11th International Conference on Modern Research in
MANAGEMENT, ECONOMICS and ACCOUNTING
18 – 20 December, 2020 Oxford, United Kingdom

Google Trends and Tourism: Regression Cluster Analysis

Miguel Ángel Ruiz Reina[1]
[1] *University of Malaga, Department of Theory and Economic History (Staff of Fundamentals), PhD. Program in Economics and Business, s/n, Plaza del Ejido, 29013 Malaga*

Abstract.

The appearance of Big Data technologies and economic development have led to a new generation of data that must be analysed. This new data structure allows identifying consumer behaviour; modelling offers and allowing countries to design Economic Policies efficiently. The service sector increasingly weights in the GDP of the nations; in particular, in Spain, it is around 12 per cent. Search engines from Google collect this data information by areas and time. In particular, the crossing of data from primary sources and secondary data is a comparative advantage over traditional analyses. We analyse datasets in this work coming from official statistical sources in Spain and data from Google Trends. The objective of this work is explaining the hotel demand trough previous searches of consumers from January 2008 to December 2019 for the Spanish case. The application of the method called Regression Cluster Analysis (RCA) represents an improvement in forecasting modelling of the literature. This model allows searching profiles to be made as a previous step to what is known in the literature as Next Best Activity. We compare three models (RCA, ADRL + Seasonality, SARIMA) thought the Ratio Theil U2, for a time horizon of six months. This methodology involves adding one more model to the debate on the use of Big Data. At the end of the work, we propose three ideas for "sustainable" development of this type of analysis with open data.

Keywords: Big Data, Forecasting, Google Trends

1. Introduction

The unstoppable development of Big Data technologies applied to the tourism sector allows making efficient decisions (Reina, 2020). Traditional Statistical forecasting techniques have been added to the Big Data environment, assuming a development for companies, consumers and institutions (Carrizosa et al., 2018; Kleinberg et al., 2018). The results of Statistical processes and Artificial Intelligence have been integrated into marketing strategies, business models and in the understanding of consumers. In the tourism industry, they have been introduced directly or indirectly through primary industries (tourism: mainly services) and secondary industry (activities dependent on tourism such as banking or the insurance sector). These methodological advances are of interest to academics and practitioners in order to anticipate market strategies for consumers (Columbus, 2019).

In the field of tourism, international flows have been providing resources of information for analysts. Studies reveal that people search before their tourist trip to find accommodations (Ruiz-Reina, 2019a). Tourism plays an increasingly important role in the Gross Domestic Products of developed countries. In particular, the study of hotel demand with the analysis of

Google Trends (GT) has provided a competitive advantage due to the information generated on the web for potential demands. This information allows adjusting supply and demand, causing efficient assignments in the market (Juul, 2015). This digital economy supposes the development of an unsuspected digital market in the beginnings of the analysis of tourism (Palos-Sanchez & Correia, 2018).

The use of GT supposes a quantitative and qualitative study over time. GT data allows studying regularity problems in the tourism market (Wu 2017; Li et al., 2018; Mariani et al., 2018). These regularities are represented by the number of repetitions of search terms in GT, and they have been temporally quantified. The ordering of these open data implies an analytical advantage over traditional Statistics.

In this article, we are interested in constructing a Regression Cluster Analysis (RCA) to reveal previous search interests of tourist accommodation consumers. In particular, the study of the demand for Hotel Accommodation in Spain (HADS) is carried out. This study exceeds the predictive capacity in comparison to other models of literature; the main reason is the inclusion of larger volumes of data. In the empirical section, we are comparing RCA for monthly predictions from January 2019 to June 2019. For the support and understanding of our RCA model, we must include seasonal deterministic variables (Reina, 2020). Based on RCA, it is possible to identify information insights to be able to make efficient decisions. These insights allow determining the propensity factors for the impact of advertising strategies according to the hotel accommodation industry. In the context of the development of our work, the data analysed are monthly and with a forecasting training period of 6 months. We have used and predicted in the medium-term horizon to make decisions; however, the same methodology is expandable in the long term.

