Big Data: Forecasting and Control for Tourism Demand

Miguel Ángel Ruiz Reina[1][0000-0001-6055-7810]

[1] University of Málaga, Department of Applied Economics (Statistics and Econometrics), PhD Program in Economics and Business, s/n, Plaza del Ejido, 29013 Málaga
ruizreina@uma.es

Abstract. In this study, innovative forecasting techniques and data source from Big Data are used for the study of Hotel Overnight Stays for Spain, from January 2018 to June 2019. The unstoppable development of the Tourism sector with the application of Big Data technologies, allow to make efficient decisions by economic agents. In this work, the use of the data collected from the Google Data Mining tools allows to obtain knowledge about Hotel Tourism Demand in Spain. The analysis carried out meets the four basic principles of Big Data analysis: volume, velocity, variety and veracity. In this setting, the methodology used corresponds to ARDL models and ECM models being developed Granger-Causality extended to seasonality. The first one explains easily when economic agents will make their decisions; while, the second one allows make forecasting for short-term and long-term. This fact means that tourist offers and demands can be perfectly adjusted at every moment of the year. As a criterion for the selection of models, the innovative Matrix U1 Theil is proposed, this allows to quantify how much a model is better than another in terms of forecasting.

Keywords: Big Data, Forecasting, Google Trends.

1 Introduction

The use of massive data in a digital environment has led to a disruptive change in the developed economies of the world. Before the appearance of the Big Data concept, the amount of data collected already exceeded the ability to process and analyse data. The generation of massive data by the millions of device users and data analysis have created an unsuspected digital economy decades ago [1].

The "Tourism Industry" [2] generates a quantity of data to be analysed. This sector increasingly has a greater weight in the Gross Domestic Products (GDP) and turn generates externalities in economic agents [3].

This paper introduces a modern unexplored analysis of the data generated on the internet network for the Spanish tourism accommodation market by country of origin. Innovative modelling of data processing from primary data sources (official sources) with secondary sources from Big Data (Google Trends - GT) are introduced following four basic principles of analysis: volume, velocity, variety and veracity. GT analyse the shift of searches throughout the time and reveal consumer intentions.

The main objective of this paper is to obtain forecasting on Hotel Overnight Demand in Spain (HODS) from January 2018 to June 2019 by establishing a causality model for monthly data. The multivariate method developed of Autoregressive Distributed Lags with seasonal variables (ARDL + seasonality) uses as an explanatory variable for HODS a search interest rate (generated by GT) and seasonal dummies variables for monthly data by country of origin. This second contribution is a very relevant fact since tourism agents will be able to make efficient decisions in the tourism market. To explain causation relations, the Granger-Causality test extended with seasonality is developed and modelling we will be able to identify when consumer interest occurs. Ultimately, a criterion for the selection of new models, such as Matrix U1 Theil, has been developed, and it will be applied in this paper [4]. The forecasting is compared with univariate techniques such as Seasonal Autoregressive Moving Average (SARIMA) and the relatively new non-parametric technique Singular Spectrum Analysis (SSA).

The remainder of this research is as follows: section 2 provides a review of the existing literature on the forecasting of Tourism Demand, influenced by the techniques of every epochs; in section 3 and 4, data analysis is initially carried out along with the methodological development and information criteria. The use of the criterion for the selection of predictive models based on Theil's index is considered a great contribution to the literature. In section 5 an empirical analysis is carried out verifying the application of the proposed methodology. Section 6 shows the conclusions and future lines of research for Data Scientists and some economics implications. Finally, there is a section for the bibliographical references used.

2 Literature review

Data science is a fundamental field for the exploitation and generation of knowledge to make decisions in efficiency. In the bibliographic research carried out the appearance of these new datasets from open data such as Google could modify the culture and business in the Tourism field [5].

Tourism Demand is caused by multiple exogenous factors and techniques have focused on obtaining robustness and dynamic modelling, scalability and granularity [6]. The variety of Big Data studies has been applied to Tourism research, making a great improvement in the area [7]. Traditionally these studies have been influenced by the techniques of the moment [8- 11]. However, researchers have found the need for greater integration between computational and scientific fields [12].

In our study we will carry out an analysis with novel techniques and will be compared with most used techniques, a contribution of this study is the use of Big Data [13] tools summarized in an index of relevance provided by GT.

2.1 Forecasting methods using Google search engines (Google Trends)

Previous researchers such as Lu & Liu [14] found correlations between Internet search behaviour and the flows produced by tourists. Shimshoni, Efron, & Matias [15]

concluded that 90% of the categories analysed are predictable, making a great contribution to the scientific literature (categories: Socio-Economics fields).

