Explanatory and Predictive Factors of Tourists' Behavior: A case from a Private Airport Transfer Company

Leonor Braga Pimentel Torres

Master's Dissertation

Supervisor: Professor Maria Teresa Galvão Dias

U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Mestrado Integrado em Engenharia e Gestão Industrial

2018-07-02

Abstract

Tourism is one of the biggest industries in the World and its growth in the past years has been considerable. The application of IT (Information Technology) in this industry is critical and includes countless opportunities due to some of its natural characteristics and to the amount of information that has been continuously generated through the past years.

This project, whose main goals are the understanding of the trends and factors that shape the company's demand and the proposal and comparison of forecast models for the annual, monthly and weekly data, is based on data from a private airport transfer company.

The methodology of this study is based on the Knowledge Discovery in Databases (KDD) process. After the understanding of the provided data, its relations and main attributes, it is necessary to undertake preprocessing and exploratory steps. There is evidence that the seasonality issues are the ones that most impact the demand, so time series of the total number of services for different frequencies are studied. Flights and airport's information are also likely to affect the demand, as they are highly correlated with the number of services, but the acquired data is not enough to perform a reliable analysis, so these are not considered in the prediction.

For the annual data, a linear regression detects the slope of the company's growth. Exponential Smoothing, ARIMA (Auto-Regressive Integrated Moving-Average) and Artificial Neural Networks are applied to both monthly and weekly number of services.

Regarding the monthly data, a Holt-Winters method, both for the raw data and for a logarithmic transformation, accurately fits the series. This achieves better results than an ARIMA and an Artificial Neural Network model, which are known for capturing complex patterns. These results might be a consequence of a limited amount of data or of the high importance of the trend and seasonal components in these series. Withal, weekly series present some distinct results as, considering the error measures, the ARIMA model outperforms the Holt-Winters and ANN. However, the plot of the observed and estimated values leads to some distinct conclusions as the high error in Holt-Winters results from a lag in time whereas in ARIMA the problem relies in the trend.

Finally, it is relevant to stress the error increase with the time frequency decrease and the augmented prediction intervals when forecasting further into the future.

Resumo

O turismo é uma das maiores indústrias mundiais, cujo crescimento nos últimos anos tem sido considerável. Devido às suas características e à grande quantidade de dados que têm sido continuamente gerados, a aplicação de tecnologias de informação neste setor é fundamental e engloba inúmeras oportunidades.

Este projeto, cujos objetivos passam pela extração das tendências e fatores que afetam a procura e pela proposta e comparação de diferentes modelos de previsão anuais, mensais e semanais, utiliza dados de uma empresa de transferes privados do aeroporto de Faro.

A metodologia deste projeto tem por base o processo KDD, processo de extração de conhecimento de bases de dados. Após total perceção dos dados fornecidos, a forma como se relacionam e os seus principais atributos, surge a necessidade de realizar um pré-processamento dos dados e análise exploratória. A sazonalidade tem uma grande influência na procura da empresa, pelo que séries temporais com os números de viagens para diferentes frequências são estudadas. Os dados referentes aos voos e ao aeroporto demonstram influência na procura pela sua alta correlação. No entanto, não são suficientes para realizar uma análise sólida e fidedigna, não sendo assim considerados para efeitos preditivos.

No que à procura anual diz respeito, uma regressão linear deteta o crescimento da empresa. Modelos de amortecimento exponencial, ARIMA e Redes Neuronais são aplicados ao agregado mensal e semanal do número de viagens realizadas. Em relação aos dados mensais, o método *Holt-Winters*, tanto para os dados em bruto quanto para a transformação logarítmica, ajusta-se com precisão à série. Assim, são alcançados melhores resultados do que no ARIMA e nas Redes Neuronais, modelos conhecidos por capturar modelos complexos. Estes resultados podem surgir como conseqência da quantidade limitada de dados ou do peso elevado da tendência e dos componentes sazonais da série. Por sua vez, as séries semanais apresentam alguns resultados distintos já que, considerando o cálculo dos erros, o modelo ARIMA apresenta melhores resultados. No entanto, aquando da observação do gráfico com os valores estimados e observados, percebe-se que o erro no Holt-Winter se deve a um desfasamento dos dados enquanto que no ARIMA deriva da captação da tendência.

Por fim, é verificado um aumento do erro com a aumento da frequência temporal e o aumento dos intervalos de previsão com o alargamento do horizonte de previsão.

Acknowledgements

After five years studying Industrial Engineering and Management, this thesis represents the end of this journey for which I am the most grateful.

First of all, I would like to express my gratitude to my supervisor, Professor Teresa Galvão, for her guidance, endless availability and effort in understanding and answering to all my problems. Regarding the understanding of some specificities of forecasting methods, I also have to thank Professor Sarsfield Cabral for his cooperation.

I would also like to thank Yellowfish Travel, LDA and especially Mr. Hortênsio Fernandes for his availability in providing the necessary data and patience to explain some of the specificities of the company's demand and operations. Also to Professor Pedro Cardoso for his advice and help.

To Catarina for the motivation and support in this project and for believing and making all the days funnier and easier. You know, we just got to keep swimming. Also to all of those who accompanied us, either in J302 or L203.

I could not go on without thanking to my family, who made it possible for me to study in Porto and supported me all the time. For being such an inspiration: to mum for her energy and care; to dad for the love for Management and for always expecting the best from me; and to my big brother, for making me believe that it is possible to do great things and have fun, to manage our time perfectly. Thank you for being always there, even when I did not make it to go home for the weekend, for the motivation and for those little things that make a huge difference in my journey, a special note to my grandma for sending the best soup every week.

To David, for his support, company, comprehension and patience. For making me believe that it was possible and worth the effort.

To my friends who went through this five year adventure with me, who lived all these amazing moments. Those who made me laugh and who made these years worth it. I will never forget the days and nights of study, the dinners, the discussions and the ideas that seemed stupid but turned out to be something incredible. We should be proud of what we did and take from here. Finally, to those who, although not present for the five years, marked this journey and taught me that everyone, regardless of the age, experience or knowledge, has something to teach you.

I would never finish if I wanted to refer everyone who, in some way, contributed to make these years so special, so I should end up thanking to all the wonderful people I met here and to the ones that never left.

Leonor Braga Pimentel Torres

"The greater danger for most of us lies
not in setting our aim too high and falling short;
but in setting our aim too low, and achieving our mark."

Michelangelo

Contents

Acronyms and Symbols

ACF	Autocorrelation Function
AIC	Akaike Information Criterion
ANN	Artificial Neural Networks
AR	Autoregressive
ARIMA	Auto-Regressive Integrated Moving-Average
BI	Business Intelligence
BIC	Bayesian Information Criterion
CSV	Comma-separated values
GOF	Golf Transfers
IT	Information Technology
KDD	Knowledge Discovery in Databases
MAE	Mean Absolute Error
MPE	Mean Percentage Error
MAPE	Mean Absolute Percentage Error
MA	Moving Average
OAL	One-way Airport-Location
OLA	One-way Location-Airport
OLL	One-way Location-Location
PACF	Partial Autocorrelation Function
RAL	Return Airport-Location
RMSPE	Root Mean Square Percentage Error
RMSE	Root Mean Square Error
RLL	Return Location-Location
SARIMA	Seasonal Autoregressive Integrated Moving Average Models
UK	United Kingdom
UNWTO	United Nations World Tourism Organization

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

According to United Nations World Tourism Organization (2018), in 2017 International Tourist Arrivals have increased 7% worldwide, which, being considerably higher than the past trend of around 4% per year, represent the best result since 2010. Coupled with the global growth, Southern and Mediterranean Europe reached an impressive growth of 13% in the same period. Consequently, tourism's importance both for the country's economy and development has increased. It is essential to deeply understand these patterns in order to provide a sustainable development and to enable touristic companies to adapt their strategy accordingly.

Furthermore, airport passengers' traffic has been subject to a considerable growth in the past years, hence this industry's data can lead to interesting insights on touristic patterns. This growth is sustained by the Airports Council International; from 2015 to 2016 the robust demand for air transport regarding passenger traffic increased 6,5% globally and 5,2% in Europe (Airports Council International, 2016). Additionally, the Global average of the passenger traffic growth rate for 2016-2040 is expected to be 4,5% per year (Airports Council International, 2017).

In addition to last years' growth, tourism has some characteristics that make it highly information intensive and lead to critical application of IT in this industry. Firstly, tourism's nature is typically heterogeneous, as in order to plan a trip tourists usually interact with different organizations. Its intangibility is also relevant, as most times it is impossible for the tourists to see, touch or feel a trip before the decision to buy it, which increases the challenge of marketing these products. Another characteristic is the perishability associated with these products; it is crucial to use high speed data communication networks to deliver information about last minute available products. Additionally, the inseparability of tourisms' consumption from the experience, the required interaction between the clients and the service providers boosts the need of information technologies to ensure an efficient and high quality production and consumption of experiences. Last but not least, tourism is a global industry that, as travelers are expanding their horizons and traveling more globally, requires access to information on diverse topics and from different geographical points

that feel the necessity for the linkage of countries, tourism organizations and travelers all around the world (Benckendorff et al., 2014).

Besides these characteristics, the use of IT in tourism is not new. In 2004, Werthner and Ricci (2004) already reported how customers were already using the Internet to look for information and book travels. By then, tourism industry was accepting e-commerce while others were still engaging with traditional methods.

Over the years, both the amount of information and Internet users has increased exponentially, leading to a huge generation of data, not only on tracking the search for information and on the purchases that are made, but also on user-generated content that provide feedback after the experience.

The motives that are referred in this section drive to a compelling industry for data mining which, by providing tools to extract valuable knowledge from large amounts of data, enables the study of touristic patterns and supports the decision-making processes.

1.2 Project Scope

The present project leads to a better understanding of touristic mobility patterns. Firstly, it is critical to perceive how the touristic activity has developed in the last years and what are the main trends that can describe it. The database of a private airport transfer company from Algarve, a touristic region in the South of Portugal, is used to understand these patterns and predict future trends of tourists' mobility in an area that deals with considerable seasonality issues.

Besides, this project enables the company to better understand their customers and respective demand. By providing prepared historic data and data analysis visualization results, the project intends to answer to questions that are capable of spotting some of these patterns and of better describing the private airport transfer company's business and operations. Additionally, time series forecasting models enable the prediction of the company's future demand which aims to avoid both the financial costs of an excess of capacity and the opportunity costs of unfulfilled demand or the costs that derive from the need of outsourcing their activities.

Therefore, the specific goals comprised in this dissertation are characterized as the following:

- To understand the trends that shape the company's demand and the main factors that might influence it;
- To propose models to forecast annual, monthly and weekly demand;
- To compare the accuracy of the different models used in the demand's forecast.

1.3 Methodology

It is important to understand which are the main phases and activities that need to be held to achieve the proposed goals of this dissertation. The chosen planning approach, presented in the next paragraph, is used to partition the activities into its main phases.

The literature review arises in the beginning of the project to explore the different studies that exist in the area of research and to gain knowledge in the methods that might be applied to the data. Then, data preparation is necessary to eliminate the chance of having dirty data biasing the results and the descriptive analysis is performed to deeply understand the company and the data provided. Additionally, data integration, selection and transformation steps are executed in order to properly prepare the data to enable the achievement of the defined objectives. These phases are of major importance for the remaining steps.

Lastly, forecasting methodologies are selected and implemented and computational tests aim to optimize the acquired results. Regarding the writing of the document, it represents a continuous task that is executed throughout the whole duration of this thesis.

1.4 Dissertation Structure

To provide a logical flow of information, this dissertation is divided into six main chapters. The present chapter concerns the motivation, scope and main goals of the project, along with its methodology and structure.

Following the introduction, Chapter 2 presents the literature review for data mining techniques and the main time series characteristics and methods. The chapter also includes a description on the process of knowledge discovery and the software used. Finally, it presents a literature review regarding tourists' arrivals forecasting or similar topics in the touristic sector.

Chapter 3 approaches the specific problem and describes some of the particularities that are related to this touristic region, its tourists' profile and facts related to the region's airport. The chapter also presents information on the company involved in this project and the main characteristics of the provided data.

A more detailed methodology is expressed in Chapter 4 that, accordingly with the process for knowledge extraction described in the literature review, presents the data preprocessing phases of the project. Additionally, this chapter presents the results of the exploratory data analysis that is conducted and reveals information that outlines the subsequent steps of the project.

In Chapter 5 the proposed models and forecasting results are displayed. Respecting each time series period, the accuracy of the models is studied and, consequently, the best model is selected.

Finally, Chapter 6 reveals the conclusions of the project and future work suggestions are exposed.

Chapter 2

Literature Review

The present chapter provides a general review on Data Analytics techniques and its application to tourism management. In Section 2.1, the main Data Mining process, applications, domains and techniques are reviewed. Time series are emphasized as they represent the technique of major interest for the case in study.

In Section 2.2, an overall study on the usage of data analytics in tourism demand analysis is executed, presenting methods used in similar data and its main advantages and limitations.

2.1 Data Mining

Data mining is defined by Han et al. (2012) as "the process of discovering interesting patterns and knowledge from large amounts of data", with data sources that might include "databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically". However, successful data mining requires to take into consideration several issues: the type of data, its quality and relationships and the preprocessing steps that make the data suitable must be explored with due importance (Tan et al., 2006).

Techniques from different domains, such as statistics, machine learning, pattern recognition visualization and algorithms, are incorporated into data mining. For example, by describing the behavior of objects from a target class, statistical models are extensively used and might lead to the outcome of a data mining task. Furthermore, machine learning refers to the learning or improvement of computers' performance based on data. This learning can be supervised, which means that the results come from the labeled examples of the training data, or unsupervised, with input that is not labeled. One of the most successful applications of data mining is Business Intelligence (BI) technology since it enables a historical, current, and predictive view of business operations (Han et al., 2012).

As explained in the next section, Data Mining is one essential step of Knowledge Discovery in Databases (KDD) process. However, sometimes it might be used as a term to refer to the whole process (Han et al., 2012).

2.1.1 The KDD Process

The KDD process includes different procedures described by Han et al. (2012) as the following phases:

1. Data Cleaning - consist in filling in missing values, smoothing noisy data, identifying or removing outliers and resolving inconsistent data as dirty data can produce unreliable results and cause confusion for the mining procedures;
2. Data Integration - combination of multiple data sources with attention to avoid inconsistencies and redundancies that might result from the integration;
3. Data Selection - retrieval of relevant data from the database;
4. Data Transformation - transformation and consolidation into forms that are appropriate for mining;
5. Data Mining - application of intelligent methods to the extraction of data patterns;
6. Pattern Evaluation - identification of interesting patterns;
7. Knowledge presentation - use of visualization and knowledge representation techniques.

The first four phases regard data preprocessing techniques that aim to improve data quality and overcome the problems caused by the huge size of data or the heterogeneous sources that originated it. Regarding data mining, its techniques are further explored in Section 2.1.2. Ultimately, for the knowledge presentation, visualization techniques play a crucial role.

2.1.2 Data Mining Techniques

Depending on the objective, data mining techniques can be divided into different groups. Firstly, a division into two major groups is suitable: predictive and descriptive tasks. Predictive tasks aim to foresee the value of one attribute, known as target or dependable variable, using explanatory or independent variables to get to this prediction. On the other hand, descriptive tasks derive patterns to better understand and summarize the relationships in the data. These tasks might include correlations, trends, clusters, trajectories and anomalies. (Tan et al., 2006)

Regarding predictive modeling it can be split into classification and regression. When the target variable is discrete classification techniques are appropriate whereas for continuous variables a regression is adequate. This distinction and its main subdivisions are characterized in Figure 2.1. Concerning descriptive methods aimed at uncovering strongly associated features in data, an association analysis that results in implication rules or feature subsets should be held. However, when the objective is to group closely related observations, a cluster analysis should be produced. Finally, anomaly detection finds observations with significantly different characteristics from the remaining data which are known as anomalies or outliers (Tan et al., 2006).

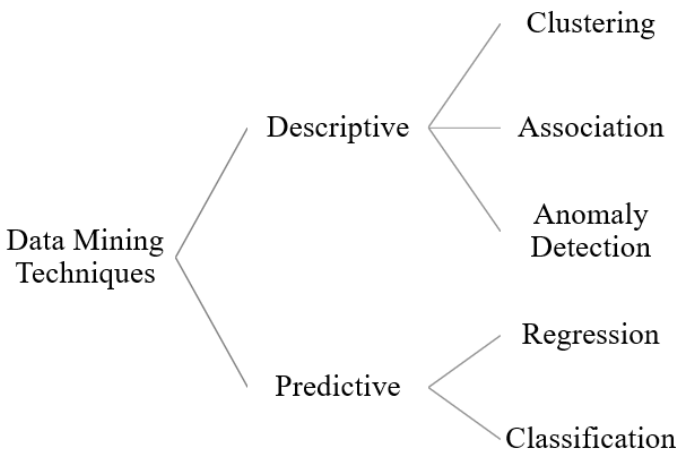

Figure 2.1: Data Mining Techniques

For each one of the tasks described in this section, numerous methods are available. Each method has its advantages and disadvantages and might be more successful depending on the data characteristics or the results needed. Considering the goal of this project of predicting the future demand of the company, along with the fact that the demand represents a continuous target attribute, methods that are applied to similar situations are further explained in this chapter. Owing to this, more information on time series data, that introduce a proper way to examine it, is presented in Section 2.1.3. These data serves as input for both the time series and artificial intelligence based methods that might be used to model and forecast the demand.

2.1.3 Time Series Analysis

Time series analysis regards the treatment of experimental data observed at successive different points in time with the primary goal of developing mathematical models that present plausible descriptions of the data. Shumway and Stoffer (2011) separate this into two different, but not necessarily mutually exclusive, approaches: the time domain, where the correlation is best explained in terms of a dependence of the current value on past values; and the frequency domain, where the interest is in periodic or systematic sinusoidal variations that are spotted in the data.

