ISBN: 978-989-8533-74-6 © 2018

FORECASTING TOURISM DEMAND FOR LISBON'S REGION THROUGH A DATA MINING APPROACH

Hugo Ricardo[1], Ivo Gonçalves[2] and Ana Cristina Costa[1]
[1]*NOVA IMS, Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal*
[2]*INESC Coimbra, DEEC, University of Coimbra Pólo II, 3030-290 Coimbra, Portugal*

ABSTRACT

Tourism stakeholders such as the government, passenger transport companies, accommodation establishments, restaurants, recreational businesses, among others, rely on tourism demand indicators' forecasts to make decisions. Most of tourism demand forecasting models are time-series and econometric based. Machine learning methods are emerging and have been proved to be quite suitable for non-linear modelling. These methods are part of an interdisciplinary field named "Data Mining" which is known by the process of knowledge discovery in databases (KDD). The core drive of this work is to enhance the available public sources of tourism forecast information and to contribute to the tourism stakeholders' strategy in Portugal. More specifically, a multivariate model to forecast international tourism demand was developed through a Data Mining approach, which assessed models derived by Regression Trees (Random Forests), Artificial Neural Networks and, Support Vector Machines (SVM). The model development was constrained to machine learning methods, publicly available data, and minimum data assumptions. The forecasted demand variable was the nights spent at tourist accommodation establishments in Lisbon's region, one of the country's main foreign tourist destinations. The objectives were achieved, as the selected model (SMOReg, support vector regression) was successful in generalization capability. The accuracy of the produced forecasts provides some evidence of the reliability of the proposed model. If institutions and decision makers have information regarding the evolution of the explanatory variables used in this model, the impact on Lisbon's tourism demand can be assessed, even in case of an emerging recession, as shown using three future plausible scenarios.

KEYWORDS

Forecast, Tourism, Machine learning, Knowledge Discovery, Lisbon

1. INTRODUCTION

International tourism has a significant weight on Portugal's main tourism destinations, namely Algarve, Madeira and Lisbon. According to OECD (2016), 70% of Portugal's tourist demand has its source in international markets. In 2014, the major tourist sources for Portugal were, by order of importance, the United Kingdom, Germany, Spain, France and the Netherlands. The nation's trade balance is positive mainly due to the export of tourism services. It represents almost half of the total service exports (Pordata, 2016). In Lisbon's region, the share of nights spent by non-residents in July'2016 was of 76.4%, according to the Regional Analysis of Turismo de Portugal. It is then clear that Lisbon's region tourism activity heavily depends of international tourist sources. Hence, tourism has a significant role in Portugal's economy, especially in the region of Lisbon which is the one that most contributed to GDP in 2014 (INE, 2016) and ranks second on top 3 national tourism destinations (Turismo de Portugal, 2016).

Tourism industry's players must deal with perishable products such as rooms, airline seats and car rentals, etc. From a macroeconomic point of view, tourism demand forecast is part of the decision-making process that aims at a positive return of the investment on infrastructures and promotion. A country such as Portugal, which depends heavily on tourism income, such forecasting estimates are a starting point for Government policy decisions. From a microeconomic perspective, those forecasts are required to plan, for instance, transportation routes, tours and hotel beds. Hence, it is crucial to forecast tourist arrivals and nights spent at tourist accommodation establishments for long, medium and short-term horizons (Frechtling, 2011).

As stated by most of econometricians, demand can be defined in general terms as the quantity of a good or service that consumers, clients, etc. are willing to pay given a specific price and time span. Hence, tourism demand is the measured desire for tourism products or services regarding a geo-location, as for instance a country, region or city. According to Witt and Witt (1995), there are several metrics that can be used to attain tourism demand, from which the tourist arrivals is the most popular one, followed by expenditure. An alternative measure is nights spent in the destination's tourist establishments. The main advantage of nights spent at tourist accommodation establishments' variable is the capacity to differentiate domestic from foreign tourism and by type of establishment (Cunha and Abrantes, 2013). It is also visible the economic relation of the total duration that tourists stay in a destination and the expenditure in the local commerce and tourism establishments. In relatively recent studies, such as those undertaken by Constantino et al. (2016), and Teixeira and Fernandes (2012) the nights spent at tourist accommodation establishments has been used as a proxy for Mozambique and Portugal's tourism demand, respectively.

