Forecasting tourist arrivals with online data: An application to the Valencian Community

Keywords: Forecasting, tourism, online data, Wikipedia, Google Trends

1. INTRODUCTION

Tourism trips are increasingly planned and organised online. This generates some digital traces correlated with the tourist movement, and thus potentially useful to improve the accuracy and timeliness of forecasts. This hypothesis is based on the fact that before booking a trip, travellers and tourists look for information about the destination on the Internet. Therefore, we expect a significant relation between the online popularity of a destination and the real, physical visits it receives. It is therefore possible to take advantage of these digital traces with the aim to improve tourism forecasting. Previous studies have shown the capacity of online data to improve the forecasts of tourism-related variables in different regions [1-2].

Assessing the potential and applicability of online behaviour data sources to support the production of official statistics is a new line in which statistical offices are working. Eurostat is performing some pilot studies on different big data sources applied to different fields, including tourism [3]. The study we present in this paper is developed as a pilot to assess to which extent online sources, which are providers of big data, can help to predict real tourist movements.

The aims of this on-going study are two-fold: Firstly, to check if online data from Google Trends and Wikipedia pageviews can help to improve the accuracy of forecasting models for tourist arrivals. Secondly, to compare the two online sources in order to assess not only which one performs better, but also to check if they are complements or, on the contrary, they are substitutes. This is fundamental for official statistics offices to decide which online sources are worth pursuing further investigation and introduction in official statistics production, e.g. to develop flash estimates and forecasting models based on these new sources of massive, timely and granular data.

2. METHODS

Provided that tourism has turned into one of the pillars of the Spanish economy, and more concretely, of the Valencian economy, where it represents about 13% of GDP, the performed analyses are focused on the Valencian Community (VC), a region on the east coast of Spain.

The period under study covers almost 6 years, from December 2011 to November 2017. It has been selected because it is the period for which Wikipedia pageviews data were available for being retrieved through a tool developed at Eurostat. Therefore, data about Google Trends and about official tourist arrivals to the VC also correspond to this period.

The research starts by the selection and retrieval of Wikipedia pageviews and Google Trends keywords hits. After that, time-series models including such online data are fitted in order to improve the predicting accuracy of tourist arrivals compared to a baseline time-series model that only includes past arrivals. All the analyses have been performed in R.

2.1. Online sources: Google Trends and Wikipedia pageviews

Google Trends (GT) is a data product made available by Google which consists of an index of the popularity that a particular term has in the search engine. The Google search engine is the entry door for many Internet activities and therefore a main source of data on online behaviour of people. To include data from this source, the first step was the selection of search terms for which time-series would be used. For that, top destinations and attractions in the VC were identified and terms based on their respective names, as well as more general tourism related terms (e.g. “beach Valencia”), were selected. Ninety GT search terms were kept after the removal of terms for which there was not enough search activity, the translation to EN, FR and DE and addition of Freebase ID codes which refer to a location or attraction. GT data were finally obtained taking as origin of the search any place in the world.

Wikipedia pageviews consist of the number of times each article is consulted in Wikipedia. The Wikimedia Foundation releases hourly pageviews data for all the wiki projects (e.g. Wikipedia, Wikivoyage, Wikibooks…) and all the existing language versions. The Wikipedia is widely used, as illustrated by the fact that 45% of individuals aged 16 to 74 years old living in the EU have consulted Wikipedia and other wikis at least once in 3 months during 2015 [4]. This makes it a relevant source of people's online behaviour which, in the context of our study, may reveal some intention to make tourism trips.

Since the data is available at the level of the Wikipedia article, a set of articles of interest was selected a priori. The starting point was the manual identification of a limited but diverse set of touristic points of interest (TPOI) in the Valencian Community (top destinations, attractions, restaurants, museums…) and corresponding articles in the Spanish version of Wikipedia. In order to enlarge the set of TPOI initially identified, the article categorisation feature of Wikipedia was used. Categories to which the initial articles belonged were selected and equivalent categories in other language versions of the Wikipedia were then selected manually (EN, DE, FR, IT, NL, PT, SV, FI, DA, GA, PL, SL, CS, ET, LT, RU). All articles belonging to the categories selected were then gathered. Finally, language versions in the 24 official languages of the EU plus Russian were obtained for all the articles gathered in the previous steps. Monthly pageviews time-series were then aggregated for all articles which refereed to the same point of interest, as registered in Wikidata. The final number of TPOI and time-series available for the forecasting models was 730. These were clustered into 20 clusters for reducing the dimensionality of the data, which is a necessary step for fitting time-series models.

2.2. Official data source

Official data about foreign tourist arrivals to the VC were obtained from the survey Statistics of Tourism Movements in the Frontiers (FRONTUR), which was developed by the General Sub-Directorate of Knowledge and Tourism studies of Spain (http://estadisticas.tourspain.es) until September 2015, and by the Spanish National Statistical Office (https://www.ine.es) since October 2015.

3. RESULTS

Firstly, the time series of tourist arrivals to the VC were plotted (see Figure 1) and decomposed in their different components. A strong seasonality and a positive trend during all the period were identified. This is an important challenge given that most of the information about the series is contained in these two components. However, some information was still in the remainder, which is the part we aim to explain with online data.