Time series, also termed as stochastic processes, are frequently characterized by a collection of random variables, $\{x_t\}$, indexed by t, which is commonly discrete and varies over integer values. For the sake of the visualization, the discrete values are regularly connected to reconstruct the hypothetical time series.

In Agrawal (2013), stationarity of a stochastic process is exposed as a form of statistical equilibrium, since statistical properties as the mean and variance of the process do not depend upon time. When it is possible to assume the stationarity of a time series, the mathematical complexity of the model is reduced, enabling the usage of some forecasting models, such as ARMA Agrawal (2013). Thus, this concept represents an indubitably important aspect to consider.

A strongly or strictly stationary process is characterized by $\{x(t), t = 0, 1, 2...\}$, where the joint probability function of $f\{x_{t-s}, x_{t-s+1}, ..., x_t, ..., x_{t+s-1}, x_{t+s}\}$ is independent of t for all s, which infers the timely independence of the joint distribution of any set of random variables (Agrawal, 2013). Shumway and Stoffer (2011) states that a strictly stationary time series is characterized by the probabilistic behavior of every collection of values which is identical to that of the time shifted set.

However, as this definition is too rigid for most applications, the weakly stationary concept arises. If the statistical moments of a process up to an order k depend only on time differences and not on time occurrences, the process is said to be weakly stationary of order k. A process $\{x(t), t = 0, 1, 2...\}$, is second order stationary if the mean and variance are time independent and the covariance values $Cov(x_t, x_{t-s})$ depend only on s. (Agrawal, 2013)

Henceforth, stationary refers to weakly stationary time series, whereas if a process is strictly stationary the complete notation is used.

Deepening the analysis, Guimarães (1988) suggests the characterization of time series based on these 4 components:

- Trend (T) - reflects the global evolution of the variable on the long-term, independently of the disturbances that might affect it on the short-term;
- Seasonal (S)- periodic fluctuation of the variable;
- Cyclical (C) - reflects non-periodic deviations from the trend that affect the series on the medium-term;
- Random (I)- Short-term fluctuations with an erratic and unpredictable character.

According to these components, time series can be divided into two different models: additive and multiplicative. An additive time series is characterized by the addition of the components expressed in the previous paragraph, characterized by Equation (2.1), while the multiplicative, as the name implies is the multiplication of the components, defined in Equation (2.2). The plot of these series presents differences in the seasonality variation. In an additive series the range imposed by the seasonal variation is independent from the time parameter, which means that this range remains constant through the whole time period. However, in the presence of a multiplicative time series, the range is proportional to the trend and cycle terms. When in a time series, an upward trend is noticeable, the variation in the seasonal component will also grow over time. (Guimarães, 1988)

$$AdditiveModel : Y(t) = T(t) + S(t) + C(t) + I(t) \tag{2.1}$$

$$MultiplicativeModel : Y(t) = T(t) * S(t) * C(t) * I(t) \tag{2.2}$$

Bearing in mind the described concepts, a series might present a non-stationary nature due to its trend or seasonal patterns. In these cases, differencing and power transformations might be used to make them stationary (Agrawal, 2013). When a trend is faced, a differencing of the data

might lead to stationarity in the mean. However, when in the presence of a multiplicative series, this might not be enough as a non-stationary in variance might still be spotted. A logarithmic transformation on the data, followed by the difference is adequate for making a series stationary on both mean and variance (Makridakis et al., 1978). According to Hyndman and Athanasopoulos (2018), differencing computes the differences between consecutive data points, if first ordered, or between an observation and the previous one from the same season, when seasonal differencing is applied. This way differencing transformations stabilize the mean by removing changes in the level of a time series, and therefore eliminating (or reducing) trend and seasonality.

In order to check if a series is stationary, tests like Dickey Fuller's are usually executed. (Agrawal, 2013)

As it is explored in more depth in the Section 2.2, time series are widely used to forecast tourism demand. Some of the methods that are applied to achieve this goal are now described.

2.1.4 Time Series Methods

2.1.4.1 Linear Regression

According to Guimarães and Cabral (1997), simple Linear Regression Model describes the relation between an independent variable, X, and a dependent variable, Y, described by

$$Y_n = \alpha + \beta \cdot (X_n - \overline{X}) + E_n \tag{2.3}$$

being X_n and Y_n the n^{th} observation of the pair, $\overline{X}$ the arithmetic mean of the observations X_n, α and β the fixed parameters of the linear relation that need to be estimated, and E_n the error associated to the value of Y_n.

The considered error affects only the variable Y_n, and is based on the assumption that its expected value is null and the variance constant and the observations are mutually independent and normally distributed.

2.1.4.2 Smoothing Methods

When a time series is the result of a constant process that is subject to random error, the mean can be used to forecast the following periods. Withal, when it involves a trend, a seasonal effect or both, the simple average is no longer a good estimator and smoothing methods seek to improve these cases. Both averaging and smoothing methods are described next, based on Makridakis et al. (1978).

Averaging methods

Averaging methods, as the concept of average indicates, refers to equally weighted observations. These methods comprise simple average of all the past data, single moving average of last n observations or double moving averages, which implies moving averages of moving averages, that turn out to be unequally weighted averages.

Simple average takes the average of all the points in the train dataset as the forecast for the next periods. It only produces good results when the observed values do not have noticeable trend or seasonality.

When the number of past observations included in the mean is specified at the beginning, a new average is computed by dropping the oldest observation and including the newest one, as each new observation becomes available. This process is named single moving average, as this moving average is used to forecast the new period. Although better than the simple average, this method is also not good for handling trend and seasonality.

Finally, with the intention of better handling trend, double moving average methods arise. An increase in the number of observations used will increase the error when using moving average along with the existence of an upward trend. To mitigate this error, a moving average of the moving average is adequate, by using the difference between a double moving average value and the single moving average value.

Exponential Smoothing

On the other hand, exponential smoothing methods apply a set of unequal weights which usually decay exponentially from the most recent to the farthest data point. All these methods imply the definition of parameters that determine these weights. The major advantage for the usage of these methods is its simplicity and low cost. Smoothing methods are usually good for forecasting a large period. These methods are thus described based on Makridakis et al. (1978).

Single Exponential Smoothing is defined by Equation (2.4), being α the weight value. When replacing F_t by its components, $\alpha X_{t-1} + (1-\alpha)F_{t-1}$, analogously F_{t-1} by $\alpha X_{t-1} + (1-\alpha)F_{t-1}$ and so forth, α reaches an exponential decrease.

$$F_{t+1} = \alpha X_t + (1-\alpha)F_t \tag{2.4}$$

This method is attractive to forecast a large amount of items as it requires little storage and computations and is adequate for data that do not comprise trend or seasonal components. As it implies the need of a parameter (α) that might be tested in order to optimize it, considering test measures as the MAPE. However, it is also possible to enable this value to vary accordingly with changes in the data pattern over time, namely *adaptive* Single Exponential Smoothing.

Double Exponential Smoothing, on the other hand, involves two exponential smoothing equations and can be applied to data with a trend component. Brown's method requires the same parameter for both equations by applying the straightforward double smoothing formula while Holt's method requires separated parameters for the distinct equations as it smooths the trend parameters separately, providing more flexibility. In Equation (2.5), S_t is directly adjusted for the trend of the previous period, b_{t-1}, which helps to mitigate the lag and brings S_t to the approximate base of the current data. In Equation (2.6) the trend, expressed as the difference between the last two smoothed values, is updated. To deal with possible remaining randomness, the trend in the last period, $S_t - S_{t-1}$, is smoothed by γ and added to the previous estimate of the trend multiplied

by $(1-\gamma)$. Finally, for forecasting Equation (2.7) is used by adding to S_t the multiplication of the trend by the number of periods ahead to predict.

$$S_t = \alpha X_t + (1-\alpha)(S_{t-1} + b_{t-1}) \tag{2.5}$$

$$b_t = \gamma(S_t - S_{t-1}) + (1-\gamma)b_{t-1} \tag{2.6}$$

$$F_{t+m} = S_t + b_t m \tag{2.7}$$

In Triple Exponential Smoothing, the same parameter might be used for the three exponential smoothing equations, Brown's Quadratic Method, or three different parameters might be used to smooth the data, the trend and the seasonal index in Holt-Winter's method. This model is adequate for directly handling seasonality.

The overall smoothing function is given by Equation (2.8) where L is the length of seasonality and I is the seasonal adjustment, calculated by Equation (2.9). As before, F_{t+m} represents the forecast for m periods ahead. The trend component is calculated by Equation (2.6).

$$S_t = \alpha \frac{X_t}{I_{t-L}} + (1-\alpha)(S_{t-1} + b_{t-1}) \tag{2.8}$$

$$I_t = \beta \frac{X_t}{S_t} + (1-\beta)I_{t-L} \tag{2.9}$$

$$F_{t+m} = (S_t + b_t m)I_{t-L} \tag{2.10}$$

Finally, Pagel's Classification of exponential smoothing methods acts as a summary by providing a framework for exponential smoothing that deals with trend and seasonal aspects separately, which might be multiplicative of additive.

2.1.4.3 ARIMA

The Box-Jenkins or ARIMA (Autoregressive Integrated Moving Average) models are more general and statistical based methods for time series analysis that have been applied in forecasting for several years. In an autoregressive model the current value is calculated as a finite, linear aggregate of the process's previous values and a shock that, in this kind of model, affects the following values. On the other hand, the Moving Average concept here differs from the one described in the smoothing methods. It implies a dependence relationship between the successive error terms and the equation. Lastly, the Integrated component refers to the non-stationary models that require a d^{th} difference (usually 0, 1 or 2) to make the process stationary (Box et al., 1994; Makridakis et al., 1978).

Thus, in order to select the appropriate model, some characteristics of the series, as stationarity and seasonality, have to be identified. The basic models, based on Makridakis et al. (1978), are presented subsequently.

A Random Model - ARIMA (0,0,0)

An ARIMA (0,0,0) is a simple random model where Y_t is composed by the overall mean, μ, and a random error, e_t, independent from period to period, as shown in Equation (2.11). As this model is stationary and Y_t does not depend on Y_{t-1} neither on e_{t-1}, there are no AR (Autoregressive), differencing or MA (Moving Average) aspects.

$$Y_t = \mu + e_t \tag{2.11}$$

A Non-stationary Random Model - ARIMA (0,1,0)

Equation (2.12) might be similar to an autoregressive model, as Y_t depends on Y_{t-1}. However, when Y_{t-1} has an unitary coefficient, it can be rewritten as the first difference of Y_t, $Y_t - Y_{t-1}$. In that case, it regards a stationary series, while Y_t is non-stationary. Thus, the model's equation is defined by Equation (2.13), representing a random model.

$$Y_t = Y_{t-1} + e_t \tag{2.12}$$

$$Y_t - Y_{t-1} = e_t \tag{2.13}$$

A Stationary Autoregressive Model of Order One - ARIMA (1,0,0)

In this model, characterized by Equation (2.14), there is already an autoregressive component as Y_t depends on Y_{t-1}, being ϕ_1 the autoregressive coefficient which lies between -1 and 1. In the Equation (2.14) and following, μ' equals to $(\mu - \phi_1\mu)$. However, it is important to notice that, when working with first differences of Y, this term is canceled out.

$$Y_t = \phi_1 Y_{t-1} + \mu' + e_t \tag{2.14}$$

A Stationary Moving Average Model of Order One - ARIMA (0,0,1)

In this model, Y_t depends on the error e_t and e_{t-1} with the coefficient $-\theta_1$ lying between -1 and 1. Equation (2.15) defines this model.

$$Y_t = \mu' + e_t - \theta_1 e_{t-1} \tag{2.15}$$

A Simple Mixed Model - ARIMA (1,0,1)

The combination of the basic elements of AR and MA might produce a great variety of models through their combination. In this case, expressed by Equation (2.16), Y_t depends on one previous error term e_{t-1} and on one previous Y_{t-1} value.

$$Y_t = \phi_1 Y_{t-1} + \mu' + e_t - \theta_1 e_{t-1} \tag{2.16}$$

Higher Order Combinations - ARIMA (p,d,q)

Combining the cases demonstrated previously, the general model is known as ARIMA(p,d,q). The autoregressive component (AR) given by p, is the order of the autoregressive process and the integrated component (I) is given by d and corresponds to the degree of differencing.The moving average component, given by q, represents the order of the moving average process. The equation for the case of ARIMA(1,1,1) is given by Equation (2.17) which derives from a combination of the previous equations. Being B the backward shift operator which has the effect of shifting the data 1 period and works as $BX_t = X_{t-1}$, $(1-B)$ represents the first difference component which is multiplied by $(1-\phi_1 B)X_t$ that comprises the AR(1) portion of the model. Additionally, $(1-\theta_1 B)e_t$ represents the part of the equation that refers to MA(1).

$$(1-B)(1-\phi_1 B)X_t = \mu' + (1-\theta_1 B)e_t \tag{2.17}$$

Seasonal Autoregressive Integrated Moving Average (SARIMA) Models

When a series exhibits a periodic behavior, with similarities after s basic time intervals, it is said that it has a seasonal pattern. In these cases a Seasonal ARIMA model is applied - $ARIMA(p,d,q)x(P,D,Q)_s$, where (P,D,Q) are the corresponding seasonal parameters of (p,d,q) (Box et al., 1994).

2.1.5 AI Based Methods

More recently, studies have proved that, under certain circumstances, Artificial Intelligence methods outperformed traditional time series forecasting methods. This section describes Artificial Neural Networks (ANN) that able to predict continuous target attributes based on time series data, and which do not require the creation of groups according to linguist rules (such as, low, middle and high) (Peng et al., 2014).

2.1.5.1 Artificial Neural Networks

As an universal function approximator, artificial neural network (ANN) models, lead to the ability to adapt and map any linear or non-linear function and contributed to its increasing interest in time series forecasting (Claveria and Torra, 2014). Additionally, ANNs have a data-driven and self-adaptive nature, which withdraws the need of specifying a particular model or making assumptions about the data's statistical distribution and are inherently non-linear, being more practical and accurate in modeling complex data patterns (Agrawal, 2013).

Artificial neural networks are originally developed based on the basic biological neural systems. Analogous to the human brain structure, an ANN is composed of both an interconnected assembly of nodes and directed links. While each node receives an input signal with the information from the previous node, it starts producing a transformed output signal that is transfered to other nodes or external outputs. ANNs try to recognize regularities and patterns in the input

data, learn from experience and then provide generalized results based on their known previous knowledge. (Agrawal, 2013; Tan et al., 2006; Zhang et al., 1998)

In Agrawal (2013), three different network architectures for forecasting problems are described. Feed forward network (FNN) performs a non-linear functional mapping from the past observations of the time series to the future value. Here, the target t_x is a function of the values $x_{t-i}, (i = 1, 2, ..., p)$ where p is the number of input nodes.

In Time Lagged Neural Networks (TLNN) the input nodes are time series values at some particular lag. Considering a typical TLNN for a time series, with seasonal period $s = 12$ the input nodes could be the lagged values at time $t-1$, $t-2$ and $t-12$. Thus, it would be possible to forecast a value at time t regarding the values at time 1, 2 and 12.

Lastly, Seasonal Artificial Neural Networks (SANN) incorporate a seasonal parameter s, used to drive the number of input and output neurons. A number of observations comprises each seasonal period. The i^{th} and $(i+1)^{\text{th}}$ seasonal period observations are used as values for the input and output neurons.

A typical multi-layer feed forward networks which is represented in Figure 2.2. The first layer is the input layer where the external information is received and in the last layer, output, the problem solution is obtained. Between these layers, the intermediate or hidden layers are found, having this example one hidden layer. Adjacent layers' nodes are usually fully connected by arcs from the input to the output layers.

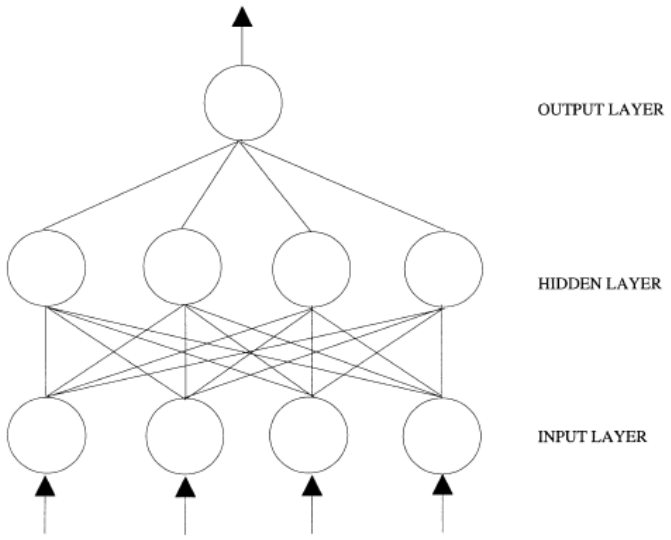

Figure 2.2: Typical feed-forward multi-layer perceptrons neural network (Zhang et al., 1998)

2.1.6 Performance Metrics

When the objective is the prediction of a target variable, models are usually divided into train and test datasets. This method enables to train the model with part of the data and then use test data to measure its accuracy. In order to compare different models, these accuracy measures that

compare estimated with observed values from the test dataset, are obtained for each model. The measures are presented and explained in this subsection based on Agrawal (2013) considering $\hat{y}$ the predicted value and y the observed:

- Root Mean Squared Error (RMSE) - standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2};$$

- Mean Absolute Error (MAE) - measures the average magnitude of the errors in a set of predictions, without considering their direction and depends on the scale of measurement and data transformations.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|;$$

- Mean Percentage Error (MPE) - computed average of percentage errors by which forecasts of a model differ from actual values of the quantity being forecast. Here the direction of error is shown and opposite signed errors cancel each other out - a small value is desirable but it is not inferable that the model performs well if the MPE is close to zero

$$MPE = \left(\frac{1}{n}\sum_{j=1}^{n}\frac{y_j - \hat{y}_j}{y_j}\right) \cdot 100;$$

- Mean Absolute Percentage Error (MAPE) - computed average of the absolute percentage errors by which forecasts of a model differ from actual values of the quantity being forecast, is independent of the scale of measurement, but affected by data transformation

$$MAPE = \left(\frac{1}{n}\sum_{j=1}^{n}\frac{|y_j - \hat{y}_j|}{y_j}\right) \cdot 100;$$

It is important that these metrics evaluate both the variability and the bias of the error. Bias refers to a systematic under or over estimating of the future value and is usually a consequence of the method, while variability regards unpredictable random errors that arise from the data. To measure the bias an average of forecast errors is sufficient while for the variability an average of the absolute or squared errors is necessary.