An extensive literature is available regarding tourism, and the academic interest in tourism demand forecast has grown (Athanasopoulos et al., 2011; Song and Li, 2008). Most studies are focused in finding the explanatory variables of tourism demand or constructing an accurate forecasting model (Li et al., 2005). Both research streams are related, and to forecast with causal or data-based methods it is advisable to previously identify the determinants of tourism demand. The need for knowledge regarding tourism demand determinants was fulfilled in scientific manner by many economists who modeled tourism demand through econometric methods (Sheldon, 1985). Systematic reviews have shown that the mostly used explanatory variables are the income, price indexes, exchange rates, travel's cost, and demographic variables (Kim et al., 2013).

Recent research reviews on tourism demand modelling and forecasting accuracy by Athanasopoulos et al. (2011) and Kim et al. (2013), show that most of the tourism demand forecasts are time-series and econometric model based. Those authors verified that Artificial Intelligence (a.k.a. machine learning) techniques are emerging. Hence, a shortage of this approach among academic literature is revealed. Traditional methods like time-series or econometric are quite suitable for data characterized by specific behavior such as trend, seasonality and cyclicality (Armstrong, 2001). On the contrary, artificial intelligence methods like the Artificial Neural Networks (ANNs) discard statistic assumptions about the data (Han et al., 2011). These techniques can deal with imperfect and irregular data or nonlinear behavior, and recent studies are demonstrating that their accuracy is satisfactory (Claveria et al., 2014; Constantino et al., 2016; Teixeira and Fernandes, 2012). Although some academic findings point out that there is not a one-fit-all forecast technique for all source markets and forecast horizons, the advantage of using machine learning algorithms is clear under the premise that the system may become chaotic, as stated by Baggio and Sainaghi, (2016).

Given that the economic decisions taken by tourism stakeholders are based on forecasts and likewise on public sources of information, considering a relative worldwide shortage of academic articles regarding Data Mining approaches (data-driven) to estimate tourism demand indicators and, an absence of a public machine learning based multivariate forecasting model in Portugal, this work sought to develop the most accurate multivariate model to forecast international tourism demand through a data-driven approach. The model development was constrained to publicly available data and machine learning methods. The forecasted demand variable was the nights spent by UK residents at tourist accommodation establishments in Lisbon's region.

2. METHODOLOGICAL FRAMEWORK

Data mining refers to the process of extracting meaningful knowledge from databases or information systems by identifying hidden but significant patterns in data. Data Mining (DM) methodologies can be KDD, SEMMA and CRISP-DM, the most complete of the last two (Azevedo and Santos, 2008). This study will follow the CRISP-DM approach (IBM, 2011). It consists in understanding the business or problem, using all the available data in repositories, prepare the data by cleaning, partitioning and modifying, modeling by applying machine learning algorithms, evaluating (assess best models) and deploying by applying the best model to new data. This will be the guideline methodology for the proposed work. The methods used for the modeling step were machine learning based, namely: Regression Trees (Random Forests), Artificial Neural

Networks and, Support Vector Machines (SVM). The multiple linear regression method was used as baseline for the comparison of above non-linear models with a linear one.

This study framework was inspired in light of previous forecasting exercises, such as those by (Cankurt and Subaşi, 2016; Constantino et al., 2016), which attempted to forecast the monthly nights spent at tourist accommodation establishments using machine learning algorithms. Here, due to practical reasons, the chosen tourism source market was the UK, which is one of the top five main tourist's source markets for the region of Lisbon and, thus considered meaningful to the region's economy.

2.1 Data

The target variable has a seasonal component (Figure 1), therefore lagged target variables were added to the original dataset. A preliminary search of explanatory variables for tourism demand in the literature revealed that macroeconomic variables, such as GDP, income and consumer price index are related (Daniel and Ramos, 2002). Most of those variables are available in databases of Eurostat, Instituto Nacional de Estatística (INE, Statistics Portugal) and the UK Office for National Statistics (ONS). Only variables with a monthly frequency were included in the dataset. Rarely used variables, such as google trends, currency exchange rates and stock market index were also included. To guarantee the reliability of data and, that any person can replicate the model, the data sources for the time-series variables used in this study were from official entities and strictly the ones available to everyone in the internet (Table 1).