Using the R programming and developing several examples in which the GT tool is used, it is worth mentioning the study of Choi and Varian [16] to analyse the tourism demand in Hong Kong. They obtained models with high explanatory capacity (on average $R^2 = 73\%$) using ARDL. Gawlik, Kabaria & Kaur [17] concluded that the GT search popularity evolution offers a useful predictor of tourism rates for series of arrivals of Hong Kong. For the Charleston region (USA), practical and interesting applications were found on the use of search engine data. The main limitation is that it was done only in one city [18].

To carry out Chinese Tourists' forecasting, Yang, Pan, Evans & Benfu [19] proposed and demonstrated the valence of the use of search engines based on web searches comparing Baidu search engines with those of GT. In this sense, with data obtained through GT, comparing purely autoregressive models with ARDL models with seasonal dummy variables, short-term results were obtained for the case of Vienna with data from images, words search or videos on YouTube [20].

Studies from the use of GT have meant an improvement in predictions for the Caribbean area. Autoregressive Mixed-Data Sampling models represent an improvement over SARIMA (Seasonal Autoregressive Integrated Moving Average) and AR for 12-months predictions [21].

The study of the tourist flows from Japan to South Korea has been examined with the construction of the Google variable combining the lowest Mean Square Error (MSE) or the absolute average of forecast errors for monthly data. Finding the best results for the model that uses Google data [22].

In the case of tourist flows from Spain, Germany, UK and France, Google data was used with the construction of indicators through Dynamic and SARIMA models [23]. For tourist arrivals in the city of Vienna [24], Google Analytics data was extracted using Bayesian methods. In the case of Puerto Rico, the volume of searches has been studied to predict the hotel demand of non-residents with a Dynamic Linear Model. The results showed improvements in forecasting for time horizons greater than 6 months [25]. Google data has been used for the flow of tourists in Portugal [26] and tourists flows in Spain [27].

Irem Önder [28] compared forecasting models with web and/or image search indices regarding two cities (Vienna and Barcelona) and two countries (Austria and Belgium). Tourist Arrivals in Prague was analysed by Zeynalov [29] with the objective to assess whether GT were useful for forecasting tourists' arrivals and overnight stays in Prague with weekly data. The results confirm that predictions based on Google searches are advantageous for policy makers and business operating in the Tourism sector.

The online behaviour of hotel consumers for the United States of America was researched with Discrete Fourier Transformation using data from GT, with empirical evidence for its use in marketing strategies [30].

In the case of Amsterdam, it has been investigated by Rödel [31] on forecasting Tourism Demand using keywords related to "Amsterdam" in GT. With the development of Big Data technology in the last decades have emerged collaborative economy

companies [32]. They have carried out studies on a vacation rental company that operates worldwide but reducing it to results from the Iberian Peninsula. In 2018, a study was published on the online and offline behaviour of consumers, for US restaurants with Google and Baidu search engine data. [33].

The data provided by Google use an index that summarizes the interest of the search words, in the case of data from Baidu. Li, Pan, Law & Hyang [34] developed an index of interest with data from Baidu. Demonstrating the forecasting capacity of Dynamic Factor Model (GDFM) to forecast tourist demand in a destination for Monthly Beijing tourist volumes from January 2011 to July 2015. A relevant study using Machine Learning algorithms is the one developed by Sun, Wei, Tsui & Wang [35], using criteria for the selection of models such as Normalized Root Squared Error (NRMSE) and MAPE, in addition to using the Diebold-Mariano criterion to determine if the prediction differences are significant.

Measures of forecasting. As observed above, the Tourist Industry has had interest in the past, in the present and in the future, and it will continue to have it. Mainly because it is an industry signal of the evolution of the service economy. So, the modelling used is very diverse, one aspect to be taken into account has been the criteria of information on the selection of models. It has been observed in the literature review the use of Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE); Theil´s index [36- 39]; Symmetric Mean Percentage Error (SMAPE) [40]. Some authors developed the RMSE ratio [41, 42] and in this article, we will develop the Matrix U1 Theil as a criterion for the selection of forecasting models [4]. This method allows quantifying the gain of the use of one methodology versus another.

To summarize the review of the literature, we can say that new models have been used in Data Science. In this work, new methodologies are developed, such as the improved Ganger-Causality test for seasonal data. Dynamic models have been developed to analyse the forecasting capacity in the short and long term. Big Data tools have been used from one of the largest search engines in the world and a decision matrix on predictive capacity has been developed for different time horizons.

3 Methodology

In this section, the scheme (see Fig. 1) of the cycle between offer and demand in tourism has been developed under four basic principles of Big Data. Specifically, in our paper, the objective is modelling and forecasting, however we will suppose ad-hoc the data from the Data Warehouse [43]. In this sense, the data will come from official sources of the INE[1] and Google[2]. So, all of the Extraction, Transformation and Loading - ETL [44] work will come from the data engineering of these entities. The main objective is to make efficiencies predictions based on knowledge to improve the user experiences of Tourism Demand and the offers of the stakeholders.

