BBVA RESEARCH

Nº 16/21
December 2016

WORKING PAPER

Forecasting travelers in Spain with Google queries

Máximo Camacho and Matías Pacce

Forecasting travelers in Spain with Google queries*

Máximo Camacho[a] and Matías Pacce[b]

December 2016

Abstract

We examine whether Google queries helps economic agents with predictions about the checking in and overnight stays of travelers in Spain in real time. 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 queries outperform models that exclude these leading indicators. In this way, we aim to contribute to the literature on the link between the Internet and the tourism market.

Keywords: Tourism, Big data analysis, Time series.

JEL classification: E32, C22, E27, Z30.

* We are thankful to Rafael Domenech, Miguel Cardoso, Camilo A. Ulloa, Alex Urcola and Miquel Moya for their comments that have greatly improved the quality of the paper. M. Camacho acknowledges support from projects ECO2012-32854-P, ECO2016-76178-P, and 19884/GERM/15 (Groups of Excellence, Fundación Séneca, and Science and Technology Agency). All remaining errors are our responsibility.
a: Universidad de Murcia and BBVA Research, mcamacho@um.es
b: BBVA Research, matias.pacce@bbva.com

1 Introduction

The Spanish economy is extremely dependent on tourism and is one of the world's top tourism destinations. In 2015, according to the World Tourism Organization, by international tourism receipts, Spain was in third position, with 56.5 billion US dollars, only behind the United States and China. By volume of international arrivals, Spain also ranked third, with 68.2 million tourists, after France and the United States. In 2014, as reported in the latest publication from the Spanish Tourism Satellite Account, the volume of tourist activity reached the amount of 10.9% 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 tourist 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 travelers. Unfortunately, in spite of these real-time monitoring requirements, data on the checking in and overnight stays of travelers, the two major measures of tourism in Spain, are published monthly with a one-month lag.

In this paper, we follow the idea that the increasingly widespread use of the Internet by travelers 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 of ongoing tourism developments. In this context, the tourist industry has been among the first to capitalize on new technology, and the number of travelers 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' behavior. 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]

Recently, several studies explored the benefits of using internet search engines to document current social trends and to predict future economic patterns. Examples are Choi and Varian

[1] In 2015, this figure rose up to 11.7% 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.

(2012), who illustrate the use of Google Trends information for making predictions about US retail sales, automotive sales, home sales and trends in travel destinations. Chamberlin (2010) finds that search terms are well correlated with disaggregated UK retail sales. McLaren (2011) shows that internet searches contain leading information for the UK housing and labor market. Vosen and Smith (2011), find that Google data contain better predictive power than that of conventional survey-based indicators of US private consumption.[3]

The ability of search query data to improve the forecasting of tourism demand has also been examined in recent years. While not claiming to be exhaustive, Pan et al. (2012) showed that including information about aggregated search volumes 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 search queries 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 queries performed from the US helps to improve forecasting accuracy on a 12-month horizon, but not for short-term forecasts. In a very specific application for the Spanish economy, Artola and Galan (2012) used searches made from the United Kingdom for the term "Spain holiday" to show some short-term improvements in forecasting British tourist inflows to Spain, although the gains depended crucially on which ARIMA model was taken as a benchmark.

We contribute to this literature in several ways. In collaboration with Google, we develop a novel web-based data set that collects information from several query indices. These provide reports on the real-time evolution of the volume of search queries related to various tourist industries in the online travel market and on the use of the Internet and e-commerce for travel. We consider that these indices constitute a reasonable source of potential indicators of what travelers are doing and what they are planning to do. This is because; largely they cover the use of the Internet as a research platform and a tourism data source. This data set of query indices

[3] For good overviews and some extensions of the related literature, the reader is referred to recent surveys by Askitas and Zimmermann (2015) and Lenaerts et al. (2016).

is available at country level for 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 queries 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).

The advantage of using the queries indices to forecast tourist data in real-time is two-fold. First, queries can use updated information up to the day before the forecast computation, which could potentially be highly valuable in this context due to the lags in the publication of the official statistics. For example, in the middle of a given month t, while the latest available monthly figure for the checking in and overnight stays of travelers refers to month $t-2$, Google data are available for month $t-1$ and an advanced view of the searches in month t. Second, lags in the query indices can also be useful in the forecast process since some tourists start planning their stays some time before they travel. This involves, among many other things, booking an airline ticket, hotel room, rental car or package tour online, to locating and compiling information on the places to visit or to stay.

The application designed in this paper requires real-time processing of high-volume data streams, which pushes the limits of traditional data processing time series models. To deal with a total of 65 series of queries from 8 different countries in real time, 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 queries 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 queries 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 travelers in real time.

Our results suggest that the model using query indices 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 find forecasting improvements from using the queries indices to forecast tourist indicators in real time.

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 queries. Section 3 analyzes the estimated factors and examines the empirical performance of Google queries 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 these 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 travelers, the target series to forecast. Let X_t be an N-dimensional vector of queries.[4] Assume that the queries 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} the forecast error.[5] The term HW is a dummy variable that takes on the value one if month t

[4]As usual, $t = 1, ..., T$, is the number of time series observations.