Table 1. Variables and Data Sources

Source	Variables	Internet address
Instituto Nacional de Estatística (INE)	Nights spent at tourist accommodation establishments in Lisbon's region Number of tourists at accommodation establishments in Lisbon's region	www.ine.pt
UK Office for National Statistics (ONS) ·	Consumer Price Index (CPI) Retail Price Index (RPI) for Travel and Air Passangers	www.ons.gov.uk/
Eurostat (European Commission)	Financial account - monthly data Consumers - monthly data Harmonised indices - monthly data Harmonised unemployment rates (%) - monthly data Interest rates - monthly data Nights spent at tourist accommodation establishments - monthly data	ec.europa.eu/eurostat
Investing.com	GBP/EUR Exchange Rate Historical Data	www.investing.com
Google Trends	Search words (ex. "Lisbon")	trends.google.pt

Figure 1. Nights Spent by UK Residents at Tourism Establishments in Lisbon's Region (2004 – 2015)

2.2 Experiments

Three experiments were executed with a data partition of 91% for training (120 instances) and 9% for testing (12 instances) to capture data behavior of most recent years and, with the same set of algorithms applied to 3 different datasets. Between the experiments, a progressive attribute selection was performed to reduce non-relevant variables and to augment the models accuracy, which is further explained.

To compare the models generated by the machine learning algorithms the baseline model was a multiple linear regression (MLR) with no prior testing of assumptions, no attribute selection, no prior check of capabilities (data type, missing values). The set of machine learning algorithms was restricted to a few parameter variation to avoid time consumption and minimize the insertion of multiple heuristics: 2 random forests (first with default parameters, second with attribute importance feature); 3 artificial neural networks (Multi Layer Percepetron – MPL) with default parameter differing in the number of hidden layers (H 2, H 3 and H 4); 4 support vector regression (SMOReg) with default parameters but distinct in the complexity parameter (C 0.5; C 1.0; C 1.5; C 2.0).

For experiment number 1 (Exp.1) the dataset used was the full attribute dataset "TourismDataset.arff" and with the already described data partition and group of ML algorithms. The results are displayed in Table 2. In this experiment, the number of heuristics is the minimum has possible, it is a start point to compare further experiments with selected attributes to improve model's accuracy.

After Exp.1 and, before a new learning process, to eliminate irrelevant variables and at the same time find the attribute interrelationships and bypass linear relationship selection (Arauzo-Azofra et al., 2004), the algorithm used for attribute selection was the Relief algorithm. In Weka software it is named as "ReliefFAttributeEval" (evaluator algorithm) and, in this work it was combined with a search algorithm called "Ranker". Both algorithms were used with the Weka's default parameter configuration. Before applying the Relief attribute selector, to find out the number of attributes to use has threshold in the Ranker's algorithm, the correlation based evaluator algorithm "CfsSubsetEval" combined with the search algorithm "BestFirst" was used. The output informed 19 relevant variables, which were not considered in the next experiment due to the linear constraints. The attribute selection with Relief restricted to 19 variables was performed using the TourismDataset.

For experiment number 2 (Exp.2) the dataset used was the full attribute dataset "TourismDataset.arff" and the dataset "dataset_19_Relief.arff" (19 selected variables) with the previously described data partition and group of ML algorithms (see results in Table 3).

After Exp.2, another attribute selection was done with the best model but, this time, with manual selection of variables to verify the impact on the model's accuracy. It was found that without the variable number of guests at tourist accommodation establishments, the model's accuracy was improved. Results of the Spearman correlation test applied to the selected variables showed strong (linear or non-linear) association between them. Hence, the next dataset was created with 18 variables and named has "dataset_18_Relief_no_hospedes.arff".

For the last experiment (Exp.3), the dataset used was the full attribute dataset "TourismDataset.arff", the "dataset_19_Relief.arff" (19 selected variables) and the "dataset_18_Relief_no_hospedes.arff" (18 selected variables). The results are displayed in Table 4.

2.3 Models Evaluation

In each single experiment the model's evaluation metrics analyzed were the mean absolute error (MAE) and the root mean squared error (RMSE):

$$\text{Mean absolute error: MAE} = \text{mean}(|e_i|),$$
$$\text{Root mean squared error: RMSE} = \sqrt{\text{mean}(e_i^2)}.$$

Additionally, the correlation coefficient (R^2) was used to know how well the predicted values change with the actual values.