[1] INE: Instituto Nacional de Estadística (Spain). The National Statistics Institute (Spain).
[2] www.google.com

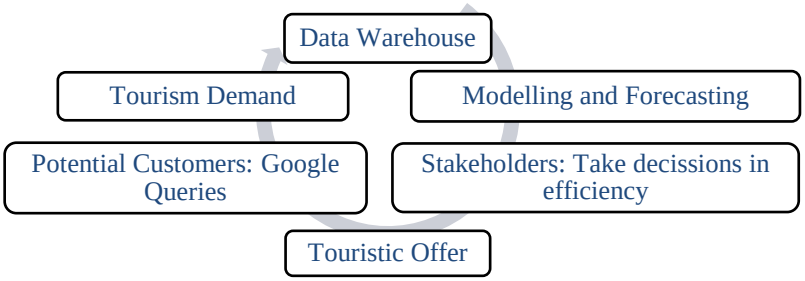

Fig. 1 Data life cycle and efficiency decision scheme. Own Elaboration.

3.1 Modelling and Forecasting evaluation.

In this paper, ARDL + seasonality model is proposed and its application with data from Big Data architectures is analysed. This modelling allows to know how HODS is generated through the searches of Google users (by country of origin). The purpose of this model is to know the causality relationship and to be able to make forecasts. To analyse the relationship between Granger-Causality and seasonality a test is developed. To evaluate the forecasting capacity is developed Matrix U1 Theil by country of origin. This matrix is developed to evaluate forecasting capabilities in order to obtain a comparative dimensionless measure among models. For a more in-depth detail of the predictions made, the reader can refer to the references of SARIMA [45] and Singular Spectrum Analysis [46]. All models are made for different scenarios and forecast comparisons are made for different time horizons h = 3, 6, 12, 18.

Granger-Causality and seasonality testing: ARDL and ECM. We develop the test proposed by Granger [47] and discussed by Montero [48] to detect the causality, since it is not observed with the simple analysis of correlation.

The model considered by Granger is for two variables (y_t, x_t). Due to the great influence of seasonality [49] in the Tourism sector, the following equation is proposed with HAC covariance method which determines the robust standard error for parameters estimated:

$$\ln(y_t) = \beta_0 \ln(x_t) + \sum_{j=1}^{m} \beta_j \ln(x_{t-j}) + \sum_{j=1}^{m} \alpha_j \ln(y_{t-j}) + \sum_{i=1}^{12} \delta_i w_i + \varepsilon'_t \quad (1)$$

Where w_i is a deterministic seasonal dummy (i = 1, …, 12) component and for monthly data is defined as follows:

$$w_1 = \text{-}1, \textit{ for others } w_i = 0$$
$$w_1 = \text{-}1, w_2 = 1 \textit{ for others } w_i = 0$$
$$w_1 = \text{-}1, w_3 = 1 \textit{ for others } w_i = 0$$
$$\vdots$$
$$w_1 = \text{-}1, w_{12} = 1 \textit{ for others } w_i = 0$$

The use of HAC covariance method guarantees efficiency of the parameters estimated. Once obtained ε'_t, this will be distributed as white noise.

The decision of causality with seasonal effects (Testing linear restrictions for parameters of x_{t-j} and w_i) is asymptotically ($T \geq 60$) as Chi-squared [50].

The most general expression of a dynamic model named ARDL[3] (m, n) with seasonal components is as follows [51, 52]:

$$\gamma(L)\ln(y_t) = \delta(L)\ln(x_t) + \sum_{i=1}^{12} \alpha_i w_i + \varepsilon_t \tag{2}$$

With the interest of evaluating the dynamic persistence of an effect on the exogenous variable at a certain moment, the Error Correction Model (ECM regression or ARDL Error Correction Regression) is constructed. The ECM[4] regression is as follows:

$$\Delta\ln(y_t) = \delta_0\Delta\ln(x_t) + \sum_{j=1}^{n} \lambda_j \Delta\ln(x_{t-j}) + \sum_{j=1}^{m} \delta_j \Delta\ln(y_{t-j}) - \gamma(L)\left[\ln(y_{t-1}) - \beta\ln(x_{t-1})\right] + \sum_{i=1}^{12} \alpha_i w_i + \varepsilon_t \tag{3}$$

In this model, short term effect is represented by parameters of first variables differentiated, while long term effects $|\gamma(L)|<1$ is represented by Correction Error term. According to Zivot [53], if long term effect is not statically significant, cointegration does not exist. Long-run multiplier is defined as $\beta = \dfrac{\delta(L)}{\gamma(L)}$