[5]For notation simplicity, the dependence of the parameters on h is suppressed

refers to the Holy Week.[6] Once the model is estimated, the forecast is then performed as

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

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

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

as a function of hypothetical values for factors, $\widetilde{F} = (\widetilde{F_1} \ldots \widetilde{F_T})$, and factor loadings, $\widetilde{\Lambda}$. When $N > T$, minimizing (4) is equivalent to maximize $tr\left[\widetilde{F}'(XX')\widetilde{F}\right]$ subject to $\widetilde{F}'\widetilde{F}/N = I_r$ where $tr(\cdot)$ 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 travelers use web search engines to planning their trips and stays. The searches performed using Google have been used to construct weekly indices that collect the relevant information on the trips and stays that travelers take and intend to take. The queries indices used to obtain all the results of this paper come from weekly reports on the volume of queries related to various tourism industries that cover the period from the first week of July 2007 to the second week of January 2016. Classified by country of origin, they show how often several traveling related topics have been searched for on Google over time. The countries where the queries 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 indexes rely on searches on travel facilities (air, ferries, bus and rail), accommodation (hotel, holiday rental and camping), vacation packages, general travel and destination

[6] The dummy variable attempts to remove remaining seasonal effects that occur on Holy Weeks.

(city and short trips, activities, weather, rent a car). All query indices start with the total query volume related to each specific term in a specific country, divided by the total number of queries in that country 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 checking in and overnight stays of travelers, which are published on a monthly basis, we compute the monthly averages of the weekly query indeces.

To examine the dynamics of travel related Google searches, 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 checking in and overnight stays of non-resident travelers in hotels. Regarding tourist indicators, the INE (National Statistics Institute) states that checked-in travelers include all people who stay one or more consecutive nights in the same collective tourist accommodation. Overnight stays include every night that a traveler spent in these establishments. In the paper, we focus on the versions of tourist indicators that only account for non-residents.[7]

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.[8] Therefore, to be compared with the annual growth rate transformation employed in the case of 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 query indices 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.

[7]In the empirical application, we examine the potential improvements of the queries to forecast tourism indicators by type of accommodation: hotels, rental apartments and the sum of the two, plus camping.

[8]It is hardly possible to compute accurate seasonal factors by employing standard techniques of seasonal adjustment since query indices are available only since 2007.

[Figure 2 about here]

Figure 2 also reveals that query indices 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 checking into hotels of travelers are up to 0.61 and 0.58, respectively. A good example of this closed relationship among queries and tourist indicators can be depicted in Figure 3, which shows how the annual growth rate of each of the travel related queries 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 queries 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 year-on-year growth rates of query indices 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 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 queries 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 queries are mainly related to "Hotels" and "Bus and Rail", while "Pure destination" continues to be relevant.[9] Regarding the third factor, queries 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 queries.

[9] "Bus and Rail" is only available for Italy.

In Figure 6, the queries 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

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

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

$$\widehat{y}_{t-1} = \widehat{\mu} + \widehat{\alpha}_1 y_{t-2} + \widehat{\alpha}_2 y_{t-3} + \sum_{i=1}^{r} \sum_{j=0}^{m} \widehat{\beta}_i^j F_{t-j-1}^i + \widehat{\gamma} HW_{t-1}. \tag{6}$$

To compute the nowcast, the model

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

is estimated with data up to $t-1$.[10] Then, the nowcast is computed as

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

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

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

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

with the extended data set up to t.[11] The forecast of $t+1$ is

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

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).[12] 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 queries, a univariate autoregressive model, which is also estimated in pseudo real-time producing iterative forecasts is included as a benchmark model.[13]

To facilitate comparisons, Table 1 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

[10] Notice that the model uses the backcast $\widehat{y}_{t-1}$ for time $t-1$.

[11] Notice that the model uses the backcast $\widehat{y}_{t-1}$ for time $t-1$ and the nowcast $\widehat{y}_t$ for time t.

[12] At month t, the nowcast at t and forecast at $t+1$ can only use queries of the first two weeks of this month.

[13] This benchmark model includes the Holy Week dummy aiming to distinguish the differences emerging when Google queries' information is incorporated in the model.

forecasts results displayed in the table is that it is beneficial to use the query indices information in forecasting the Spanish tourism. However, the relative gains from the model that uses the query 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 traveler 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 travelers 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 1 about here]

This result confirms the leading forecasting ability of tourism indicators by query indices, which is clearly achieved when the early available search data is accounted for by the model. The promptly published information of query indices 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 queries 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.[14]

[Figure 8 about here]

4 Conclusions

The Internet has radically changed the manner in which tourists and travelers obtain travel-related information. The evidence presented in this paper, based on the performance of tourism query indices provided by Google over a real-time exercise, has provided very promising support

[14]In 2015, Easter occurred entirely during April, while in 2016 it took place in March.

for using search information to predict checked-in and overnight stays of travelers in Spain. As in any big data setup, the first step is to capture the big amount of information provided by the query indices. For this purpose, we assume that the queries admit a factor model decomposition, in which each query 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 stay travelers.

The main conclusion that follows from the paper is that using query indices can be useful as the basis for computing timely short-term forecasts of tourism developments in Spain. From the vantage point of an early warning system, the results are encouraging in that the signals from search data occur sufficiently early to allow for preemptive actions. Therefore, the analysis can be viewed in the line of some recent studies that explored the benefits of using an internet search engine and social media activity to document current social trends and predict future economic patterns.

