Forecasting travellers in Spain with Google's searches volumes indices[+]

Maximo Camacho[*]
Universidad de Murcia and BBVA
mcamacho@um.es

Matías José Pacce
BBVA Research
matias.pacce@bbva.com

Abstract

We examine whether Google's searches volume indices helps economic agents with real-time predictions about the checked in and overnight stays of travellers in Spain. Using a dynamic factor approach and a real-time database of vintages that reproduces the exact information that was available to a forecaster at each particular point in time, we show that the models including Google's queries volumes indices outperform models that exclude these leading indicators. In this way, we are the first in finding conclusive evidence that tourism related queries helps to improve tourism forecast in Spain. Our finding is of significance in this literature since Spain is one of the world's top tourism destinations and extremely depends on tourism.

JEL Classification: E32, C22, E27, Z30.

Keywords: Tourism, Big data analysis, Time series.

+ We are thankful to R. Domenech, M. Cardoso, C. Ulloa, A. Urcola, M. Trias, M. Moya, the editor, and the anonymous reviewers for their comments that have greatly improved the quality of the paper. M. Camacho acknowledges support from projects ECO2013-45698-P, ECO2016-76178-P, and 19884/GERM/15 (Groups of Excellence, Fundación Séneca, and Science and Technology Agency). All remaining errors are our responsibility.

[*] Corresponding Author: Universidad de Murcia, Facultad de Economia y Empresa, Departamento de Metodos Cuantitativos para la Economia y la Empresa, 30100, Murcia, Spain. E-mail: mcamacho@um.es

1. Introduction

The Spanish economy is extremely dependent on tourism and is one of the world's top tourism destinations. In 2016, according to the World Tourism Organization, by international tourism receipts, Spain was in second position, with 60.3 billion US dollars, only behind the United States. By volume of international arrivals, Spain ranked the third best, with 75.6 million tourists, after France and the United States. In 2015, as reported in the latest publication from the Spanish Tourism Satellite Account, the volume of tourism activity reached the amount of 11.1% of GDP.[1]

In accordance with these magnitudes, having accurate previsions about the dynamism of current and upcoming tourism is of primary importance for policy authorities in assessing overall economic developments. In addition, having timely information about the evolution of tourism is also crucial in the previsions of the hospitality and tourism industry, which need to find and develop new means to distribute travel and hospitality products and services, to manage marketing information for consumers, and to provide comfort and convenience to travellers. Unfortunately, in spite of these real-time monitoring requirements, data on the number of travellers checked in and on the number of overnight stays, the two major measures of tourism in Spain, are published monthly with a one-month lag, which difficult the previsions.

In this paper, we follow the idea that the increasingly widespread use of the Internet by travellers has led to the creation of a potentially useful data source of leading tourism indicators that could help both policy authorities and the tourist industry to perform early assessments on tourism performance. In this context, the tourist industry has been among the first to capitalize on new technology, and the number of travellers that use the Internet to plan and book their business and pleasure trips has significantly grown during the last decade. In line with those developments, recent literature has focused on exploiting the valuable information search query data provided about tourists' behaviour. Google's dominance in the field of search engines makes this web search engine a reliable representative from which to examine the forecasting contents of search results.[2]

While not claiming to be exhaustive, Pan et al. (2012) showed that including information about aggregated search trends improved the weekly forecast accuracy of demand for hotel rooms in South California. Jackman and Naitram (2015) found that air passenger arriving in Barbados from Canada and UK could be better predicted one week ahead, by including a Google Trends series with queries performed from those two countries. Li et al. (2017), used a generalized dynamic factor model to extract a weekly

"search index" based on Google Trend data to obtain out-of-sample improvements in forecast accuracy of tourist arrivals in Beijing. Yang et al. (2015) examined the predicted power of the queries entered into search engines on the number of visitors in Hainan (China). Bangwayo-Skeete and Skeete (2015) used query trends from Canada, the US and UK to forecast values 12 months prior to monthly tourist arrivals in five Caribbean countries. Rivera (2016) found that including information about query trends from the US helps to improve forecasting accuracy on a 12-month horizon, but not for short-term forecasts.

In spite of these promising results for other countries, the potential ability of the amount of information from internet searches to forecast tourism in Spain has been underestimated. To the knowledge of the authors, only Artola and Galan (2012) presented a very specific application for the Spanish economy, namely British tourists (the Spanish tourist industry's main clients) visiting Spain. Although they computed an adjusted indicator of the flow of British tourists with a lead of almost one month, the improvement in forecasting tourism provided by their short-term models is very limited. Therefore, they suggested exploring in further research the information available from other countries to compute leading indicators of incoming tourists.

This study pretends to fulfil this gap by contributing to the literature in several ways. In collaboration with Google, we develop a novel data-set that collects information on the volume of queries associated with different specific tourism-related terms from some specific countries. This Google volume searches data-set provides reports on the real-time evolution of queries related to various tourism industries in the online travel market and on the use of the Internet and e-commerce for travel.

The data-set based on volume searches departs from Google trends data in two main aspects. First, the volume of search data-set is related with the total number of queries of a set of terms from a specific country while Google trends refers to the popularity that a specific term reaches with respect to the total searches performed at an specific time range and geography. Second, while Google trends data comes from a periodic random sample of searches data, volume searches are always collected from a larger, but fixed, sample of queries regardless of the moment when the data is extracted.[3]

In particular, the Google searches volume data set is built at a country level for queries done from Austria, Germany, France, Ireland, Italy, Switzerland, the United States and United Kingdom, which accounted for almost two-thirds of the total non-resident overnights stays in Spanish hotels during 2015. In addition, the query volumes are related to

travel facilities (air, ferries, bus and rail), accommodation (hotel, holiday rental and camping), vacation packages, and general matters about travel and destination (city and short trips, activities, weather, rent a car). This amounts to a total of 65 series of searches volumes from 8 different countries in real time.[4]

To deal with this large amount of information, we rely on Dynamic Factor Models (Stock and Watson, 2011). Within this framework, the goal is to explain the maximum amount of variance in the searches volumes with the fewest number of common factors. Therefore, we allow all the information contained in the series to be potentially valuable in order to extract the relevant signals on the query volumes dynamics in a small number of common components. Then, we examine the usefulness of this information to improve the accuracy of short-term forecasts of the checking in and overnight stays of travellers in real time.