Forecasting Evaluation: Theil´s measures. To verify the forecasting accuracy of different models, we adopted an evaluation criterion to compare the out-sample forecasting performance. We will work with the inequality index of Theil [36]:

$$U_1 = \frac{\left[\frac{1}{h}\sum_{h=1}^{18}\left(y_{T+h} - \hat{y}_{T+h}\right)^2\right]^{1/2}}{\left[\frac{1}{h}\sum_{h=1}^{18}\left(y_{T+h}\right)^2\right]^{1/2} + \left[\frac{1}{h}\sum_{h=1}^{18}\left(\hat{y}_{T+h}\right)^2\right]^{1/2}} \tag{4}$$

Ratio Theil´s (RT´s) is designed to comparisons between predicted variables with horizons h= 3, 6, 12 ,18.

$$RT's_{y_{it}, y_{jt}} = \frac{U_1^{y_{it}}}{U_1^{y_{jt}}} \tag{5}$$

In the mathematical interpretation of the RT´s, three situations are described according to the predictive capacity of models: if the RT´s is equal to one, both models have the same explanatory capacity; if the ratio is greater than one, this would indi-

[3] m is the number of endogenous variables y_t (HODS); n is the number of exogenous variables x_t (Google Queries). ln is the Natural Logarithm. (L) is the Lag operator. Stability conditions: if inverted roots are $|\gamma(L)|<1$.

[4] Granger-Engle representation theorem and parameters are estimated in two stages. Consistency and Efficiency of estimators are fulfilled.

cate that the denominator's model has a better explanatory capacity than that of the numerator; if the ratio is less than one, the numerator´s model has better predictive results than the denominator.

4 Data

The Data of the number of HODS has been collected by INE. For the number of tourists in Spain by country of origin, dataset from the first month of 2010 to June of 2019 were obtained. In the grouping of nationalities, the name of "Resident abroad" should be noted. This includes all foreigner nationalities except for the 5 main nationalities described in the table (Germany, France, Italy, Netherlands, UK, USA).

According to the data represented in the Fig.2, the average of Residents Abroad was 16,180,005.75 in the period cited. The maximum number of hotel occupancy was recorded in August 2017 with 29,594,071 and the minimum 11.887.105 in January 2010.

To obtain data from Google, the Big Data tool called GT has been used. Previously GT tools have been used to make forecasts as is cited in the literature review. The lowest interest occurred in December of the year 2010. Analysing the data obtained of interest for the keyword or Google Query (GQ) "visit Spain", the greatest worldwide interest of the word was in May 2017, just with three periods of advance to the maximum historical overnight stays in Spain.

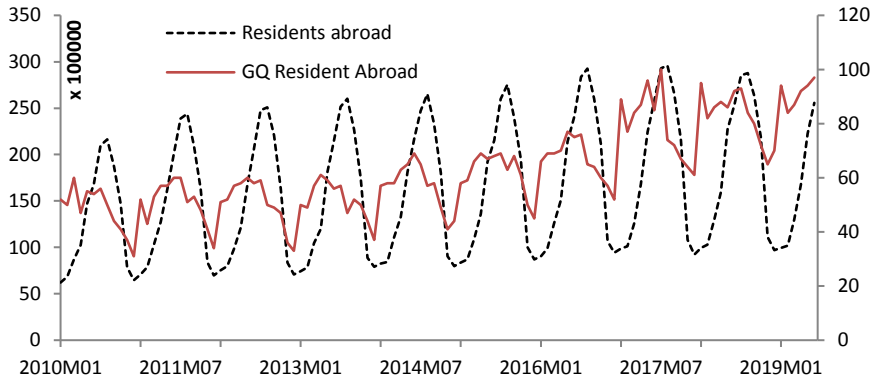

Fig. 2 number of HODS and keyword “visit Spain” for Resident abroad (Jan. 2010 to June 2019). Own Elaboration.

With the observation of the maximum and minimum values of both series analysed, it is observed graphically that searches on the Internet are made with at least one period in advance.

Table 1 displays a summary of variables selected by nationalities: Hotel demand and GQ. According to the two series selected, it is worth mentioning that only the

variable "Google Queries" in the case of Residents abroad (and USA HODS) meets the hypothesis of normality at 95% confidence (Jarque-Bera). As for stochastic trends (ADF test), all nationalities have unitary roots in Hotel demand and only three cases have been found in which there is evidence of unit root: they are the Google Queries of the Residents abroad, UK and USA. Regarding the stationarity in variance (KPSS), a more stationary behaviour is observed in the Hotel Demand variable for all nationalities including Residents abroad. On the other hand, in the Google queries variable, there is a clearly non-stationary behaviour in the series of Residents Abroad, UK and USA.

Table 1 Mean and Stationary Analysis of HODS and Keyword "visit Spain" sample period Jan. 2010 to December 2017. P-values in brackets. Own Elaboration.