The reference models to compare our contribution are the traditional Seasonal Autoregressive Integrated Moving Average Models - SARIMA (Box et al., 2013) and Autoregressive Distributed Lags extended to Seasonality ARDL + Seasonality (Reina, 2020). As a criterion of accuracy, we have applied the Ratio Theil U2 (Ruiz-Reina, 2019b). For predictions with a 6-month time horizon, we have quantified improvements in the RCA model compared to SARIMA and ADRL + Seasonality.

The rest of the research is ordered as follows: Section 2 provides a review of the existing literature on the forecasting of Big Data applied to tourism; in section 3, the theoretical methodology is described; in section 4, data analysis of Official Spanish Statistics and data from GT is done; section 5 is dedicated to analysing empirical results. Once the empirical analysis is finished, we will discuss the conclusions of this work and future lines of research.

2. Literature review

The problem of data grouping has been addressed by researchers from different branches such as Economics, Finance, Statistics, among others. Clustering methods are methods that allow simplifying reality to find conclusions (Liao, 2005; Aghabozorgi et al., 2015; Caiado et al., 2015). Clustering methods with cross-section data are widely studied in the literature. In the scope of this study, we could mention about clustering applied to Time Series (Kakizawa et al., 1998; Alonso et al., 2006; Scotto et al., 2010; Vilar et al., 2010; D'Urso et al., 2017; Alonso & Peña, 2019; Alonso et al., 2020).

The novelty of this study lies in the use of alternative data sources to Official Statistics such as those obtained by GT. Technological development, commercial development, international mobility flows of people and data collection support this modelling. They allow the grouping of data according to RCA. The integration of this analysis can be applied to data from other primary or secondary branches of the tourism industry. Assuming a contribution to the recommendation systems recognized in Big Data such as Next Best Activity (NBA) marketing or next-best offers to Customers (Davenport et al., 2020).

In the scientific literature applied to tourism and forecasting, articles related to GT have been published (Li et al., 2018; Mariani et al., 2018). The applied techniques have been very varied, being roughly summarized in Econometric methods, Time Series, Machine Learning or Artificial Intelligence (Wu et al., 2017). RCA contributes to the literature an analysis tool with the use of GT with time series (Bokelmann & Lessmann, 2019).

The use of keywords in the literature with GT has allowed increasing the predictive capacity of the ADRL + Seasonality models. The use of the keyword "visit Spain" has allowed improving the previous models significantly (Reina, 2020). In our RCA model, in addition to increasing the predictive capacity, there is a value added since it allows finding customer profiles for future NBA tasks.

For the analysis of the goodness of accuracy measures, we have focused on the Ratio Theil U2. The objective of this measure is to be able to quantify the improvement that occurs from the use of one model concerning another (Ruiz-Reina, 2019b). The Ratio Theil U2 is a relative measure of accuracy between models, in the forecasting literature, the benefits of the criteria have been discussed (Hyndman & Koehler, 2006).

In the literature review, the most modern and close references to the RCA technique have been cited. The authors are aware of the limitations of this analysis since the same idea can be expressed with different words in each language around the world. But we consider that the study of “broad matching” will allow us to obtain a well predictive capacity with data from GT In the next sections, we will continue with the theoretical and empirical analysis.

3. Methodology

The objective of this methodological section is to describe a univariate technique for Time Series forecasting and two theoretical causal models. SARIMA models are widely developed in the literature for univariate time series (Box et al., 2013). According to causal modelling, we will describe the ADRL + Seasonality models (Reina, 2020) y and we develop RCA. Once this theoretical section has been described, we will work with the empirical model under the parsimonious criterion. Parsimony implies finding the simplest model with the most significant explanatory power.