Our results suggest that the model using searches volumes yields significant forecasting improvements over benchmark predictions computed from standard autoregressive specifications. To show the advantages of our proposal, we develop a pseudo real-time forecasting exercise, which is carried out over from 2014.09 until 2016.01, in a recursive way. With every new vintage of data, the model is re-estimated and the forecasts for different horizons are computed. The vintages are constructed by taking into account the lag of synchronicity in data publication that characterizes the real-time data, by mimicking the pattern of the actual chronological order of the data releases. In each forecasting day in month t, the model predicts the tourism data in month t-1 (backcast), in month t (nowcast) and in month t+1 (forecast). Although the gains depend on the forecasting horizon, we do find forecasting improvements from using the query volumes to forecast tourist indicators in real time for all the forecasting horizons.

The structure of this paper is as follows. Section 2 outlines the dynamic factor model, which relates the tourism indicators to be forecast to the set of Google searches volumes. Section 3 analyses the estimated factors and examines the empirical performance of Google query volumes in forecasting tourism indicators in Spain. Section 4 concludes and proposes several future lines of research.

2. Dynamic Factor Models

Models that manage large sets of indicators typically suffer a trade-off between the data reduction requirements and the cost of discarding relevant information. Factor models are traditional dimensionality reduction techniques that try to mitigate this problem by

summarizing the whole cross-section dynamic in a few common factors (Geweke, 1977; Sargent, 1977). Then, the estimated factors can be used to provide efficient forecasts of a target variable in a simple linear regression. Significant examples can be found in Stock and Watson (2002a, 2002b), Bai (2003) and Forni et al. (2005).

The forecast problem can be described using two basic equations. Let y_t be either the checking in or overnight stays of travellers, the target series to forecast. Let X_t be an N-dimensional vector of searches volumes.[5] Assume that the query volumes admit a factor model representation, i.e., the evolution of the time series can be decomposed as the sum of r common unobserved factors, F_t, and their respective idiosyncratic dynamics, e_t,

$$X_t = \Lambda F_t + e_t, \tag{1}$$

where Λ is an $N \times r$ matrix of the factor loadings, and e_t is an $N \times 1$ vector of independent idiosyncratic disturbances. Provided that F_{t+h} is available, the h-horizon forecast equation is described by the forecasting equation

$$y_{t+h} = \mu + \beta(L) F_{t+h} + \alpha(L) y_{t+h-1} + \gamma HW_{t+h} + \varepsilon_{t+h}, \tag{2}$$

where μ is a constant, $\beta(L)$ is a vector lag polynomial, $\alpha(L)$ is a scalar lag polynomial and ε_{t+h} is the forecast error.[6] The term HW_t is a dummy variable that takes on the value one if month t refers to the Holy Week.[7] Once the model is estimated, the forecast is then performed as

$$\hat{y}_{T+h} = \hat{\mu} + \hat{\beta}(L) \hat{F}_{T+h} + \hat{\alpha}(L) \hat{y}_{T+h-1} + \hat{\gamma} HW_{T+h}, \tag{3}$$

where the $\hat{\mu}$, $\hat{\beta}(L)$, $\hat{\alpha}(L)$, $\hat{\gamma}$, $\hat{F}_{T+h}$, and $\hat{y}_{T+h-1}$ are the estimated coefficients, the estimated factors up to $T+h$, and the estimated dependent variable up to $T+h$-1.

In order to estimate the unobserved common factors, we follow the lines suggested by the influential contribution by Stock and Watson (2002a). Skipping details, the methodology is based on estimating the dynamic factors through principal components. Following their notation, it is possible to write the nonlinear least square function,

$$V(\tilde{F}, \tilde{\Lambda}) = (NT)^{-1} \left(X - \tilde{\Lambda}\tilde{F}\right)' \left(X - \tilde{\Lambda}\tilde{F}\right), \tag{4}$$ as

a function of hypothetical values for factors, $\tilde{F} = \left(\tilde{F}_1, ..., \tilde{F}_T\right)$, and factor loadings, $\tilde{\Lambda}$. When $N>T$, minimizing (4) is equivalent to maximize $tr\left[\tilde{F}'(XX')\tilde{F}\right]$ subject to $\tilde{F}'\tilde{F}/N = I_r$, where $tr(\bullet)$ denotes the trace of the matrix. This problem is solved by writing down the

principal component estimator $\hat{F}$ as the matrix that contains the eigenvectors associated with the r largest eigenvalues of XX'.

3. Empirical Results

3.1. Data description

Due to the widespread popularity of the Internet, a growing number of travellers use web search engines to planning their trips and stays. The anonymised searches made with Google have been used to construct weekly indices that collect the relevant information on the trips and stays that travellers take and intend to take. The searches volumes used to obtain all the results of this paper come from weekly reports on indexed volumes of different search term baskets related to various tourism industries that cover the period from the first week of July 2007 to the second week of January 2016.

This data set, based on searches volumes, differs from the data sets collected from Google trends in two main aspects. First, Google trend is an index of the popularity of a specific term with respect to the total searches performed at a specific time range and geography. In this sense, Google trend data is typically scaled on a range of 0 to 100 while searches volumes are referred to a value of 100 at the first observation of the sample. The second distinctive feature of our data set with respect to Google trends data sets has to do with randomization issues. While Google trends data comes from a periodic random samples of searches data that change every week, volume searches are always collected from a larger, but fixed, sample of queries regardless of the moment when the data is extracted.