Hotel demand	Mean	Jarque- Bera	ADF	KPSS
Residents abroad	16,180,005.75	10.03 (0.01)	-1.50 (0.52)	0.49
Germany	3,846,629.63	13.23 (0.00)	-2.35 (0.15)	0.08
France	1,231,000.87	16.41 (0.00)	-1.36 (0.59)	0.51
Italy	711,484.83	76.14 (0.00)	-1.01 (0.74)	0.10
Netherlands	645,451.61	8.25 (0.02)	-0.60 (0.86)	0.43
UK	4,113,511.96	11.72 (0.01)	-1.55 (0.50)	0.35
USA	437,373.41	5.22 (0.07)	1.19 (0.99)	0.67
Google Queries (GQ)	Mean	Jarque- Bera	ADF	KPSS
Residents abroad	62.62	4.98 (0.08)	1.53 (0.99)	1.07
Germany	47.02	7.71 (0.02)	-3.70 (0.00)	0.53
France	44.29	8.41 (0.01)	-10.67 (0.00)	0.65
Italy	28.28	29.93 (0.00)	-9.51 (0.00)	0.49
Netherlands	38.04	25.15 (0.00)	-9.16 (0.00)	0.49
UK	41.54	11.38 (0.00)	-0.76 (0.81)	1.14
USA	57.11	8.23 (0.01)	1.29 (0.99)	0.95

5 Empirical results

The empirical results obtained from the application of the previously proposed methodology section are briefly summarized in the following text. In this paper of predictive techniques, we will focus expressly on the dynamic model with explanatory variables of Internet searches ("visit Spain") and seasonal factors. The Granger-Causality test extended to seasonality confirms this hypothesis at least within 95 per cent of confidence. As usual in the literature, the forecasting is carried out for time horizons h=3, 6,12,18 months. Moreover, this article considers the training period January 2010-December 2017 and out-sample period January 2018-June 2019.

The results obtained through the Granger-Causality test including seasonal factors have determined that the number of HODS could be explained by the number of searches generated on the internet and by a systematic seasonality.

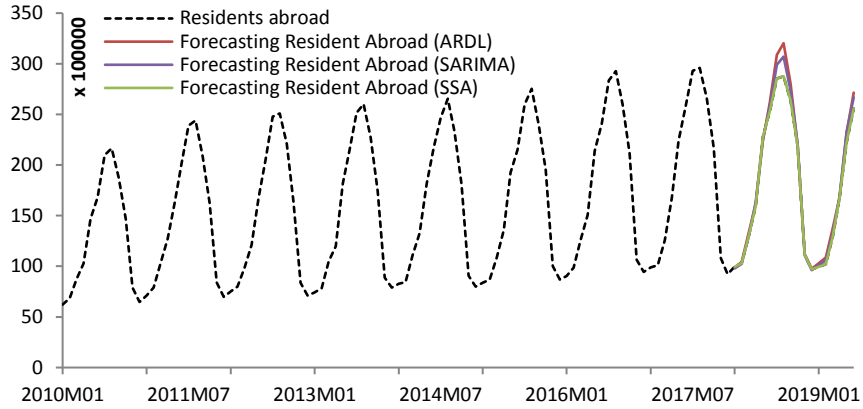

Fig. 3 Out-sample forecast HODS h= 18 (Jan. 2018 to Jun. 2019). Own Elaboration.

The ECM with seasonality obtained for residents abroad is as follows (lags selected under Akaike Info Criterion):

$$\Delta \ln(\hat{y}_t) = \underset{(0.00)}{-0.28}\Delta \ln(x_t) - \underset{(0.03)}{0.13}\left[\ln(y_{t-1}) - \underset{(0.00)}{0.55}\ln(x_t)\right] + \sum_{i=1}^{12}\hat{\alpha}_i w_i + \hat{\varepsilon}_t$$

$$Sample:\ 2010M1\ 2017M12\ R^2 = 0.9888$$

$$\sum_{i=1}^{12}\hat{\alpha}_i w_i = \underset{(0.03)}{-22.41}w_1 + \underset{(0.02)}{1.86}w_2 + \underset{(0.01)}{2.05}w_3 + \underset{(0.01)}{2.08}w_4 + \underset{(0.00)}{2.23}w_5 + \underset{(0.01)}{2.13}w_6 + \underset{(0.01)}{2.11}w_7 + \underset{(0.02)}{2.01}w_8 + \underset{(0.04)}{1.81}w_9 + \underset{(0.06)}{1.65}w_{10} + \underset{(0.18)}{1.16}w_{11} + \underset{(0.08)}{1.48}w_{12}$$

In the model defined for the HODS resident abroad variable, two aspects stand out (p-values in brackets): firstly, the existence of a cointegration relationship; second, the strong influence of seasonality. Table 2 shows models and results for HODS by country of origin.