2.2 Data Analytics in Tourism Demand Analysis

In accordance with the information provided in Subsection 2.1.2, some methods might be more adequate than others depending on the case in study, its data features and several other factors, such as the desired forecasting horizon. This section presents data mining studies, encompassing forecasting or descriptive tasks, that are explored in tourists behavior and destinations' demand analysis.

The importance of forecasting tourists' arrivals has been stressed out for the past decades and has increased its importance for destination and tourism companies' management. A review of

works published between 1980-2011 (Peng et al., 2014) partitions the quantitative tourism forecasting methods in basic and advanced time-series, static and dynamic econometric models and artificial intelligence methods. Regarding artificial intelligence, the models presented are Rough Sets Approach, Fuzzy Time Series Method, Grey Theory, Artificial Neural Networks and Support Vector Regression, of which only the last two do not rely on linguistic variables that categorize the continuous attributes. The review additionally reveals that the strength of a model is dependent on factors as the destination being forecast, the data frequency, the number of explanatory variables, the forecasting horizon or the existence of seasonality.

Another review that focuses on studies only from the XXIth century, but for a smaller period, only until 2007, was performed by Song and Li (2008). It emphasizes the increasing diversity in techniques applied and also concludes that the performance varies according to the situation, leading to the non-existence of a model that outperforms the remaining models in all situations. The review considers 121 studies, of which 72 used time series techniques and 71 used econometric models. A smaller percentage of studies was already entering the artificial intelligence category. It is also relevant to highlight that more than two-thirds of the studies that use time series techniques apply different versions of the ARIMA model. Most of the studies in the analysis used historical data and regarding the time frequency of the data used, about half of them rely on annual data. Monthly and quarterly data are used in the same proportion and only one study focus daily counts.

Regarding daily data, it is also used for the estimation of the growth rate and volatility of daily tourism in Peru. Divino and McAleer (2010) uses the daily international tourist arrivals at the only International Airport in Peru. Data with daily frequency is advantageous as it is able to capture day-of-the-week effects and week-ends' arrival patterns. The estimation might additionally enable to understand the growth rate of the daily spending per tourist arrival as it is virtually identical to the growth rate being predicted. The alternative ARMA models are chosen based on the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) measures.

As an example, when forecasting the US demand for travel to Durban, South Africa, Burger et al. (2001) found that the neural network approach performed better than models as ARIMA, Naïve and Single Exponential Smoothing. These models, together with multiple regression, genetic regression, moving average and decomposition, are based on aggregated monthly data over the time period of 1992 - 1998 with a high degree of nonlinearity, seasonality and upward trend. Additionally, the innate model complexity performs far better when handling non-linear behavior. The results also concluded that longer forecasting periods lead to worse predictions but, when it covers the repetitions of expected similar seasonal patterns, the network performs fairly well. Though with different data and counting on six explanatory variables, Law (1998) also concluded that the forecasting efficiency of a neural network outperformed multiple regression and naïve extrapolation when predicting the room occupancy rate based on the number of tourists, average length of stay, number of hotels, number of rooms, tourists per room and percentage of hotel accommodation. On the other hand, Claveria and Torra (2014) concluded that, for the prediction of overnight stays and tourists arrivals from different countries to Catalonia, and especially regarding shorter horizons, ARIMA models prevail over self-exciting threshold auto regressions and artifi-

cial neural network models. For this study monthly data of tourist arrivals and overnight stays from foreign countries to Catalonia between 2001 and 2009 was used. In order to eliminate both linear trends as well as seasonality we obtained the trend-cycle component of the series and used year-on-year growth rates. As Neural Networks are more suitable to deal with nonlinearity in the data, the accuracy of the forecasts might be influenced by the degree of pre-processing, since the information loss that resulted from the filtering process lowers the accuracy of neural network forecasts compared to those of linear models.

The accuracy measures that are widely used in tourism demand forecasting are MAPE and RMSE. One of the reasons why Chu (2004) uses MAPE is the fact that positive errors do not cancel the negative ones. Marković et al. (2016), on the other hand, uses R^2 for a comparison of the different models modeled by a varied amount of scenarios, and RMSE for the comparison of the prediction accuracy. To compare eight different models for daily forecasting of a casino buffet, Hu et al. (2004) uses MAPE and RMSPE (Root Mean Square Percentage Error), which differs from the metrics described in Subsection 2.1.6 by adding the percentage term to the RMSE.

Regarding other areas of tourism demand analysis besides forecasting, methods for the implementation of a Destination Management Information Systems are addressed by the validation, testing and implementation of critical concepts as the definition of industry's knowledge requirements, data extraction, data warehousing and user-interfaces. This leads to a Destination Management Information System prototype which comprises web-search, booking and feedback data, mainly generated by customers. (Fuchs et al., 2014)

2.2.1 Big Data in Tourism

When the amount of data is so big that it is not possible to capture, manage or process by general computers, it is called big data. Song and Liu (2017), in its works on big data characterization, added one more V to the 4 V's characteristics of big data in the literature, resulting in the following concepts: Volume, regarding the huge quantities of data; the existing Variety of formats; Velocity for the speed of storage and retrieval; Veracity for the truthfulness and accuracy of the data; and finally, Value, indicates the application of big data.

When predicting tourists' arrivals, both adding big data information sources to past arrivals and using data mining techniques, as k-nearest-neighbor, instead of statistical approaches like linear regression, have proved to be advantageous to the accuracy of the prediction and to the identification of non-linear relationships (Wolfram et al., 2017).

Miah et al. (2017) presented a method that is able to extract, rank, locate and identify meaningful tourist information. The method combines four computational techniques such as text processing, geographical data clustering, visual content processing and time series modeling and uses geotagged photos and related details from the case of Melbourne, Australia, to transform unstructured big data sets to support the decision-making process.

2.2.2 Practical Application of Forecasting Methods

Hu et al. (2004) integrates the forecasting with capacity, demand, duration, price and queuing management practices of the Casino Buffet Restaurant. An accurate forecasting allows the anticipation of the demand which comprises possible improvements in all of these management fields. Regarding capacity, a flexible strategy might be applied, allowing to schedule the servers tasks accordingly with the number of guests. Secondly, to enable a better demand management, incentives might be provided in the slow time to attract more price-sensitive customers, while for the duration management the length of the meal might be adjusted according to the demand patterns. Considering the price, proper strategies might be implemented to take advantages of patterns and price elasticity estimations. Finally, the estimation of customers waiting time might be used to provide ways to distract or call them.

Chapter 3

A Private Airport Transfer Company

This chapter describes the region's characteristics, the company that, by providing a large volume of data on its customers and operations, enabled this study and the provided data.

Section 3.1 exposes an introduction on the evolution of Tourism in Algarve, describing the Algarve tourists' profile accompanied by a short presentation with some numbers and trends that characterize Faro airport. In Section 3.2 the private airport transfer company is presented, followed by the dataset characterization presented in Section 3.3.

3.1 Tourism in Algarve

Algarve is an area where the economy relies heavily on the tourism industry. In 2017, with an increase of over 5% relatively to the previous year, Algarve achieved the first place in Portugal for overnight stays. The average stay for this region was 5 nights for foreigners and 3,6 nights for natives (Turismo de Portugal, 2018). Algarve is also the region of Portugal with the greatest amount of visits per motives of leisure, recreation or holidays, according to Instituto Nacional de Estatística (2017).

3.1.1 Algarve Tourist's Profile

A great part of tourists that visit Algarve might be described by some typical characteristics. According to a study by (Universidade do Algarve, 2017) that inquired tourists in the summer of 2016, over 85% of them come to this region by virtue of holiday and leisure motives, which reiterates what is referred in the previous section, and over 50% travel with family. As a consequence of being a popular beach destination the same source indicates that more than 60% of Algarve tourists choose to travel in the Summer, over 40% of them do it once a year, and over 80% want to stay in seaside areas.

Regarding the country of residence of the international tourists, the United Kingdom is the most important origin country. 94% of United Kingdom tourists' that visit Algarve arrive by airplane, the british tourists plan their holidays 122 days ahead, in average, and most of them do it online. Algarve relies on a long and well-known relationship with British tourists that is sustained

by the statistics. 82% of the inquired tourists had already visited Algarve before and over 30% had their first visit before the year of 2000. (Universidade do Algarve, 2017)

These facts indicate that Algarve is a destination with high seasonality that is sought by its beaches and good climate. It deals with a high clients' repetition rate, as a large percentage of them goes back to the region at least once a year.

3.1.2 Faro Airport

Faro Airport, located in Faro, serves the region of Algarve, in the south of Portugal, and some parts of southern Spain. It is currently the third biggest portuguese airport regarding passenger traffic, after Lisbon and Porto. However, not long ago it was the second biggest airport in Portugal, as until 2010 the annual passenger traffic in this airport was greater than in Porto. (Francisco Manuel dos Santos Foundation, 2017)

The main routes of Faro airport consist of connections with airports located in the United Kingdom and Ireland followed by the Portuguese capital, Lisbon. As an example, by the end of May 2018 the top route in Faro Airport, according to Flightradar24 AB (2018), is London Gatwick with 60 flights per week. This airport is followed by the airports of Dublin (with 50 flights), Manchester and Lisbon, having the last one 32 flights per week.

3.2 Company Description

3.2.1 Airport Transfer Companies

An airport transfer company provides a booked transportation between the airport and the client's final destination and is usually prepaid. This transportation solution is offered in most major airports, which frequently arrange a number of shuttle services options (Kelly, 2017). However, the company under study is a private transfer company, which means that besides the mentioned characteristics, it provides a private and exclusive service to each customer.

It is important to emphasize that there are numerous companies performing similar services in Faro Airport which intensifies the concern about granting an excellence service.

3.2.2 YellowFish

The company is located in Albufeira, in Algarve (Portugal) and performs private transfer trips between two different places with only one client at once. Its activity started in 2010 with 5 cars and has experienced a big growth counting, in the beginning of 2018, with a fleet of 106 vehicles, comprising estate cars, people carriers, convertibles (cabriolets) and minibuses, all with a capacity to transport up to eight people (Yellowfish Travel, 2018).

Most of the trips are airport transfers, which means that the clients are transported between the airport and other location, mostly hotels and resorts. These are principally placed in the Algarve region but, occasionally the company serves Lisbon, Alentejo or some regions in southern Spain. Besides the airport services, the company also performs services between different locations or

golf courses. The majority of their clients book the service through the online platform in the company's website but can also rely on the company's helpdesk service anytime.

The main goal of the company is to perform a high quality service, without delays and according to each client special needs. One way to achieve this excellence is through the drivers' skills and attitude. All drivers should have good English level and knowledge about the areas where they are operating. It is also crucial for them to be friendly and professional and to keep a dynamic and motivated attitude.

The referred business copes with high seasonal variations and the demand might depend on some complex and unpredictable factors, as the weather, the country economic situation or even one-off events that might occur. Moreover, the company has to deal with fast changing drivers' availability and to be able to efficiently allocate the services to the drivers and the cars. The full-time drivers, who provide their availability weekly, are assigned a car and usually take it to their home. If they do not have full availability, there might be part time drivers that share the vehicle with them.

3.3 Data Characteristics

Data concerning the booking details, vehicles, drivers, car usage, locations, clients and feedback were provided by the company. In order to extract knowledge from the data, it is crucial to understand the database and its characteristics.

Due to the fact that the data is provided in different CSV (Comma-separated values) files, the relationship between them is not directly withdrawn. For the sake of the absolute understanding of the intrinsic relations in the data, an UML diagram is created applying reverse engineering techniques. The diagram, presented in Figure 3.1 aims to represent the different datasets that were provided by the company and to disclose the relationships that are believed to connect them.

To properly present the wide amount of data acquired, the UML diagram is now explained. In the center of the diagram, the service represents each trip, which represents the client's need that the business satisfies.

The upper part of the diagram is related to operational data, regarding availability of cars and drivers, expenses and allocation of trips. Starting with each car, the respective expenses are aggregated as gas, toll or parking receipts along with the information on the value and the date in which this cost is undertaken. The available cars that might serve the clients are allocated to the drivers (one full-time and possibly one part-time) considering their availabilities stored in *car_usage* that also informs the day and hours at which the car is assigned to each driver. However, when the number of bookings exceeds the capacity, the service is allocated to an external supplier, instead of a Yellow Fish solution provided by a *car_usage* combination. Another possibility is that the service is booked for a number of passengers that exceeds the company's maximum number of seats per car. This is solved by splitting the booking of that specific trip in more than one service to enable the allocation to different *car_usage* combinations.

Additionally, each service is characterized by two locations: the *pickup* and *dropoff* and, for the most accurate determination of the route, the resort or accommodation might even be specified. Moreover, a service corresponds to a wage that, considering the *pickup* and *dropoff*, is subsequently attributed to the driver who performed the service. Finally, each service is connected to the booking in which it was reserved.

Through the *book_id*, the client's personal data might be extracted and consequently, his country of origin. After the service, the experience's feedback might be provided by the client, quantifying some aspects, such as the drivers' punctuality and courteously or the quality of the trips, and comments.

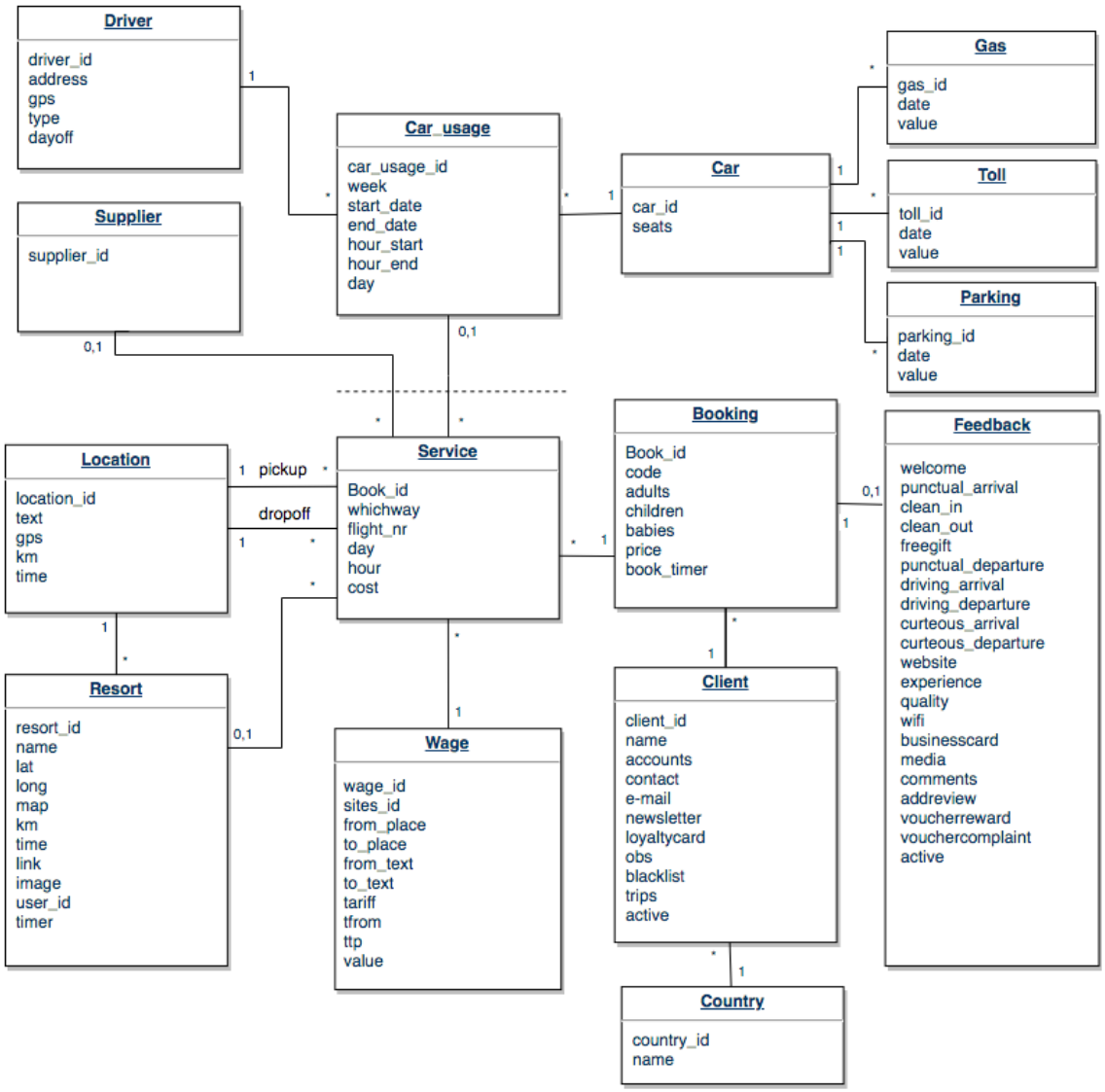

Figure 3.1: UML Diagram of the services database

The main dataset provided regards services and booking data, the company provided 317413 records that represent their operation between 2012 and 2017. This dataset is described by the parameters in Table 3.1, where each attribute is described and the data type is characterized and

each rows regards one different service. In its initial state the data only takes the types integer and categorical. Not all the data provided by the company is available for the entire six years of operation; data respecting the availability of drivers and their allocation to each vehicle and the *price* of each trip performed are only accessible for 2017.