[Table 1 about here]

Table 1 summarizes the searches related with tourism, the country of origin, and the availability of the data. Classified by country of origin, searches volumes show how often several traveling related topics have been searched for on Google over time. The countries where the searches were collected from are Austria, Germany, France, Ireland, Italy, Switzerland, the United States and United Kingdom, which accounted for 62% of the total non-resident overnight stays in Spanish hotels during 2015.

The query volume indices rely on searches on travel facilities (air, ferries, bus and rail), accommodation (hotel, holiday rental and camping), vacation packages, general travel and destination (city and short trips, activities, weather, rent a car).[8] As previously said, all search volume indices start with a large sample of the total query volume related to each

specific term in a specific country divided by a constant at a point in time. The resulting figures are then normalized so that they start at 100 in the first week of July 2007. Finally, to be compared with the checked-in and overnight stays of travellers, which are published on a monthly basis, we compute the monthly averages of the weekly indices.

To examine the dynamics of travel related Google search, Figure 1 shows a weighted average of all query indices, which although not used in the empirical analysis, is obtained for reasons of presentation. In addition, the figure also plots two official tourism statistics, the overnight stays and the number of non-resident travellers checked in hotels. Regarding tourist indicators, the INE (National Statistics Institute) states that checked-in travellers include all people who stay one or more consecutive nights in the same collective tourist accommodation. Overnight stays include every night that a traveller spent in these establishments. In the paper, we focus on the versions of tourist indicators that only account for non-residents.[9]

The figure shows a high correlation between short-term movements in the tourist indicators and the weighted query index, in both cases showing the same strong seasonal pattern. Moreover, the averaged query index appears to start growing a few months before the beginning of each summer season, which could be related to people planning ahead for their holidays.

[Figure 1 about here]

To remove seasonal patterns, we use year-on-year growth rates instead of monthly growth rates of seasonally adjusted data.[10] Therefore, to be compared with the annual growth rate transformation employed in the case of the query indices, we also use year-on-year growth rates for the tourist indicators in the model. According to Figure 2, the evolution of tourist indicators in Spain showed a phase of deep decline during the Great Recession followed by a period of steady growth thereafter. In light of the severity of the 2008 downturn and the rapid recovery in 2009 suffered in the tourism sector, the relevant question is whether queries volumes can help to anticipate the current and short-term evolution of tourist developments, to allow policy makers and the tourist industry to adopt preemptive measures.

[Figure 2 about here]

Figure 2 also reveals that search volumes and tourist indicators cohere strongly across time during the sample period. In fact, the in-sample correlation between total travel related Google queries and non-resident overnight stays or the checked-in into hotels are up to 0.61 and 0.58, respectively. A good example of this closed relationship among searches volumes and tourist indicators can be depicted in Figure 3, which shows how the annual growth rate of each of the travel related query from Italy correlates with the annual growth rate of Italian overnight-stays in Spanish hotels. In particular, we show a two-year rolling window of that correlation for each of the query volume index specified. According to the figure, the correlations are close to one in most of the cases and along the complete period (vintages from 2010.07 to 2015.12).

[Figure 3 about here]

3.2. In-sample analysis

A total of 65 series of year-on-year growth rates of query volumes are used to estimate the common factor model by principal components. The first three estimated factors are plotted in Figure 4.

[Figure 4 about here]

In order to give an interpretation of the estimated unobserved components, we follow Stock and Watson (2002a) and we compute the R^2 of the regression of the 65 query volumes series against each of the first three factors estimated over the full sample period. These R^2 are plotted in Figures 5 and 6 as bar charts with one chart for each factor. In Figure 5, the search volumes are grouped by category, starting from those which have a larger R^2 with respect to the first factor.

[Figure 5 about here]

The figure shows that the first factor loads primarily on "Pure Destination", where the R^2 is above 0.3 in seven out of eight cases. For the second factor, the query volumes are mainly related to "Hotels" and "Bus and Rail", while "Pure destination" continues to be relevant.[11] Regarding the third factor, query volumes related to "Hotels", "Air" and

"Activities at destination" are the most significant, although the R^2 is bigger than 0.1 in only 6 out of 65 search volumes series.

In Figure 6, the query volumes indices are grouped by countries to examine the importance of the country searches on the formation of factors. The figure shows high correlations between the first factor and the country searches, which implies that the first factor is representative for all countries. However, searches from Italy and the United States seem to play a prominent role in the formation of the second factor while the first third rests on the United Kingdom, Germany and Ireland.

[Figure 6 about here]

3.3. Simulated real-time analysis

The results obtained in the in-sample analysis are in practice only of limited usefulness. In monitoring the tourist sector, the analysis is developed in real time, where data are subject to differences in publication lags, which we need to take account of when computing the forecasts. Accordingly, we propose a forecast evaluation exercise that is designed to replicate the typical situation where the model manages real-time data flow. For this purpose, we construct a sequence of data vintages from the final vintage data set that tries to mimic the actual real-time vintages, in the sense that the delays in publication are incorporated.

Without losing generality, we assume that the forecasts are computed on the 15th of each month *t*. According with the publication lags, in month *t* the data set used in the forecasts is updated with the tourist indicator up to month *t*-2. However, query indexes are available to compute monthly averages up to month *t*-1 and the average of the first two weeks of month *t*. Figure 7 shows that the latter are accurate proxies of the monthly query averages of month *t*.