Table 2 Summary of ARDL + seasonality models by country of origin for HODS. Sample Jan. 2010- December 2017. The table shows no relevant seasonality (months). Own Elaboration.

Hotel demand	ARDL	EC term (Prob)	seasonality	R^2
Germany	(2,0)	-0.34 (0.00)	-	0.97
France	(4,0)	-0.06 (0.03)	2, 10 ,11, 12	0.97
Italy	(2,1)	-0.11 (0.01)	9, 10, 11	0.97
Netherlands	(4,2)	-0.12 (0.01)	11, 12	0.96
UK	(1,1)	-0.07 (0.09)	7, 8 ,9 ,10 ,11, 12	0.99
USA	(3,0)	-0.10 (0.05)	2, 8, 10, 11, 12	0.97

It emphasizes, on the one hand, that all models show a long-term relationship (except for the UK) with a 95 per cent confidence level (USA with 90 per cent). On the other hand, all models are affected by the monthly seasonality, highlighting the fact that the German country of origin every month is significantly different to zero.

Once the results of the three forecasting models cited in the methodology section have been obtained by nationalities of tourists who visit Spain, the RT´s can be applied to quantify which model is better in predictive terms.

The results of the forecasting accuracy (see Table 3) depend on the time horizon used and the country of origin analysed.

Table 3 Matrix U1 Theil forecasting evaluation (Jan. 2018 to June 2019): RT´s by country of origin. Own Elaboration.

h	Ratio Theil	Residents Ab.	Ger.	France	Italy	Net.	UK	USA
3	SSA/SARIMA	236.84	5.57	5.08	2.33	9.86	5.88	3.48
	ARDL/SARIMA	2.44	0.83	0.98	1.51	1.41	0.84	0.88
6	SSA/SARIMA	76.62	1.47	2.71	1.73	1.75	4.64	2.92
	ARDL/SARIMA	1.46	1.01	0.74	1.30	0.48	1.47	0.85
12	SSA/SARIMA	33.45	1.55	3.53	4.10	0.73	1.58	1.39
	ARDL/SARIMA	1.66	0.79	1.14	2.24	0.43	1.15	0.63
18	SSA/SARIMA	33.50	1.33	3.79	1.96	0.69	1.79	1.00
	ARDL/SARIMA	1.57	0.67	1.11	1.71	0.38	1.05	0.37

In general, we can say that SARIMA models have obtained better results than SSA models (except the Netherlands with h = 12, 18). On the other hand, when comparing with the ARDL causal models with seasonality, the diversity of the results does not allow us to conclude which model has the best forecasting capacity. With a time horizon of 3 months, SARIMA presents the best results in three nationalities of origin (Residents abroad, France, UK), for the rest they have obtained better results of forecasting with ARDL seasonally. For a 6-month time horizon, the best results of ARDL with seasonality have been obtained for France and the Netherlands against SARIMA. For the 12-month and 18-month time horizons, the gains from using ARDL models with seasonality are observed in the German and Netherlands nationalities. For the rest of the cases, the SARIMA models are superior to those analysed in this paper.

6 Conclusions

In this paper, the importance of Forecasting modelling and historical analysis carried out in the literature review has been highlighted. The four dimensions of Big Data have been discussed: *volume*, the technologies coming from Google tools for data ETL have allowed analysing the main markets of origin tourism in Spain; *velocity*, related to the volume of data, the data engineering provided by Google technologies allow us to monitoring the Tourism Demand search intentions of the main nationalities who visit Spain; variety, the use of primary data source (INE) and secondary

(Google) have allowed build knowledge based on the data. This last one is a novel aspect in the analysis since the users show their interest through the search of information on the Internet; *veracity* of the data verified through the cointegration contrasts carried out. They have allowed modelling the forecasts of Spanish hotel demand by country of origin.

In addition, this article has used more common techniques (SARIMA or ARDL) with a novel technique named SSA. The contribution, in particular, can be divided into the following points:

1. A Granger-Causality test extended to seasonality has been developed. In the literature it was usual to perform only the contrast between endogenous and exogenous variable.
2. A criterion of model's selection based on the predictive capacity of the models has been developed (RT´s). In previous literature work, the gain in the use of models has not been quantified. Theil ratio quantifies the gain between pairs of models.
3. Related to the previous point, Econometric modelling with data from Big Data technologies does not guarantee an improvement in forecasting capacity. It has been demonstrated by main nationalities who visit Spain.
4. Concerning the dynamic models with seasonality, we have empirically demonstrated that hotel demand decisions are made with at least a period in advance.
5. Cointegration relationship has been revealed expressed in the ECM model.

