PASOS
Revista de Turismo y Patrimonio Cultural
www.pasosonline.org

Vol. 15 N.º 4. Págs. 943-958. 2017
https://doi.org/10.25145/j.pasos.2017.15.064

Forecasting tourism arrivals with an online search engine data: A study of the Balearic Islands

Óscar García Rodríguez*
Universitat de les Illes Balears(España)

Abstract: This study explores issues related to the forecasting in revenue management in the prediction of tourism arrivals for the Balearic Islands. Specifically, the study uses queries from a web search data (Google Trends) in order to demonstrate the forecasting power of such measures compared to traditional methods. I developed a database formed by the two main tourist volumes and then, I compared each model with its corresponding baseline to figure out whether the Google Trends indicator can increase accuracy of the prediction. Consequently, Granger causality test indicated a positive causality between variables suggesting good estimating results. Besides, I calculated the Mean Absolute Percentage Errors (MAPE) for each model and the results showed a considerable improvement of the Google Trends models compared to baseline models. The results provide some hints for increasing company efficiency and enhance policy maker decision making.

Keywords: Google Trends, Forecasting, Revenue Management, ARMAX, Balearic Island

Predicción de la llegada de turistas con los datos de un buscador online: Un estudio para las Islas Baleares

Resumen: Este estudio analiza diferentes temas relacionados con la predicción en el área del revenue management sobre el número de llegadas turísticas para las Islas Baleares. Específicamente, el estudio utiliza búsquedas de un buscador online (Google Trends) para demostrar su poder predictivo en comparación con los métodos tradicionales. He desarrollado la base de datos en base a dos principales volúmenes de llegadas turísticas, y después he comparado cada modelo con su correspondiente modelo de referencia para descubrir si el indicador de Google Trends puede mejorar la precisión de la predicción. Consecuentemente, el test de causalidad de Granger indicó una causalidad positiva entre las variables indicando unos buenos resultados de la estimación. Además, calculé el porcentaje de error absoluto medio (MAPE) para cada modelo y los resultados mostraron una mejora considerable en los modelos que incluyen Google Trends respecto al modelo de referencia. Los resultados muestran algunas pistas para mejorar la eficiencia de las compañías y realzar la toma de decisiones de los legisladores.

Palabras clave: Google Tendencias, Pronósticos, Gestión de Ingresos, ARMAX, Islas Baleares

1. Introduction

Throughout the last fifty years the boom of the tourism activity affected the Balearic Islands significantly. Nowadays, this sector far from being considered as declining is on the stagnation phase according to the expanded version of the tourism lifecycle of Butler (1980).

In fact, within the Balearics the tourism industry is the main source of the regional GDP. The Balearics GDP has grown 3.2% during the last year (http://www.ibestat.cat/ibestat/inici) and it is mainly boosted by the influence of the tourism activity. This enormous growth implied a widespread increased of the number of tourism arrivals, and hence it demonstrates the urgency for proper prediction, especially near-term prediction, in order to correctly allocate resources and managing tourist flows.

* E-mail: oscargr93@hotmail.com

Thus, most of the current literature of the Balearic Islands has focused on analyzing the environmental problems directly derived from the tourism activity and how to solve them by implementing "Eco taxes" or other instruments (e.g. Aguiló, Riera and Rosselló, 2005; Palmer and Riera, 2003). Meanwhile others have studied the environmental innovation as a source of increase competitiveness (e.g. Jacob, Florido and Aguiló, 2010). However, Alvarez-Diaz, Mateu-Sbert and Rosselló-Nadal (2009) sought how to forecast the monthly tourism demand on UK and Germany arrivals for the Balearic Islands but none of them did it by using search engine data. In addition, if companies and governments are better able to forecast the number of tourists, they will be willing to assess more efficiently resources.

Actually, the tourist volume forecasting is based on a bundle of techniques such as statistical or econometric models that rely on historical data to forecast future tourist activities by assuming ceteris-paribus the economic environment. Therefore, these methods might not be so accurate since they primarily focus on long-term horizons such as yearly or quarterly, instead of monthly either weekly data (Yang *et al.,* 2015).

The globalization jointly with the fast evolution of the ICT technologies has led hundreds of millions of different search queries by tourists. In fact, these queries reflect the possible customers' trends, but also offering the possibility to forecast their future behavior. This study will use a Web search query to generate data useful for forecasting the numbers of visitors coming to the Balearic Islands. Specifically, Google will be the search engine to be used in the project. Particularly, query data on visitors will be generated by using Google Trends.

Therefore, in this study we proposed a baseline project based on the search engine, named Google Trends, which will be compared with the actual forecasting models with their equivalent time counterpart in order to assess the validity of forecasting the tourism arrivals to the Balearic Island in shorter periods.

The implications of our study contribute to the existent literature of the revenue management in two ways. First, confirming the statistical significance and accurateness of search engine tools such as Google Trends, in predicting the tourism arrivals to a well-known tourism destination like the Balearic Islands. Second, the managerial decision that companies either destination management organizations (DMO) could perform whether they better forecast the tourism demand, in terms of higher efficiency and effectiveness of resources.

The structure of the paper is divided as follows. The next section will deal with a brief description of the latest articles related to the management and the revenue management. The third part will pose the methodology, thus the dataset, a description of the variables to analyze and the different models to be implemented. Fourth section will pose and discuss the main research results. Lastly, the conclusions and recommendations are going to be exposed.

2. Literature Review

The revenue management appeared during the 70's in the U.S. travel industry. However, its importance and relevance has dramatically increased due to the higher competition among companies. Since, companies compete more; they need to be more efficient in allocating their resources (i.e. physical, human and financial resources) and thus increase their market share in order to survive.

Indeed, the way for increasing efficiency of allocated resources comes from the forecasting of the potential number of customers or the expected demand. If a company can assess adequately its potential number of customers, it will be possible to enhance company's efficiency. This issue is closely analyzed within the revenue management or yield management. Some authors posed "*To remain competitive, the practice of revenue management is of strategic and tactical importance in improving hotel revenues and profitability*" (Vinod, 2004: 178)

The revenue management techniques are appropriated under a certain circumstances (1) when a firm is operating with a relatively fixed capacity, (2) when demand can be clearly segmented into different customer groups, (3) the inventory is perishable, (4) the product is sold in advance, (5) when the demand fluctuates, (6) when marginal sales costs and production costs are low, yet capacity change costs are high. Generally, these are the characteristics needed for the proper implementation of revenue management, such they might happen in the tourism industry. For instance, within the hotel industry the hotel room's number are fixed in the short-term, moreover the demand can be segmented into different groups (e.g. leisure vs. business), also the inventory is perishable since whether the hotel does

not sell the room the hotel does not earn the marginal revenue obtained from the room. Besides, the product is sold beforehand because normally the tourism demand occurs in advance. Therefore, the hotel industry presents several features that suits with the application of revenue management techniques.

Often, the revenue management has to deal with several issues in order to correctly adopt it. As Vinod (2004) mentioned in his article, there exists three major components of the revenue management (i) pricing, (ii) revenue management and (iii) product distribution. For instance Areone, Van der Rest and Kattuman (2015) analyzed how to determine the equilibrium prices in the hotel industry. They concluded that each hotel implement their own revenue management strategy by maximizing their individual revenue function conditioned to non-price characteristics of all competing hotels.