[Figure 7 about here]

In each month *t*, using the generated sequence of data vintages the models compute inferences of the tourist indicators in month *t*-1 (backcast), in month *t* (nowcast) and in month *t*+1 (forecast) in a recursive way. Starting with the backcasts, the model

$$y_{t-2} = \mu + \alpha_1 y_{t-3} + \alpha_2 y_{t-4} + \sum_{i=1}^{r}\sum_{j=0}^{m} \beta_i^j F_{t-j-2}^i + \gamma HW_{t-2} + \varepsilon_{t-2}, \qquad (5)$$

where *r* refers to the number of factors and m to the number of factor lags, is estimated using data up to *t*-2. Then, the backcasts of *t*-1 are computed as

$$\hat{y}_{t-1} = \hat{\mu} + \hat{\alpha}_1 y_{t-2} + \hat{\alpha}_2 y_{t-3} + \sum_{i=1}^{r}\sum_{j=0}^{m} \hat{\beta}_i^j F_{t-j-1}^i + \hat{\gamma} HW_{t-1}, \quad (6)$$

To compute the nowcast, the model

$$y_{t-1} = \mu + \alpha_1 y_{t-2} + \alpha_2 y_{t-3} + \sum_{i=1}^{r}\sum_{j=0}^{m} \beta_i^j F_{t-j-1}^i + \gamma HW_{t-1} + \varepsilon_{t-1}, \quad (7)$$

is estimated with data up to *t*-1.[12] Then, the nowcast is computed as

$$\hat{y}_t = \hat{\mu} + \hat{\alpha}_1 \hat{y}_{t-1} + \hat{\alpha}_2 y_{t-2} + \sum_{i=1}^{r}\sum_{j=0}^{m} \hat{\beta}_i^j F_{t-j}^i + \hat{\gamma} HW_t, \quad (8)$$

where we use the backcast $\hat{y}_{t-1}$.

Finally, the forecasting equation is re-estimated to compute forecasts

$$y_{t-2} = \mu + \alpha_1 y_{t-3} + \alpha_2 y_{t-4} + \sum_{i=1}^{r}\sum_{j=0}^{m} \beta_i^j F_{t-j-3}^i + \gamma HW_{t-2} + \varepsilon_{t-2}, \quad (9)$$

with the extended data set up to *t*. The forecast of *t*+1 is

$$\hat{y}_{t+1} = \hat{\mu} + \hat{\alpha}_1 \hat{y}_t + \hat{\alpha}_2 \hat{y}_{t-1} + \sum_{i=1}^{r}\sum_{j=0}^{m} \hat{\beta}_i^j F_{t-j}^i + \hat{\gamma} HW_{t+1}, \quad (10)$$

where $\hat{y}_{t-1}$ is the backcast and $\hat{y}_t$ is the nowcast.

The first data vintage of this experiment refers to data as it would be known on October 15, 2014. According to the three-month blocks of forecasts computed from the model, the models produce forecasts of the tourist indicators in September 2014 (backcast), October 2014 (nowcast), and November 2014 (forecast).[13] Following this updating scheme, the data set is updated each month up to January 15, 2016, leading to 15 different vintages.

We are now in a condition to assess the extent to which the searches in Google data help tourism prediction. For this purpose, we compute the Root Mean Squared Error (RMSE), which is the average of the deviations of the predictions from the latest releases of the tourist indicators available in the data set. In addition to the model that incorporates the information coming from Google searches volumes, a univariate autoregressive model, which is also estimated in pseudo real-time producing iterative forecasts is included as a benchmark model.[14]

To facilitate comparisons, Table 2 reports the RMSEs relative to the univariate autoregressive model. Hence, an entry of less than one indicates that the factor model forecast is superior to the autoregressive univariate forecast. The immediate conclusion obtained when comparing the forecasts results displayed in the table is that it is beneficial to

use the query volumes information in forecasting the Spanish tourism. However, the relative gains from the model that uses the search volumes indices depends on the number of factors and lags for the factors included in the model. Regarding the backast and nowcast ability of the model, major gains are obtained when two factors and three lags for those factors are included in equation (3), both in the case of predicting overnight-stays and checked-in traveller variables. In the former, the RMSEs fall, in general, by at least 7% (in the case of rental apartments major gains are found when three factors and one lag for those factors are included). Regarding checked-in travellers the gains are relatively lower, being in general between 6% and 10%. When the focus is on forecasts, the higher gains are found when a model with 3 factors and 0 lag for the factors is used. In that case, the relative RMSEs are, depending on the target variable, between 13% and 24% lower than in the case of an AR(2).

[Table 2 about here]

This result confirms the leading forecasting ability of tourism indicators by query volumes indices, which is clearly achieved when the early available search data is accounted for by the model. The promptly published information of search volumes series is relatively much richer and more valuable in forecasting than in the backcasting and nowcasting exercises.

As a final remark, we point out that this model can be used to compute backcasts, nowcasts and forecasts on any day of the month, which implies using information on query volumes updated until the day before the forecast computation. As an example of how the model produces inferences Figure 8 shows the backcast, nowcast and forecast of overnight-stays in hotel that were obtained on February 15, 2016, along with the prediction errors. It should be noticed that the remarkable increase expected for March, is associated with a base effect related to Easter.[15]

[Figure 8 about here]

4. Conclusions

The Internet has radically changed the manner in which tourists and travellers obtain travel-related information. The evidence presented in this paper, based on the performance of tourism search volumes provided by Google over a real-time exercise, has provided very promising support for using search information to predict checked-in and overnight stays of

non-resident travellers in Spain. Our finding is of significance in this literature since Spain is extremely dependent on tourism and is one of the world's top tourism destinations.

As in any big data setup, the first step is to capture the big amount of information provided by the volume of searches. For this purpose, we assume that the queries volume indices admit a factor model decomposition, in which each query volume series is the sum of a small set of common factors and an idiosyncratic component. Then, common factors are used to forecast checked-in and overnight stays of travellers. Within this framework, we find that the promptly published information of search volumes series is relatively much richer and more valuable in forecasting than in the backcasting and nowcasting exercises.

Despite these promising results, it is important to recognize that the conclusions regarding the performance of searches volume series examined in this paper are necessarily tentative, mainly because of the limited number of observations that are available for the query volume indices. As more data become available, future work on the help of search volumes series in the forecasting of tourism indicators could include using additional tourism indicators, extracting seasonal components from the time series with seasonal adjustment techniques, and using nonlinear forecasting methods.