Despite these promising results, it is important to recognize that the conclusions regarding the performance of query indices examined in this paper are necessarily tentative, mainly because of the limited number of observations that are available for the query indexes. As more data become available, future work on the help of query indices 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.

References

[1] 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.

[2] Askitas, N., and Zimmermann, K. 2015. The Internet as a data source for advancement in social sciences. *International Journal of Manpower* 36: 2-12.

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

[4] 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

[5] Chamberlin, G. 2010. Googling the present. *Economic and Labour Market Review* 4: 59-95.

[6] Choi, H., and Varian, H. 2012. Predicting present with google trends. *Economic Record* 88: 2-9.

[7] 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.

[8] 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.

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

[10] Lenaerts, K., Beblavý, M., Fabo, B. 2016. Prospects for utilisation of non-vacancy Internet data in labour market analysis-an overview. *IZA Journal of Labor Economic* 5: 1-18.

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

[12] McLaren, N. 2011. Using internet search data as economic indicators. *Bank of England Quarterly Bulletin* 51: 134-140.

[13] 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.

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

[15] 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.

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

[17] 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.

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

[19] Vosen, S., and Schmidt, T. 2011. Forecasting private consumption: Survey-based indicators vs. Google Trends. *Journal of Forecasting* 30: 565-578.

[20] 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: Predictive accuracy: Enlarged AR (values relative to an AR model)

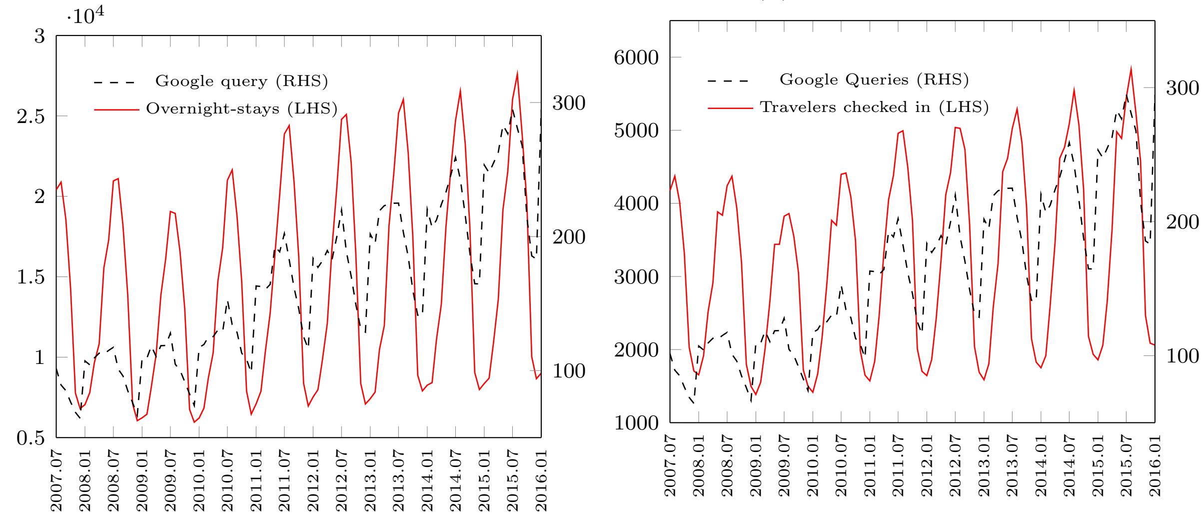
		Non-residents overnight-stays								
		Total			Hotels			Rental Apartments		
k	m	$t-1$	t	$t+1$	$t-1$	t	$t+1$	$t-1$	t	$t+1$
AR(2)		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0	1.00	.95	.98	.99	.97	.98	1.01	.95	.96
	1	.98	.96	.98	.99	.97	.98	1.01	1.00	.97
1	2	.99	.97	1.05	1.00	.97	1.03	1.01	1.01	1.01
	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	.88	.92	.95	.87	.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-residents traveled 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 $t+1$ refers to the backasting, 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) Travelers checked in