Moreover, another relevant implication of the yield management lies in assessing the optimal quantity of room, airplane seats, restaurant seats, and so on which allow the company to obtain the maximum revenue possible. Schwartz, Stewart and Backlund (2012) focused on determining the revenue management strategy for the Grand Canyon National Park in order to generate a greater income by keeping constant the capacity rate. Finally, they concluded that implementing a fee could, indeed, potentially increase revenues, however, these structure modifications are likely to enlarge the exclusion gap of determined user groups.

Nevertheless, our analysis will be focused on the revenue management area. One of its main components is the forecasting because during the whole process the demand, the supply and the overbooking controls must be predicted at different levels. For example, forecasting the tourism demand would imply to figure out the number of rooms that are going to be booked for the next week. Moreover, forecasting the supply, namely late or early checkouts, is useful to assess the number of rooms that can be sold. And lastly, to correctly predict the number of cancellations and no-shows is vital in order to set an adequate overbooking policy. Again, Vinod (2004:183) said "*high forecast errors will result in conservative inventory controls and increase the likelihood of revenue dilution*".

Thus, we have seen that whether companies or DMOs want to improve their efficiency they have to accurate their forecasts. Hence, increasing the accuracy of the predictions means to enhance the actual statistical methods or to find out new ways to better define the tourism behaviour.

2.1. Predicting tourism demand

Researchers during the last years, and thanks to the development of computer technology, have adopted two main frameworks for predicting tourist volumes and tourism demand. The most common used is based on statistical techniques or time series, such linear regression, exponential smoothing, and autoregressive models (e.g. Gounopoulos, Petmezas and Santamaria, 2012). While the other type employs artificial intelligence methods such grey theory, fuzzy theory among others (Canestrelli and Costa, 1991). Nevertheless, in previous studies investigators found out that there is no a particular method that outperforms other forecasting methods in terms of accuracy (e.g. Li, Song and Witt, 2005). In fact, Shen, Li and Song (2011) sought if the combination of different forecasting methods can improve the predicting performance in the context of tourism demand. Their study is based on the idea that combining the information included within different individual forecasts can yield to a greater accuracy. Finally, they found out that combination forecasts can enhance forecast accuracy and particularly, the more sophisticated combination forecasts (i.e. MSFE and VACO methods) perform better than other combination forecasts.

Song and Li (2008) reviewed the published studies on tourism demand forecasting since the 2000. They analyzed the different dependent variables that were commonly used for predicting, for instance, tourist arrivals, tourist expenditure, number of days spent, etc. Then, they concluded that the development of quantitative techniques can be encompassed into three main categories: time-series models, the econometric approach and other methods such as artificial intelligence techniques, although none of them presents better results than the rest. Moreover, they presented that data disaggregation (i.e. nationality of tourist arrivals, purpose of the trip, and level of income) might enhance forecasting accuracy.

Traditionally, since its origin the usual framework for predicting tourism demand was to implement time series model and its variations. Actually, these models are well-established and, in fact, are somehow better than other methods (Song, Witt and Li, 2008). In the Greek area Gounopoulos *et al.* (2011) compared different time-series models and econometric models such as ARIMA and Holt's exponential smoothing model for tourism forecasting by analyzing the impact of random macroeconomic shock in

the short-run. The result showed that ARIMA model outperforms other models as a forecasting tool (e.g. ease of use, cost of producing the forecast, the speed that the forecast can be produced), nevertheless focusing on its accuracy Holt's exponential smoothing model with trend generate better estimations. Furthermore, Chu (2009) investigated the accurateness of the ARMA in forecasting tourism demand for Asian-Pacific destinations such as Hong Kong, Japan, Korea or Australia. He used monthly and quarterly data in three different ARMA models (ARAR, SARIMA and ARFIMA) and he concluded that the ARFIMA model is outperforming the rest, and therefore having a greater accuracy observed from the MAPE.

The development of ICT technology has increased the use of non-conventional methods for predicting tourism demand, such as artificial intelligence methods. In fact, one interesting case is the study performed by Alvarez-Diaz *et al.* (2009). They used Genetic Program (GP) to predict the monthly arrivals of UK and Germany to the Balearic Islands. Thus, the study compared the performance of the GP model against different univariate models such as no-change model, Moving Average and ARIMA. Finally, they concluded that GP can be used as a source for forecasting tourism arrivals since the GP gets better estimations for the case of the German demand.

However, every forecasting methodology has their limitations. For instance, time series data analyses rely too much on a consistent historical pattern and a stable economic structure (Yang *et al.*, 2015). Then, any change in the economic activity perhaps a dramatic change such the financial crisis suffered on 2007 or large scale effects might decrease their accuracy. Moreover, artificial intelligence methods are relatively new, and they both require a lot of training data and effort in order to accurate the predictions.

2.2. Predicting tourism demand with search engine data

Search engine data has been recently used as a source for gathering information and forecasting socioeconomic activities in different knowledge branches such as medicine (Althouse, Ng and Cummings, 2011; Ginsberg, Mohebbi, Patel, Brammer, Smolinski and Brillian, 2009).Nevertheless, the adoption of this source of information is becoming widespread among the economic activity to predict future tourist behavior. Within this framework researchers have sought whether these sources of information can increase forecast accuracy mainly using two search engines, Google Trends and Baidu.

In fact, the most common search engine used is Google Trends due to its potential to inform strategic decision making in tourism destinations. A lot of research is recently been carried out, for instance, Jackman and Naitram (2015) analyzed whether the inclusion of Google Trends index within a support vector regressions (SVR) could enhance accuracy of forecasting for the Barbados Island tourist arrivals. Indeed, they refer to forecasting as nowcasting based on Castle, Fawnett and Hendry (2009), "*nowcasting, in its most basic form, can be summarized as predicting the present and sometimes the recent past*" as a way for predicting the incoming future. Lastly, they both concluded that including the Google Trends index increase the accuracy of estimations by means of a lower MAPE for the predictions of UK and Canadian tourists, whilst not for the US tourist arrivals. Rivera (2016) considered the information gathered by Google Trends as a source for predicting the number of hotel nonresident registrations in Puerto Rico. He showed a positive result in conducting the analysis; however, some problems arose from the model and finally concluded that for the short term the Holt-Winter forecast resulted in smaller forecast errors, while for the long term periods their model using Google Trends index outperformed other models. Bangwayo-Skeete and Skeete (2015) sought whether the introduction of a new indicator based on Google Trends could improve the accurateness efficiency of current forecasting models for Caribbean destinations using data on US, UK and Canadian travellers. Thus, they developed three different models, AR-MIDAS, SARIMA and AR the last two of them used as a benchmark for comparison. Finally, both concluded that AR-MIDAS model, which includes the queries in Google, gave superior predictions to SARIMA and AR models in terms of accuracy (e.g. RMSE and MAPE criteria).

However, Google Trends is not the unique tool used in research. Huang, Zhang and Ding (2016) focused on forecasting which would be the total number of tourist arrivals for the Chinese "Golden Week". They compare three different models first, an ARIMA model without the Baidu index, second an ARIMA model including the index and lastly, an auto regression distributed lag model. Furthermore, they positively proved the improvement in the prediction for the ARIMA model including the Baidu

index. Moreover, Yang *et al.* (2015) used two search engines data to improve the forecasting power of an ARMA model; indeed, they compare the predictive power of Google Trends versus Baidu in the Hainan Province in China. Finally, they pointed out that the Baidu results performed better because its larger market shares in China, although both search engines reduced significantly the forecasting errors. Besides, their study also contributed to the existing literature by proposing a method for selecting the queries that we will discuss in further sections.

