A study on the optimal aircraft location for human organ transportation activities. *Transp. Res. Procedia* **2018**, *30*, 314–323. [CrossRef]
10. Siler, K.F. Predicting demand for publicly dispatched ambulances in a metropolitan area. *Health Serv. Res.* **1975**, *10*, 254.
11. Aboueljinane, L.; Sahin, E.; Jemai, Z. A review on simulation models applied to emergency medical service operations. *Comput. Ind. Eng.* **2013**, *66*, 734–750. [CrossRef]
12. Nehme, Z.; Andrew, E.; Smith, K. Factors influencing the timeliness of emergency medical service response to time critical emergencies. *Prehospital Emerg. Care* **2016**, *20*, 783–791. [CrossRef]
13. Eastwood, K.; Morgans, A.; Smith, K.; Hodgkinson, A.; Becker, G.; Stoelwinder, J. A novel approach for managing the growing demand for ambulance services by low-acuity patients. *Aust. Health Rev.* **2015**, *40*, 378–384. [CrossRef] [PubMed]
14. Swalehe, M.; Aktas, S.G. Dynamic ambulance deployment to reduce ambulance response times using geographic information systems: A case study of Odunpazari District of Eskisehir Province, Turkey. *Procedia Environ. Sci.* **2016**, *36*, 199–206. [CrossRef]
15. Ong, M.E.; Ng, F.S.; Overton, J.; Yap, S.; Andresen, D.; Yong, D.K.; Lim, S.H.; Anantharaman, V. Geographic-time distribution of ambulance calls in Singapore: Utility of geographic information system in ambulance deployment (CARE 3). *Ann. Acad. Med. Singap.* **2009**, *38*, 184. [CrossRef] [PubMed]
16. Hanchate, A.D.; Paasche-Orlow, M.K.; Dyer, K.S.; Baker, W.E.; Feng, C.; Feldman, J. Geographic variation in use of ambulance transport to the emergency department. *Ann. Emerg. Med.* **2017**, *70*, 533–543. [CrossRef] [PubMed]
17. Sasaki, S.; Comber, A.J.; Suzuki, H.; Brunsdon, C. Using genetic algorithms to optimise current and future health planning-the example of ambulance locations. *Int. J. Health Geogr.* **2010**, *9*, 4. [CrossRef]
18. Lin, A.X.; Ho, A.F.W.; Cheong, K.H.; Li, Z.; Cai, W.; Chee, M.L.; Ong, M.E.H. Leveraging machine learning techniques and engineering of multi-nature features for national daily regional ambulance demand prediction. *Int. J. Environ. Res. Public Health* **2020**, *17*, 4179. [CrossRef]
19. Vile, J.L.; Gillard, J.W.; Harper, P.R.; Knight, V.A. Predicting ambulance demand using singular spectrum analysis. *J. Oper. Res. Soc.* **2012**, *63*, 1556–1565. [CrossRef]
20. Wang, Z.; Xia, T.; Jiang, R.; Liu, X.; Kim, K.S.; Song, X.; Shibasaki, R. Forecasting ambulance demand with profiled human mobility via heterogeneous multi-graph neural networks. In Proceedings of the 2021 IEEE 37th International Conference on Data Engineering (ICDE), Chania, Greece, 19–22 April 2021; pp. 1751–1762. [CrossRef]
21. Chen, A.Y.; Lu, T.Y.; Ma, M.H.M.; Sun, W.Z. Demand forecast using data analytics for the preallocation of ambulances. *IEEE J. Biomed. Health Inform.* **2015**, *20*, 1178–1187. [CrossRef]

22. Bermúdez, J.D.; López, M.A.; Segura, J.V.; Vercher, E. Análisis estadístico de la evolución de la demanda asistencial en una consulta de atención primaria. In *Proceedings of the II Jornada Científica SEE-SEB "In Memoriam Juan Ferrándiz"*; Análisis espacial y temporal; Universitat de València: València, Spain, 2004.

23. Grekousis, G.; Liu, Y. Where will the next emergency event occur? Predicting ambulance demand in emergency medical services using artificial intelligence. *Comput. Environ. Urban Syst.* **2019**, *76*, 110–122. [CrossRef]

24. Bélanger, V.; Ruiz, A.; Soriano, P. Recent optimization models and trends in location, relocation, and dispatching of emergency medical vehicles. *Eur. J. Oper. Res.* **2019**, *272*, 1–23. [CrossRef]

25. Nicoletta, V.; Guglielmi, A.; Ruiz, A.; Bélanger, V.; Lanzarone, E. Bayesian spatio-temporal modelling and prediction of areal demands for ambulance services. *IMA J. Manag. Math.* **2022**, *33*, 101–121. [CrossRef]

26. Westgate, B.S.; Woodard, D.B.; Matteson, D.S.; Henderson, S.G. Travel time estimation for ambulances using Bayesian data augmentation. *Ann. Appl. Stat.* **2013**, *7*, 1139–1161. [CrossRef]

27. Argiento, R.; Guglielmi, A.; Lanzarone, E.; Nawajah, I. Bayesian joint modelling of the health profile and demand of home care patients. *IMA J. Manag. Math.* **2017**, *28*, 531–552. [CrossRef]

28. Zhou, Z.; Matteson, D.S.; Woodard, D.B.; Henderson, S.G.; Micheas, A.C. A spatio-temporal point process model for ambulance demand. *J. Am. Stat. Assoc.* **2015**, *110*, 6–15. [CrossRef]

29. Zhou, Z.; Matteson, D.S. Predicting ambulance demand: A spatio-temporal kernel approach. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, 10–13 August 2015; pp. 2297–2303. [CrossRef]

30. Channouf, N.; L'Ecuyer, P.; Ingolfsson, A.; Avramidis, A.N. The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta. *Health Care Manag. Sci.* **2007**, *10*, 25–45. [CrossRef] [PubMed]

31. Matteson, D.S.; McLean, M.W.; Woodard, D.B.; Henderson, S.G. Forecasting emergency medical service call arrival rates. *Ann. Appl. Stat.* **2011**, *5*, 1379–1406. [CrossRef]

32. Setzler, H.; Saydam, C.; Park, S. EMS call volume predictions: A comparative study. *Comput. Oper. Res.* **2009**, *36*, 1843–1851. [CrossRef]

33. Li, M.; Vanberkel, P.; Zhong, X. Predicting ambulance offload delay using a hybrid decision tree model. *Socio-Econ. Plan. Sci.* **2022**, *80*, 101146. [CrossRef]

34. Hann, J.; Kamber, M. *Data Mining: Concepts and Techniques*; Morgan Kaufman: Burlington, MA, USA, 2000.

35. Villavicencio, J. *Introducción a Series de Tiempo*; Instituto de Estadísticas de Puerto Rico: San Juan, Puerto Rico, 2010.

36. Novales, A. *Series Temporales. Estacionariedad, Raíces Unitarias*; Editorial en Madrid: Madrid, Spain, 2016.

37. Calvo, V.; Luis, J. Funciones de Autocorrelación Robustas, Test Estacionarios y Test de Raíces Unitarias. Master's Thesis, Universidade Aberta de Portugal, Lisboa, Portugal, 2010.

38. Montenegro García, A. La función de autocorrelación y su empleo en el análisis de series de tiempo. *Rev. Desarro. Soc.* **2017**, *23*, 117–132. [CrossRef]

39. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

40. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13 August 2016; pp. 785–794. [CrossRef]