From these attributes some require deeper explanations, due to the complexity of options or specificities of the company. *Book_id* might refer to one or more trips as every return service regards two trips and one client is capable of booking different trips in the same reservation This leads to replicated values in the attribute in discussion. Considering the *whichway*, it takes the value of *departure* when the destination of the trip is 'Faro Airport" and, on the other hand, if the origin is the airport it takes the value of *arrival*. Similarly, *code* refers to the type of service performed, taking the values of RAL, OAL, OLA, OLL, RLL or GOF. Respectively, these acronyms stand for return airport location, one-way airport to location, one-way location to airport, one-way location to location, return location to location and golf services. As explained previously, a return trip means a round trip where both arrival and departure trips are booked simultaneously. Airport refers to a 'Faro Airport' *pickup* or *dropoff* and location regards any other place besides it. Thereby, RAL is a round trip to or from the airport, OAL and OLA are one-way trips from and to the airport, respectively. OLL and RLL are, respectively, one-way and round trips between two locations and GOF regards golf services. Finally, *operators_id* identifies the agent to whom the reservation belongs.

Table 3.1: Attributes description

Attributes	Description	Data Type
book_id	identification number of the booking	Integer ($\mathbb{Z}_{>0}$)
whichway	direction	Categorical
flight_nr	flight identification number	Categorical
clients_id	client identification number	Integer ($\mathbb{Z}_{>0}$)
code	type of service performed	Categorical
day	day of the service	Categorical
hour	hour of the flight	Categorical
pickup	pickup location	Categorical
pickup_gps	geographic coordinates of the pickup location	Categorical
dropoff	dropoff location	Categorical
dropoff_gps	geographic coordinates of dropoff location	Categorical
adults	number of adults (> 13 years old)	Integer ($\mathbb{Z}_{\geq 0}$)
children	number of children (3 to 12 years old)	Integer ($\mathbb{Z}_{\geq 0}$)
babies	number of babies (< 3 years old)	Integer ($\mathbb{Z}_{\geq 0}$)
car_id	identification number of the car	Integer ($\mathbb{Z}_{\geq 0}$)
driver_id	identification number of the driver	Integer ($\mathbb{Z}_{\geq 0}$)
supplier_id	identification number of the supplier	Integer ($\mathbb{Z}_{\geq 0}$)
book_timer	date and time at which the booking is done	Categorical
pickup_place_resort	hotel or apartment name	Categorical
dropoff_place_resort	hotel or apartment name	Categorical
operators_id	identification number of the agent	Integer ($\mathbb{Z}_{>0}$)

3.4 Summary

This chapter's main goal is to describe the problem and its specificities.

From the information provided in this chapter, it is critical to emphasize the high percentage of tourists that prefer to enjoy this region in the Summer and who visit it at least once a year. Moreover, a considerable number of these tourists come from the United Kingdom and almost all of them arrive by airplane. Accordingly, the main routes are established between Faro Airport and British and Irish airports.

The case is studied with data from a private airport transfer company that provides transportation services, exclusive to each client, being most of them between the airport and other locations in Algarve. The data consists of several datasets and the relations between them are represented in an UML diagram. Finally, the main dataset attributes are explained with more detail.

The next chapter approaches the first steps of the methodology tested for this case, preprocessing and exploratory efforts, and its results.

Chapter 4

Preprocessing and Exploratory Analysis

This chapter arises in accordance with the KDD process, presented in Chapter 2.

Firstly, the Data Cleaning process is conducted and the methods and results are presented followed by the integration of data. Secondly, to better understand the intrinsic characteristics of the data, the exploratory analysis is added to the process (Han et al., 2012) and the main conclusions are revealed in this chapter. These results lead to conclusions on the data mining techniques that should be used. The remaining steps of the process, which consist of Data Selection and Transformation, are also described in this chapter. The steps and results presented in the chapter are crucial to derive conclusions on the attributes to be studied in more detail and the models to be used. Consequently, the necessary input for Data Mining, Pattern Evaluation and Knowledge Presentation steps, which represent a great deal for the project and are presented in more detail in Chapter 5, is created. Finally, this chapter ends with a brief overview of the software that was used in this project and a summary of the chapter's conclusions.

4.1 Data Cleaning

The files with the booking and operational details described previously demand that data cleaning processes are carried out. The provided data represents a long period in which the operations experienced some changes which lead to some differences in the data structure. Additionally, a considerable part of these data is filled by customers throughout the reservation process and consequently, mistakes and different notations. Besides, some reservations might be done by direct contact, where the website is not used and as the form is filled by the client. To perform a reliable analysis the quality of the data is crucial and was achieved by the process described in the next paragraphs.

The process starts with the deletion of duplicated rows and of entries in which the *vehicle_id*, *driver_id* and *supplier_id* are all equal to zero. These report services that did not actually exist as all the parameters are duplicated or no car, driver or supplier is allocated should, consequently, not be considered. This step results in a reduction of 9201 entries, leading to a dataset of 308212 rows.

Subsequently, all the columns are studied, one at each time, to find possible mistakes, their explanation and the best correction action.

In categorical variables as *code* and *whichway*, and where a small range of options is available, spelling mistakes or null values are corrected based on the *pickup* and *dropoff* and the number of rows with the same *book_id* that indicate the correct way and type of service. As an example, if the *code* is null, it might be filled based on the verification of the uniqueness of the *book_id*, and the *pickup* and *dropoff* values.

Regarding integer variables, such as the number of passengers (*adults*, *children* and *babies*), negative values that were found and do not comprise the range of accepted values, as it is impossible to transport a negative number of passengers, are replaced by the mode. The existence of entries with zero adults is also intriguing but explained by the division of big groups in different cars or by the transport of only luggage, which is also possible. Finally, the *flight_nr*, *pickup_gps* and *dropoff_gps* are processed in order to standardize the entries and remove redundant levels.

Additionally, cases as the *client_id* equal to zero, *book_timer* posterior to the service date, *pickup* and *dropoff* represented by acronyms are derived from the operational changes referred previously and only appear in the first years of operations, before 2014. Due to the big obstacle that might become to the analysis to be performed, the solution to these problems is given by the separation of the data into two different datasets with the following description:

- Although the first dataset concerns data from the whole period, it mainly allows the reliable study of the evolution of the demand, number of clients or bookings over the time and the reservation date, resulting in 308211 observations of 8 attributes, *book_id*, *whichway*, *client_id*, *code*, *day*, *hour*, *book_timer* and *operators_id*;

- Regarding the remaining attributes, the second dataset presents data from 2015 to 2017, resulting in 206454 rows and 23 columns. This allows a more detailed study of the clients and their origins, the number of passengers, the length of stay, the drivers and vehicles utilization or the need of suppliers and the most common clients' destinations in Algarve.

The result of these cleaning processes, regarding number of rows and columns, is more explicit in Figure 4.1.

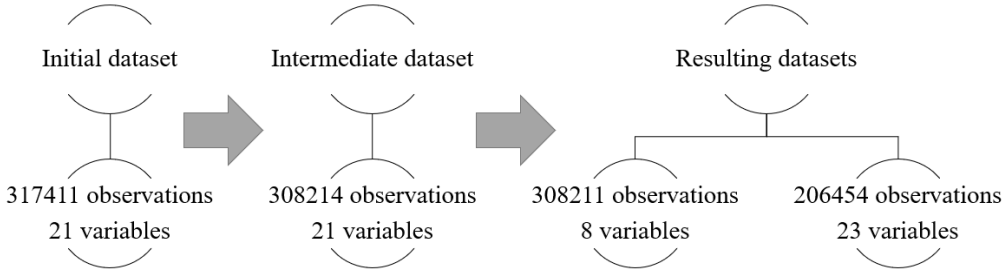

Figure 4.1: Cleaning process transformation's output

4.2 Data Integration

Regarding data integration, it is pertinent to stress out the various operations that need to be conducted in order to successfully join information from different tables.

Both Exploratory Analysis and the preparation of data for mining benefit from these integrations. Based on the UML diagram presented in Chapter 3, any analysis that requires information from different classes needs data integration efforts to relate the classes and properly extract the information.

For the analysis of the clients, it is necessary to join the table that comprises the bookings and trips to the table with the clients' information. This is achieved with a left join where each *client_id* from the booking information is joined to the corresponding client information.

Regarding data that needed integration for the analysis performed in Chapter 5, the proper extraction of the kilometers associated with each service performed, which requires the combination of each service with its location, is required. However, as location's kilometers are measured between each location and the airport it is also necessary to filter only the airport services and to extract only the *pickup* or *dropoff*, dependently on the direction of the service. For the correct integration of different data is crucial to understand and carefully examine the referred data, to avoid mistakes or redundant data.

The exploratory studies of the data, which are based on the two datasets that arise from the data cleaning processes and rely on integrated data, are presented in the next section.

4.3 Exploratory Data Analysis

Through the extraction of insightful information from the data, an exploratory analysis is conducted, which is essential to understand the trends, the company's operations and to prove some of the previously described touristic data. The analysis is divided into Booking details, clients, locations and operational details, which regard cars and drivers availability.

Booking details

Firstly, the services performed in the last six years of operations (2012 to 2017) are studied. In Figure 4.2, where the number of services performed in each week throughout each year of operations is displayed, the growth of the company in the last years is clear. However, it is important to emphasize that this growth is much more evident in the summer months, as the difference between years increases. Consequently, the figure also highlights the annual seasonality of the company's demand, which was already foreseeable when the tourists' preference for the summer months was stressed out in Subsection 3.1.1.

In the next stage, the differences in the number of services are spotted in the day of the week. In order to understand this trend the percentage that each day of the week represents in the total amount of services per year is explored along with total values for the six years in study. As it is seen in Figure 4.3, Saturday is the day with the highest demand, satisfying 18.34% of the

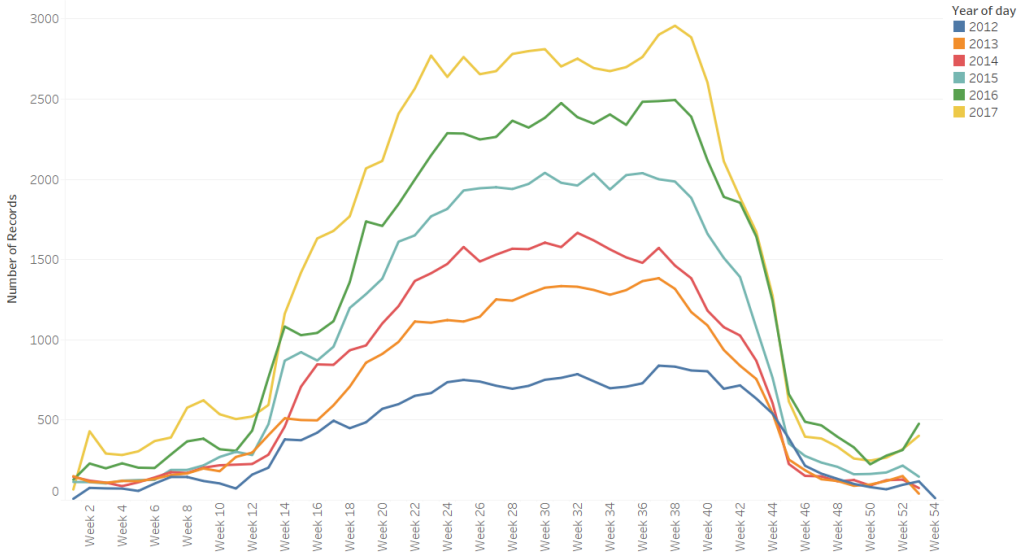

Figure 4.2: Number of weekly services over the years

total amount of trips, followed by Sunday, Thursday and Friday. The less common weekday is Wednesday. This information displays the increase in the demand that is noticeable towards the weekend. However, it is also conclusive that the difference in the number of services and increase during the weekend is more evident in the first years of operation, the range between the most and the less busy day in 2012 is 10.77%, which is high compared to the difference of 5.37% in 2017. Thus, along with the annual seasonality, the company has to deal with slightly different daily demand throughout each week.

	Monday	Tuesday	Wednesd..	Thursday	Friday	Saturday	Sunday
2012	12.02%	11.83%	9.62%	14.68%	11.87%	20.39%	19.59%
2013	12.77%	11.52%	11.18%	15.97%	12.51%	19.18%	16.87%
2014	13.04%	12.53%	11.59%	15.92%	13.49%	17.94%	15.50%
2015	12.97%	12.48%	12.16%	14.66%	14.25%	17.38%	16.10%
2016	13.20%	12.78%	12.12%	14.20%	14.28%	18.73%	14.69%
2017	13.32%	12.55%	12.93%	14.13%	14.38%	17.92%	14.77%
Grand Total	**13.03%**	**12.41%**	**11.97%**	**14.74%**	**13.81%**	**18.34%**	**15.70%**

Figure 4.3: Weekday percentage per year

In Figure 4.4 a difference in number of airport services according to the hour of day at which they occur and the type of service that represents which way the service takes is noteworthy. This figure, where only airport services are considered, represents the hours of the day with most services performed according to the direction of the trip and the month of the year. It is noticeable that departure trips start earlier than arrivals, as clients need transport to the airport for early flights but there are no arrivals early in the morning. Analogously, clients do not need to be transported to the airport after 8PM as there are no more departures, but night evening flights are still arriving to the airport, demanding more arrival services from the company. This represents one of the company's operational problems as the existence of successive multiple trips in the same direction

implies the need of journeys with an empty vehicle. Additionally, as well as in Figure 4.2, the increment in the demand present in the Summer months is remarkable.

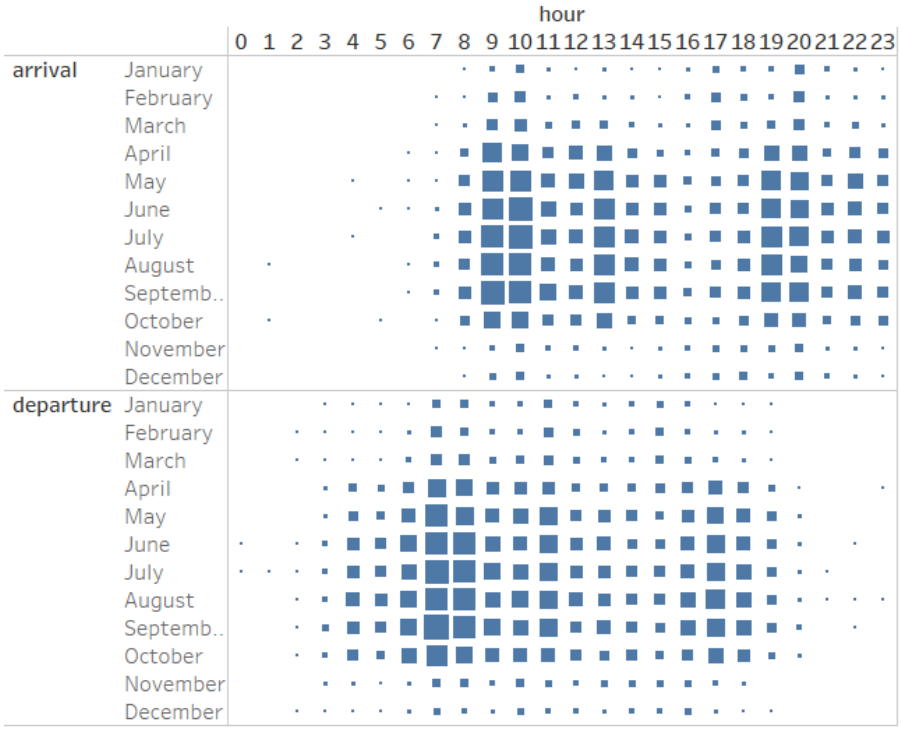

Figure 4.4: Map of demand per type of service and month vs hour of day

It is also important to draw attention to the different types of services. The first graph in Figure 4.5 represents the monthly aggregate of the totality of the services performed in the six years by each different code, while the second graph presents only the three codes that have a lower dimension in order to detect their variations in a proper scale. As it is shown in Figure 4.5 the large majority of bookings are for airport codes. RAL, means return airport-location - regarding bookings that encompass an arrival from the airport and subsequent departure - is the most common, achieving almost six thousand monthly bookings in the summer of 2017. Following RAL, with similar dimension, OLA and OAL - one-way codes to and from the airport, respectively - achieved about 1500 monthly bookings each, in 2017. On the other hand, when it comes to services that do not serve the airport (RLL and OLL - return and one way between two locations - or GOF - Golf Transfers) the second graphic reaches lower proportions, with a maximum of 115 bookings for OLL in July 2017. Finally, Golf Transfers only started in 2014 and result in a tiny segment of the overall services.