Overall, the accuracy of the forecast becomes a key issue in determining company's and DMO strategy in order to promote a sustainable tourism development. As mentioned before, enhancing the estimators' prediction power would yield to a higher efficiency, and thus greater welfare. Therefore, testing the accuracy of current models versus new techniques such Google Trends source could yield to a better comprehension of the revenue management and, management in general. In addition, the limitation of the traditional predictive methods lies on the time lag between the collection and publication of the data and also, databases can be based on insufficient samples, hence providing useless predictions (Huang *et al.*, 2016). I therefore propose to compare the fitness and prediction of the current forecasting models versus Google search data in forecasting the tourist volume for the Balearic Island. Additionally, based on Yang *et al.* (2015), I would implement a systematic way to better select queries for predicting.

3. Methodology

3.1. Empirical testing

The Balearic Islands was chosen as destination for testing our empirical model. It is well known that since the middle 60s the attractive of the Balearic Islands had increased dramatically. Especially during these last years the number of tourist arrivals has boosted, in 2014, the Balearic Island received over 13.5 million overnight tourists, with 11,348,000 tourists (almost 84%) of them being foreigner visitors. In fact, most of the tourist arrivals are concentrated in two major countries Germany and United Kingdom with 4.1 and 3.4 millions of tourist arrivals respectively. Overall, these countries represent 7.52 million of arrivals implying more than 55% of the whole tourist demand for the year 2014 (Ibestat, 2016).

3.2. Data source

Google Trends is a search engine tool that provides an index of the different Google queries sorted by geographic location and category. Thus, the database will be directly extracted from Google Trends since it represents the search engine with the largest market share in Germany and in the U.K., particularly 97.09% and roughly 90% respectively (Kennedy and Hauksson, 2012).

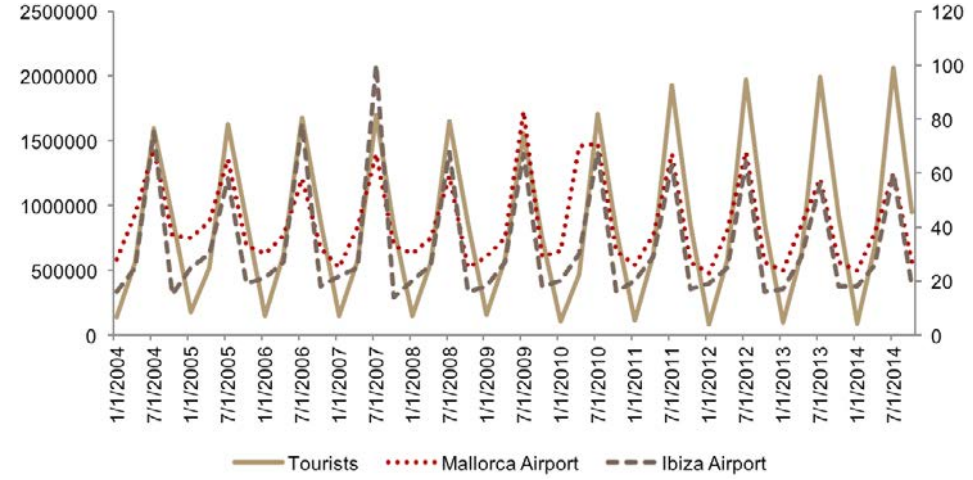
Moreover, Google Trends reports a query index which displays frequently how often a particular query has been searched compared to the total search volume from different countries and languages; actually it does not report the raw level of total queries. In Choi and Varian (2009) explain the entire process of how Google Trends creates the index, and how it began at January 2004. Since that year, the numbers indicate the percentage of variance from the query share in 2004.

The forecasted variable Balearic Islands' monthly visitor volume data were obtained from Ibestat database, derived directly from the Balearic Islands national statistical center (Ibestat, 2016). Moreover, Fig. 1 represents the correlation between Balearic Island's quarterly visitor volumes and two airport queries volumes for German and U.K. population. I chose airport route because it is the main entrance gate to the Balearic Islands, in 2014 13.5 million tourists were arriving and 9.5 million were coming to Mallorca, over 2.6 million to Ibiza both of them by plain. Hence, the aviation represents the primarily way of entrance to the Islands (Ibestat, 2016).

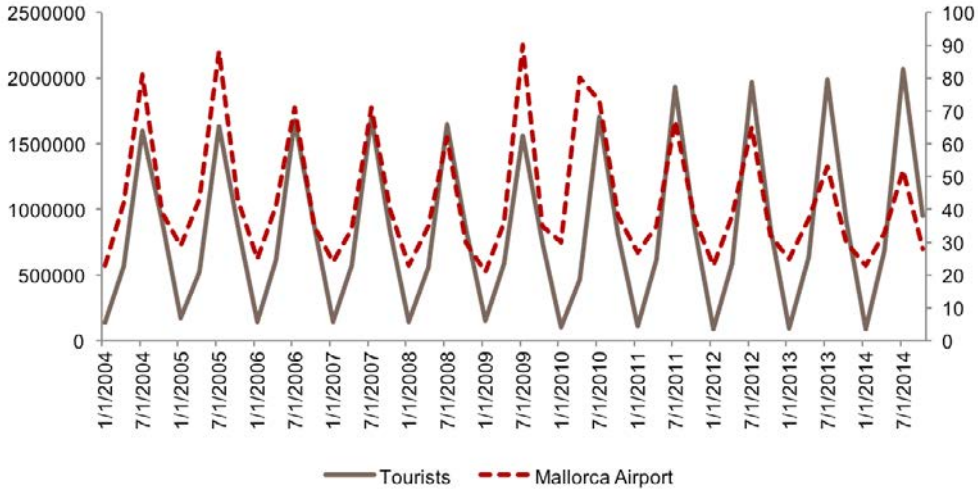
In the case of U.K. visitors I selected the queries "Mallorca Airport" and "Ibiza Airport" since these places are the only ones which tourists can travel by plane, excluding Menorca due to the information provided by Google Trends was not large enough. And for the German population I selected "Mallorca *Flughafen*" and "Ibiza *Flughafen*"; however the Ibiza search was not large enough to include it on the graph so it was removed from the sample. In both graphs it can be appreciated a high concordance between visitors number and query volume. Nevertheless there exist thousands of potential queries to the Balearic's tourism industry.

Fig. 1. Search data from Google Trends and tourist arrivals for the Balearic Islands

Search data for U.K. population



Search data for German population



Source: Own elaboration with Google Trends and Ibestat data

3.3. Selection of search queries

In the selection of search queries, I followed various researchers such as Pan, Litvin and Goldman (2006) or Yang *et al.* (2015). Firstly, in Pan *et al.* (2006) the investigators figured out, by means of search engine tools (Excite), which were the most common terms whenever tourists plan to travel. In fact, they segmented their database by using two main criteria; the first was based into trip's aspects (e.g. destination, hotels, restaurants, transportation, attractions and activities) while the second was related to their generalization levels, for instance the term "Restaurant" was classified as general term, "Miami" was sorted at the city level. In my study, I will follow this classification for the different search queries.