We can conclude that the models used in this work improve the explanatory capacity of causality (R^2 close to 1) and cointegration relationships have been demonstrated, provide seasonal knowledge in decision making for the Spanish Tourism Demand. According to the results obtained, it is not possible to conclude that there is a gain in terms of forecasting by the use of tools from Big Data engineering; in contrast to what some authors claims [35]. The econometric interpretation of causality models and the economic interpretation can facilitate an adjustment of the offer in terms of prices or even advertising to the agents interested in visiting Spain. This article has been the basis of future research in which data from Big Data technologies are used to make efficiency decisions. The theoretical framework could be developed in fields where online markets are relevant. The preferred frameworks for this type of analysis could be Finance, Automotive, Insurance or any sort of market which implies searches on the internet network and this is translated into a quantification of the final decision of the consumer.

References

1. J. García, J. M. Molina, A. Berlanga, M. Á. Patricio, Á. L. Bustamante and W. R. Padilla: Ciencia de Datos: Técnicas Analíticas y Aprendizaje Estadístico. Un enfoque práctico. Alfaomega, Tarragona (2018).
2. M. Juul: Tourism and The European Union: Recents Trends and Policy Developments.

European Parliamentary Research Service, (2015).

3. S. Pegg, I. Patterson and P. Vila Gariddo: The impact of seasonality on tourism and hospitality operations in the alpine. International Journal of Hospitality Management 31, 659-666 (2012).
4. M. Á. Ruiz-Reina: Big Data: Does it reallly improve Forecasting Techniques for Tourism Demand in Spain?. In: ITISE 2019: International Conference on Time Series and Forecasting on Proceedings of Papers, pp. 694-706. Godel Impresiones Digitales S.L. Granada (2019).
5. B. J. Jansen: Review of "The search: how google and its rivals rewrote the rules of business and transformed our culture". Information Processing and Management: an International Journal 2(5), 1399-1401 (2006).
6. D. C. Wu, H. Song and S. Shen: New developments in tourism and hotel demand modeling and forecasting. International Journal of Contemporary Hospitality Management 29(1), 507-529 (2017).
7. J. Li, L. Xu, L. Tang, S. Wang and L. Li: Big data in tourism research: A literature review. Tourism Management 68, 301-323 (2018).
8. C. Li, H. Song and S. Wit: Recent Developments in Econometric Modeling and Forecasting. Journal of Travel Research 44(1) (2005).
9. H. Song and G. Li: Tourism demand modelling and forecasting- A review of Recent research. Tourism Management 29(2), 203-220 (2008).
10. B. Peng, H. Song and G. I. Crouch: A meta-analysis of international tourism. Tourism Management 45, 181-183 (2014).
11. E. Xiaoying Jiao and J. Li Chen: Tourism forecasting: A review of methodological developments over the last decade. Tourism Economics XX (X), 1-24 (2018).
12. M. Mariani, R. Baggio, M. Fuchs and W. Höepken: Business intelligence and big data in hospitality and tourism: a systematic literature review. International Journal of Contemporary Hospitality Management, (2018).
13. E. S. Silva, H. Hassani, S. Heravi and X. Huang: Forecasting tourism demand with denoised neural networks. Annals of Tourism Research 74, 134-154 (2019).
14. Z. Lu and N. Liu: The guiding effect of information flow of Australian tourism website on tourist flow: process, intensity and mechanism. Human Geography 22(5), 88-93 (2007).
15. https://www.researchgate.net/publication/238115677_On_the_Predictability_of_Search_Trends, last accesed 2019/11/06.
16. https://static.googleusercontent.com/media/www.google.com/es//googleblogs/pdfs/google_predicting_the_present.pdf , , last accesed 2019/11/06.
17. http://cs229.stanford.edu/proj2011/GawlikKaurKabaria-PredictingTourismTrendsWithGoogleInsights.pdf, , last accesed 2019/11/06.
18. B. Pan, D. C. Wu and H. Song: Forecasting hotel room demand using search engine data. Journal of Hospitality and Tourism Technology 3(3), 196-210 (2012).
19. X. Yang, B. Pan, J. A. Evans and L. Benfu: Forecasting Chinese Tourist volume with

search engine data. Tourism Management, (2015).