3.1.1 Automatic SARIMA

The relevance of the developments in programming allows estimating automatically the models proposed by Box-Jenkins in the 70s (Box et al., 2013). In this case, we will perform forecasting tasks with the ARIMA R automatic package (Hyndman & Khandakar, 2008). A generic scheme for modelling could be:

$$\phi_p(B^p)\Phi_P(B^P)\nabla^d\nabla_s^D Y_t = \theta_q(B^q)\Theta_Q(B^Q)\varepsilon_t \quad (1)$$

Where p and q represents nonseasonal-ARIMA order, and d is the number of regular differences (P, D, Q represent seasonal order of ARIMA); B is the backshift operator and could be defined $Y_t(B^p) = Y_{t-p}$; $(\phi_p, \Phi_P, \theta_q, \Theta_Q)$ are parameters to be estimated; $\nabla^d\nabla_s^D$ describe regular and seasonal differences; ε_t is the random term that is distributed as white noise. In our work, we express the endogenous variable (Y =HADS) as natural logarithms.

3.1.2 Autoregressive Distributed Lag + Seasonality (ADRL + Seasonality)

According to Reina (2020), ADRL + Seasonality model is expressed as follows:

$$\gamma(L)\ln(y_t) = \mu + \delta(L)\ln(x_t) + \sum_{i=2}^{12}\alpha_i w_i + u_t \quad (2)$$

Where $u_t \sim white\ noise$. The exogenous variables are x_t and for endogenous y_t. The natural logarithm is $\ln$ and (L) is the lag operator. Stability conditions are found if inverted roots are $|\gamma(L)| < 1$. To avoid the dummy variable trap, we do not consider $i = 1$ (Draper & Smith, 1998) and the seasonal dummy variables are defined as monthly data ($i = 1, 2, \ldots, 12$). In our research, we use the ADRL + Seasonality (1,0) model where the endogenous variable y_t is HADS, and the exogenous x_t is the keyword "visit Spain" (Ruiz-Reina, 2019a).

3.1.3 Regression Clustering Analysis (RCA)

Big Data applications mainly present problems of human-computer interpretation and vice versa. Each language has its jargons, which makes keyword analysis more difficult by researches. To facilitate our task of finding Keywords which reveal interest of HADS, GT presents geographically related search term with correlated words, related topics, or related queries. This correlation allows us to create clustering of information, which is given in regression as follows:

$$y_t = f(cluster_{kt}) + \sum_{i=1}^{12}\alpha_i w_i + u_{1t} = \sum_{m=1}^{j}\beta_{mk} keyword_{mkt} + \sum_{i=2}^{12}\alpha_i w_i + u_{1t} \quad (3)$$

$$\left\{keywords_{mkt} \in \square^{+};\ m = 1, 2, 3\ldots; t \in T = 1, 2, 3\ldots T\right\}$$

In our model, $keywords_t$ are words that future consumers search on the internet before their tourist demand. According to the GT coding, the least interest is shown when the index is 0 and the maximum when it is 100. For instance, "hotels Spain", "surfing Spain" or "culture in Spain" among others. The selection of the variables is taken into account based on the possible collinearity; to solve this, we apply the Klein technique to solve the deficiencies (Christie et al., 1984). With the use of HAC covariance method, to make inference consistently about the parameters (Newey & West, 1987).

Under the parsimony criterion, the model is simplified, seeking the fewest number of significant explanatory variables with explanatory capacity. The interpretation of coefficients are elasticities for the keywords, and in the seasonal dummy variable are semi-elasticities (Peng et al., 2015).

3.1.4 Comparison of forecasting and evaluation

As a model selection criterion, we will use the Ratio Theil U2. It is a dimensionless relative measure designed for the decision to select predictive models (Ruiz-Reina, 2019a). We initially define the index Theil U2 (Christ & Theil, 1962):

$$U_2 = \frac{\sqrt{\sum_{h=1}^{H}\left(\frac{\hat{y}_{t+h}-y_{t+h}}{y_t}\right)^2}}{\sqrt{\sum_{h=1}^{H}\left(\frac{y_{t+h}-y_t}{y_t}\right)^2}} \quad (4)$$