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

Figure 2: Comparison of yearly growth rates

(a) Overnight Stays

(b) Travelers checked in

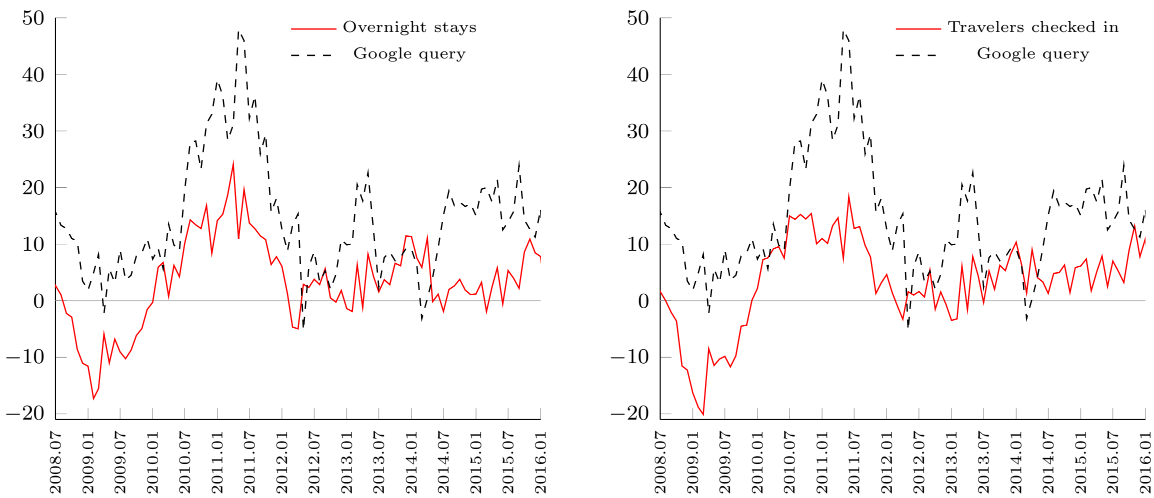
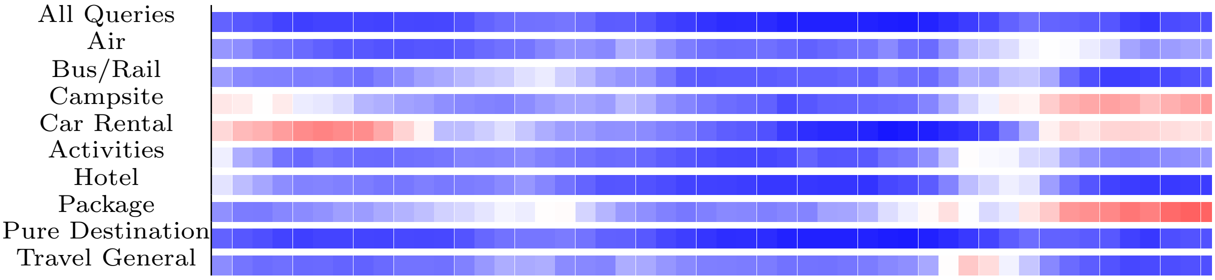
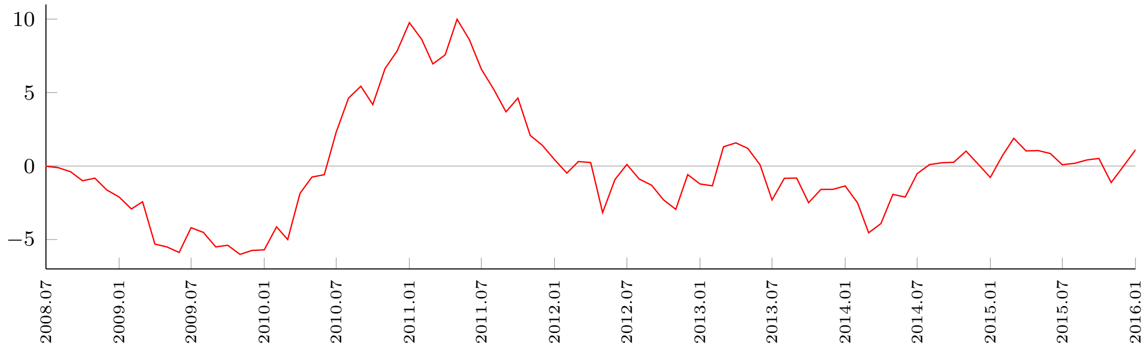


Figure 3: Correlations between Italian overnights stays (Spanish hotels) and travel related Google query

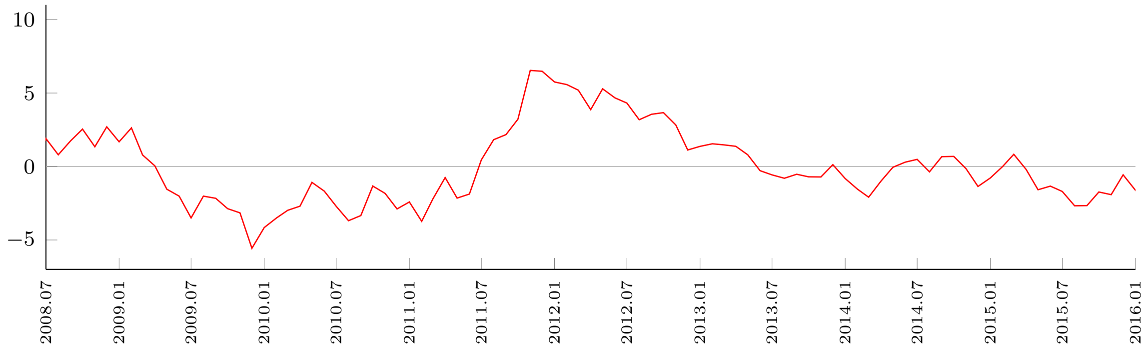


Note: Two years rolling windows correlations. A deeper blue indicates proximity to 1 while a deeper red to -1. Windows from 2010.07 to 2016.01