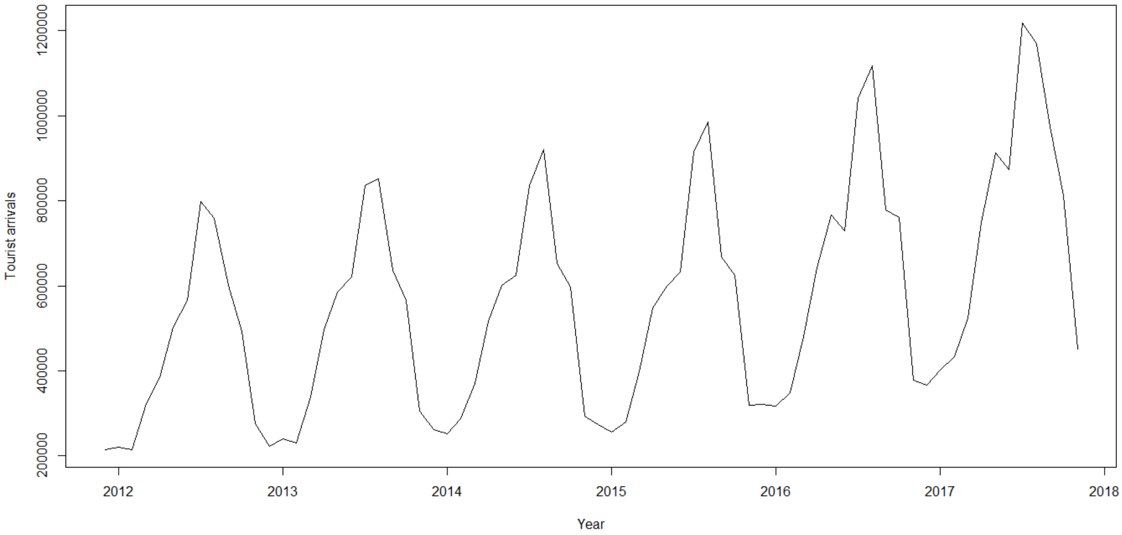

Figure 1. Monthly tourist arrivals to the Valencian Community

The next step was to find the best baseline model to forecast tourist arrivals only based on past information in the series (a classic time-series approach). This consisted in a Seasonal Autoregressive Integrated Moving Average (SARIMA) model with the parameters specified in Table 1. This table shows results regarding some measures of goodness-of-fit (the small-sample-size corrected version of Akaike Information Criterion, AICc), and of out-of-sample accuracy (Mean Absolute Error, MAE, and Root Mean Squared Error, RMSE, computed with a rolling window, one-step ahead prediction of the last year of the sample).

Table 1. Results for the forecasting models

Forecasting models	AICc (all data)	Out-of-sample accuracy		% improvement RMSE
		MAE	RMSE	
Baseline: SARIMA (0,1,1)x(0,1,1)12	1,419.83	49,614.8	59,207.69	-
Model 1: SARIMAX (selection of GT terms)	1,402.37	42,709.41	48,835.85	17.51%
Model 2: SARIMAX (selection of Wikipedia pageviews clusters)	1,419.88	47,512.31	56,467.25	4.63%
Model 3: SARIMAX (GT and Wikipedia)	1,405.67	41,045.82	47,890.48	19.11%

Once the baseline was established, it was enriched by adding different selections of online data, as Models 1, 2 and 3 reflect in Table 1. These are three SARIMA models with additional explanatory variables (SARIMAX), in which the explanatory variables are a selection of GT terms, in the first model; a selection of clusters of Wikipedia pageviews, in the second model; and a combination of both online data, in the third model. All models outperform the baseline, with an improvement in the RMSE of 17.51% when using GT data and of 4.63% when using Wikipedia data. These preliminary results verify the usefulness of online data to improve the forecasts of tourist arrivals, and suggest that GT data are more helpful to improve the forecast of tourist arrivals to the VC when compared to Wikipedia pageviews. However, the third model in which GT and Wikipedia data are combined is the one that brings best results, with an improvement on the forecasting accuracy above 19%, showing that both sources are to some extent, complementary. Further analyses performing different preprocessing to the data and using different models are necessary to obtain more insights and confirm these findings.

4. CONCLUSIONS

Online data both from Google Trends and Wikipedia pageviews have been shown to improve the accuracy of the one-step ahead forecasts of tourist arrivals to the Valencian Community. These preliminary results shed some light on our initial hypothesis and encourages us to continue with the research, in which we will work on two important challenges identified for these sources of data in relation to the aim of the study: first, the dimensionality of the online data retrieved is too high to directly fitting time-series models with them, which has led us to perform a manual selection of the GT terms (out from the total 90 GT time-series) and Wikipedia data (out from the 20 clusters) that in fact improved the forecasts; second, online data reflect interest in a destination at a much lower level of aggregation than the one for which we want to do forecasts, which makes it more difficult to identify the signal that reveals real visits to a place. Checking additional models using each online source separately, and combining both, as well as using alternative machine learning techniques to the classic time-series approach in order to automate the selection of online data are still work in progress. Issues such as including lags in the online data are going to be considered too.

REFERENCES

[1] Bangwayo-Skeete, P. F., & Skeete, R. W. Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. Tourism Management, 46, 454-464, (2015).

[2] C. M. Alis, A. Letchford, H. S. Moat & T Preis. Estimating tourism statistics with Wikipedia page views. In Proceedings of the ACM Web Science Conference (2015), p. 33, ACM.

[3] S. Signorelli, F. Reis and S. Biffignandi. Virtual vs. real visits: an analysis of three cities through Wikipedia page views and tourism data. In Proceedings of NTTS2017 (2017).

[4] Eurostat. Individuals using the internet for consulting wiki. Available at: https://ec.europa.eu/eurostat/tgm/table.do?tab=table&tableSelection=1&labeling=labels&footnotes=yes&layout=time,geo,cat&language=en&pcode=tin00128&plugin=1 (accessed 10th October, 2018).