Finally, we can define the Ratio Theil U2 for comparison between two models v and w as follows:

$$RTh\acute{U}_{v/w;2} = U_{v;2}/U_{w;2} \quad (5)$$

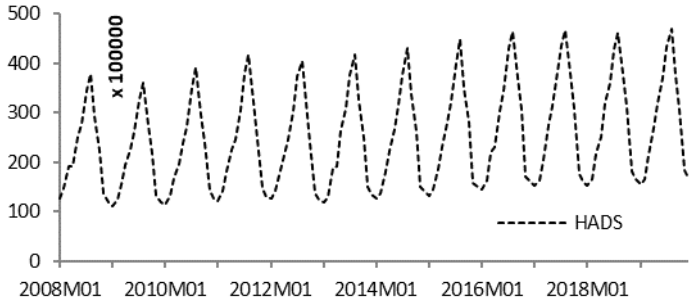
The results can be three: when it is equal to 1 (both models have the same predictive capacity); when it is more than one, the denominator model is better in terms of forecasting; when it is less than one, the numerator model is a better predictor.

4. Data

HADS data has been collected for the training period (January 2008 to December 2019) from the INE[1]. Concerning the exogenous variables, they are obtained from the GT for groups of variables the same frequency, and there are no missing observations.

In Figure 1, HDAS presents some relevant characteristics for the analysis of time series, it is worth noting a high seasonality and a growing upward trend throughout the period analysed.

Fig. 1 Number of HADS (Jan. 2008 to December 2019).



Source: INE. Own Elaboration.

[1] Spanish National Statistics Institute (Instituto Nacional de Estadística) https://ine.es/

From a descriptive point of view, it should be noted that the maximum values for each year occur in the summer season (every August). The highest value in August 2018 with 46,657,187 hotel overnights in Spain, and the lowest value in January 2009 with 11,203,819. For the 144 observations analysed (Table 1) the Std. Dev. = 9,978,231 and Skewness = 0.46. Kurtosis is 2.05.

Table 1 Descriptive Statistics Analysis of HADS (Jan. 2008 to December 2018). Observations=144.

	Mean	Maximum	Minimum	Std. Dev.	Skewness	Kurtosis
HADS	24,661,970	46,657,187	11,203,819	9,978,231	0.46	2.05

Source: INE. Own Elaboration.

The most common secondary data information service are those that come from Google; this is due to the large number of devices that use their services. GT began offering information in 2004, and its information classification algorithms have been improved. Keywords and their trends are summarized in indexes between 0 and 100, with 0 being the least interest and the maximum 100. Concerning the broad matching obtained by GT, we have studied a sample which involved 18,000 observations contemporary to HADS and that the summary of most correlated keywords appears in the following table (Choi & Varian, 2012):

Table 2: Summary of clusters and keywords (broad matching) relevant for HADS. Sample Jan. 2008- Dec. 2018.

cluster	relevant keywords
Sports	sport
Laws	visa
Transport	car, flight
Seasonality	summer, winter
Social	Airbnb, Youtube, English, Tripadvisor, Twiter
Tech.	Android, Xiaomi, Apple, Samsung
Searches	low-cost, Spain Tourism, visit Spain
Culture	alcohol, city breaks, monuments, architecture
Places	Beach, Canary Island, Alhambra, Plaza de España, Sagrada Familia

Source: Google Trends. Own Elaboration

In the next section on empirical results, we will estimate the models described in the methodological section. The results of the best model with the highest predictive capacity will be determined by the lowest Ratio Theil U2 with a time horizon h = 6.

5. Empirical results

In the results section, we are showing the results of the estimated models with forecasting purpose. The sample studied is 132 observations between January 2008 to December 2018. The model finally estimated for RCA and corrected by HAC is the following:

$$\hat{y}_t = \underset{(0.00)}{15.84} + \underset{(0.04)}{0.02}\, Airbnb_t + \underset{(0.00)}{0.07}\, Apple_t - \underset{(0.00)}{0.12}\, car_t + \underset{(0.00)}{0.03}\, city\_breaks_t + \underset{(0.00)}{0.08}\, flight_t \quad (6)$$

$$- \underset{(0.00)}{0.08}\, Samsung_t + \underset{(0.01)}{0.03}\, sport_t + \underset{(0.00)}{0.16}\, visa_t + \sum_{i=2}^{12} \vartheta_i w_i$$

The RCA estimation results show a high descriptive and being the parameters interpreted as elasticities. The results include clusters related to transport (car, flight), accommodation/social (Airbnb), legal aspects (visa), technology (Apple, Samsung), cultural (city breaks) and sports. This reduction of a universe of words implies a high explanatory capacity (0.99).