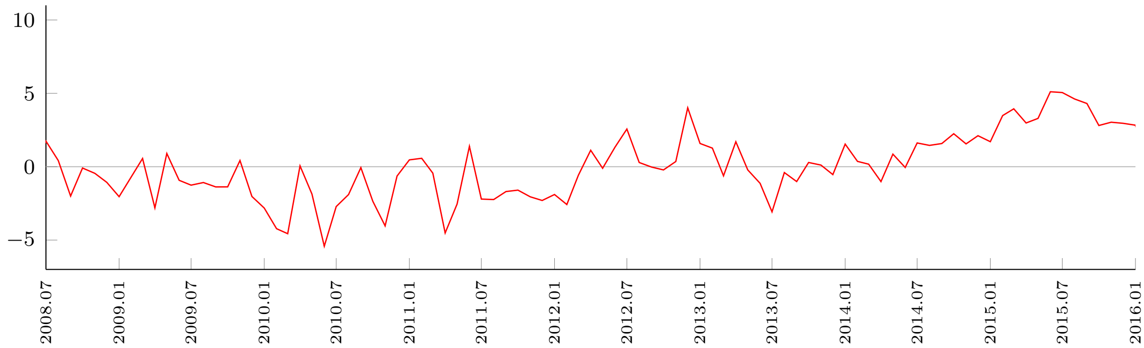
Figure 4: Estimated common factors



(a) Factor 1



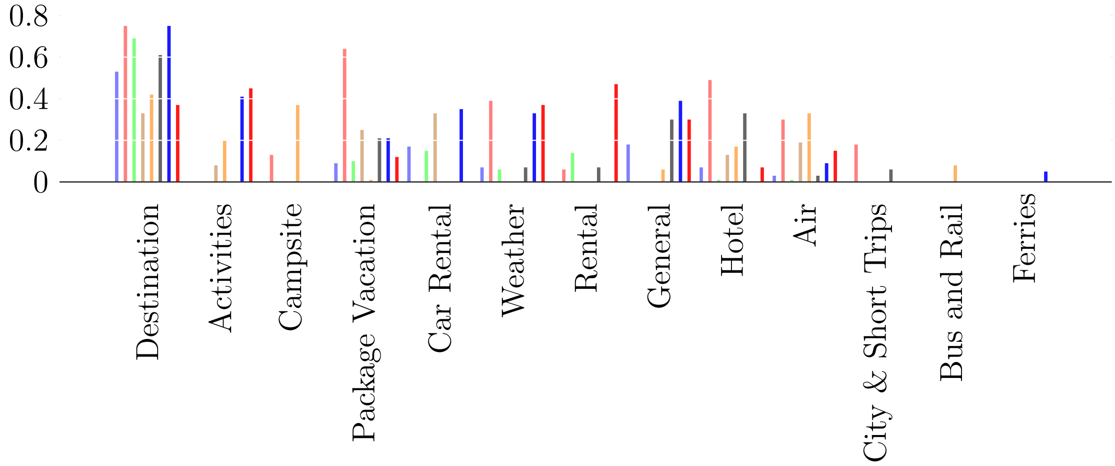
(b) Factor 2



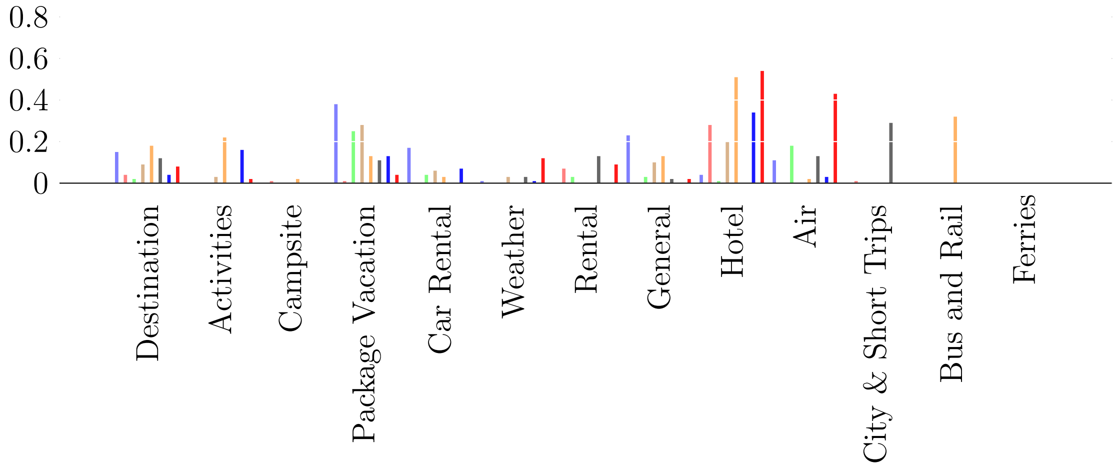
(c) Factor 3

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

(a) Factor 1