Notwithstanding, it is meaningful to stress the increase in the variance of the daily services. The boxplots in Figure 4.6 present the distribution of the number of daily services per year, which shows that, in addition to the undeniable yearly growth, the maximum number of daily services has been increasing while the minimum number does not follow a similar trend. This range has been expanding which enhances the seasonal effect. Over the last few years, the company had to

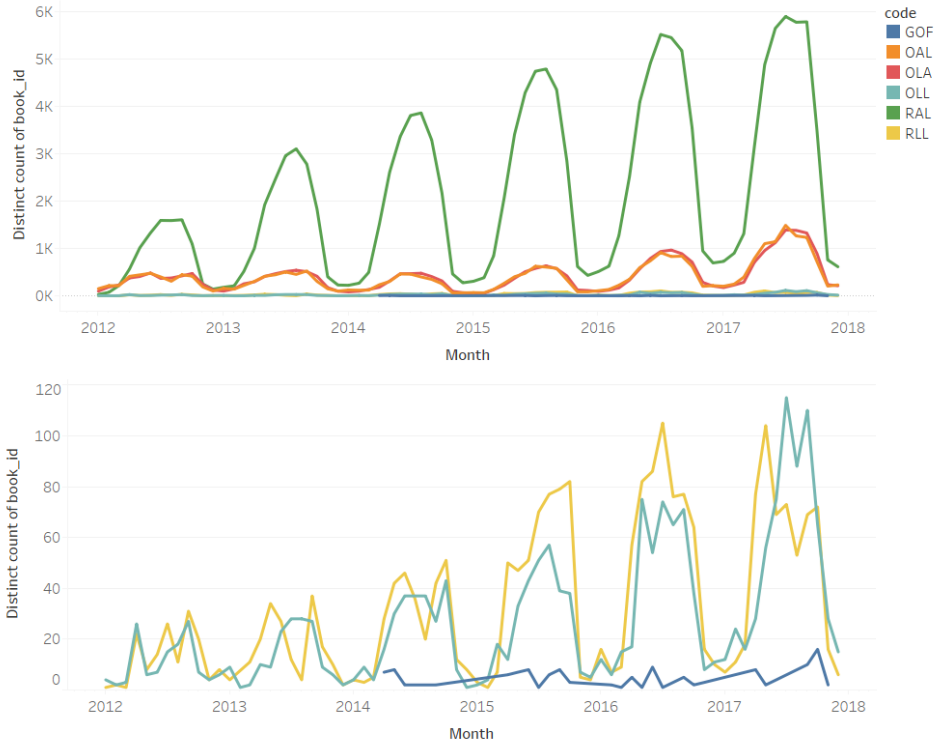

Figure 4.5: Number of bookings per code

deal with a larger range in the number of daily services, which imply more flexibility and more operational differences between the Winter and Summer months.

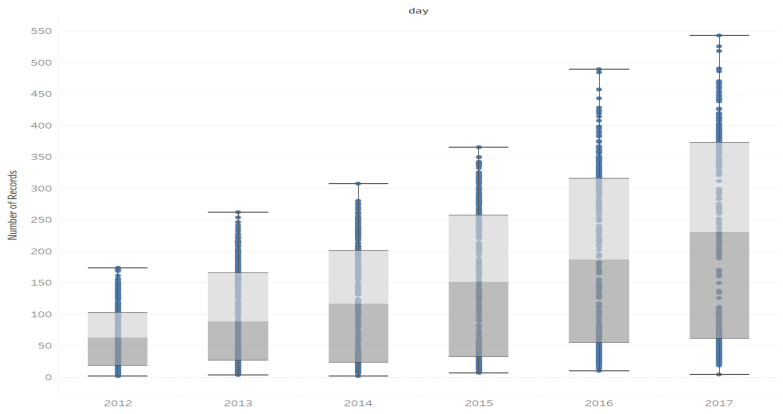

Figure 4.6: Boxplots of number of daily services per year

Finally, in order to understand the antecedence of bookings, the *book_timer* is subtracted to the day of the service. This integer attribute is named *difference*. In this analysis some services that exhibited a difference of zero or less and resulted from reservations that are not made in the website and might be added to the database later on, were filtered. Thus, this analysis regards only bookings that are made until the day before the trip.

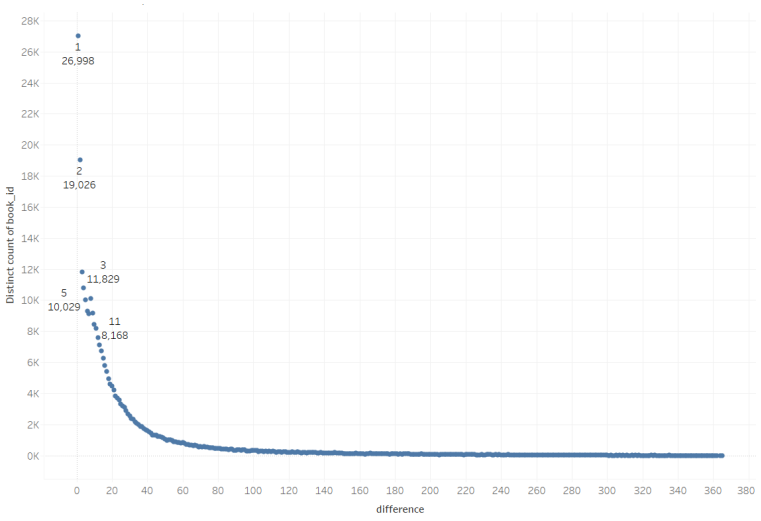

Figure 4.7: *Difference* scatter plot

difference	
1	9.269%
2	6.414%
3	3.952%
4	3.601%
5	3.340%
6	3.085%
7	3.035%
8	3.358%
9	3.039%
10	2.803%
11	2.713%
12	2.518%
13	2.359%
14	2.247%
15	2.078%
16	1.923%
17	1.797%
18	1.644%
19	1.532%
20	1.484%

Figure 4.8: *Difference* percentage

In Figure 4.7 the antecedence of booking is plotted against the number of services of each of these differences. Here, an approximation to the exponential negative distribution is depicted. The most common value is one, with 26998 bookings that are booked only in the day before the transfer. Even though some transfers might be reserved about one year before, it is not common as the number of bookings in advance decreases when the number days rises.

Figure 4.8, displays the percentage of the bookings that are done with less than twenty days prior the trip. Here, once again, the trend of less trips booked with greater advance is observable. About 33% of the services are booked during the week before the service, although the booking of a transfer eight days before the need, is more common than between five to seven.

Understanding the Clients

When it comes to the clients it is interesting to understand which are the most common countries of origin. This attribute is based on the country code, which comes from the indicative number of each country for mobile phone calls. Figure 4.9 displays the relative size of the different countries according to the total amount of bookings. This shows that 65% of bookings are made by clients from the United Kingdom (UK), 24% from Ireland, almost 4% from Portugal, being the 31 and 49 codes representing the Netherlands and France respectively. Even though, according to the country code, Portugal is the third country with more clients, when looking at the clients' name it is noticeable that most do not correspond to Portuguese names which might be caused by foreign clients who already have a house or a Portuguese phone number. There is also a small percentage of clients, less than 1%, who do not provide information about their country of origin.

Regarding the number of passengers, an interesting point is the distribution of the number of children and babies. Figure 4.10 shows the total number of passengers per category for each month. This refers to the total sum of each month in the six years. The categories are adults,

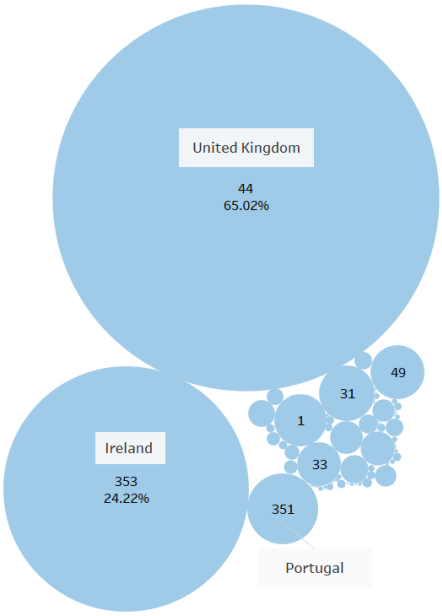

Figure 4.9: Packed bubbles of *country_id*

children and babies, combined with the number of passengers, which is the sum of the first three. The number of children and babies increases significantly in the months of July and August, most likely by virtue of the school holidays, while the number of adults decreases relatively to the previous months. During the rest of the year it is less common to see entire families together.

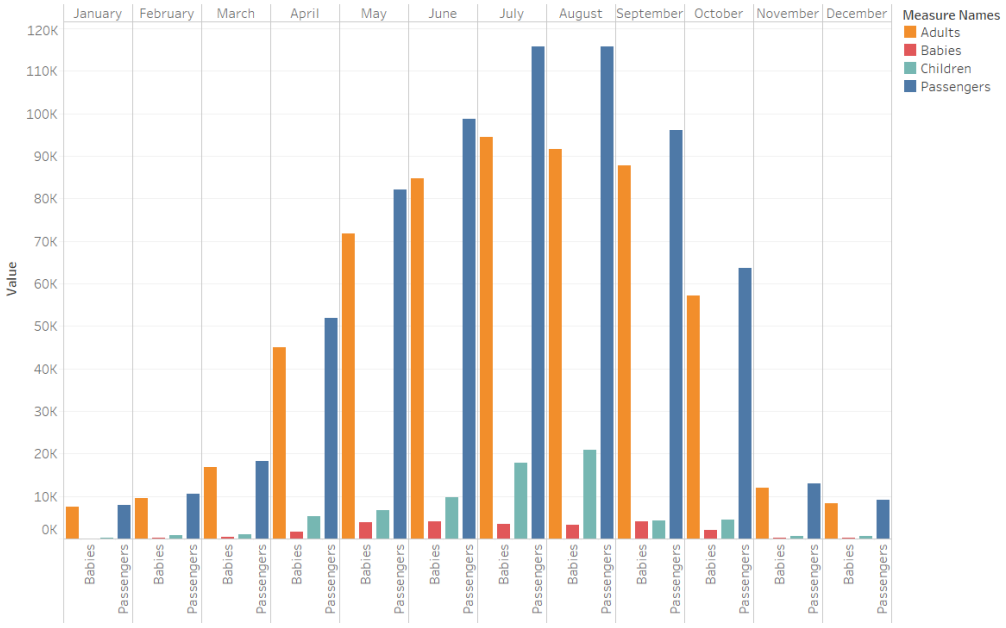

Figure 4.10: Number of passenger per month and category

Understanding locations

Figure 4.11 displays a heat map of the most frequent locations of airport services. There is evidence that the area between Albufeira and Vilamoura, where the red zone is seen, is the most frequently visited by tourists. There are also some green and blue areas that are commonly visited. For a further analysis, Appendix A presents the most frequent places used in airport services with the percentage of each of them per year and *whichway*. In the appendix it is observable the equity between both ways and the difference between Albufeira and Vilamoura and the remaining locations. Albufeira is the main location, displaying about 30% of the totality of airport services and Vilamoura comes right after it with less than half the amount, with about 14% of services. Additionally, *Albufeira* regards the city but more beaches, as *Praia da Oura*, *Gale* and *Falesia* are considered different locations for the company but are also located in Albufeira, which intensifies the importance of this area. Moreover, it is perceptible that, independently of the growth that the company is subject to, the proportion of each location is similar throughout the years.

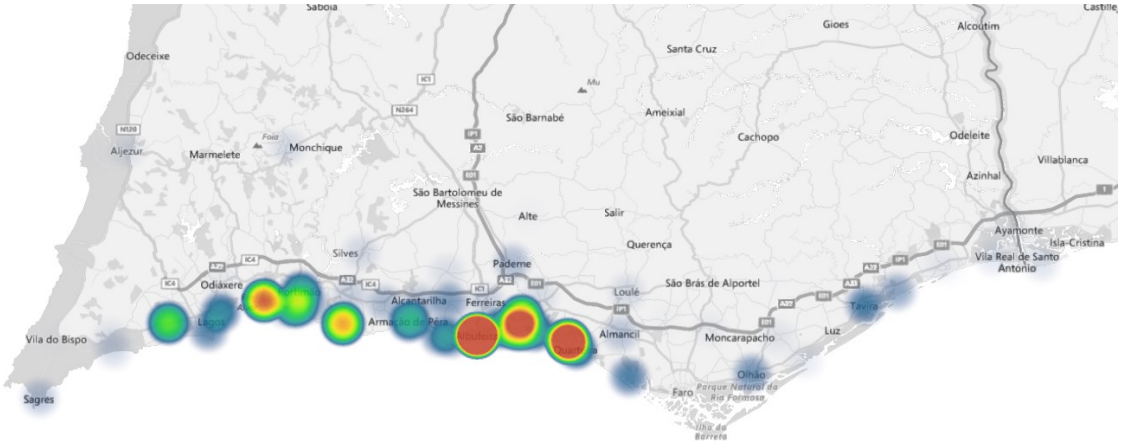

Figure 4.11: Heat map of most frequent locations in Algarve

Operational Details

Finally, a count of distinct drivers that execute a service in each day by type of work, is presented in Figure 4.12. Even though the amount of full-time drivers prevails, both types follow similar distributions.

Regarding the drivers, it is also relevant to state that the driver's average number of daily services is 4.787. However, this is quite different in months as December, where this same average stands for 2.996 or September with an average of 5.697 daily services per driver.

The vehicles demonstrate similar results. Even though the average of daily services performed by each car is 5.37, in the Winter this average is much smaller than in Summer months. These

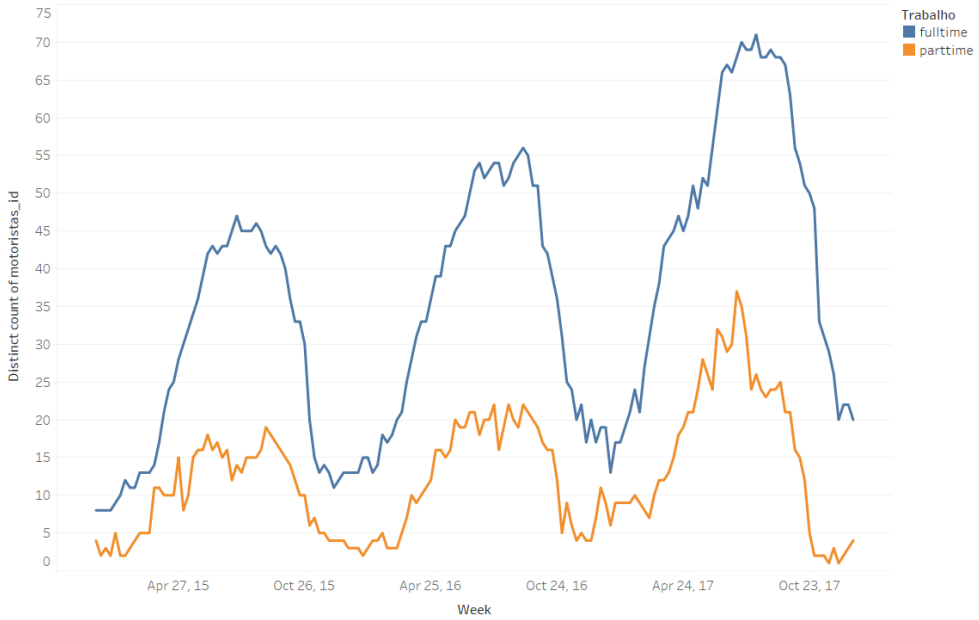

Figure 4.12: Distinct count of *driver_id* per day and type of work

values are, withal higher than the ones regarding the drivers which is originated by the possibility of the usage of one car by both a full-time and a part-time driver in the same day.

Taking into account these conclusions, the methodology for the patterns recognition, prediction and evaluation is defined and leads to different needs for the next phases of the process. Considering the most influencing factors the next sections will guide the reader through the remaining steps of the path that forwards to the prepared data.

Flights and Airport's analysis

As the big majority of the services performed involve the airport, the *flight_nr* attribute is of great interest. However, despite the cleaning efforts, there is no access to the flights number of all the flight from Faro Airport. The company provided some data on these flights to enable the analysis but, after some data preprocessing it is concluded that the acquired data only regards the period between April and October of 2016.

To extract a possible relation, these data are aggregated into daily number of flights and compared to the daily number of services performed by the company for the exact same period. The correlation between these values is, approximately, 86.13%.

Additionally, data on the Faro Airport's passenger traffic is acquired in www.faroairport.net (2018) in a monthly basis for the whole period, between 2012 and 2017. The comparison between these data and the total number of services in each month gets to a correlation of, approximately, 94.25%.

This analysis is limited as there is not sufficient data to extract more insightful information. By comparing the number of flights to the number of services it should also be taken into consideration

the size of each plane as the number of services might depend not only on the number of flights but on the number of transported passengers.

4.4 Data Selection

The selection of the most useful data, namely data that is believed to be explanatory of the target variable is of major importance for the continuity of a clean and organized study. With this knowledge it is possible to select the data that deserves further exploration and to avoid the existence of counterproductive information in the following steps.

In this case, where the goal is to predict the future demand of the company, the study of time-series data reveals to be the most effective field. This leads to the need of selecting the first data set to study the frequency of services and different clients, which is believed to adequately demonstrate and recognize the evolution of the demand throughout the six years. Additionally, the number of kilometers might enable the detection of different patterns in the size of the services, which would lead to a big impact in the company's revenue and operations. If this difference is confirmed this variable represents compelling input for the models. However, the kilometers might only be extracted for the last three years of operation and only regard airport services, which represents a limitation for its use.

To extract these data, the urge to aggregate into different time horizons arises and additional integration efforts are crucial. For the annual data the services and different clients are aggregated into years. The number of services is achieved by counting the number of trips performed, while for the clients it is necessary to extract the year and the client_id of each trip, eliminate the duplicates and count the remaining entries. Regarding monthly or weekly data, which would enable the detection of the seasonality along with the trend, similar processes are conducted but require the transformation of the date indicating the variable in need. Additionally, for these time periods, the aggregated sum of the kilometers is also extracted from the dataset.

4.5 Data Transformation

Finally, and in order to examine the time series data with decreased variance, a logarithmic transformation is performed. As it is further explored in the next chapter, the monthly and weekly series concern a multiplicative series with trend and seasonality, where the variance increases over the year. By applying the logarithmic transformation, as stated by Nogales et al. (2002), the data is normalized and a more homogeneous variance is achieved.

4.6 Software

The main software used in this project is RStudio® for data cleaning and other preprocessing stages, for the statistical analysis and for creation of the predictive models. Regarding visualization, Tableau® is used for the exploratory analysis of data and to create tools that enable the

understanding of the main characteristics from the case study.