Secondly, Yang *et al.* (2015) developed a framework to systematically select queries from Google Trends and Baidu. Indeed, they set up a four stage process which will be implemented in my article in order to find out the terms with higher correlation.

The aim of my study is to analyze whether the inclusion of the Google Trends index can enhance the accuracy of the traditional forecasting models. Hence, considering a mature destination such as the Balearic Island, mainly visited by German and English tourism, it is interesting to dig into the different search queries by nationalities. Therefore, two different models will be set up according to the major outbound countries.

In order to facilitate the analysis, I concentrated the search queries database in the main tourism destination of the Balearic Islands, which is Mallorca. In 2014 Mallorca gathered 9.6 million of tourist arrivals, who came principally from Germany and U.K., 3.7 and roughly 2.2 million respectively (Ibestat, 2016). So, as it is stated above, from Yang *et al.* (2015) results the four stage process used in my study to select the different search queries and to create the index that will be used in the different models.

(1) Initially 20 basic search queries were selected based on the different trip aspect posed by Pan *et al.* (2006) including 10 search queries in German and another 10 search queries in English. The translated searches are sorted in Table 1 with their corresponding categories.

(2) Afterwards, the 20 queries were introduced in Google Trends as seed queries and retrieved the related queries. Then, I chose the related queries and repeat the procedure for second and third rounds. The total number of queries converged to 134, however only 100 queries remained in the database after duplications were removed. In fact, the UK database is composed by 49 search queries and the German database is formed by 51 search queries.

Table 1. Basic search queries related to tourists

	German Queries	*U.K. Queries*
Destination	Mallorca Tourism Travel Agency	Mallorca Travel Palma de Mallorca
Hotels	Hotel Mallorca	Hotel Mallorca Hotel Magaluf
Restaurants	Mallorca Restaurant	Mallorca Restaurant
Transportation	Flight Mallorca Mallorca Airport Mallorca Tickets	Flight Mallorca Mallorca Airport
Attraction	Party Mallorca	Party Mallorca
Activities	Mallorca Beach Mallorca Weather Mallorca Activities	Mallorca Beach Mallorca Weather

Source: Own elaboration with Stata11

(3) I have calculated the Pearson correlation coefficient between Mallorca monthly visitor volume and each of the search queries with different lag periods. In fact, for each query I calculated eight correlation coefficients of 0-7 months ahead respectively. Moreover, from the two datasets (i.e. Germany and UK) I selected 17 queries from UK, and 17 queries from Germany represented in Table 2 and Table 3 respectively. In order to calculate an appropriate number of search queries, I used 0.65 as the threshold for the correlation between visitors' volume and Google Trends data (Yang *et al.,* 2015). The reason of that threshold is the following. First, if we set a threshold below 0.65, let us imagine that I choose a threshold of 0.65, then 43 keywords for the UK and 38 keywords for Germany would be selected and it might reduce the parsimony and generalizability of the models. Second, if I increase the threshold to 0.70 the number of keywords would be 10 for the UK and 9 for Germany and therefore, it will result in a low forecasting accuracy. Hence, the selection of the threshold represents the trade-off between forecasting accuracy and model parsimony.

Table 2. Maximum correlation coefficient of search queries from Germany

Maximum correlation coefficient of search queries from Germany

Search query	*Lag order*	*Search query*	*Lag order*
Mallorca beach	1	*Mallorca airport arrival*	0
Alcudia Mallorca	1	*Palma airport arrival*	0
Alcudia beach	1	*Palma airport*	0
Playa Palma	1	*Palma Mallorca airport*	0
Mallorca holiday ballermann	1	*Mallorca weather*	0
Mallorca airport	1	*Weather in Mallorca*	0
Most beautiful beach Mallorca	1	*Weather Mallorca April*	0
Mallorca rental car	1	*Mallorca departures*	0
		Weather Mallorca Alcudia	0

Source: Own elaboration with Stata11

(4) Since the purpose is to forecast the future tourist volume, based on Yang *et al.* (2015) I only select queries with, at least, one lag previous to the arrival. Therefore, 8 queries with 1 lag were chosen for Germany and 11 queries with 1 and 7 lags were chosen for UK as Google Trends predictors.

Table 3. Maximum correlation coefficient of search queries from UK

Search query	*Lag order*	*Search query*	*Lag order*
Mallorca weather	7	Mallorca airport	1
Palma weather	7	Flight Palma	1
BCM Magaluf	1	*Weather in Alcudia*	0
Alcudia beach	1	*Weather in Majorca*	0
Mallorca travel	1	*Weather in Mallorca*	0
Palma de Mallorca	1	*Flight to Palma*	0
Calas de Mallorca	1	*Palma Airport arrivals*	0
Palma Mallorca	1	*Palma departures*	0
Hotel Mallorca	1		

Source: Own elaboration with Stata11

The data reveals interesting information about the travel behaviour of both countries. First, Table 2 shows that for the German tourists the lags of the maximum correlation coefficient varied mainly from 0 to 1, distributed similarly across the sample. Indeed, they are concentrated in the transportation and destination such as Mallorca airport arrival, Mallorca departures, Alcudia Mallorca, Mallorca beach. Nevertheless, it is important to point out that the weather queries were also relevant for the German tourists. Second, Table 3 represents the highest correlated lags for the UK travelers. It varies from 7 to 0 lags, being 1 the most common. Furthermore, it gathers information about destination (e.g. Calas de Mallorca, Palma de Mallorca, and Palma Mallorca), transportation (e.g. Mallorca travel, Mallorca airport, Palma departures) and weather information (e.g. Weather in Majorca, Weather in Alcudia).

3.4. Search data index

In order to develop the different models I aggregated search data by using PCA analysis. Firstly, I computed Cronbach's Alpha to validate and testing the internal consistency of the PCA analysis. In fact, I obtained a value of 0.805 and 0.951 for the UK and German data respectively (recommended value of 0.7). Then, a PCA analysis was calculated for each database.

Table 4 shows German data results of the PCA analysis. The main indexes confirm the appropriateness of using this analysis (Delgado-Verde *et al*., 2011), since the KMO index had a value of 0.88 (higher than 0.6 that is the value recommended); the Bartlett test was significant at a level lower than 0.05 (0.000); the extraction column of the commonality showed high values between 0.70 and 0.91 that can be interpreted as the factor analysis adjust level. In addition, I applied Varimax orthogonal rotation, all items that had a load higher than 0.85 (items with a load lower than 0.4 were excluded from the

table), finding one factor that deeply capture the extent and the degree of the relationship. Finally, the percentage of accumulated explained variance for the factor was 83.26 percent, being higher than the proposed value for social science: 60 percent (Hair, Anderson, Tatham, and Black, 2004).

I further calculated the PCA analysis for the UK data, which is represented on Table 5, and the results confirmed the appropriateness of implementing this framework. The KMO test showed a value of 0.893; the Barlett test was significant at a level lower than 0.05 (0.000); besides the communalities' extraction column showed high values between 0.705 and 0.942 indicating the factor adjusting level. Furthermore, Varimax orthogonal rotation was implemented obtaining item loads higher than 0.70, finding two factors that can be used to capture most of the information of the variables. Indeed, the accumulated explained variance by the two factors was 81.35 percent.