(b) Factor 2



(c) Factor 3

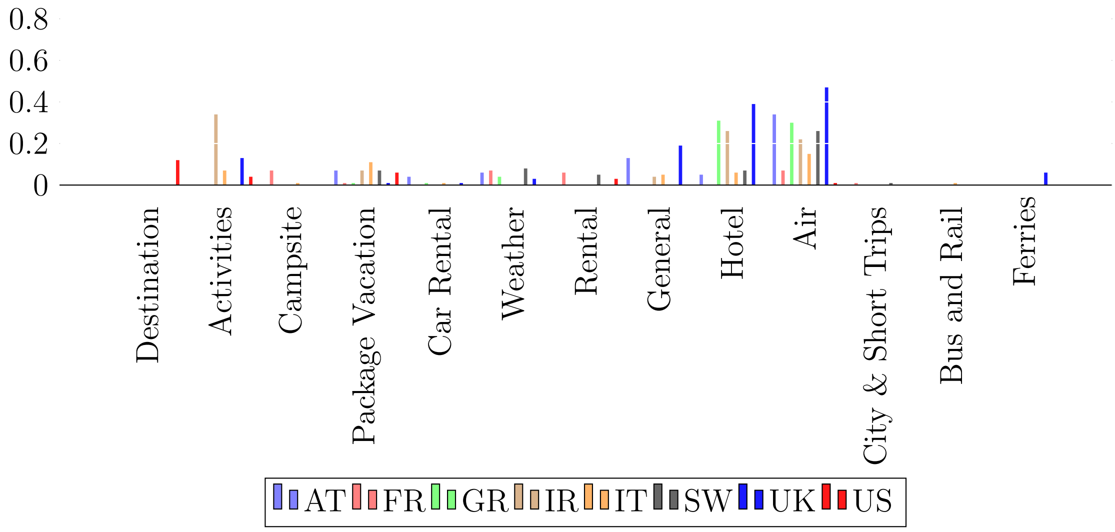


Figure 6: R^2 beetween factors and individual query (grouped by country)

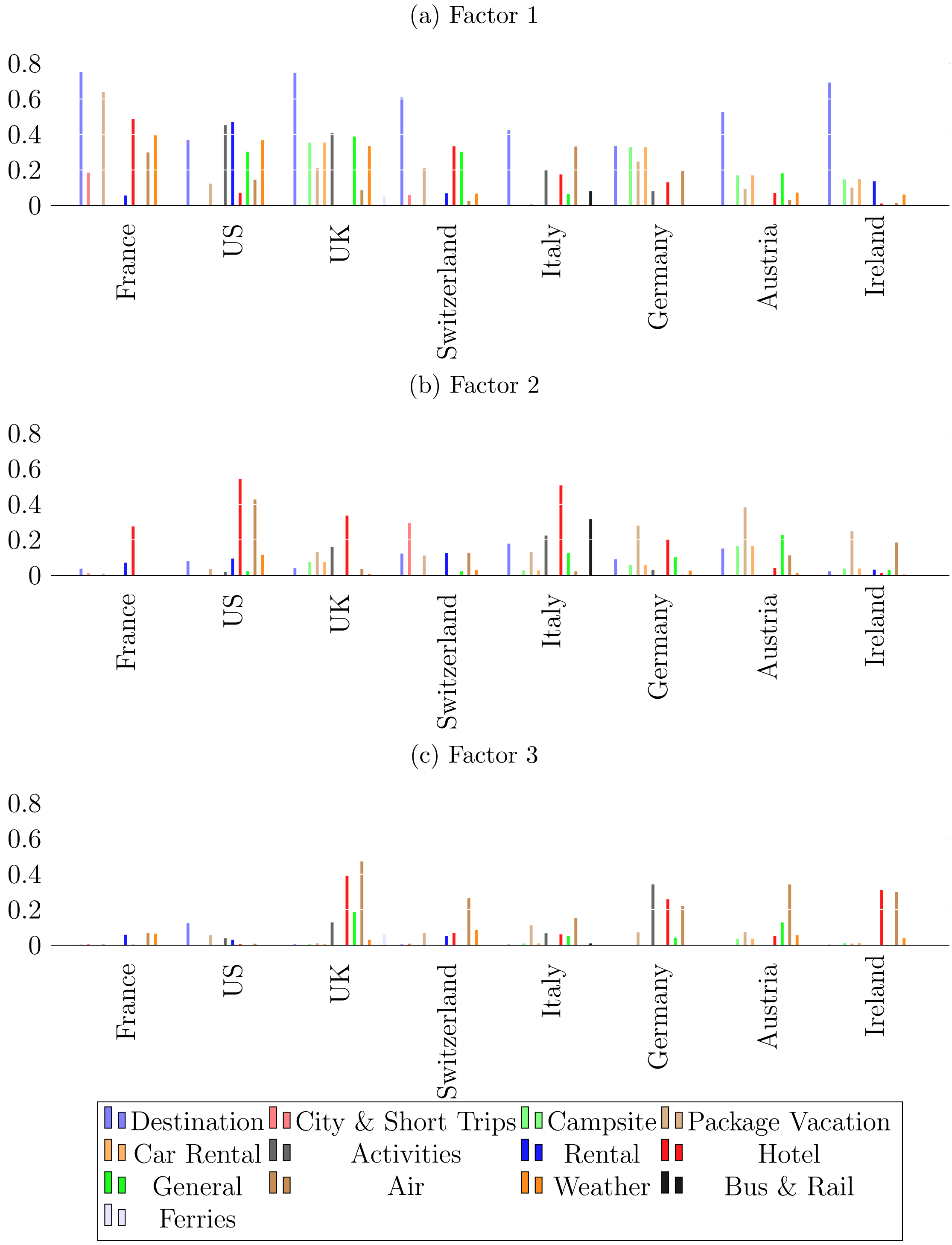
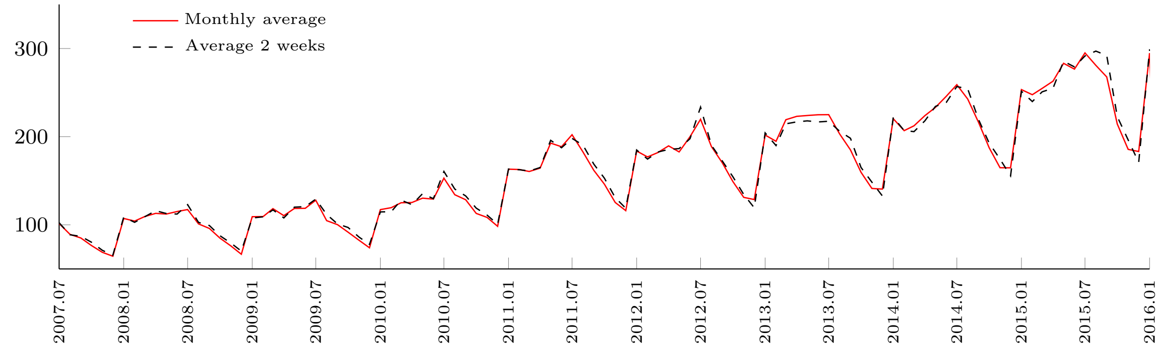
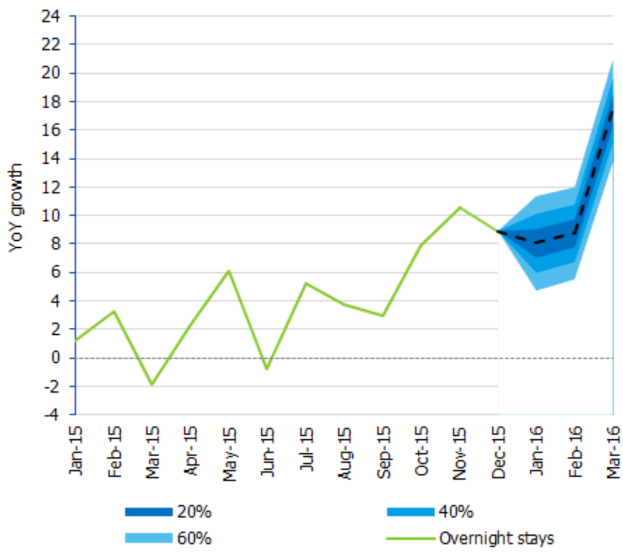