3. RESULTS AND DISCUSSION

Analyzing the test results of Exp.1, the Linear Regression outstand the rest of the algorithms in RMSE (and R^2), followed very closely by the MLP H2 which had the best MAE of all (Table 2). Outlier values seem to have similar impact in MLR and MLP, since RMSE is almost equal on both models. The MLR correlation coefficient is quite high, which indicate a relative good fit of the model's test predictions but, the bias measure (MAE) is much lower on MLP, which points toward a better accuracy. MLP showed to have the best performance in modeling an ordered dataset of a high dimension, which may indicate a "presence" of non-linearity relationship. Additionally, has explained before, MLR model cannot be considered reliable for estimation since its assumption were not tested. Nevertheless, MLR works as a benchmark for the experiments.

Table 2. Exp.1 (Top 3 Models)

Dataset	Algorithm	Parameter's options	MAE	RMSE	R^2
TurismDataset	MLR	- default	9.522	11.060	0.964
TurismDataset	MLP	- default with H 2	8.704	11.325	0.939
TurismDataset	SMOreg	- default with C 1.0	10.724	12.106	0.965

After the first attribute selection, Exp.2 was conducted to assess the best trade-off between full and attribute selected datasets and algorithms. In Exp.2, it is evident the superior performance with the attribute selected dataset ("dataset_19_Relief.arff") and the support vector regression algorithm (SMOReg C 2.0) in all evaluation metrics (Table 3). SMOReg C2.0 was then used for the second attribute selection evaluation, where it was achieved a reduction of 3.169 on MAE and of 3.114 on RMSE.

Table 3. Exp. 2 (Top 3 Models)

Dataset	Algorithm	Parameter's options	MAE	RMSE	R^2
dataset_19_Relief	SMOreg	- default with C 2.0	9.255	10.387	0.954
dataset_19_Relief	SMOreg	- default with C 1.5	9.597	10.700	0.953
dataset_19_Relief	SMOreg	- default with C 0.5	9.635	10.858	0.950

The last experiment (Exp. 3) was run to assess the best model with the last dataset "dataset_18_Relief_no_hospedes.arff" (18 selected variables). The SMOReg was still the best model achieved after the attribute selection (Table 4), especially in the RMSE metric, followed by the MLR and MLP (H2). However, the version SMOReg with a C of 1.0 outperformed the version with a C 2.0 by a difference of less 0.06 on MAE. Therefore, it is the selected model for the forecasting exercise. Table 5 displays the model's summary results, and Figure 2 shows the test prediction outputs.

Table 4. Exp. 3 (Top 3 Models)

Dataset	Algorithm	Parameter's options	MAE	RMSE	R^2
dataset_18_Relief_no_hospedes	SMOreg	- default C 1.0	6.024	7.234	0.943
dataset_18_Relief_no_hospedes	SMOreg	- default C 1.5	6.041	7.259	0.942
dataset_18_Relief_no_hospedes	SMOreg	- default C 2.0	6.086	7.274	0.942

Table 5. Summary of the SMOReg C 1.0 Modelling Results

=== Summary ===	
Correlation coefficient	0.9425
Mean absolute error	6.0237
Root mean squared error	7.2338
Relative absolute error	0.2283
Root relative squared error	0.2329
Total Number of Instances	12

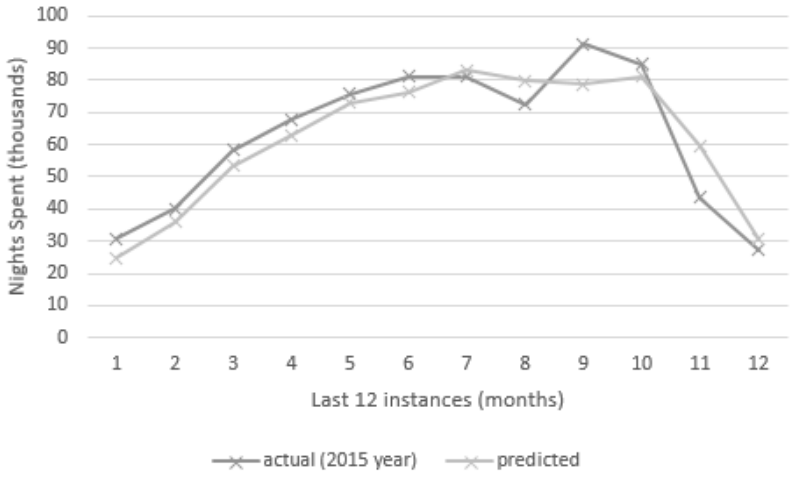

Figure 2. Test Predictions with SMOReg C1.0 (Exp. 3)

Test results for SMOReg C1.0 in Table 5 show a relative absolute error (MAE in percentage) of 22.83% and a Root Relative Absolute Error (RMSE in percentage) of 23.28%. Such results are not so promising of a high forecasting accuracy. However, since the model's generalization capability was roughly tested, due to small number of dataset instances, the forecast exercise was proceeded with the best achieved model.