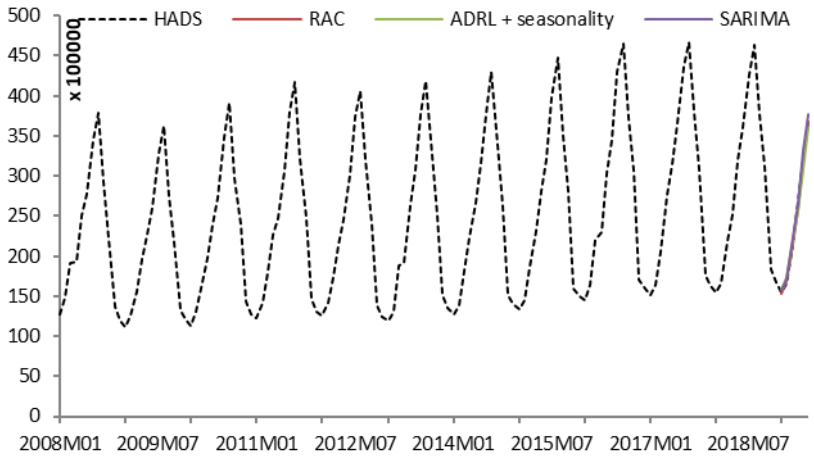
As we observed in the data section, a fundamental component in the tourist accommodation variables in Spain is seasonality. In our estimated model, it would be expressed as follows:

$$\sum_{i=2}^{12} \hat{\vartheta}_i w_i = \underset{(0.00)}{0.10}\, w_2 + \underset{(0.00)}{0.34}\, w_3 + \underset{(0.00)}{0.50}\, w_4 + \underset{(0.00)}{0.69}\, w_5 + \underset{(0.00)}{0.82}\, w_6 \quad (7)$$

$$+ \underset{(0.00)}{1.05}\, w_7 + \underset{(0.00)}{1.16}\, w_8 + \underset{(0.00)}{0.90}\, w_9 + \underset{(0.00)}{0.69}\, w_{10} + \underset{(0.00)}{0.18}\, w_{11} + \underset{(0.00)}{0.10}\, w_{12}$$

The prediction of the final RCA model is compared to other models cited in the Introduction section. The comparative graph of the forecasting Time Series can be seen in Figure 2.

Fig. 2 Out-sample forecast HADS h= 12 (Jan. 2019 to June 2019).



Source: INE. Own Elaboration.

Table 3 below shows the comparison between the RCA model and the other predictive models (ADRL + Seasonality, SARIMA) using Ratio Theil U2. We can see that the RCA model presents a better forecasting capacity than ADRL + Seasonality and SARIMA. This is because the numerator of the ratio is less than the denominator. The estimated value for h = 6 implies a minimal difference compared to the actual value. Looking at Table 3, we can

conclude the following order of accuracy: first RCA, second SARIMA and third ADRL + Seasonality.

Table 3 Summary of forecasting accuracy. Out-Sample training Jan. 2019- Dec. 2019.

	RCA	ADRL + Seasonality	SARIMA
RCA (*h=6)*	1.00	0.50	0.69

Source: Own Elaboration.

Theoretical development has proposed a methodology to provide knowledge about search interests with the RCA model. In the empirical section, we have confirmed the high explanatory and predictive capacity in comparison with the two models used in the literature. We will proceed to the drafting of the conclusions.

6. Conclusions

The tourism industry in Spain represents approximately 12 per cent of the Gross Domestic Product in recent years (before to COVID-19). The studies of their predictions are of great interest due to their economic-social repercussion. The RCA methodology analysed in this work has allowed us to know the previous interests of the clients regarding hotel reservations. The main reason for this research has been to provide an econometric tool to quantify and predict customer search interests. Understanding the market can influence marketing strategies and consumer behaviour. The data sets have been from official sources and the data provided by GT for the analysis; we have established a training period between January 2008 and December 2018. This period implies being able to validate our model as a previous step to the forecasting task between January 2019 to June 2019. In a context, such as the current one due to the crisis of the coronavirus. The keywords of interest would be modified, and the most relevant would be “covid 19” or “coronavirus”. Being aware of the relevance and structural change, researchers have decided to formulate a general theoretical framework whose application implies knowledge for decision-making.