Figure 7: Query indices with partial information



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 15, 2016



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

Working Papers

2016

16/21 **Máximo Camacho and Matías Pacce:** Forecasting travelers in Spain with Google queries.

16/20 **Javier Sebastian Cermeño:** Blockchain in financial services: Regulatory landscape and future challenges for its commercial application.

16/19 **Javier Alonso, Alfonso Arellano, David Tuesta:** Factors that impact on pension fund investments in infrastructure under the current global financial regulation.

16/18 **Ángel de la Fuente:** La financiación regional en Alemania y en España: una perspectiva comparada.

16/17 **R. Doménech, J.R. García and C. Ulloa:** The Effects of Wage Flexibility on Activity and Employment in the Spanish Economy.

16/16 **Ángel de la Fuente:** La evolución de la financiación de las comunidades autónomas de régimen común, 2002-2014.

16/15 **Ángel de la Fuente:** La liquidación de 2014 del sistema de financiación de las comunidades autónomas de régimen común: Adenda.

16/14 **Alicia García-Herrero, Eric Girardin and Hermann González:** Analyzing the impact of monetary policy on financial markets in Chile.

16/13 **Ángel de la Fuente:** La liquidación de 2014 del sistema de financiación de las comunidades autónomas de régimen común.

16/12 **Kan Chen, Mario Crucini:** Trends and Cycles in Small Open Economies: Making The Case For A General Equilibrium Approach.

16/11 **José Félix Izquierdo de la Cruz:** Determinantes de los tipos de interés de las carteras de crédito en la Eurozona.

16/10 **Alfonso Ugarte Ruiz:** Long run and short run components in explanatory variables and differences in Panel Data estimators.

16/09 **Carlos Casanova, Alicia García-Herrero:** Africa's rising commodity export dependency on China.

16/08 **Ángel de la Fuente:** Las finanzas autonómicas en 2015 y entre 2003 y 2015.

16/07 **Ángel de la Fuente:** Series largas de algunos agregados demográficos regionales, 1950-2015.

16/06 **Ángel de la Fuente:** Series enlazadas de Contabilidad Regional para España, 1980-2014.

16/05 **Rafael Doménech, Juan Ramón García, Camilo Ulloa:** Los efectos de la flexibilidad salarial sobre el crecimiento y el empleo.

16/04 **Angel de la Fuente, Michael Thöne, Christian Kastrop:** Regional Financing in Germany and Spain: Comparative Reform Perspectives.

16/03 **Antonio Cortina, Santiago Fernández de Lis:** El modelo de negocio de los bancos españoles en América Latina.

16/02 **Javier Andrés, Ángel de la Fuente, Rafael Doménech:** Notas para una política fiscal en la salida de la crisis.

16/01 **Ángel de la Fuente:** Series enlazadas de PIB y otros agregados de Contabilidad Nacional para España, 1955-2014.

2015

15/33 **Shushanik Papanyan:** Digitization and Productivity: Where is the Growth? Measuring Cycles of Technological Progress.

15/32 **Alfonso Arellano, Noelia Cámara, David Tuesta:** Explaining the Gender Gap in Financial Literacy: the Role of Non-Cognitive Skills.

15/31 **Ángel de la Fuente:** Series enlazadas de Contabilidad Regional para España, 1980-2014. Parte II: Empleo asalariado, rentas del trabajo y salarios medios.

15/30 **Jingnan Cai, Alicia García-Herrero, Le Xia:** Regulatory arbitrage and window-dressing in the shadow banking activities: evidence from China's wealth management products.