20. I. Onder and U. Gunter: Forecasting Tourism Demand with Google Trends: The Case of Vienna. Tourism Analysis (2015).
21. P. Bangwayo-Skeete and R. W. Skeete: Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. Tourism Management 46, 454-464 (2015).
22. S. Park, J. Lee and W. Song: Short-term forecasting of Japanese tourist inflow to South Korea using Google trends data. Journal of Travel and Tourism Marketing 34(3), 357-368 (2017).
23. C. Artola, F. Pinto and P. de Pedraza: Can internet searches forecast tourism inflows. International Journal of Manpower 36(1), 103-116 (2015).
24. U. Gunter and I. Onder: Forecasting city arrivals with Google Analytics. Annals of Tourism Research 61, 199-212 (2016).
25. R. Rivera: A dynamic linear model to forecast hotel registrations in Puerto Rico using Google Trends data. Tourism Management, 57, 12-20 (2016).
26. G. Dinis, C. Costa and O. Pacheco: The Use of Google Trends Data as Proxy of Foreign Tourist Inflows to Portugal. International Journal of Cultural and Digital Tourism 3(1), 66-75 (2016).
27. M. Camacho and M. J. Pacce: Forecasting travellers in Spain with Google's search volume indices. Tourism Economics 24(4), 434-448 (2017).
28. I. Önder: Forecasting tourism demand with Google trends: Accuracy comparison of countries versus cities. International Journal of Tourism Research 19(6), 1-39 (2017).
29. A. Zeynalov: Forecasting Tourist Arrivals in Prague: Google Econometrics. Munich Personal RePEc Archive (2017).
30. J. Liu, X. Li and Y. Guo: Periodicity analysis and a model structure for consumer behavior on hotel online search interest in the US. International Journal of Contemporary Hospitality Management 29(5), 1486-1500 (2017).
31. E. Rödel: Forecasting tourism demand in Amsterdam with Google Trends. Master Thesis (2017).
32. P. R. Palos-Sanchez and M. B. Correia: The Collaborative Economy Based Analysis of Demand: Study of Airbnb Case in Spain and Portugal. Journal of Theoretical and Applied Electronic Commerce Research 13(3), 85-98 (2018).
33. H. Tang, Y. Qiu and J. Liu. Comparison of Periodic Behavior of Consumer Online Searches for Restaurants in the U.S. and China Based on Search Engine Data. IEEE Access (2018).
34. X. Li, B. Pan, R. Law and X. Hyang: Forecasting tourism demand with composite search index.Tourism Management 59, 57-66 (2017).
35. S. Sun, Y. Wei, K.-L. Tsui and S. Wang: Forecasting tourist arrivals with machine learning and internet search index. Tourism Management 70, 1-10 (2019).
36. H. Theil, Economic Forecasts and Policy, (1958).

37. H. Theil, Applied Economic Forecasting, (1966).
38. F. W. Bliemel: Theil's Forecast Accuracy Coefficient: A Clarification. Journal of Marketing Research 10(4), 444-446 (1973).
39. D. A. Ahlburg: Forecast evaluation and improvement using theil's decomposition. Journal of Forecasting 3(3), 345-351 (1984).
40. C. Tofallis: A Better Measure of Relative Prediction Accuracy for Model Selection and Model Estimation. Journal of the Operational Research Society 66 (8), 1352-1362 (2015).
41. H. Hassani, A. Webster, E. Simiral Silva and S. Heravi: Forecasting U.S. Tourist arrivals using optimal Singular Spectrum Analysis. Tourism Management 46, 322-335 (2015).
42. Hassani, E. S. Silva, N. Antonakakis and G. Filis: Forecasting accuracy evaluation of tourist arrivals. Annals of Tourism Research 63, 112-127 (2017).
43. Dedić and Stanier : An Evaluation of the Challenges of Multilingualism in Data Warehouse Development. 18th International Conference on Enterprise Information Systems - ICEIS 2016 (2016).
44. T. Dunning and E. Friedman, Time Series Databases: New Ways to Store and Access Data. O'Reilly Media (2014).
45. G. E. P. Box, G. M. Jenkins and G. C. Reinsel. Time Series Analysis, Forecasting and Control. Wiley, United States of America (2008).
46. N. Golyandina, A. Korobeynikov and A. Zhigljavsky. Singular Spectral Analysis with R, Springer (2018).
47. C. Granger: Investigating causal relations by econometric models and cross spectral methods. Econometrica 37(3), 424-438 (1969).
48. R. Montero: Test de Causalidad. Documentos de Trabajo en Economía Aplicada. Universidad de Granada. España, (2013).
49. Vergori: Forecasting tourism demand: the role of seasonality. Tourism Economics 18 (5), 915-930 (2012).
50. A. Buse: The Likelihood Ratio, Wald, and Langrange Multiplier Test: An Expository Note. The American Statitician 36(3), 153-157 (1982).
51. S. Hylleberg, R. Engle, C. Granger and B. Yoo: Seasonal integration and cointegration. Journal of Econometrics 44, 215-238 (1990).
52. E. Nkoro and K. Uko: Autoregressive Distributed Lag (ARDL) cointegration technique: application and interpretation. Journal of Statistical and Econometric Methods 5(4), 63-91 (2016).
53. E. Zivot: The Power of Single Equation Tests for Cointegration When the Cointegrating Vector is Prespecified. Econometric Theory 16(3), 407-439 (2000).