Endnotes

[1] In 2016, this figure rose up to 11.2% according to the Spanish group Exeltur (Alliance for Tourism Excellence).
[2] According to StatCounter, Google has roughly 90 percent of the global search market in 2016, though precise share varies by country.
[3] In contrast to data based on volume searches, data based on trends could miss the cases where foreigners are increasingly planning traveling to Spain but are searching by using non-trending topics.
[4] The complete data base is available from the authors upon request.
[5] As usual, $t=1,...,T$, is the number of time series observations.
[6] For notation simplicity, the dependence of the parameters on h is suppressed.
[7] The dummy variable attempts to remove remaining seasonal effects that occur on Holy Weeks.
[8] The information on queries is analysed on the local language used at the country from which the query was originated.
[9] In the empirical application, we examine the potential improvements of the query volume indices to forecast tourism indicators by type of accommodation: hotels, rental apartments and the sum of the two, plus camping.
[10] It is hardly possible to compute accurate seasonal factors by employing standard techniques of seasonal adjustment since searches volume indices are available only since 2007.
[11] "Bus and Rail" is only available for Italy.
[12] Notice that the model uses the backcast $\hat{y}_{t-1}$ for time t-1.
[13] At month t, the nowcast at t and forecast at t+1 can only use query volumes series of the first two weeks of this month.
[14] This benchmark model includes the Holy Week dummy aiming to distinguish the differences emerging when Google query volumes' information is incorporated in the model. Note that our proposal becomes the benchmark when all the parameters that refer to the Google query volumes are set to zero.
[15] In 2015, Easter occurred entirely during April, while in 2016 it took place in March.

References

Artola, C., and Galan, E. (2012). Tracking the future on the web: construction of leading indicators using Internet searches. Banco de España Occasional Paper Series N. 1203.

Bai, J. 2003. Inferential theory for factor models of large dimensions. *Econometrica* 71: 135-171.

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

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. 2005. The generalized dynamic factor model: one-sided estimation and forecasting. *Journal of the American Statistical Association* 100: 830-40.

Geweke, J. 1977. The dynamic factor analysis of economic time series. In *Latent variables in socio-economic models*, D. Aigner and A. Goldberger (eds). Amsterdam: North-Holland.

Jackman, M., and Naitram, S. 2015. Nowcasting tourist arrivals to Barbados. Just Google It! *Tourism Economics* 21: 1309-1313.

Li, X; Pan, B; Law, R; and Huang, X. 2017. Forecasting tourism demand with composite search index. *Tourism Management* 59: 57-66.

Pan, B., Wu, C., and Song, H. 2012. Forecasting hotel room demand using search engine data. *Journal of Hospitality and Tourism Technology* 3: 3-13.

Rivera, R. 2016. A dynamic linear model to forecast hotel registrations in Puerto Rico using Google Trends data. *Tourism Management* 57: 12-20.

Sargent, T., and Sims, C. 1977. Business cycle modeling without pretending to have too much a-priori economic theory. In *New Methods in Business Cycle Research*, C. Sims et al. (eds). Minneapolis: Federal Reserve Bank of Minneapolis.

Stock, J., and Watson, M. 2002a. Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics* 20: 2: 147-162.

Stock, J., and Watson, M. 2002b. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97: 1167-1179.

Stock, J., and Watson, M. 2011. Dynamic Factor Models. In *Oxford handbook of forecasting*, M. Clements and D. Hendry (eds). Oxford: Oxford University Press.

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

Table 1. Query volume series available per countries

	Austria	France	Germany	Ireland	Italy	Switzerland	UK	US
Air	a	a	a	a	a	a	a	a
Bus & Rail	na	na	na	na	a	na	na	na
Camping	na	a	a	na	a	na	na	na
Rent a car	a	na	na	a	a	na	a	na
Activities at destination	na	na	na	a	a	na	a	a
Ferries	na	na	na	na	na	na	a	na
Travel in General	a	na	a	a	a	a	a	a
Hotels	a	a	a	a	a	a	a	a
Vacation Package	a	a	a	a	a	a	a	a
Pure destination	a	a	a	a	a	a	a	a
Holiday Rental	na	a	a	na	na	a	na	a
City & Short trips	na	a	na	na	na	a	na	na
Weather	a	a	a	a	na	a	a	a

Note: The symbol a (na) means that the query volume was (not) available for that country.

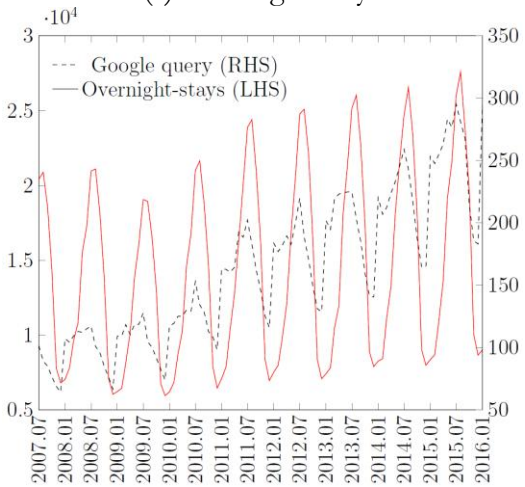
Table 2. Predictive accuracy: Enlarged AR (values relative to an AR model)