Table 4. German data results from PCA analysis

Item	*Weight Index 1*
Mallorca beach	0.910
Alcudia Mallorca	0.953
Alcudia beach	0.914
Most beautiful beach Mallorca	0.923
Mallorca holiday ballermann	0.930
Playa de Palma	0.900
Mallorca airport	0.910
Mallorca rental car	0.859
Cronbach Alpha	0.951
KMO	0.888
Explained variance (%)	83.26
Accumulated (%)	83.26

Note: Extraction method: Principal Component; Rotation method: Varimax normalization with Kaiser; Only one component extracted, solution cannot be rotated

Source: Own elaboration with Stata11

The results showed in Table 4 indicate that for the German data the index captures the information regarding all the variables, namely destination and transportation variables. However, the UK sample variables, represented in Table 5, are gathered into two main factors. The first index captures mainly information regarding the destination and transportation, while the second index is focused on weather information. The following analyses were based on the PCA index created for each model.

Table 5. UK data results from PCA analysis

Item	*Weight Index 1*	*Weight Index 2*
Mallorca weather		0.726
BCM Magaluf	0.797	
Palma weather		0.790
Alcudia beach	0.898	
Hotel Mallorca	0.966	
Palma de Mallorca	0.839	
Calas de Mallorca	0.845	
Palma Mallorca	0.822	
Mallorca Airport	0.950	
Flight Palma	0.884	
Mallorca travel	0.885	
Cronbach Alpha	0.803	
KMO	0.893	
Explained variance (%)	69.72	11.63
Accumulated (%)	81.35	

Note: Extraction method: Principal Component; Rotation method: Varimax normalization with Kaiser; Rotation has converged after three iterations

Source: Own elaboration with Stata11

3.5. Co-integration analysis of search index and Mallorca visitors

Considering the PCA conclusions I constructed two different time series models based on Google Trends data for Germany and UK. The dependent variable in the three models is T_{t1} T_{t1} which denotes the Mallorca monthly tourist volume, from August 2004 to June 2016.

$$\text{Log}\, T_{t1} = c_0 + \beta_1 \,\text{Log}\, T_{t1}\,(-12) + u_t \quad (1)$$

$$\text{Log}\, T_{t1} = c_0 + \beta_1 \,\text{Log}\, Ger_{t1} + u_t \quad (2)$$

$$\text{Log}\, T_{t1} = c_0 + \beta_1 \,\text{Log}\, UK_{t1}^1 + \beta_2 \,\text{Log}\, UK_{t1}^2 + u_t \quad (3)$$

Eq. (1) represents the baseline model which uses historical tourist volume data to predict the actual tourism arrivals. Since, the tourism demand presents seasonality features I decided to use as predictor the 12 periods T_{t1} T_{t1}. Moreover, Eq. (2) showed the first model with the PCA component derived from the German database. In addition, Eq. (3) represents the second model with PCA components using UK data from Google Trends, being $UK_{t1}^1 UK_{t1}^1$ the first weight index and $UK_{t1}^2 UK_{t1}^2$ the second weight index. Furthermore, the forecasting models of each country have been compared with its corresponding baseline model.

Due to the seasonality of the dependent variable, I decided to apply a logarithmic transformation to the tourist arrivals for Germany and for UK; meanwhile the idiosyncratic term is represented by u_t u_t which captures the residual series.

With equations (1)-(3) I analyzed the correlogram and the partial correlogram of each independent variable in order to figure out whether the variable follows a stationary process or not. Moreover, augmented Dickey-Fuller (ADF) unit-roots test were applied to all independent variables in each model (see Table 6). Only the two dependent variables of the baseline models (i.e. Log$T_{t1}^{GER} T_{t1}^{GER}$for Germany and Log$T_{t1}^{UK} T_{t1}^{UK}$ for UK) needed to be modified in order to transform to a stationary process. Moreover, I computed the ADF tests for the residuals of each regression in order to check for any co-integration relationship. The evidence shows that there exists a positive co-integration relationship between the exogenous and endogenous variable of each model represented in Table 6. In fact, the co-integration relationship implies that there exists a long-term relation between variables, so that when both variables grow in time T, they both do it in a totally synchronized form such that the error term between the variables does not increase.

On the one hand this might support the Granger causality analysis and, on the other hand my modelling process based on ARMAX models (Autoregressive Moving Average with External Variables). In all the four models, I employed data from August 2004, to December 2015, I omitted the last six periods in order to out-sample forecast. I created a baseline model for each database, namely I developed a baseline model for Germany $L_{b1} L_{b1}$ and a baseline model for UK $L_{b2} L_{b2}$. Then, I estimated an ARMAX model for Germany and UK and I compared them with their corresponding baseline model.

3.5.1. ARMAX results for Germany

First of all I will briefly comment the baseline model results, and secondly the German ARMAX model estimation results. Indeed, the best baseline model for Germany is shown at Table 6. In $L_{b1} L_{b1}$ all variables were significant at a 0.01 level, although the exogenous variable, in this case Log$T_{t1}(-12)$ $T_{t1}(-12)$, presented a coefficient close to zero (-0.0001). In addition, the constant, moving averages of order 1 and 12, and autoregressive of order 1 and 2 were statistically significant at a 0.01 level. The expression of the model is represented at equation 4:

$$\begin{cases} \text{Log}\, T_{t1} = 12.375 - 0.0001 \,\text{Log}\, T_{t1}\,(-12) + u_t \\ u_t = 1.429\, u_{t-1} - 0.752\, u_{t-2} + \varepsilon_t - 0.466\, \varepsilon_{t-1} + 0.701\, \varepsilon_{t-12} \end{cases} \quad (4)$$

The ARMAX model for Germany that best fitted the data was an ARMAX (1, 12) meaning that the error term presented an autoregressive part of order 1 and a moving average of order 12. The independent variable and all the autoregressive and moving average coefficients were statistically significant at a 0.01 level. The positive coefficient of the PCA index suggested a correlation between web search data and Mallorca tourist volumes. The formal expression of the model is represented at equation 5:

$$\begin{cases} \text{Log}\, T_{t1} = 0.209\, Ger_{t1} + u_t \\ u_t = -0.696\, u_{t-1} + \varepsilon_t + 0.899\, \varepsilon_{t-12} \end{cases} \quad (5)$$

Moreover, unit root-test for the residuals indicated that both models were stationary at a 0.01 level. Thus, it confirmed the co-integration relationship that was embedded between the principal component factor for Germany ($Ger_{t1}Ger_{t1}$) and German visitor volume for Mallorca.

In order to select a model I based the conclusions on different indicators such as the Log-likelihood, and the information criteria such as AIC (Akaike Information Criteria) and BIC (Bayesian Information Criteria). The results lead to the selection of the baseline model since, the lower the information criteria the better 55.793 vs. 99.572 (AIC) and 75.920 vs. 114.386 (BIC), and also the Log-likelihood indicates the same results, however, in this last case the larger the indicator the better the model, -20.896 vs. -44.786 (Log likelihood).