3.1 Scenario Forecasting

Since the data mining model predicts through interpolation, it is necessary to have accurate observations of explanatory variables in a future period and, in the absence of official data, it would be necessary to undertake several studies for all the input variables to find reasonable estimates for them, which is not viable. In order to forecast the target variable for a 12 month horizon (2016 year) using real data, three future plausible scenarios defined as unfavorable (Scenario 1), moderated (Scenario 2), and favorable (Scenario 3) were established. To obtain the input variables, the averages of the attribute values were calculated in the following way for each scenario.

Scenario 1 (unfavorable) – for each independent variable, it was computed the average of each month of the 2005-2009 period (5 years prior to the financial crises);

Scenario 2 (moderate) – for each independent variable, it was computed the average of each month of the 2011-2015 period (5 years after financial crises), the year 2010 was excluded has it is the inflexion point of the dependent variable;

Scenario 3 (favorable) – for each independent variable, the average of each month of the 2014-2015 (2 years of great growth of the dependent variable);

For the 12 months lagged target variable, the real values of the year 2015 were used and, for the 1 month lagged variable the value of the last instance of the target variable (December's 2015) was used to begin the model's predictions. Every time an estimate was computed for each month, it was used as input value for the 1 month lagged variable attribute. The results are shown in the graphical representation (Figure 3). Due to data availability, the predicted total value was compared to the total actual value (Table 6). The scenario's estimates with highest forecasting accuracy were achieved in scenario 3, which is not surprising because this scenario used the most recent data.

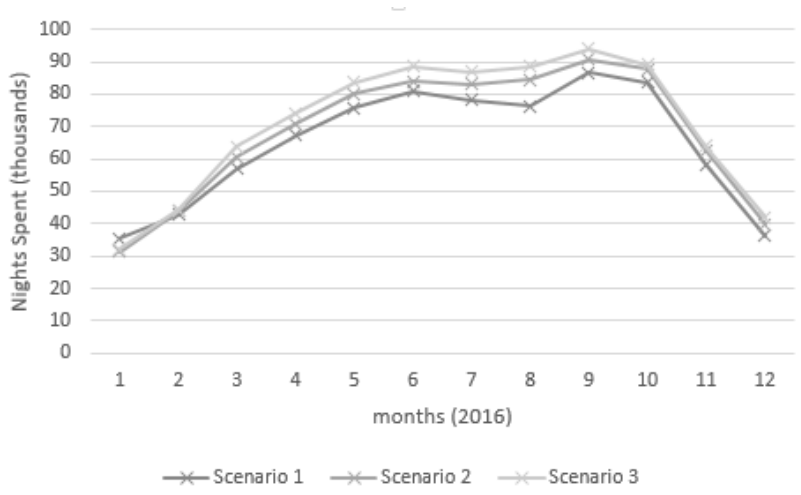

Figure 3. Scenario Forecast h = 12m (year 2016)

Table 6. Forecast Error Evaluation (*Total in Thousands)

Scenarios	Predicted Values (2016)*	Actual Values (2016)*	Absolute Error (%)**
Scenario 1	778.6	845.6	7.92
Scenario 2	818.6	845.6	3.19
Scenario 3	851.1	845.6	0.65

**Absolute Percentage Error formula: $\frac{|\text{Approximate Value} - \text{Exact Value}|}{|\text{Exact Value}|} \times 100\%$

4. CONCLUSION

The objectives of the current study were to enhance the available public sources of tourism forecast information and contribute to the tourism stakeholder's strategy in Portugal. More specifically, to develop a multivariate model to forecast international tourism demand through a Data Mining approach. The forecasted variable was the nights spent at tourist accommodation establishments in Lisbon's region, one of the country's main foreign tourist destinations. The model development was constrained to publicly available data, machine learning methods, and minimum data assumptions.

The selected model (SMOReg) was successful in generalization capability, in resemblance to a similar case study in Turkey by Cankurt and Subasi (2016). Despite of having a high relative absolute error and a high Root Relative Absolute Error in the test phase, the model produced quite accurate forecasts, especially in scenario 3.