		Non-resident overnight-stays								
		Total			Hotels			Rental Apartments		
k	*m*	*t*-1	*t*	*t*+1	*t*-1	*t*	*t*+1	*t*-1	*t*	*t*+1
AR(2)		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0	1.00	.95	.98	.99	.97	.98	1.01	.95	.96
	1	.98	.96	.98	.99	.97	.99	1.01	1.00	.97
1	2	.99	.97	1.05	1.00	.97	1.03	1.01	1.01	1.01
	3	.92	.93	1.08	.92	.92	1.07	1.02	1.05	1.01
	4	.93	.97	1.07	.93	.97	1.07	1.02	1.05	1.02
	0	.96	.98	.95	.87	.89	.89	1.01	.95	.94
	1	.94	.91	.92	.93	.88	.89	1.03	1.00	.98
2	2	.92	.88	.98	.92	.86	.94	1.00	1.06	1.03
	3	.89	.89	1.04	.89	.86	1.01	1.00	1.13	1.07
	4	.92	.95	1.01	.93	.93	.98	.98	1.20	1.08
	0	1.02	.91	.81	1.01	.90	.80	1.07	1.07	.78
	1	.98	.97	.85	.97	.94	.85	.89	.84	.81
3	2	.98	.98	.94	.99	.97	.94	.87	.90	.88
	3	.95	.99	1.05	.97	.98	1.07	.96	1.04	.92
	4	1.00	1.09	1.06	1.04	1.13	1.12	.90	1.00	.81
		Non-resident travelled checked-in								
		Total			Hotels			Rental Apartments		
k	*m*	*t*-1	*t*	*t*+1	*t*-1	*t*	*t*+1	*t*-1	*t*	*t*+1
AR(2)		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0	1.00	1.01	1.00	1.00	1.01	.99	1.00	.98	.99
	1	1.02	1.01	.99	1.04	1.02	.99	.99	1.00	.99
1	2	1.03	1.01	1.04	1.04	1.02	1.03	.99	.98	1.05
	3	.97	.98	1.05	.99	.98	1.04	.98	1.05	1.05
	4	.96	.97	1.02	.98	.98	1.01	.97	1.04	1.03
	0	.95	.91	.93	.94	.90	.93	.97	.93	.93
	1	.96	.92	.92	.96	.92	.92	.97	.96	.93
2	2	.98	.92	.97	.98	.91	.96	.96	.95	1.00
	3	.98	.89	.99	.99	.89	.98	.97	1.02	.99
	4	.97	.90	.88	.98	.91	.88	.97	1.02	.99
	0	1.01	.97	.91	1.00	.97	.93	.99	.95	.82
	1	1.03	.99	.93	1.02	1.00	.94	.93	.87	.84
3	2	1.04	1.01	.99	1.04	1.01	.98	.95	.87	.94
	3	1.04	.97	1.03	1.04	.98	1.03	.97	.93	.88
	4	1.08	1.03	.94	1.08	1.07	.97	1.02	.98	.80

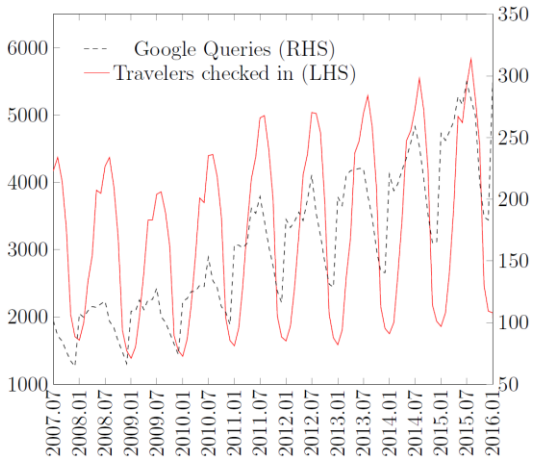
Note: *t*-1, *t* and *t*+1 refer to the backcasting, nowcasting and forecasting exercises; *k* and *m* refers to the number of factors and lags (for those factors) included in the model. The forecasting sample is 2014.09-2016.01, which implies comparisons over 17 forecasts. Entries are the relative (to an AR model) Root Mean Squared Errors (RMSE) of an autoregressive model that is enlarged with the first *k* common factors extracted from a principal component for travel related query.

Figure 1: Query index and non-resident tourism indicators

(a) Overnight Stays



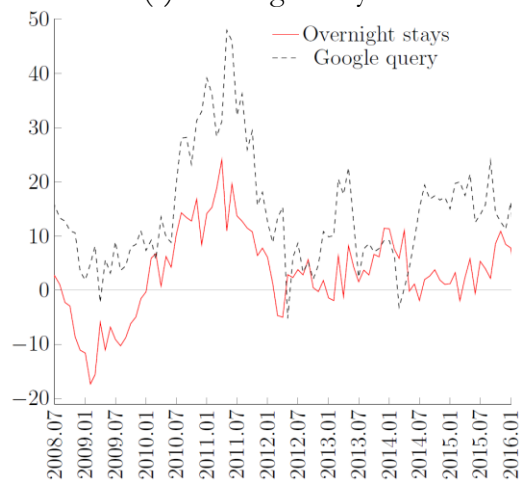
(b) Travellers checked in



Note: Travellers checked in and overnight stays are expressed in thousands. Both tourism indicators are obtained from the National Statistics Institute. The query index is from Google