4.7 Summary

All the steps, from data cleaning to the discovery of the relevant data and its preparation are described in detail in this chapter. Data cleaning removes duplicated or impossible entries, spelling or data entry errors and fills out missing data and data integration that relates the different datasets. Distinctively, exploratory analysis presents insights that help to better understand the company's demand and operations. From this study, the seasonality issues are emphasized and its influence on the demand is conclusive. At that point, the selection and transformation steps are explained. Finally, the software used in the project and its applications are presented. Chapter 5 continues to the presentation of the models that indicate the patterns, their evaluation, the prediction and respective accuracy.

Chapter 5

Pattern Recognition and Prediction

The analysis exhibited in Chapter 4 concludes that the main factors that might affect the company's demand are the seasonality and the flights' schedules. Owing to this reason, along with the fact that the flights data is not sufficient to perform a reliable study, the best way to forecast the future values is through univariate time series analysis. This chapter presents the modeling of the series and the comparison between different forecasting methods.

The models are estimated using the volume of annual, monthly and weekly number of trips. The annual data is additionally studied based on the annual number of different clients but it is concluded that these grow at a similar pace. The aggregations of the number of different clients and of the sum of kilometers traveled, for each of the time frequencies in analysis, are proved to be highly correlated with the number of services, 99.81% and 99.72% respectively. Thus, the variable of greatest interest to the company and the project is the number of services being this the focus of this chapter.

5.1 Annual Demand

Firstly, the annual number of trips performed is extracted from the data, resulting in the aggregated data that is presented in Table 5.1 and plotted in Figure 5.1.

Table 5.1: Number of records per year

Year	2012	2013	2014	2015	2016	2017
Number of records	23213	35863	42680	54938	69489	82028

5.1.1 Proposed Model

As the plot of the number of trips performed per year seems to follow a linear distribution and the correlation between the years and the number of services is superior to 99%, a linear model is fitted in the data presented in Figure 5.1.

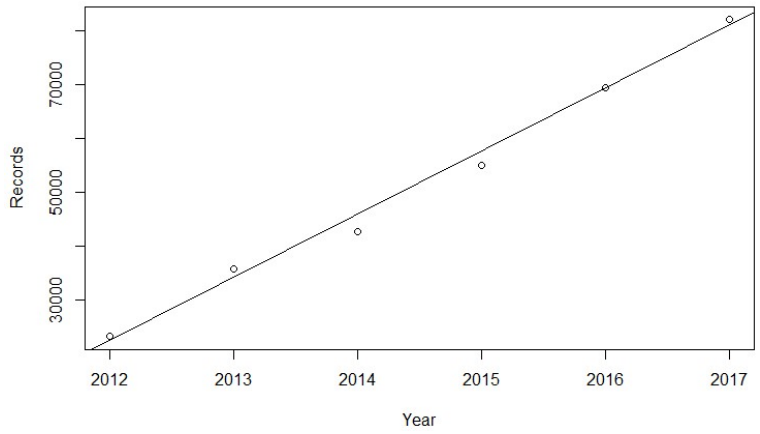

Figure 5.1: Annual records plot with regression line

In order to do so, the dataset is randomly divided between train and test data, allocating 20% of data to the test dataset, being the remaining data points used for training the model. This separation results in four non consecutively years for the train dataset which train the model that is subsequently compared to the test values. After applying the *lm()* function in R®, for fitting linear models, the resulting model is expressed by the Equation 5.1.

$$Y_n = -23517134.71 + 11699.64 \cdot X_n \tag{5.1}$$

5.1.2 Analysis of the Errors

The results demonstrate that the model fits the data considerably well. The R^2, which measures proximity of the data to the fitted regression line, is 99.24%. Table 5.1 presents the actual values, from the test dataset, and the estimated number of records.

Table 5.2: Observed vs estimated values

Year	Observed	Estimated
2015	54938	57645.64
2016	69489	69345.29

Besides, the number of different clients that booked with YellowFish is studied. Despite the high correlation, 99.90% that these values have with the annual number of services, both are studied. Likewise, this second model presents good results, as it is noticeable in Table 5.3, where the MAPE (Mean Absolute Percentage Error) of both predictions is exposed.

Table 5.3: Mean Absolute Percentage Error

Target Attribute	MAPE
Number of services	2.57%
Number of different clients	3.27%

5.1.3 Interpretation of the Results

The slope of the regression line from the number of different clients, 5266.64, is smaller than the one from the number of services, which demonstrates that the growth of the number of services performed has been more accentuated. Half the slope of the number of services would be expected. However, a value slightly above the double might indicate that there is also a growth in the number of trips booked by the same client throughout the time period in study, meaning that besides getting new clients, some clients have increased needs.

In conclusion, both results demonstrate that the annual demand has followed a linear trend which is useful to understand the growth of the company through the last years of operation. However, these are based only on six data entries and enable a prediction of long periods of time. To provide more complex and complete information, the time horizon in study is divided. This also allows a better understanding of the seasonality and the provision of a more complete forecast.

5.2 Monthly Demand

In order to deepen the study, aggregated monthly data of the number of services over the time period 2012-2017 is obtained. This data is converted into time series and the plot is presented in Figure 5.2. The aggregation results in a series with 72 observed values throughout the six concerned years.

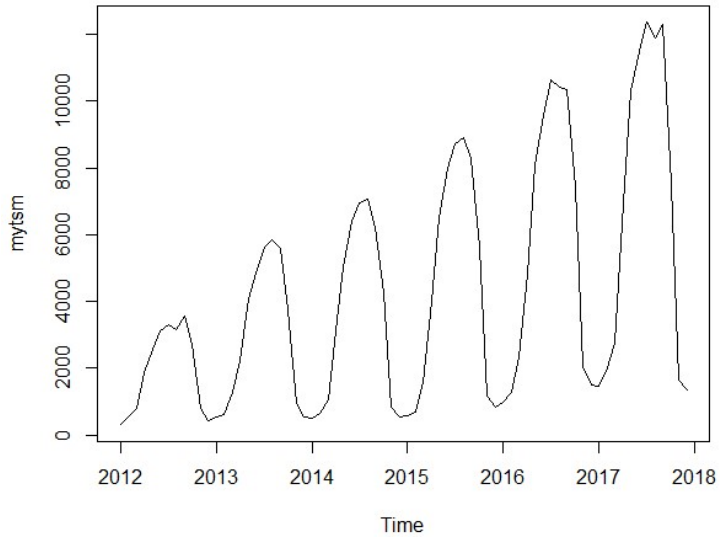

Figure 5.2: Monthly records time series plot

5.2.1 Time Series Analysis

As expressed in the Literature Review, a time series is composed of 4 components: trend, seasonality, cycle and randomness. Through the analysis of Figure 5.2, an upward trend and seasonality are easily spotted. In order to understand the behavior of the series without these

components, the random component is analyzed. Figure 5.3 presents the observed values together with the trend, seasonal and random components. As a consequence of its multiplicative character, the equation that characterizes the time series, given in Equation (5.2), allows the extraction of the random component by firstly detecting the trend, detrending the time series, averaging the seasonality and leading to the calculation and examination of the remaining random noise.

$$TimeSeries = Trend \cdot Seasonality \cdot Random \Leftrightarrow Random = \frac{TimeSeries}{Trend \cdot Seasonality} \tag{5.2}$$

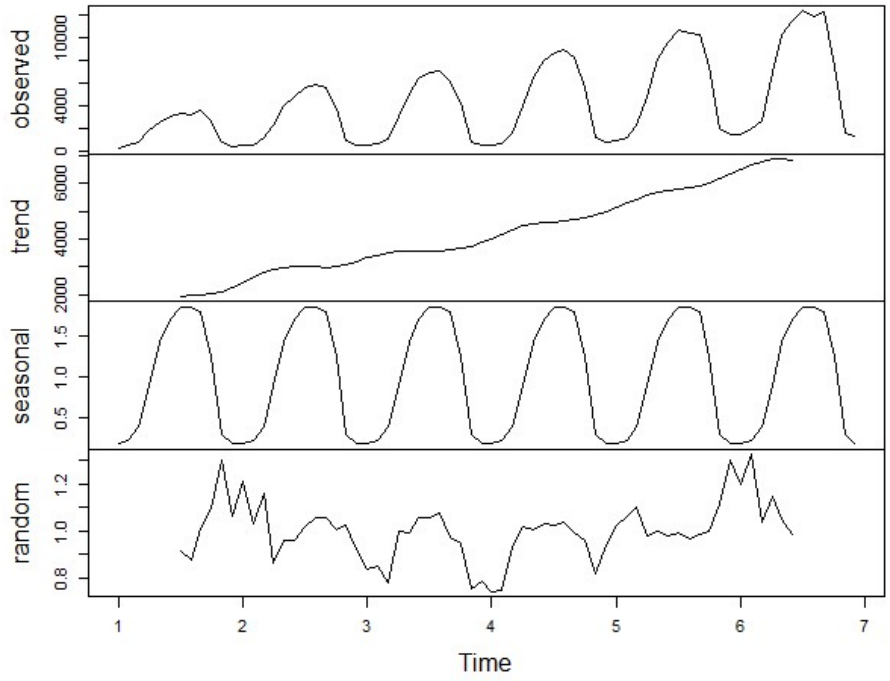

Figure 5.3: Monthly time series decomposition

5.2.2 Proposed Models

In order to model the series both Time Series and Artificial Intelligence models are fitted. As this is a complex series with both trend and seasonality it was decided to compare the results of an exponential smoothing model with an ARIMA and an ANN, in order to choose the models that achieve a better performance for the provided data. For the proposed models the data is divided into train and test series. The train dataset includes the data from 2012 to 2015, regarding the first four years of operation, while the test regards the last two years. This division results in a test dataset that covers two seasonal periods, making the accuracy metrics more reliable. Thus, in this monthly analysis the train dataset is composed by 48 observations and the test by the remaining 24. The errors are measured using the formulas in Chapter 2 by the comparison between the estimated and observed values.

Holt-Winters

As it is concluded that the data has a well defined seasonality and trend and the random component does not seem to have a big impact, the Exponential Smoothing might be a good method to fit these data. The data is multiplicative and there are 4 seasonal periods in the train dataset. Hence a model is fitted through the *HoltWinters()* function from the R® package *stats*. By providing the seasonality type (additive or multiplicative) the model is fitted automatically. The prediction is achieved by using the function *forecast()* from the R® package *forecast*, with the multiplicative *type*.

Logged Holt-Winters

A logarithmic transformation transforms a multiplicative in an additive time series, which is easier to fit for an Holt-Winters model. By logging the data it is transformed into an additive series, which means that the *type* is now additive and the $start.periods$ parameter is added. It represents the number of periods used for the auto detection of the start values and a simple linear regression on the trend component is used for starting level and trend. This value should be at least two and four is found to be the best value.

ARIMA

As the data comprehends a seasonal component, a SARIMA model is adequate. This is a complex model, given by $ARIMA(p,d,q)x(P,D,Q)_s$, and involves the selection of the parameters to find an optimal model. The seasonal period, s, was already defined and equals to 12. The process of selecting the remaining parameters is not direct and includes the following steps, defined by Makridakis et al. (1978):

1. Make the series stationary - An initial visualization of the data gives insights on the series stationarity. First or second order differencing (seasonal and/or nonseasonal) is useful to make the series stationary in the mean and a logarithmic transformation of raw data provides a series that is stationary in the variance;

2. Consider nonseasonal aspects - The examination of autocorrelation function (ACF) and partial autocorrelation function (PACF) plots reveals the existence of AR or MA components in the nonseasonal aspects of the data, which regard parameters p and q;

3. Seasonal aspects - The examination of ACF and PACF plots for lags multiple of the seasonal period, indicates the existence of AR and MA properties in the seasonality of the data and indicates possible P and Q parameters.

As it is noticeable in Figure 5.2, raw monthly data is non-stationary. In order to make it stationary both a first and a seasonal difference are necessary along with the logarithmic transformation. Only after these steps, and using the Dickey Fuller Test of Stationarity with R® function *adf.test()*

the null hypothesis of non-stationarity is rejected with a p-value smaller than 0.05. This indicates that the ARIMA model should have both d and D equal to one.

Regarding the AR and MA components, these are more difficult to estimate. Both Figures 5.4 and 5.5 have a difficult interpretation that leads to unclear conclusions. According to Makridakis et al. (1978) an AR(1) model is characterized by exponentially decaying autocorrelations and one significant partial, MA(1) is characterized by exponentially decaying partials and one significant autocorrelation, while AR(1,1) should demonstrate exponentially decaying autocorrelations and partials. When facing an AR(2) or MA(2), damped sine wave decays of autocorrelations or partials, respectively, and two significant spikes.

Considering non-seasonal aspects, an MA(1) would be adequate as the PACF follows an approximate negative exponential decay and one significant spike is noticed in ACF plot. On the other hand, the respecting seasonal aspects, an AR(1) model might be fitted as the autocorrelation has an exponential decay in lags 12 and 24.

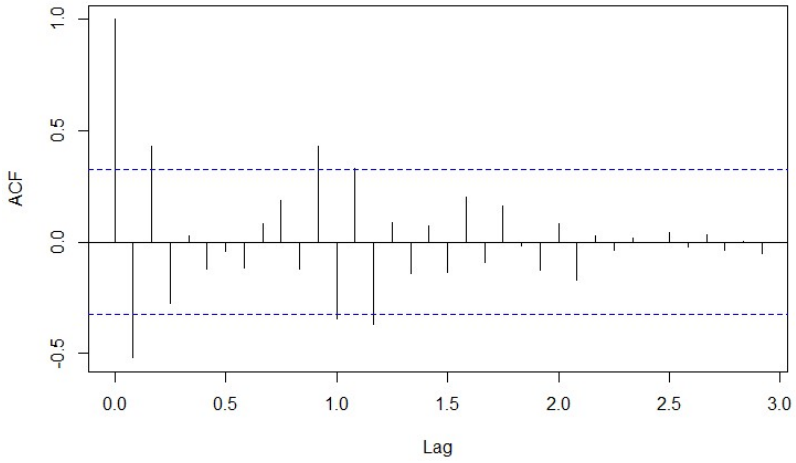

Figure 5.4: ACF plot of monthly data after transformations

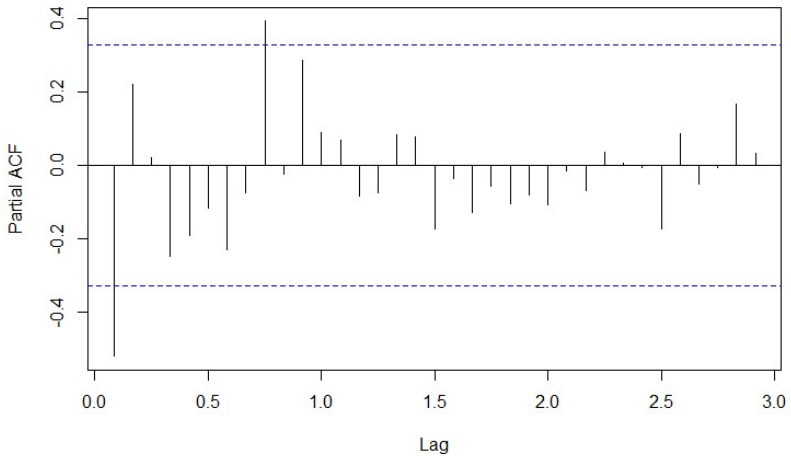

Figure 5.5: Partial ACF plot of monthly data after transformations

Therefore, the applied model is an $ARIMA(0,1,1)x(0,1,1)_{12}$. It is important to emphasize that this only indicates a possible model and not necessarily the optimal one. Other models could be fitted to these data as the estimation of parameters does not lead to clear and unquestionable conclusions.

Artificial Neural Networks

ANN model is fitted by using the package *forecast* with the function *nnetar()* which stands for Neural Network Time Series Forecasts and performs a feed-forward neural network with one single hidden layer and lagged inputs for forecasting univariate time series.

The model parameters are P, which represents the number of seasonal lags used as inputs; *size* which stands for the number of nodes in the hidden layer; and λ, the Box-Cox transformation parameter, that when taking the value of zero corresponds to a logarithmic transformation. According to Zhang and Qi (2005), the number of seasonal lags included in a neural networks model might be fairly small, even though, theoretically, many seasonal lagged observations may be included. It is also known that the value of P for the used function should be at least two. Both the values of P and *size* are defined by an iterative process. After 3 runs, with different values of P between 2 and 4, the value that fits a best model is $P = 3$. Regarding the number of nodes in the hidden layer *size* $= 35$ is chosen after several iterations. This is more than half the number of inputs, which is the default used by R®. Finally, $\lambda = 0$ which performs a logarithmic transformation and makes the series easier to fit, as proved in the previous methods.

5.2.3 Models Comparison

This subsection presents the analysis of the results achieved with the different models. It is given in terms of MAPE, RMSE and MPE, which enable the evaluation of the difference between the forecast and actual values, the standard deviation of the residuals, how far data points are from the model, and the difference between forecasting and test dataset but considering the direction of the error.

MPE provides insightful information on the bias of the error, even though opposite direction of the errors might cancel each other out. On the other hand, MAPE and RMSE are measures of variability. It is important to consider that MAPE is independent of the scale of measurement, but affected by data transformation.

Table 5.4: Table presenting monthly forecasting accuracy for different models

Methods	MAPE	RMSE	MPE
Holt-Winters	8.21%	466.97	2.68%
Logged Holt-Winters	1.69%	0.18	1.24%
ARIMA	14.42%	1202.87	7.62%
ANN	15.26%	532.85	10.90%

Considering the errors, the Holt-Winters methods are clearly the best models as they outperform the other two methods in the three metrics compared. Here, both the raw data and the logged are interesting to explore as they represent different scales that lead to distinct conclusions. The logarithmic transformation normalizes the data conducting to a different interpretation of the results.

In the literature review that was performed artificial intelligence based models often outperformed traditional time series forecasting methods. One reason that might lead to the worst results in this specific case might be the limited amount of data, as ANN methods perform better with larger amounts of data. Exponential Smoothing methods might achieve better performance due to the weight that trend and seasonality have on the data. Exponential Smoothing is also easier to interpret and to explain which might be an advantage for the practical applicability of these models.