As previously discussed, both the European economy and the interest of UK residents in Lisbon's region kept growing slightly in 2016. The "matching" of the total forecasted values of the favorable scenario with the total actual values of nights spent at tourist accommodation establishments in Lisbon's region in 2016, provides evidence of the reliability of the proposed forecasting model.

The integration of this forecasting system with public information systems, such as the recent platform of tourism knowledge management (TravelBI) from the Instituto do Turismo de Portugal, INE databases, Eurostat and other European organizations, would provide up-to-date forecast information to tourism stakeholders.

In summary, if institutions and decision makers have information regarding the evolution of the explanatory variables used in this model, the impact on Lisbon's tourism demand can be assessed, even in case of an emerging recession.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the financial support of "Fundação para a Ciência e Tecnologia" (FCT), Portugal, through the MagIC research center (Centro de Investigação em Gestão de Informação).

This work was also financed through the Regional Operational Programme CENTRO2020 within the scope of the project CENTRO-01-0145-FEDER-000006.

REFERENCES

Arauzo-Azofra, A., Benitez, J. M., & Castro, J. L. (2004). A feature set measure based on relief. In *Proceedings of the fifth international conference on Recent Advances in Soft Computing*(pp. 104-109).

Armstrong, J. Scott (2001). *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Norwell, MA

Athanasopoulos, G., Hyndman, R. J., Song, H., & Wu, D. C. (2011). The tourism forecasting competition. *International Journal of Forecasting*, 27(3), 822–844.

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADIS European Conference Data Mining*, (January), 182–185.

Baggio, R., & Sainaghi, R. (2016). Mapping time series into networks as a tool to assess the complex dynamics of tourism systems. *Tourism Management*, 54, 23–33.

Cankurt, S., & Subasi, A. (2016). Tourism demand modelling and forecasting using data mining techniques in multivariate time series: a case study in Turkey. *Turkish Journal of Electrical Engineering & Computer Sciences*, 24(5), 3388-3404.

Claveria, O., Monte, E., & Torra, S. (2014). A multivariate neural network approach to tourism demand forecasting.

Constantino, H. A., Fernandes, P. O., & Teixeira, J. P. (2016). Tourism demand modelling and forecasting with artificial neural network models: The Mozambique case study. *Tékhne*.

Cunha, L. & Abrantes, A. (2013). *Introdução ao turismo* (5ª ed.). Lidel, Lisboa, Portugal

Daniel, A. C. M., & Ramos, F. F. R. (2002). Modelling inbound international tourism demand to Portugal. *The International Journal of Tourism Research*, 4(3), 193.

Frechtling, D. C. (2011). *Forecasting Tourism Demand: Methods and Strategies*. Routledge, New York, USA

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

IBM. *(2011). IBM SPSS Modeler CRISP-DM Guide, 53.* Retrived from: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_17.1.0/modeler_crispdm_ddita/modeler_crispdm_ddita-gentopic1.html

Kim, N., & Schwartz, Z. (2013). The accuracy of tourism forecasting and data characteristics: a meta-analytical approach. *Journal of Hospitality Marketing & Management*, 22(4), 349-374.

Li, G., Song, H., & Witt, S. F. (2005). Recent Developments in Econometric Modeling and Forecasting. *Journal of Travel Research*, 44(1), 82–99.

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

Teixeira, J. P., & Fernandes, P. O. (2012). Tourism Time Series Forecast -Different ANN Architectures with Time Index Input. *Procedia Technology*, 5, 445–454.

Witt, S. F., & Witt, C. A. (1995). Forecasting tourism demand: A review of empirical research. *International Journal of Forecasting*, 11, 447–475.

OECD (2016). *OECD Tourism Trends and Policies 2016*. Retrieved from: http://www.oecd.org/cfe/tourism/oecd-tourism-trends-and-policies-20767773.htm

Pordata (2016). Table: Exportações de serviços: total e por tipo – Portugal. Retrieved from: http://www.pordata.pt/Portugal/Exporta%C3%A7%C3%B5es+de+servi%C3%A7os+total+e+por+tipo-2352

INE (2016). Table: Produto interno bruto (B.1*g) a preços correntes (Base 2011 - €) por Localização geográfica (NUTS - 2013); Anual. Retrieved from: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrCod=0008836

Sheldon, P. J., & Var, T. (1985). Tourism forecasting: a review of empirical research. Journal of Forecasting, 4(2), 183-195.