Figure 2: Comparison of yearly growth rates

(a) Overnight Stays



(b) Travellers checked in

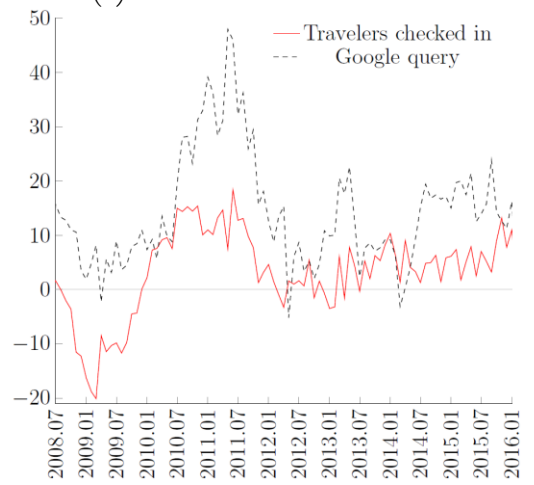
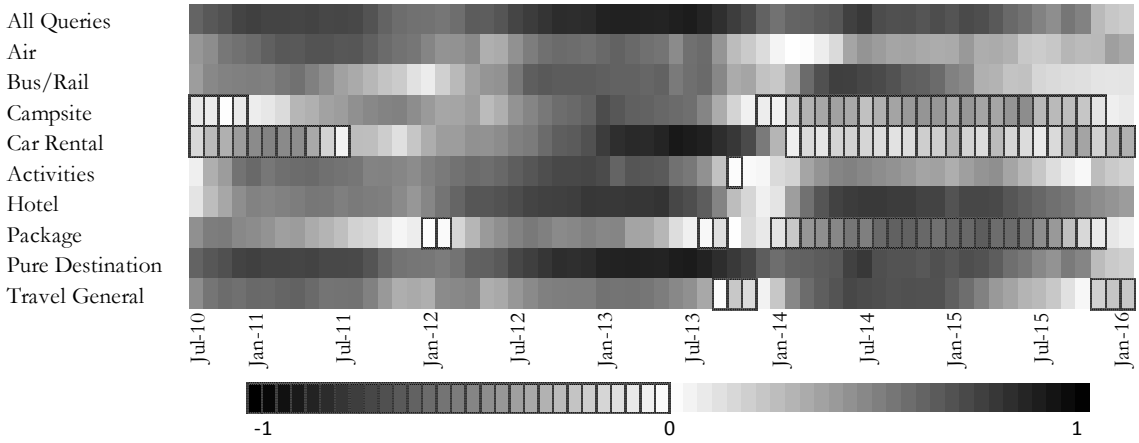


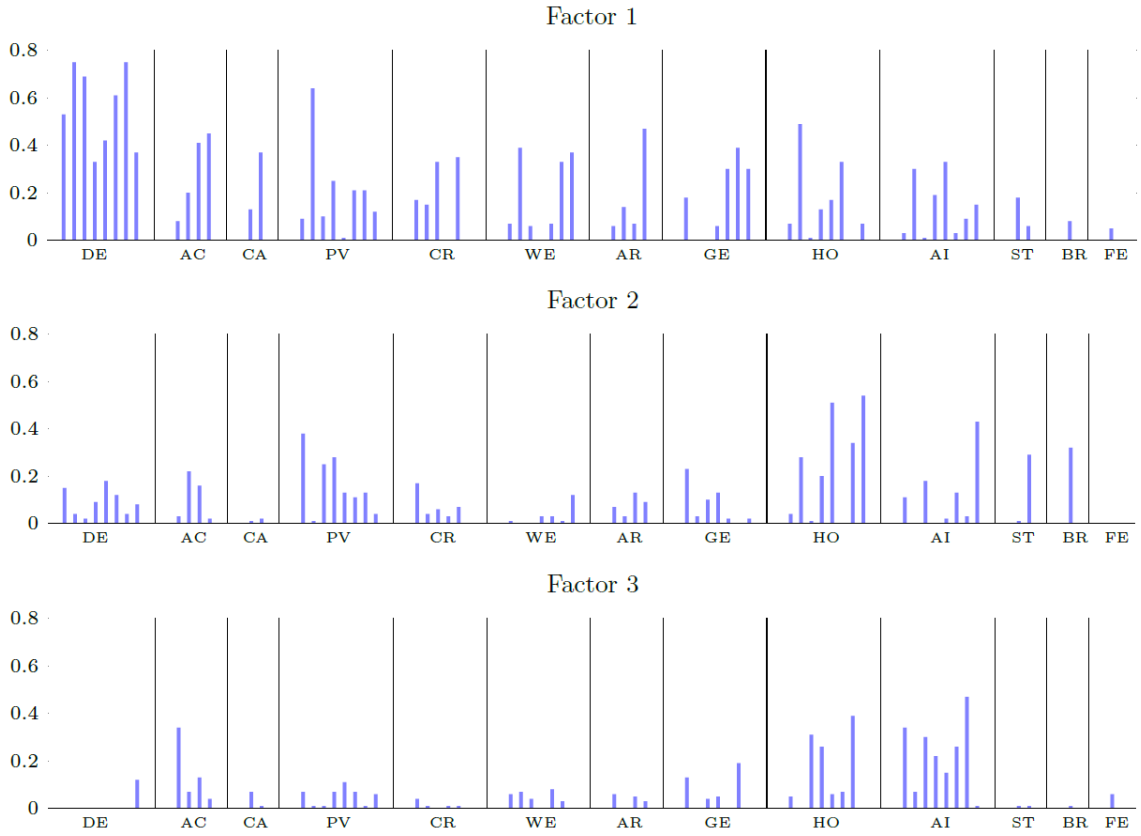
Figure 3: Correlations between Italian overnights stays (Spanish hotels) and travel related Google query