In interpreting the estimated RAC model, it should be remembered that the estimated parameters are interpreted as elasticities. They are highlighting a word of interest belonging to the cluster of laws that is "visa". According to the analysed data, an increase of one per cent of the word “visa” will imply an increase of approximately 13 per cent ceteris paribus. In second-order of importance, the percentage variations of the name "car" indicate variations of roughly 12 per cent. The rest of the words have an interpretation of around 6 and 8 per cent. The terms such as “Airbnb”, “city breaks” and “sport” have less relevance comparing to other explanatory variables. Presenting predictive advantages observed in the Ratio Theil U2 and also providing more information in the knowledge of the future client. These identification tasks can have an impact on what we have called the NBA. Knowing the search words would imply a better understanding of the market and improve the correlations with the NBA. The positive implications for the tourist accommodation market are direct, and secondary industries could benefit from this modelling.

Finally, we can conclude that the modelling of this work can be used in industries such as Finance, Airlines, Insurance or any other in which it is desired to improve decision-making.

Big Data technologies provide open data that can be crossed and analysed to improve the well-being of the market (supply and demand). To examine the global impact of marketing strategies, we propose three significant areas of development: first, to analyse how market strategies will change as a function of algorithms and technologies; second, to identify changes in the interests of consumers according to their available technologies; third, a broad consensus on legality and data privacy. These are the bases for future research and "sustainable" development of this technological industry applied to Big Data analysis.

Acknowledgement:

The author wishes to acknowledge the support given by the University of Malaga. PhD. Program in Economics and Business, effective from July 16, 2013. Especially to Associate Professor Antonio Caparrós Ruiz from the Department of Statistics and Econometrics of the University of Malaga, for reviewing this work. Group of research: "SEJ157-INIDICADORES SOCIALES".

7. References

Aghabozorgi, S., Seyed Shirkhorshidi, A., & Ying Wah, T. (2015). Time-series clustering - A decade review. *Information Systems*, *53*, 16–38. https://doi.org/10.1016/j.is.2015.04.007

Bokelmann, B., & Lessmann, S. (2019). Spurious patterns in Google Trends data - An analysis of the effects on tourism demand forecasting in Germany. *Tourism Management*, *75*, 1–12. https://doi.org/10.1016/j.tourman.2019.04.015

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2013). Time series analysis: Forecasting and control: Fourth edition. In *Time Series Analysis: Forecasting and Control: Fourth Edition*. https://doi.org/10.1002/9781118619193

Caiado, J., Ann Maharaj, E., & D'Urso, P. (2015). Time-series clustering. In *Handbook of Cluster Analysis* (pp. 241–264). https://doi.org/10.1201/b19706

Carrizosa, E., Guerrero, V., & Romero Morales, D. (2018). Visualizing data as objects by DC (difference of convex) optimization. *Mathematical Programming*, *169*, 119–140. https://doi.org/10.1007/s10107-017-1156-1

Choi, H., & Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record*, *88*, 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x

Christ, C. F., & Theil, H. (1962). Economic Forecasts and Policy. *Econometrica*. https://doi.org/10.2307/1911296

Christie, A. A., Kennelley, M. D., William King, J., & Schaefer, T. F. (1984). Testing for incremental information content in the presence of collinearity. *Journal of Accounting and Economics*, *6*, 205–217. https://doi.org/10.1016/0165-4101(84)90025-9

Columbus, L. (2019). 10 Charts That Will Change Your Perspective Of AI In Marketing. In *Forbes*. https://www.forbes.com/sites/louiscolumbus/2019/07/07/10-charts-that-will-

change-your-perspective-of-ai-in-marketing/amp/

D'Urso, P., Maharaj, E. A., & Alonso, A. M. (2017). Fuzzy clustering of time series using extremes. *Fuzzy Sets and Systems*, *318*, 56–79. https://doi.org/10.1016/j.fss.2016.10.006

Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*, *48*, 24–42. https://doi.org/10.1007/s11747-019-00696-0

Draper, N. R., & Smith, H. (1998). "Dummy Variables." *Applied Regression Analysis*, 2999–326. ISBN 0-471-17082-8

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, *27*(3). https://doi.org/10.18637/jss.v027.i03

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Juul, M. (2015). Tourism and the European Union. Recent trends and policy developments. *European Paliamentary Research Service*. https://www.europarl.europa.eu/RegData/etudes/IDAN/2015/568343/EPRS_IDA(2015)568343_EN.pdf

Kakizawa, Y., Shumway, R. H., & Taniguchi, M. (1998). Discrimination and clustering for multivariate time series. *Journal of the American Statistical Association*, *93*, 328–340. https://doi.org/10.1080/01621459.1998.10474114

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *Quarterly Journal of Economics*, *133*(1), 237–283. https://doi.org/10.1093/qje/qjx032

Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. *Tourism Management*, *68*, 301–323. https://doi.org/10.1016/j.tourman.2018.03.009

Mariani, M., Baggio, R., Fuchs, M., & Höepken, W. (2018). Business intelligence and big data in hospitality and tourism: a systematic literature review. In *International Journal of Contemporary Hospitality Management*. https://doi.org/10.1108/IJCHM-07-2017-0461

Newey, W. K., & West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Author(s): Whitney K. Newey and Kenneth D. West Source: *Econometrica*, *55*(3), 703–708. https://doi.org/10.2307/1913610

Palos-Sanchez, P. R., & Correia, M. B. (2018). The collaborative economy based analysis of demand: Study of airbnb case in Spain and Portugal. *Journal of Theoretical and Applied Electronic Commerce Research*. https://doi.org/10.4067/S0718-18762018000300105

Peng, B., Song, H., Crouch, G. I., & Witt, S. F. (2015). A Meta-Analysis of International Tourism Demand Elasticities. *Journal of Travel Research*, *54*(5), 611–633. https://doi.org/10.1177/0047287514528283

Reina, M. Á. R. (2020). Big Data: Forecasting and Control for Tourism Demand. In R. I. Valenzuela O., Rojas F., Herrera L.J., Pomares H. (Ed.), *Theory and Applications of Time*

Series Analysis. ITISE 2019 (pp. 273–286). Springer, Cham. https://doi.org/https://doi.org/10.1007/978-3-030-56219-9_18

Ruiz-Reina, M. Á. (2019a). Big Data: Does it really improve Forecasting techniques for Tourism Demand in Spain? In I. Rojas (Ed.), *International Conference on Time Series and Forecasting* (pp. 694–706). Godel Impresiones Digitales SL https://itise.ugr.es/ITISE2019_Vol1.pdf

Ruiz-Reina, M. Á. (2019b). Forecasting using Big Data: The case of Spanish Tourism Demand. *International Conference on Time Series and Forecasting*, 782–789. https://itise.ugr.es/ITISE2019_Vol2.pdf

Scotto, M. G., Alonso, A. M., & Barbosa, S. M. (2010). Clustering time series of sea levels: Extreme value approach. *Journal of Waterway, Port, Coastal and Ocean Engineering*, *136*, 215–225. https://doi.org/10.1061/(ASCE)WW.1943-5460.0000045

Warren Liao, T. (2005). Clustering of time series data - A survey. *Pattern Recognition*, *38*, 1857–1874. https://doi.org/10.1016/j.patcog.2005.01.025

Westfall, P. H. (2014). Kurtosis as Peakedness, 1905–2014. RIP *American Statistician*, *68*(3), 191–195. https://doi.org/10.1080/00031305.2014.917055

Wu, D. C., Song, H., & Shen, S. (2017). New developments in tourism and hotel demand modeling and forecasting. *International Journal of Contemporary Hospitality Management*, *29*(1), 507–529. https://doi.org/10.1108/IJCHM-05-2015-0249