5.2.4 Interpretation of the Results

The results of the best models are presented in Appendix B. Table B.1 presents the values without transformation and is useful to see the prediction of the number of monthly services and so have an overall idea on the number of trips and the evolution of the company's demand. Regarding Table B.2 it presents a more accurate prediction but at a different scale. Here it is important to bear in mind that the results do not represent the real demand values but a relative change on the number of services.

Regarding the prediction intervals, as seen in Figure 5.6, the Holt-Winters method applied to the data at a normal scale presents large intervals for a 95% and 80% error, whereas pursuant to the logarithmic transformation, these intervals are smaller. These are observed in Figure 5.7 and illustrate a more accurate prediction. It is also meaningful to mention the increase in the range of the prediction interval with the increase in the forecasting time horizon. It is noticeable that, for both cases, the interval is higher in 2017 than in 2016, which demonstrates that the further into the future it is intended to predict, the bigger might be the error.

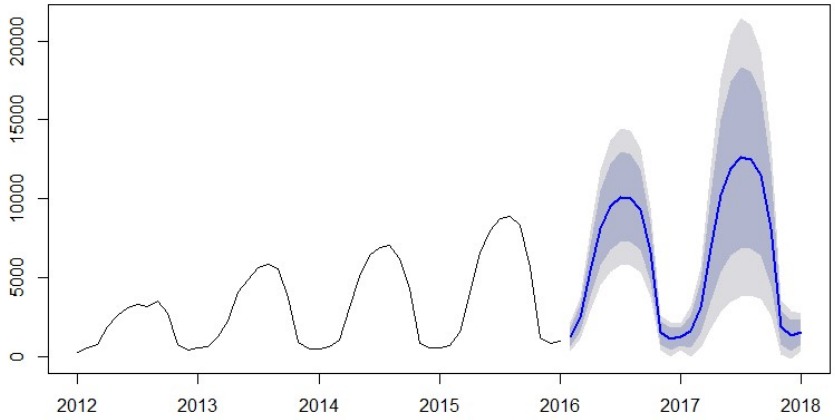

Figure 5.6: Prediction results with error bounds of 95% and 80%

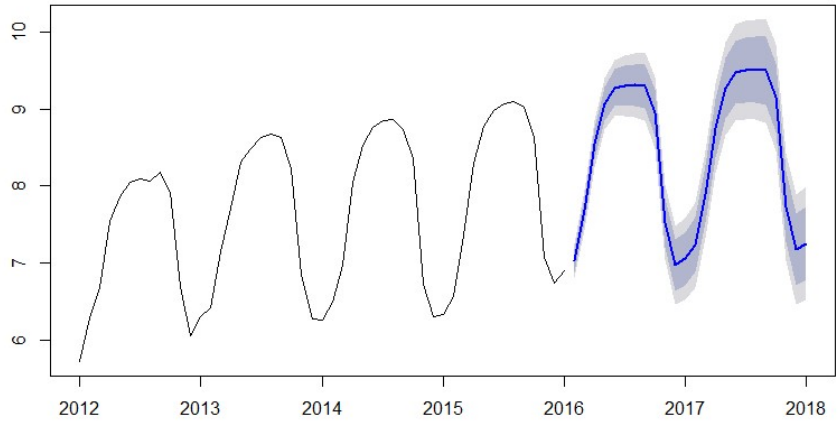

Figure 5.7: Prediction results of logged data with error bounds of 95% and 80%

5.3 Weekly Demand

Finally, weekly aggregated data is studied. Weekly data represents a highly important data frequency for the company as these predictions might be compared to the weekly availability provided by the drivers and car allocation. Differently from the previous time frequency, this data aggregation provides more data points, with a total of 312 observations, and consequently bigger variability. However, the processes are similar to the ones described in Section 5.2 and it will be described with less detail in this section.

5.3.1 Time Series Analysis

The weekly data is plotted in Figure 5.8 where the first segment presents the raw series and where an upward trend and seasonality are again spotted being decomposed in the following parts of the figure. The series is multiplicative, being the Equation 5.2 also applicable to the weekly series.

In this case the random component is more significant than in the monthly analysis, and while the trend and seasonality are similar, randomness, presented in the last segment of Figure 5.8, is interesting to analyze. However, when the stationarity test is performed on the random component, the stationarity is also assumable, as the p-value is again smaller than 0.05.

5.3.2 Proposed Models

Similarly to the monthly division, the train dataset includes data from 2012 to 2015 that is subsequently tested in 2016 and 2017 data. The train dataset of the weekly data is composed by 208 observations and the test dataset by 104. The models proposed, as well as the methodology to find the best parameters, are the same as in the previous section, being the main change the seasonal period from 12 to 52.

Regarding the remaining changes in the parameters the *size* of the ANN is raised to 50 and the new ARIMA model is given by $ARIMA(0,1,2)x(0,1,0)_{52}$. Once again, these models present

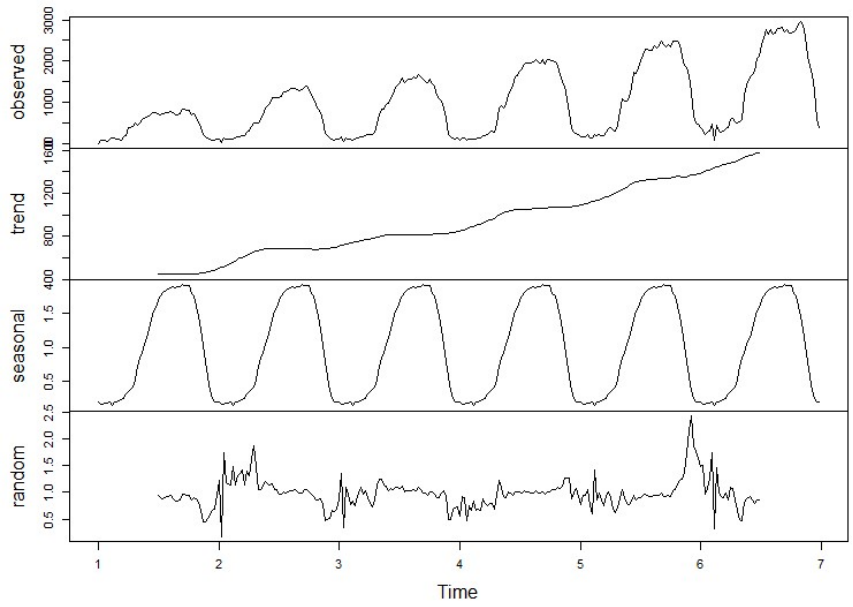

Figure 5.8: Weekly time series decomposition

good results but this does not mean that are the chosen parameters represent unique possibilities. As the parameters estimation is difficult and not exact, especially regarding the ARIMA model, different interpretations of the ACF and PACF plots might lead to some different parameters that might as well be acceptable.

5.3.3 Models Comparison

In line with the error analysis methodology from Section 5.2, MAPE, RMSE and MPE are also compared in this analysis and the errors are presented in Table 5.5.

Table 5.5: Weekly forecasting accuracy for different models

Methods	MAPE	RMSE	MPE
Holt-Winters	32.62%	474.50	-8.45%
Logged Holt-Winters	7.68%	0.59	-4.57%
ARIMA	25.25%	420.04	5.45%
ANN	29.48%	356.45	-8.07%

Regarding the weekly data predictions, and possibly due to the more accentuated random component that is observable, the Holt-Winters gets an higher error than the ARIMA and ANN. These might result precisely from one of the possible causes of accuracy loss in the previous analysis. ANN is now fitted in a bigger dataset which improves its performance and the series is more complex which makes it more adequate for ARIMA than the monthly data.

However, when looking at the plotted results of Holt-Winters and ARIMA, presented in Figure 5.9 and 5.10, respectively, it is noticeable that the higher error in Holt-Winters might be due to the lag that is observable, as the ARIMA does not capture the trend so accurately.

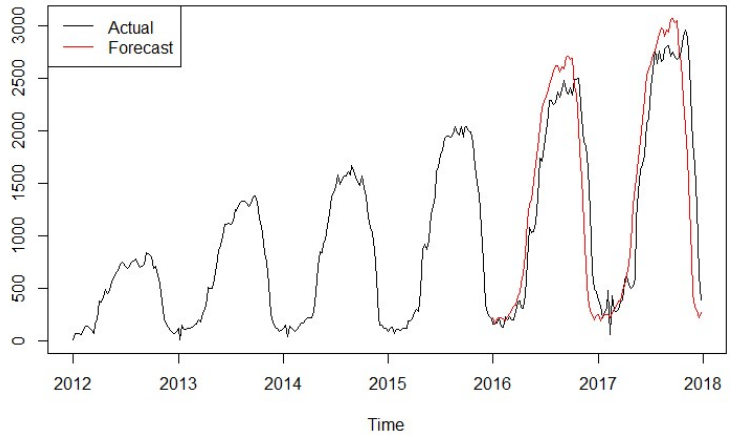

Figure 5.9: Holt-Winters observed and predicted values

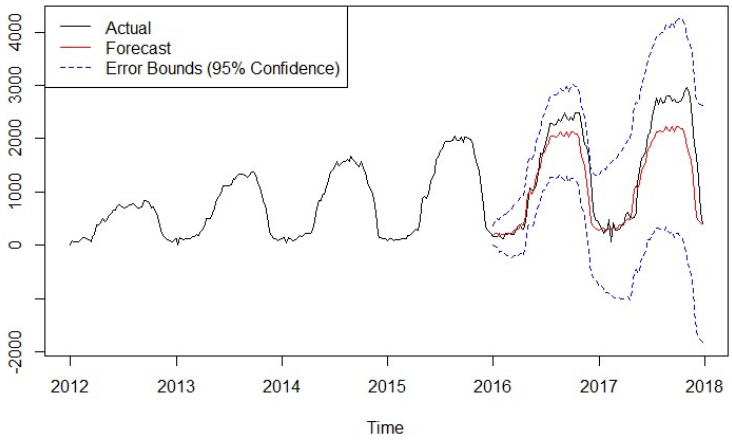

Figure 5.10: ARIMA observed and predicted values

5.3.4 Interpretation of the Results

The results that arise from this prediction, presented in Appendix C, might be useful when compared with the *carusage* information described in Chapter 3. By the comparison of the percentage of used cars and drivers with the number of services performed and considering a safety margin, the number of necessary cars and drivers for the future is estimated.

Appendix C presents the observed and estimated values when predicted using the ARIMA and Holt-Winters for both raw and transformed data. The interpretation of these results is similar to the previous section.

Chapter 6

Conclusions and Future Work

This study allows a deeper understanding of a company's demand and the comparison of different forecasting methods. As presented in Chapter 1, the main objectives are the extraction of the factors that affect the demand and the proposal and comparison of forecasting methods for annual, monthly and weekly demand.

The first objective is derived from the preprocessing steps and exploratory analysis that provides insightful information to the company about its demand and operations. During this process the main difficulties are the integration of the different datasets provided by the company and the understanding of the main questions that would be interesting to explore. It corroborates the acquired information about the region and its tourists and enables the company to get an overall idea of how the business has evolved and its specificities.

Regarding the proposal and comparison of models, the study provides different methods to model the data in analysis for different time frequencies: annual, monthly and weekly. The increase in error from the annual to weekly analysis is noteworthy. Even though the forecasts are made based on more data, this might be a consequence of the increased variability as the data is aggregated in smaller periods. Another important aspect to discuss is the performance of the ANN. The worst results in this case might be caused by the limited amount of data. However, this also means that better results might be achieved in the future. Besides, there might be changes that decrease the growth rate of the company, modifying the trend and paving the way for Neural Networks to improve its performance.

Furthermore, the weekly prediction might be compared with the weekly availability of drivers and vehicles. Thus, this study offers the necessary tools for the company to continuously analyze its data and predict the demand. Based on the prediction, the necessary number of vehicles and drivers can be estimated.

Finally, it is crucial to address the limitations of this analysis and of any forecasting solution. Agrawal (2013) presents the definition for time series forecasting as "the act of predicting the future by understanding the past". Here the assumption that the future will behave like the past is explicit, and it is important to be conscious that it might always be changed by unpredictable factors that may arise. Moreover, it is important to bear in mind that nothing can be predicted with

complete accuracy and that if you try to forecast further into the future, the error will increase.

In this specific case, as the majority of the clients are from the UK, the influence of external and unpredictable factors might considerably impact the company. Additionally, there is a risk of devaluation of the pound due to the UK's different currency and the political changes have also been significant with *Brexit*. The increase in the error due to attempting to forecast further into the future is also proven by the analysis as both in the monthly and weekly forecasts the prediction error increases in the second year.

6.1 Future Research

As stated in Chapter 1, the touristic sector is highly information intensive and there is an enormous amount of sources and data being generated every second. Thus, in this sector the prospects of future studies are endless. However, this section focuses on the data that was provided for this project and future studies that might be more advantageous for the company and for the improvement of this analysis.

Firstly, and regarding the conclusions of the exploratory analysis, it would be interesting to shorten the time frequency, to daily or even hourly predictions, to ensure that, besides the annual seasonality, the particularities of the weekly and daily demand would be captured. This would also allow a better and more precise integration with an allocation algorithm that is being developed in parallel to this study.

Another conclusion of the exploratory analysis is related with the flights and airport data. By exploring the few available data, a relation to the airport's traffic is expected and very likely to influence the company's demand. With a more careful data gathering, this could lead to a more detailed solution that, apart from the historic data, could incorporate the future flights.

Finally, a deeper analysis on the clients' data could improve the prediction and enable a more personalized solution. The creation of clients' clusters could allow a better analysis of factors such as the country of origin, length of stay and number of passengers and understand if the demand patterns differ according to these. Also feedback and data that presents behavior of tourists' on the website are continuously being generated by the customers and might be really interesting to analyze.

Bibliography

Agrawal, R. R. A. (2013). An Introductory Study on Time Series Modeling and Forecasting. 1302.6613:1–68.

Airports Council International (2016). Annual World Airport Traffic Report. http://www.aci.aero/Data-Centre/Airport-Statistics-Infographics, Last accessed on 2018-05-20.

Airports Council International (2017). Annual World Airport Traffic Forecasts 2017– 2040. http://www.aci.aero/Data-Centre/Airport-Statistics-Infographics, Last accessed on 2018-05-20.

Benckendorff, P. J., Sheldon, P. J., and Fesenmaier, D. R. (2014). *Tourism Information Technology*. CABI Tourism Texts, 2nd edition.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*. Prentice-Hall, Inc, 3rd edition.

Burger, C., Dohnal, M., Kathrada, M., and Law, R. (2001). A practitioners guide to time-series methods for tourism demand forecasting — a case study of Durban, South Africa. *Tourism Management*, 22(4):403–409.

Chu, F. L. (2004). Forecasting tourism demand: A cubic polynomial approach. *Tourism Management*, 25(2):209–218.

Claveria, O. and Torra, S. (2014). Forecasting tourism demand to Catalonia: Neural networks vs. time series models. *Economic Modelling*, 36:220–228.

Divino, J. A. and McAleer, M. (2010). Modelling and forecasting daily international mass tourism to Peru. *Tourism Management*, 31(6):846–854.

Flightradar24 AB (2018). Faro Airport (FAO/LPFR) | Arrivals, Departures & Routes | Flightradar24. https://www.flightradar24.com/data/airports/fao, Last accessed on 21/05/2018.

Francisco Manuel dos Santos Foundation (2017). Passenger traffic at major airports: Lisbon, Oporto and Faro. https://www.pordata.pt/en/DB/Portugal/Search+Environment/Chart, Last accessed on 21/05/2018.

Fuchs, M., Höpken, W., and Lexhagen, M. (2014). Big data analytics for knowledge generation in tourism destinations - A case from Sweden. *Journal of Destination Marketing and Management*, 3(4):198–209.

Guimarães, R. C. (1988). *Introdução aos métodos estatísticos de previsão*.

Guimarães, R. C. and Cabral, J. A. S. (1997). *Estatística*. McGraw-Hill Portugal Lda.

Han, J., Kamber, M., and Pei, J. (2012). *Data Mining: Concepts and Techniques*.

Hu, C., Chen, M., and McCain, S.-L. C. (2004). Forecasting in Short-Term Planning and Management for a Casino Buffet Restaurant Clark. *Journal of Travel & Tourism Marketing*, 16(2-3):79–98.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts, 2nd edition.

Instituto Nacional de Estatística, I. P. (2017). *Estatísticas do Turismo 2016*. 2017 edition.

Kelly, D. (2017). Tripsavvy. `https://www.tripsavvy.com/what-are-airport-transfers-468260`, Last accessed on 2018-05-22.

Law, R. (1998). Room occupancy rate forecasting: a neural network approach. *International Journal of Contemporary Hospitality Management*, 10(6):234–239.

Makridakis, S., Wheelwright, S. C., and McGee, V. E. (1978). *Forecasting: Methods and Applications*. John Wiley & Sons, second edition.

Marković, N., Kim, M. E., and Schonfeld, P. (2016). Statistical and machine learning approach for planning dial-a-ride systems. *Transportation Research Part A: Policy and Practice*, 89:41–55.

Miah, S. J., Vu, H. Q., Gammack, J., and McGrath, M. (2017). A Big Data Analytics Method for Tourist Behaviour Analysis. *Information and Management*, 54(6):771–785.

Nogales, F. J., Contreras, J., Conejo, A. J., and Espínola, R. (2002). Forecasting next-day electricity prices by time series models. *IEEE Transactions on Power Systems*, 17(2):342–348.

Peng, B., Song, H., and Crouch, G. I. (2014). A meta-analysis of international tourism demand forecasting and implications for practice. *Tourism Management*, 45:181–193.

Shumway, R. H. and Stoffer, D. S. (2011). *Time Series Analysis and Its Applications With R Examples*.