3.5.2. ARMAX results for the UK

Again, I developed two models to allow comparison, first the baseline model for UK represented by $L_{b2}L_{b2}$ and then the ARMAX model including the PC indicators (UK1). In Table 6 there are shown the coefficients and their significance level. The baseline model presented all significant coefficients at a 1% confidence level, yet the exogenous variable had a coefficient close to zero (-0.0003). The rest of the variables (i.e. Moving Average of order 1 and 2) had a positive coefficient and were statistically significant at a 0.01 level. In equation 6 is represented the baseline model:

$$\begin{cases} \text{Log}\, T_{t1} = \ 11.739 - 0.0003\ \text{Log}\, T_{t1}\,(-12) + \ u_t \\ \qquad u_t = \varepsilon_t + 1.453\varepsilon_{t-1} + 0.861\ \varepsilon_{t-2} \end{cases} \qquad (6)$$

The ARMAX model with the UK principal component factors is defined in equation 7. In the model, all coefficients were statistically significant at a 1% level, although the first component was not (i.e. UK^1_{t1} UK^1_{t1}) neither at a 10% of significance. The second component was statistically significant and presented a positive coefficient indicating a positive correlation between the weather search information and the Mallorca visitor arrivals.

$$\begin{cases} \text{Log}\, T_{t1} = 11.812 - 0.069\ UK^1_{t1} + 0.064\ UK^2_{t1} + \ u_t \\ u_t = 0.989\ u_{t\text{-}12} + \varepsilon_t + 0.414\ \varepsilon_{t-1} + 0.180\ \varepsilon_{t-2} - 0.279\ \varepsilon_{t-12} \end{cases} \qquad (7)$$

I tested for unit-root implementing an ADF in the residuals and both models lead to a stationary conclusion at a 0.01 level. This confirmed the co-integration association between dependent and independent variables, and the idea that there is a long-term association between variables. Based on the various tests like Log likelihood, AIC and BIC criteria I selected the UK model since it presents better estimation results.

Table 6. Regression comparison of model G1, UK1 with baseline models L_{b1} and L_{b2} (2004.8 – 2015.12)

Data source		Germany visitor data		Germany ARMAX model		U.K. visitor data		U.K. ARMAX model	
Model		Baseline L_{b1}		Model G1		Baseline L_{b2}		ModelUK1	
Indepenent variables		LogTe1(-12)	-0.0001***	$Gere_{t1}$	0.209***	$LogT_{t1}$ (-12)	-0.0003***	UK^1_{t1}	-0.069
			(-2.94)		(3.62)		(-9.69)		(-1.04)
		C	12.375***	MA(12)	0.889***	C	11.739***	UK^1_{t2}	0.064***
			(106.69)		(9.30)		(46.84)		(22.65)
		MA(1)	-0.466***	AR(1)	0.696***	MA(1)	1.453***	C	11.812***
			(-2.55)		(10.48)		(29.89)		(22.65)
		MA(12)	0.701***			MA(2)	0.861***	AR(12)	0.989***
			(7.71)				(16.59)		(118.40)
		AR(1)	1.429***					MA(1)	0.414***
			(14.42)						(6.74)
		AR(2)	-0.752***					MA(2)	0.180***
			(-7.69)						(2.60)
								MA(12)	-0.219***
									(-3.61)
Log likelihood		-20.896		-44.786		-142.341		16.417	
AIC		55.793		99.572		294.682		-16.834	
BIC		75.920		114.386		309.058		6.868	
Residual stationary	ADF	-12.327***		-9.935***		-10.946***		-9.782***	
	1% Crit.	-3.500		-2.594		-3.500		-2.353	
	5% Crit.	-2.888		-1.950		-2.888		-1.656	
	10% Crit.	-2.578		-1.613		-2.578		-1.288	
	Conclude	Stationary		Stationary		Stationary		Stationary	
Conclusion		Co-integration		Co-integration		Co-integration		Co-integration	
Adjusted observations		130		142		130		142	

Note: *, ** and *** denotes significance at the 10%, 5% and 1 % level.

To sum up, it is interesting to see that in all the models, but especially on the G1 and UK1 models, the coefficient of $u_t u_t$ were statistically significant, and it implies that the exogenous variable (i.e. Ger_{t1} Ger_{t1}, $UK^1_{t1}UK^1_{t1}$ and $UK^2_{t1}UK^2_{t1}$) could not explain all the variability in the tourists' volume. In fact, other factors might influence visitors' short-term variations.

3.6. Granger causality analysis

The existence of a correlation, either positive or negative, among two variables does directly imply causality between them. Indeed, it does mean that one variable causes the fluctuations of the other. Theses causes and consequences might come from a spurious origin. Thus, the Granger causality tests allow us to analyze whether a variable X causes variable Y. In fact, under null hypothesis there is no causality relationship between variables and under alternative hypothesis otherwise.

Consequently, I tested Granger causality for our two ARMAX models with principal component indexes (i.e. G1 and UK1). Yet, in the case of Germany it was preferable to select the baseline model I wanted to analyze the relationship between variables to figure out the possible existence of causality. Due to the great sensitivity to the lag order, I previously considered five test criteria for the selection of the lag order: LR (Likelihood Ratio Test), FPE (Final Prediction Error Criterion Minimum), AIC (Akaike Information Criterion), SBIC (Schwarz Information Criterion) and HQ (Hannan-Quinn Information Criteria) which are represented in Table 7.

Table 7. Lag order selection criteria for Granger causality test

Lag	Lool	LR	FPE	AIC	SBIC	HQ
0	-294.254	NA	0.335	4.584	4.627	4.601
1	-222.938	150.63	0.112	3.495	3.626	3.548
2	-167.819	110.24	0.051	2.714	2.934	2.803
3	-152.353	30.932	0.043	2.539	2.847	2.664
4	-134.055	36.596	0.034	2.321	2.716	2.481
5	-105.126	57.858	0.023	1.940	2.423	2.137
6	-66.990	76.273	0.014	1.419	1.990	1.651
7	-34.122	65.735	0.009	0.978	1.637	1.246
8	-19.602	29.04	0.007	0.818	1.564	1.121
9	-6.922	25.36	0.006	0.685	1.519	1.024
10	-3.666	6.512	0.006	0.697	1.619	1.071
11	25.402	58.137	0.004	0.314	1.324	0.724
12	67.8751	84.946*	0.002*	-0.272*	0.824*	0.173*

*Indicates optimal lag order selected by the criterion

Germany

Lag	Lool	LR	FPE	AIC	SBIC	HQ
0	-422.619	NA	2.241	6.482	6.526	6.500
1	-339.892	165.45	0.673	5.280	5.412	5.334
2	-304.469	70.846	0.417	4.801	5.020	4.890
3	-292.397	24.145	0.368	4.677	4.985	4.802
4	-271.15	42.493	0.283	4.414	4.809	4.575
5	-238.158	65.985	0.182	3.971	4.454	4.168
6	-192.821	90.675	0.096	3.340	3.911	3.572
7	-171.463	42.715	0.074	3.075	3.734	3.343
8	-161.782	19.362	0.068	2.989	3.735	3.292
9	-144.462	34.64	0.055	2.785	3.619	3.124
10	-142.623	3.678	0.057	2.818	3.740	3.193
11	-130.45	24.345	0.051	2.693	3.703	3.104
12	-61.699	137.5*	0.019*	1.705*	2.802*	2.151*

*Indicates optimal lag order selected by the criterion

UK

Source: Own elaboration with Stata11

The different criteria lead to the selection of the same lag order of 12. This order was used in the Granger causality test which is represented in Table 8. Furthermore, the results of the analysis showed that for the G1 model $Ger_{t1}Ger_{t1}$ and $\text{Log}\, T_{t1}\text{Log}\, T_{t1}$ Granger caused each other, that means, Google Trends data can predict Mallorca tourism volume and vice versa. However, for the UK1 model $UK^2_{t1}UK^2_{t1}$ and $\text{Log}\, T_{t1}\text{Log}\, T_{t1}$ I only found out a positive Granger causality for the UK Google Trends data implying that Google Trends can be used as a predictor of UK tourism volumes, although I did not find support for the opposite causal relationship, since the p.value is larger than 0.05 (0.7303).