Note: Two years rolling windows correlations. Windows from 2010.07 to 2016.01

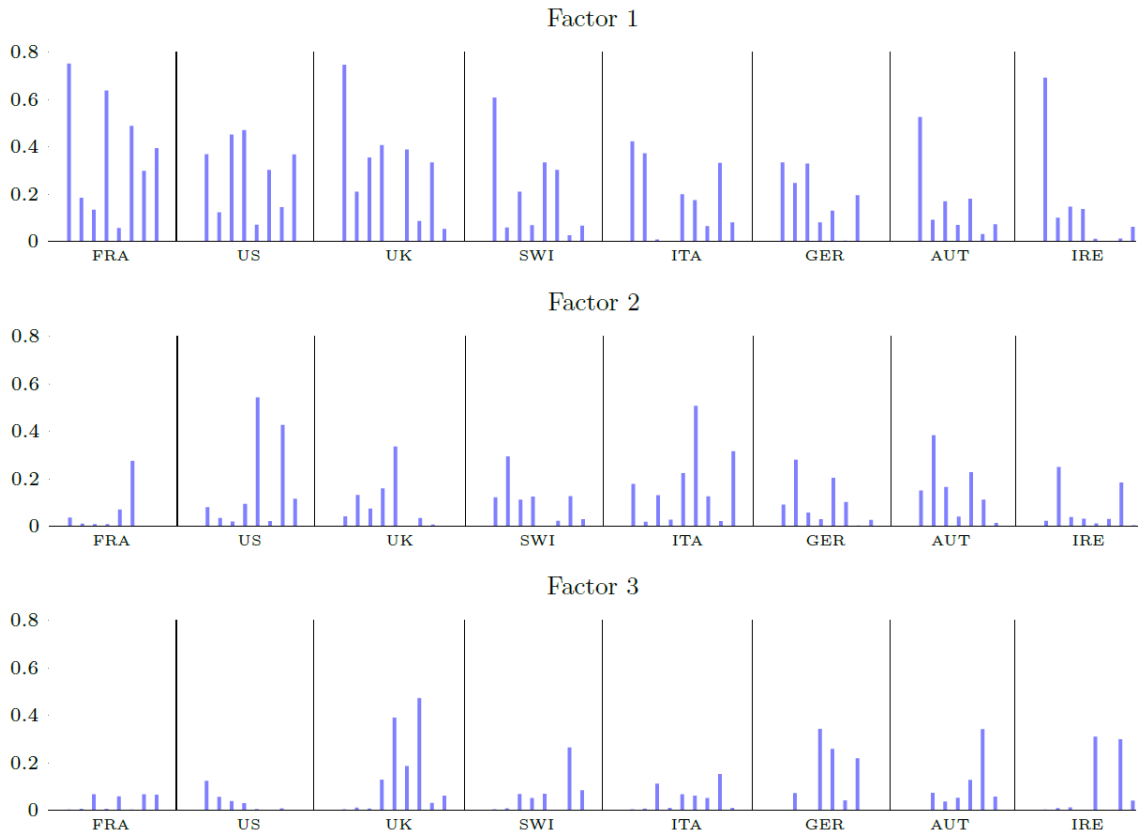
Figure 4: Estimated common factors

Figure 5: R^2 between factors and individual query (grouped by query)



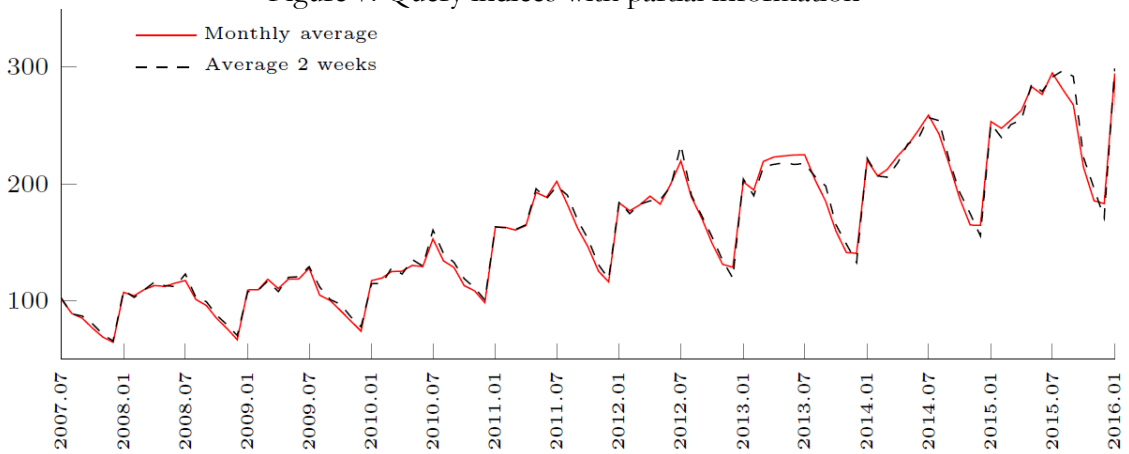
Note: DE=Pure Destination, AC=Activities at destination, PV=Vacation Package, CR=Car Rental, WE=Weather, AP=Holiday Rental, GE= Travel in General, HO=Hotels, AI=Air, ST= City & Short Trips, BR=Bus & Rail, FE=Ferries.

Figure 6: R^2 between factors and individual query (grouped by country)



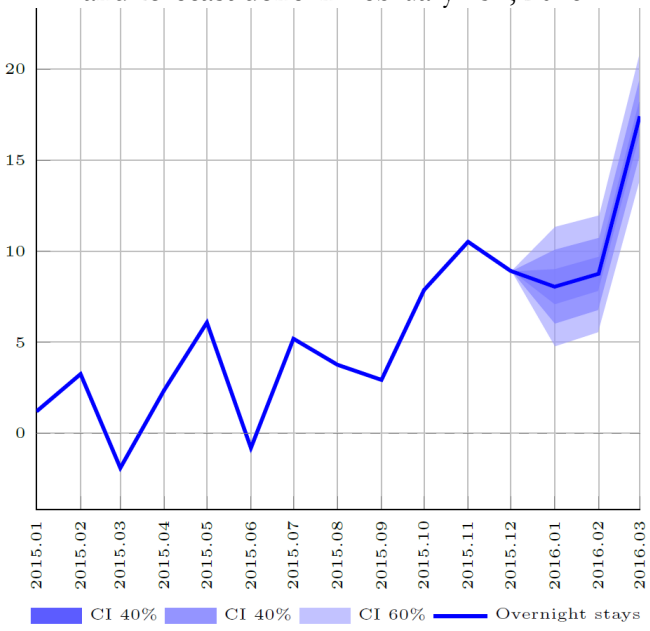
Note: FRA=France, US=United States, UK=United Kingdom, SWI=Switzerland, ITA=Italy, GER=Germany, AUT=Austria, IRE=Ireland.

Figure 7: Query indices with partial information



Note: "Monthly average" refers to averages over all the weeks of the month the weekly index is available. "Average 2 weeks" refers to averages over the first two weeks of each month.

Figure 8: Overnight stays in hotels. Backast, nowcast and forecast done in February 15th, 2016



Note: 20%, 40% y 60% refers to prediction error bands. Estimated values refers to the point estimate for backcast, nowcast and forecast in February 15th, 2016