Song, H. and Li, G. (2008). Tourism demand modelling and forecasting — A review of recent research. 29:203–220.

Song, H. and Liu, H. (2017). Analytics in Smart Tourism Design. pages 13–30.

Tan, P.-N., Steinbach, M., and Vipin Kumar (2006). *Introduction to data mining*.

Turismo de Portugal (2018). Análise Regional | 2017. `http://travelbi.turismodeportugal.pt/pt-pt/Documents/Análises/Alojamento/analise-regional-2017.pdf`, Last accessed on 20/05/2018.

United Nations World Tourism Organization (2018). 2017 International Tourism Results: the highest in seven years. `http://media.unwto.org/press-release/2018-01-15/2017-international-tourism-results-highest-seven-years`, Last accessed on 2018-02-18.

Universidade do Algarve (2017). O Perfil do Turista que visita o Algarve (2016). `https://issuu.com/turismo_algarve/docs/perfil_do_turista_2016_relatorio_fi`, Last accessed on 2018-05-20.

Werthner, H. and Ricci, F. (2004). E-commerce and tourism. *Communications of the ACM*, 47(12):101–105.

Wolfram, H., Ernesti, D., Fuchs, M., Kronenberg, K., and Lexhagen, M. (2017). Big Data as Input for Predicting Tourist Arrivals. In *Information and Communication Technologies in Tourism 2017*, pages 187–199.

www.faroairport.net (2018). Faro Airport Passenger Numbers. `https://www.faroairport.net/passenger-statistics.shtml`, Last accessed on 04/06/2018.

Yellowfish Travel, L. (2018). YellowFish Transfers. `https://www.yellowfishtransfers.com/en/home`, Last accessed on 2018-05-21.

Zhang, G., Patuwo, B. E., and Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, 14:35–62.

Zhang, G. P. and Qi, M. (2005). Neural network forecasting for seasonal and trend time series. *European Journal of Operational Research*, 160(2):501–514.

Appendix A

Most Frequent Locations

local	2015		2016		2017	
	arrival	departure	arrival	departure	arrival	departure
ALBUFEIRA	36.33%	36.94%	27.22%	27.53%	27.08%	27.43%
VILAMOURA	14.94%	15.11%	13.82%	13.82%	13.00%	13.20%
ALVOR	6.18%	6.00%	6.46%	6.43%	6.40%	6.50%
LAGOS	6.07%	5.96%	5.44%	5.30%	6.57%	6.37%
CARVOEIRO	5.06%	4.98%	5.33%	5.34%	5.31%	5.36%
OLHOS de AGUA	4.01%	4.03%	4.01%	3.93%	3.45%	3.46%
PRAIA da LUZ	2.80%	2.77%	2.78%	2.75%	2.80%	2.84%
PRAIA da ROCHA	2.05%	2.02%	2.71%	2.68%	2.36%	2.24%
PRAIA OURA /ALBUFEIRA	2.03%	1.93%	6.16%	6.08%	5.52%	5.52%
GALE /ALBUFEIRA	1.78%	1.81%	1.51%	1.53%	1.33%	1.39%
ARMACAO de PERA	1.42%	1.44%	1.69%	1.69%	1.48%	1.44%
BURGAU	1.36%	1.33%	1.32%	1.31%	1.29%	1.27%
BALAIA /ALBUFEIRA	1.20%	1.11%	2.16%	2.07%	2.09%	2.05%
SÃO RAFAEL /ALBUFEIRA	1.13%	1.13%	1.05%	1.09%	1.08%	1.12%
PORTIMAO	1.09%	1.07%	1.32%	1.36%	1.72%	1.79%
QUARTEIRA	1.07%	1.04%	1.33%	1.34%	1.77%	1.76%
CORCOVADA /ALBUFEIRA	1.01%	0.98%	1.26%	1.30%	1.09%	1.07%
FALESIA /ALBUFEIRA	0.90%	0.92%	1.35%	1.36%	1.54%	1.46%
SANTA EULALIA /ALBUFEIRA	0.87%	0.85%	1.57%	1.58%	1.65%	1.61%
MONTECHORO /ALBUFEIRA	0.86%	0.79%	1.37%	1.34%	1.25%	1.22%
TAVIRA	0.49%	0.46%	0.51%	0.54%	0.80%	0.68%
SESMARIAS /ALBUFEIRA	0.48%	0.47%	0.51%	0.55%	0.64%	0.68%
VALE de PARRA /ALBUFEIRA	0.48%	0.44%	0.46%	0.46%	0.48%	0.49%
SALGADOS /ALBUFEIRA	0.47%	0.55%	0.59%	0.62%	0.70%	0.68%
QUINTA do LAGO	0.45%	0.45%	0.62%	0.64%	0.61%	0.61%
VALE de LOBO	0.42%	0.45%	0.58%	0.57%	0.59%	0.61%
PORCHES	0.40%	0.40%	0.20%	0.23%	0.17%	0.16%
FERRAGUDO	0.38%	0.39%	0.39%	0.42%	0.42%	0.44%
OLHAO	0.36%	0.37%	0.42%	0.43%	0.50%	0.45%
CABANAS TAVIRA	0.34%	0.31%	0.35%	0.34%	0.36%	0.37%
GUIA	0.30%	0.31%	0.35%	0.33%	0.51%	0.52%
PRAIA do VAU	0.26%	0.25%	0.41%	0.40%	0.26%	0.27%
ALMANCIL	0.25%	0.25%	0.17%	0.18%	0.18%	0.17%
SAGRES	0.22%	0.22%	0.24%	0.24%	0.16%	0.19%
PADERNE	0.21%	0.22%	0.22%	0.23%	0.24%	0.24%
LAGOA	0.17%	0.16%	0.16%	0.15%	0.15%	0.12%
SALEMA	0.15%	0.15%	0.12%	0.10%	0.12%	0.11%

Figure A.1: Most frequent locations by year and *whichway*

Appendix B

Monthly Prediction Results

Table B.1: Observed vs Estimated values for monthly Holt-Winters model

	Observed	Estimated
Jan 2016	989.000	1256.805
Feb 2016	1276.000	2452.437
Mar 2016	2339.000	5434.976
Apr 2016	4812.000	8130.749
May 2016	8112.00	9514.02
June 2016	9580.00	10107.17
July 2016	10655.00	10031.81
Aug 2016	10421.000	9302.183
Sep 2016	10346.000	6605.678
Oct 2016	7428.000	1523.096
Nov 2016	2045.000	1104.744
Dec 2016	1486.000	1268.125
Jan 2017	1469.000	1603.362
Feb 2017	2003.000	3113.493
Mar 2017	2741.000	6867.792
Apr 2017	6942.00	10228.17
May 2017	10300.00	11916.62
June 2017	11501.00	12606.95
July 2017	12359.00	12462.86
Aug 2017	11894.00	11511.79
Sep 2017	12300.00	8144.31
Oct 2017	7512.000	1871.109
Nov 2017	1657.000	1352.451
Dec 2017	1350.00	1547.25

Table B.2: Observed vs Estimated values for monthly Holt-Winters model

	Observed	Estimated
Jan 2016	6.896694	7.034004
Feb 2016	7.151485	7.692207
Mar 2016	7.757479	8.546516
Apr 2016	8.478868	9.071038
May 2016	9.001100	9.277701
June 2016	9.167433	9.303746
July 2016	9.273785	9.311491
Aug 2016	9.251578	9.297986
Sep 2016	9.244355	8.948535
Oct 2016	8.913012	7.531065
Nov 2016	7.623153	6.976657
Dec 2016	7.303843	7.052595
Jan 2017	7.292337	7.237029
Feb 2017	7.602401	7.895232
Mar 2017	7.916078	8.749542
Apr 2017	8.845345	9.274064
May 2017	9.239899	9.480727
June 2017	9.350189	9.506771
July 2017	9.422140	9.514517
Aug 2017	9.383789	9.501012
Sep 2017	9.417355	9.151561
Oct 2017	8.924257	7.734091
Nov 2017	7.412764	7.179682
Dec 2017	7.20786	7.25562

Appendix C

Weekly Prediction Results

Table C.1: Observed vs Estimated values for weekly Holt-Winters model for 2016

2016	Observed	Estimated		Observed	Estimated
1	162.0000	221.4634	27	1999.000	2326.792
2	164.0000	165.7394	27	2150.000	2391.767
3	173.0000	215.0374	29	2288.000	2436.937
4	216.0000	220.0511	30	2286.00	2500.85
5	147.0000	217.7056	31	2249.000	2568.957
6	128.0000	216.8393	32	2265.000	2622.694
7	229.0000	196.9587	33	2366.000	2614.163
8	199.0000	227.9909	34	2323.000	2556.697
9	230.0000	263.1898	35	2384.000	2611.529
10	203.0000	287.6723	36	2476.000	2591.645
11	201.0000	334.8583	37	2388.000	2699.161
12	285.0000	338.9158	38	2348.000	2706.706
13	367.0000	421.8157	39	2405.000	2679.858
14	384.0000	464.4621	40	2340.000	2691.139
15	318.0000	553.5103	41	2484.000	2418.813
16	308.0000	645.0893	42	2488.00	2344.07
17	434.0000	821.0353	43	2495.000	2036.305
18	766.000	1136.256	44	2391.000	1807.195
19	1083.000	1295.543	45	2119.000	1551.072
20	1029.000	1340.275	46	1891.000	1124.014
21	1043.000	1522.987	47	1855.0000	712.1679
22	1116.000	1679.317	48	1644.0000	404.6404
23	1360.000	1843.859	49	1243.0000	287.7578
24	1738.000	2034.019	50	662.0000	256.3523
25	1710.000	2222.308	51	489.0000	201.4201
26	1846.000	2281.089	52	467.0000	238.2944

Table C.2: Observed vs Estimated values for weekly Holt-Winters model for 2017

2017	Observed	Estimated		Observed	Estimated
1	395.0000	253.6286	27	2410.000	2641.854
2	330.0000	189.7442	28	2566.000	2714.786
3	224.0000	246.0958	29	2772.000	2765.204
4	278.0000	251.7456	30	2639.000	2836.856
5	313.0000	248.9757	31	2763.000	2913.224
6	477.0000	247.8991	32	2656.000	2973.259
7	67.0000	225.0934	33	2675.000	2962.691
8	430.0000	260.4692	34	2782.000	2896.692
9	291.0000	300.5798	35	2800.000	2957.931
10	282.0000	328.4291	36	2812.000	2934.534
11	305.0000	382.1714	37	2704.000	3055.369
12	369.0000	386.6725	38	2753.000	3063.005
13	391.0000	481.0931	39	2694.000	3031.732
14	577.0000	529.5566	40	2675.000	3043.604
15	623.0000	630.8765	41	2700.000	2734.815
16	536.0000	735.0142	42	2763.00	2649.54
17	506.0000	935.1809	43	2902.000	2301.005
18	522.000	1293.804	44	2958.000	2041.527
19	594.000	1474.7	45	2886.000	1751.694
20	1162.000	1525.126	46	2604.000	1269.038
21	1419.000	1732.481	47	2114.0000	803.8263
22	1632.000	1909.707	48	1887.0000	456.5905
23	1679.000	2096.156	49	1673.0000	324.6107
24	1770.000	2311.607	50	1277.0000	289.1025
25	2069.000	2524.798	51	617.0000	227.0895
26	2115.000	2590.769	52	396.0000	268.5888

Table C.3: Observed vs Estimated values for weekly logged Holt-Winters model for 2016

2016	Observed	Estimated		Observed	Estimated
1	5.087596	5.607488	27	7.600402	8.034157
2	5.099866	5.048929	28	7.673223	8.052088
3	5.153292	5.437719	29	7.735433	8.071464
4	5.375278	5.593796	30	7.734559	8.098125
5	4.990433	5.649072	31	7.718241	8.124304
6	4.852030	5.641667	32	7.725330	8.144736
7	5.433722	5.537895	33	7.768956	8.141332
8	5.293305	5.702709	34	7.750615	8.119583
9	5.438079	5.845041	35	7.776535	8.141867
10	5.313206	5.935347	36	7.814400	8.137498
11	5.303305	6.112849	37	7.778211	8.180324
12	5.652489	6.140417	38	7.761319	8.181374
13	5.905362	6.341792	39	7.785305	8.171577
14	5.950643	6.450441	40	7.757906	8.178314
15	5.762051	6.595467	41	7.817625	8.072411
16	5.730100	6.722688	42	7.819234	8.045241
17	6.073045	7.045698	43	7.822044	7.907023
18	6.641182	7.388065	44	7.779467	7.788243
19	6.987490	7.521092	45	7.658700	7.624872
20	6.936343	7.553769	46	7.544861	7.264067
21	6.949856	7.675266	47	7.525640	6.805711
22	7.017506	7.773595	48	7.404888	6.301419
23	7.215240	7.861591	49	7.125283	5.961954
24	7.460490	7.953145	50	6.495266	5.853112
25	7.444249	8.039304	51	6.192362	5.616493
26	7.520776	8.061010	52	6.146329	5.780570

Table C.4: Observed vs Estimated values for weekly logged Holt-Winters model for 2017

2017	Observed	Estimated		Observed	Estimated
1	5.978886	5.810447	27	7.787382	6.048000
2	5.799093	5.251888	28	7.850104	6.138306
3	5.411646	5.640678	29	7.927324	6.315808
4	5.627621	5.796756	30	7.878155	6.343376
5	5.746203	5.852031	31	7.924072	6.544752
6	6.167516	5.844627	32	7.884577	6.653401
7	4.204693	5.740855	33	7.891705	6.798426
8	6.063785	5.905669	34	7.930925	6.925647
9	5.673323	6.048000	35	7.937375	7.248658
10	5.641907	6.138306	36	7.941651	7.591025
11	5.720312	6.315808	37	7.902487	7.724052
12	5.910797	6.343376	38	7.920447	7.756728
13	5.968708	6.544752	39	7.898782	8.374537
14	6.357842	6.653401	40	7.891705	8.381273
15	6.434547	6.798426	41	7.901007	8.275371
16	6.284134	6.925647	42	7.924072	8.248200
17	6.226537	7.248658	43	7.973155	8.109983
18	6.257668	7.591025	44	7.992269	7.991203
19	6.386879	7.724052	45	7.967627	7.827832
20	7.057898	7.756728	46	7.864804	7.467027
21	7.257708	5.640678	47	7.656337	7.008671
22	7.397562	5.796756	48	7.542744	6.504379
23	7.425954	5.852031	49	7.422374	6.164913
24	7.478735	5.844627	50	7.152269	6.056071
25	7.634821	5.740855	51	6.424869	5.819453
26	7.656810	5.905669	52	5.981414	5.983530

Table C.5: Observed vs Estimated values for weekly ARIMA model for 2016

2016	Observed	Estimated		Observed	Estimated
1	162.000	184.529	27	1999.000	1865.602
2	164.0000	220.6021	28	2150.000	1911.602
3	173.0000	224.6021	29	2288.000	2026.602
4	216.0000	171.6021	30	2286.000	2040.602
5	147.0000	209.6021	31	2249.000	2046.602
6	128.0000	208.6021	32	2265.000	2035.602
7	229.0000	200.6021	33	2366.000	2067.602
8	199.0000	217.6021	34	2323.000	2136.602
9	230.0000	221.6021	35	2384.000	2074.602
10	203.0000	222.6021	36	2476.000	2057.602
11	201.0000	284.6021	37	2388.000	2132.602
12	285.0000	284.6021	38	2348.000	2032.602
13	367.0000	312.6021	39	2405.000	2122.602
14	384.0000	365.6021	40	2340.000	2134.602
15	318.0000	397.6021	41	2484.000	2096.602
16	308.0000	377.6021	42	2488.000	2082.602
17	434.0000	569.6021	43	2495.000	1980.602
18	766.0000	965.6021	44	2391.000	1756.602
19	1083.000	1018.602	45	2119.000	1606.602
20	1029.000	966.6021	46	1891.000	1487.602
21	1043.000	1052.602	47	1855.000	1171.602
22	1116.000	1294.602	48	1644.0000	859.6021
23	1360.000	1380.602	49	1243.0000	449.6021
24	1738.000	1476.602	50	662.0000	371.6021
25	1710.000	1707.602	51	489.0000	330.6021
26	1846.000	1746.602	52	467.0000	303.6021

Table C.6: Observed vs Estimated values for weekly ARIMA model for 2017

2017	Observed	Estimated		Observed	Estimated
1	395.0000	280.1311	27	2410.000	1961.204
2	330.0000	316.2042	28	2566.000	2007.204
3	224.0000	320.2042	29	2772.000	2122.204
4	278.0000	267.2042	30	2639.000	2136.204
5	313.0000	305.2042	31	2763.000	2142.204
6	477.0000	304.2042	32	2656.000	2131.204
7	67.0000	296.2042	33	2675.000	2163.204
8	430.0000	313.2042	34	2782.000	2232.204
9	291.0000	317.2042	35	2800.000	2170.204
10	282.0000	318.2042	36	2812.000	2153.204
11	305.0000	380.2042	37	2704.000	2228.204
12	369.0000	380.2042	38	2753.000	2128.204
13	391.0000	408.2042	39	2694.000	2218.204
14	577.0000	461.2042	40	2675.000	2230.204
15	623.0000	493.2042	41	2700.000	2192.204
16	536.0000	473.2042	42	2763.000	2178.204
17	506.0000	665.2042	43	2902.000	2076.204
18	522.000	1061.204	44	2958.000	1852.204
19	594.000	1114.204	45	2886.000	1702.204
20	1162.000	1062.204	46	2604.000	1583.204
21	1419.000	1148.204	47	2114.000	1267.204
22	1632.000	1390.204	48	1887.0000	955.2042
23	1679.000	1476.204	49	1673.0000	545.2042
24	1770.000	1572.204	50	1277.0000	467.2042
25	2069.000	1803.204	51	617.0000	426.2042
26	2115.000	1842.204	52	396.0000	399.2042