Table 8. Granger causality test results for search data and volume visitors

Null hypothesis	F-statistics	P-value
Model G1		
Ger_{t1} does not Granger cause Log T_{t1}	6.5021	0.0120
Log T_{t1} does not Granger Cause Ger_{t1}	7.5962	0.0067
Mode/UK1		
UK^2_{t1} does not Granger cause Log T_{t1}	4.3695	0.0386
Log T_{t1} does not Granger cause UK^2_{t1}	0.1193	0.7303

Source: Own elaboration with Stata11

4.7. Forecasting with web search data

In order to test the predictive power and accuracy of the different models, I dropped from the training set the last six months from January 2016 until June 2016 for testing and compare the results. I compared the actual value of the dependent variable with its corresponding predicted value, and then calculated the percentage of error and lastly, the MAPE (Mean Absolute Percentage Error) shown in Table 9.

The results indicated that both models, G1 and UK1, predicted 6 months of the Mallorca visitors' volume more accurately than their corresponding baseline models. In fact, the G1 model improves the results in a roughly 2%, although in the UK case the UK1 model enhances the prediction by reducing the error in a 5%.

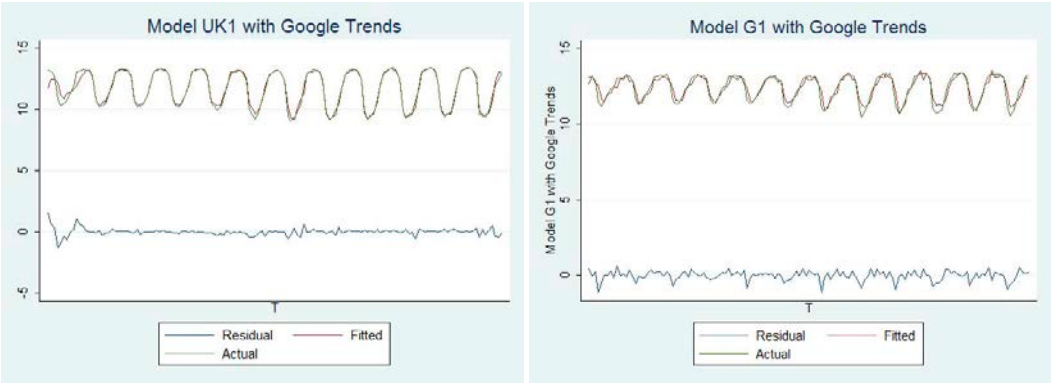
Table 9. Forecast comparison of G1 and UK1 models with baseline models L_{b1} and L_{b2}

Data source		German Visitar history data		Google Trends Germany		UK visitar history data		Google Trends UK	
Model		Baseline L_{b1}		Model G1		Baseline L_{b2}		Model UK1	
Month	Actual	Fitted	Error%	Fitted	Error%	Fitted	Error%	Fitted	Error%
2016M01	9.39	10.77	14.69%	11.21	16.89%	11.45	21.93%	9.60	1.83%
2016M02	9.82	11.51	17.20%	11.48	14.42%	9.53	2.95%	9.69	1.36%
2016M03	11.27	12.31	9.22%	11.73	3.73%	10.69	5.14%	10.76	4.77%
2016M04	11.88	12.76	7.40%	12.40	4.07%	12.37	4.12%	12.42	4.36%
2016M05	12.55	13.19	5.09%	13.03	3.63%	12.02	4.22%	13.04	4.07%
2016M06	12.88	13.13	1.94%	13.05	1.29%	11.72	9.00%	12.97	0.76%
MAPE		9.26%		7.34%		7.89%		2.86%	

Source: Own elaboration with Stata11

In addition, I plotted on Fig. 2 the residuals, the actual value of the dependent variable (i.e. the natural log of the tourism arrivals) and also, the fitted value of the two estimated models. Hence, the results pointed out that model UK1 outperforms model G1 in fitting the data, since it seemed to present less variance.

Fig. 2 Residual, actual and fitted value of model G1 and UK1



Source: Own elaboration with Stata11

4. Conclusions

This article analyzed all the previous literature about forecasting, focusing especially in forecasting with search engine data. I followed the methodology and tried to replicate the study performed by Yang *et al.* (2015) using Google Trends data for the Balearic Island context.

The focus of my study was framed on the most recent part of the forecasting literature, in the nowcasting framework. Nowcasting as Castle *et al.* (2009) said implies to forecast the present or the recent past, that means that with quite recent data (i.e. weekly or monthly) we are able to make accurate predictions for a determined variable. Indeed, what I wanted to show in this article is the nowcasting power of Google Trends as a source for forecasting German and UK tourism volumes.

The database was directly extracted into weekly data from Google Trends. Initially, I chose 10 main keywords for each nationality, and then I selected the related queries from the second, third and fourth iteration. In total, 49 queries for Germany and 51 for UK were selected. Afterwards, I transformed the data from weekly into monthly data by implementing dynamic tables, since the tourist arrivals are represented into monthly data. Due to the seasonality of the queries I created lags 0-7 for each indicator, and then I calculated the Pearson correlation coefficient and select those which presented the highest correlation value. Later, I constructed a principal component analysis in order to gather the maximum information in different indexes. I found very good results and, indeed one index was created for Germany and two indexes were created for UK.

Thereafter, the comparison of the baseline model with alternative models using ARMAX framework was implemented. I compared each country with their corresponding baseline model, and I concluded that according to the different information criteria the best models were the baseline model for Germany, and the alternative model for the UK.

Our article makes two principal contributions in the existing forecasting literature. First, this is the first study that forecast tourism arrivals in the Balearic Islands by implement search engine methods such as Google Trends, and the results supported the idea that the Google Trends index enhances forecasting accuracy. In fact, this can pose the base for further research on that topic since it is interesting to reveal the structure of the "nowcasting" models. Second, I supported the methodology for query selection to better fit and predict visitors' volume suggested by Yang *et al.* (2015), however, I applied slightly variations to the initial methodology.

4.1. Implications

Our results reveal relevant implications for managerial decisions and for policy maker decisions. This new forecasting approach can influence managerial decisions mainly in the tourism and hospitality services by developing a new framework for monitoring and tracking short-term information about the consumers' demand. For instance in the case of a hotel, the capacity is limited for the short-term, since the hotel cannot enlarge the capacity from one day to another. Thus a room which is not occupied implies directly a loss to the hotelier, and therefore a decrease in the profits generated. Whether the hotel company is able to better adapt the demand, by increasing the forecasting accuracy including web search data into their forecasting models as I have shown in this paper, then the hotel will be able to improve its efficiency.

Indeed, our results pose some hints on how to solve some problems that revenue management tries to answer. For example, the forecasting part of the revenue management is crucial for determining overbooking strategies, pricing strategies, and so on. Hence, compared to traditional models of monitoring visitor numbers, the predictive power of our models based on web search data is much higher.