15/29 **Javier Alonso, Alfonso Arellano:** Heterogeneity and diffusion in the digital economy: spain's case.

15/28 **Javier Alonso, Alfonso Arellano:** Heterogeneidad y difusión de la economía digital: el caso español.

15/27 **Angel de la Fuente**: Series enlazadas de Contabilidad Regional para España, 1980-2014.

15/26 **Carlos Casanova, le Xia and Romina Ferreira:** Measuring Latin America's export dependency on China.

15/25 **Nathaniel Karp, Boyd W. Nash-Stacey:** Embracing the Financially Excluded in the U.S.: A Multi-Dimensional Approach to Identifying Financial Inclusion Across MSAs and Between Cohorts.

15/24 **Alicia Garcia-Herrero, K.C. Fung:** Determinants of Trade in Parts and Components: An Empirical Analysis.

15/23 **Mariano Bosch, Angel Melguizo, Enith Ximena Peña, David Tuesta:** El ahorro en condiciones formales e informales.

15/22 **Antonio Villar:** Crisis, households' expenditure and family structure: The Palma ratio of the Spanish economy (2007-2014).

15/21 **Andrés Hernández, Bernardo Magnani, Cecilia Posadas, Jorge Redondo, Gonzalo Robles, Juan M. Ruiz y Enestor Dos Santos:** ¿Cuáles son los sectores con mayor potencial para aprovechar la Alianza del Pacífico?

15/20 **Gonzalo de Cadenas, Alicia Garcia-Herrero, Alvaro Ortiz and Tomasa Rodrigo:** An Empirical Assessment of Social Unrest Dynamics and State Response in Eurasian Countries. / *Published in Eurasian Journal of Social Sciences, 3(3), 2015, 1-29.*

15/19 **Mariano Bosch, Angel Melguizo, Enith Ximena Peña and David Tuesta:** Savings under formal and informal conditions.

15/18 **Alicia Garcia-Herrero, K.C. Fung, Jesus Seade:** Beyond Minerals: China-Latin American Trans-Pacific Supply Chain.

15/17 **Alicia Garcia-Herrero, Le Xia, Carlos Casanova:** Chinese outbound foreign direct investment: How much goes where after round-tripping and offshoring?

15/16 **Diego José Torres Torres:** Evaluando la capacidad predictiva del MIDAS para la Eurozona, Alemania, Francia, Italia y Portugal.

15/15 **Alicia Garcia-Herrero, Eric Girardin, Arnoldo Lopez-Marmolejo:** Mexico's monetary policy communication and money markets.

15/14 **Saidé Salazar, Carlos Serrano, Alma Martínez, Arnulfo Rodríguez**: Evaluation of the effects of the Free Trade Agreement between the European Union and Mexico (EU-MX FTA) on bilateral trade and investment.

15/13 **Saidé Salazar, Carlos Serrano, Alma Martínez, Arnulfo Rodríguez:** Evaluación de los efectos del Tratado de Libre Comercio entre la Unión Europea y México (TLCUEM) en el comercio bilateral y la inversión.

15/12 **Alicia Garcia-Herrero, Eric Girardin and Enestor Dos Santos:** Follow what I do, and also what I say: Monetary policy impact on Brazil's financial markets.

15/11 **Noelia Cámara, David Tuesta, Pablo Urbiola:** Extendiendo el acceso al sistema financiero formal: el modelo de negocio de los corresponsales bancarios.

15/10 **Noelia Cámara, David Tuesta, Pablo Urbiola:** Extending access to the formal financial system: the banking correspondent business model.

15/09 **Santiago Fernández de Lis, José Félix Izquierdo de la Cruz y Ana Rubio González:** Determinantes del tipo de interés del crédito a empresas en la Eurozona.

15/08 **Pau Rabanal and Juan F. Rubio-Ramírez:** Can International Macroeconomic Models Explain Low-Frequency Movements of Real Exchange Rates?.

15/07 **Ándel de la Fuente y Rafael Doménech:** El nivel educativo de la población en España y sus regiones: 1960-2011.

15/06 **Máximo Camacho and Jaime Martínez-Martín:** Monitoring the world business cycle. / *Published in Economic Modelling 51 (2015) 617–625.*

15/05 **Alicia García-Herrero and David Martínez Turégano:** Financial inclusion, rather than size, is the key to tackling income inequality.

15/04 **David Tuesta, Gloria Sorensen, Adriana Haring y Noelia Cámara:** Inclusión financiera y sus determinantes: el caso argentino.

15/03 **David Tuesta, Gloria Sorensen, Adriana Haring y Noelia Cámara:** Financial inclusion and its determinants: the case of Argentina.

15/02 **Álvaro Ortiz Vidal-Abarca and Alfonso Ugarte Ruiz:** Introducing a New Early Warning System Indicator (EWSI) of banking crises.

15/01 **Alfonso Ugarte Ruiz:** Understanding the dichotomy of financial development: credit deepening versus credit excess.

Click here to Access the Working Paper published

Spanish

and English

Contact details:
BBVA Research
Azul Street, 4
La Vela Building - 4 and 5 floor
28050 Madrid (Spain)
Tel.: +34 91 374 60 00 and +34 91 537 70 00
Fax: +34 91 374 30 25
bbvaresearch@bbva.com
www.bbvaresearch.com