Moreover, policy makers could use web search queries of a particular region to release a forecasted tourists' index and companies can use it as a benchmark for local tourism and hospitality companies to measure their performance. For instance, whether the tourists' index showed a 30% of increase for a certain area in a month, yet if the reservation volume of that area only increases a 15%it should not be considered as a good result since there is still a 15% gap of reservation volume that will be lost. Furthermore, enhancing the accuracy of the estimating models will lead to a better prediction of the tourism demand and hence an increase in the efficiency of the resources managed by policy makers. For instance, in the Balearic Islands the overcrowded seasonal months (i.e. June, July and August) imply the need for better handle the tourism activity, such as the environmental management. Indeed, Álvarez-Díaz and Rosselló-Nadal (2010) estimated three different models (i.e. ARIMA, transfer function model and causal artificial neural network) and included meteorological explanatory variables in the analysis. Finally, they found out that including this weather factors enhanced the forecasting accuracy. This is a clear example of how weather factor, as in the case of the UK1 model, can improve the forecasting results of an estimating model. Lastly, policy makers can use the forecasted demands to predict the possible income generated by imposing a tourism tax such as the Ecotax.

4.2. Limitations and further research

Nevertheless our study presents some limitations that can pose a milestone for future research in this topic. The first one is that I only focused in Balearic Islands tourism arrivals; therefore the ability to generalize the conclusions is limited. Another important limitation that I ad hoc experimentation reveals that there exists a gap between Google Trends data and travel agency queries, at least, for the two countries that I selected. In fact, this might pose a threat to our results since they do not gather information regarding travel agency companies that play a key role in the travel and plan process of these two tourism countries. Further research in this last issue must be studied in order to clarify the significance of the relationship, and how it can influence the estimation results.

Bibliography

Aguiló, E., Riera, A. and Rosselló, J.
2005 "The short-term price effect of a tourist tax through a dynamic demand model: The case of the Balearic Islands". *Tourism Management,* 26(3):359-365.

Althouse, B.M., Ng, Y. Y. and Cummings, D.A.
2011 "Prediction of dengue incidence using search query surveillance". *PLoS Neglected Tropical Diseases,* 5(8): e1258.

Alvarez-Diaz, M., Mateu-Sbert, J. and Rosselló-Nadal, J.
2009 "Forecasting tourist arrivals to Balearic Island using genetic programming". *International Journal of Computational Economics and Econometrics,*1(1):65-75.

Álvarez-Díaz, M. and Rosselló-Nadal, J.
2010 "Forecasting British tourist arrivals in the Balearic Islands using meteorological variables". *Tourism Economics,* 16 (1):153-168.

Arenoe, B., Van der Rest, J.P. and Kattuman, P.
2015 "Game theoretic pricing models in hotel revenue management: An equilibrium choice-based conjoint analysis approach". *Tourism Management,* 51: 96-102.

Bangwayo-Skeete, P.F. and Skeete, R.W.
2015 "Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach". *Tourism Management,* 46: 454-464.

Butler, R.
1980 "The concept of a Tourist Area Cycle of Evolution: Implications for Management of Resources". *Canadian Geographer,* 24: 5-12.

Canestrelli, E. and Costa, P.
1991 "Tourist capacity: A fuzzy approach". *Annals of Tourism Research,* 18 (2): 295-311.

Castle, J.L., Fawcett, N.W. and Hendry, D.F.
2009 "Nowcasting is not just contemporaneous forecasting". *National Institute Economic Review*, 210 (1):71–89.

Choi, H. and Varian, H.
2009 "Predicting the Present with Google Trends". *Economic Record,* 88(1):2-9

Chu, F.L.
2009 "Forecasting tourism demand with ARMA-based methods". *Tourism Management,* 30(5):740-751.

Delgado-Verde, M., Martín-de-Castro, G., Navas López, J.E. and Cruz-González, J.
2011 "Capital social, capital relacional e innovación tecnológica. Una aplicación al sector manufacturero español de alta y media-alta tecnología". *Cuadernos de Economía y Dirección de la Empresa*, 14:207-221.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. and Brilliant, L.
2009 "Detecting influenza epidemics using search engine query data". *Nature,* 457(7232):1012-1014.

Gounopoulos, D., Petmezas, D. and Santamaria, D.
2012 "Forecasting tourist arrivals in Greece and the impact of macroeconomic shocks from the countries of tourists' origin". *Annals of Tourism Research,* 39 (2):641-666.

Hair, J.F. Jr, Anderson, R.E., Tatham, R.L. and Black, W.C.
2004 *Análisis multivariante*, 5th ed.,Pearson-Prentice Hall, Spain:Madrid.

Huang, X., Zhang, L. and Ding, Y.
2016 "The Baidu Index: Uses in predicting tourism flows: A case study of the Forbidden City". *Tourism Management*, In press.

Ibestat, Institut d'Estadística de les Illes Balears (2016). *Estadísticas Economía, turismo, flujo de turistas.* Retrieved from http://www.ibestat.cat/ibestat/estadistiques/economia/turisme/fluxe-turistes-frontur/043d7774-cd6c-4363-929a-703aaa0cb9e0. Accessed 26.07.16

Jackman, M. and Naitram, S.
2015 "Nowcasting tourist arrivals in Barbados- just Google it!". *Tourism Economics,* 21 (6):1309-1313.

Jacob, M., Florido, C. and Aguiló, E.
2010 "Environmental Innovation as a competitiveness factor in the Balearic Islands". *Tourism Economics,* 16(3):755-764

Kennedy, A.F. and Hauksson, K.M.
2012 *Global Search Engine Marketing: Fine-tuning Your International Search Engine Results*. Pearson, USA: Indiana

Li, G., Song, H. and Witt, S. F.
2005 "Recent developments in econometric modelling and forecasting". *Journal of Travel Research*, 44: 82–99.

Palmer, T. and Riera, A.
2003 "Tourism and environmental taxes. With special reference to the "Balearic Ecotax". *Tourism Management,* 24(6):665-674.

Pan, B., Litvin, S.W. and Goldman, H.
2006 *Real Users, Real Trips, and Real Queries: An analysis of Destination Search on a Search Engine*, Annual Conference of Travel and Tourism Research Association, Ireland.

Schwartz, Z., Stewart, W. and Backlund, E.
2012 "Visitation at capacity-constrained tourism destinations: Exploring revenue management at a national park". *Tourism Management,* 33 (3):500-508.

Shen, S., Li, G. and Song, H.
2011 "Combination forecast of international tourism demand". *Annals of Tourism Research,* 38(1):72-89.

Song, H. and Li, G.
2008 "Tourism demand modelling and forecasting- A review of recent research". *Tourism Management,* 29(2):203-220.

Song, H., Witt, S.F. and Li, G.
2008 *The advanced econometrics of tourism demand.* New York: Routledge

Vinod, B.
2004 "Unlocking the value of revenue management in the hotel industry". *Journal of Revenue and Pricing Management,* 3(2):178-190.

Yang, X., Pan, B., Evans, J. and Lv, B.
2015 "Forecasting Chinese tourist volume with search engine data". *Tourism Management,* 46:386-397.

Recibido *29/12/2016*
Reenviado *12/03/2017*
Aceptado *12/03/2017*
Sometido a evaluación por pares anónimos