Annals of Tourism Research 61 (2016) 199–212

Contents lists available at ScienceDirect

Annals of Tourism Research

journal homepage: www.elsevier.com/locate/atoures

Forecasting city arrivals with Google Analytics

Ulrich Gunter *, Irem Önder

MODUL University Vienna, Department of Tourism and Service Management, Am Kahlenberg 1, A-1190 Vienna, Austria

ARTICLE INFO

Article history:
Received 29 December 2015
Revised 5 October 2016
Accepted 10 October 2016

Coordinating editor: Haiyan Song

Keywords:
Bayesian analysis
Big data
City tourism
Factor analysis
Forecast combination
Vector autoregression

ABSTRACT

The ability of 10 Google Analytics website traffic indicators from the Viennese DMO website to predict actual tourist arrivals to Vienna is investigated within the VAR model class. To prevent overparameterization, big data shrinkage methods are applied: Bayesian estimation of the VAR, reduction to a factor-augmented VAR, and application of Bayesian estimation to the FAVAR, the novel Bayesian FAVAR. Forecast accuracy results show that for shorter horizons (h = 1, 2 months ahead) a univariate benchmark performs best, while for longer horizons (h = 3, 6, 12) forecast combination methods that include the predictive information of Google Analytics perform best, notably combined forecasts based on Bates–Granger weights, on forecast encompassing tests, and on a novel fusion of these two.

Introduction

There has been dramatic growth in Internet usage, especially for travel information search purposes. Each time an individual uses a website he or she leaves traces on that site. These traces can be collected and used for different purposes such as tracking user behavior, recommending products to the customer on their next visit to the website and optimizing website usability. Google Analytics accounts make it possible for businesses to collect these traces from their websites. Although many destination management organizations (DMOs) are collecting these types of information from their websites, this information is usually not used for making managerial decisions, but merely by IT departments to enhance website usability. Often, the interpretation of website traffic indicators such as Google Analytics is not clear to DMO managers: such as what it means to the DMO to have one million website visitors. However, website traffic data can be very informative: showing, for instance, from which countries users originate. This information can then be combined to see if there is a correlation between the country of origin of website visitors and the country of origin of the actual arrivals to the destination.

A primary objective of this article is therefore to show how website traffic data can be used by DMO managers to forecast tourism demand. Tourism demand analysis and forecasting is one of the core areas of tourism economic research since tourism demand ultimately is the basis of all business decisions in tourism (Song, Witt, & Li, 2009), which includes the business decisions of DMOs. These business decisions require accurate tourism demand forecasts in order to reduce the many risks that may occur during the decision making process (Frechtling, 2001).

The theory explaining the search behavior of the visitors to the DMOs' websites is called *information foraging theory*, which is derived from behavioral ecology and which is similar to food foraging theories in anthropology. It was developed

* Corresponding author.
E-mail addresses: ulrich.gunter@modul.ac.at (U. Gunter), irem.onder@modul.ac.at (I. Önder).

http://dx.doi.org/10.1016/j.annals.2016.10.007
0160-7383/

by Pirolli and Card (1999) and states that, when possible, the information search process evolves toward maximization of relevant information per unit cost. In the same sense, the optimal information forager would be the one who maximizes the rate of relevant information gained according to the task environment, the profitabilities of different sources, and the cost of finding and accessing them (Pirolli & Card, 1999). Thus, foraging information from a DMO's website comes at a comparably low unit cost. In a tourism demand forecasting context, the information foraged from the DMOs' websites and expressed by various Google Analytics website traffic indicators can therefore constitute an important predictor for actual arrivals to a destination if the foraged information proves relevant. Relevance of this foraged information, in turn, is given if considering it in appropriate forecast models results in comparatively more accurate forecasts.

Information search on search engines such as Google, Yahoo, or Baidu is one of the first steps in planning a vacation, with travel-related searches comprised of city, country, and region names representing approximately 60% of all travel related searches (Jansen, Ciamacca, & Spink, 2008). From the search engine results, the DMOs' websites can typically be found on the first page of search results and thus potential travelers are immediately directed to the DMOs' websites. Google Analytics in this case would provide researchers with quantitative measures for potential travelers to the destination that can be used for in-depth analysis.

Google Analytics is a free service by Google to the owner of a website that provides website traffic data including all user behavior on the website, such as the number of visitors, the time spent on the website, the number of actions taken on the site, where the users come from, etc. Google Analytics sends the website traffic data to the analytics server by means of a snippet (tracking code) that is included on the website and activated when a visitor views a page on somebody's website (Boswell, 2011). Overall, Google Analytics has a nearly 83% share of the website tracking tools market (W3Techs, 2015), yet this service is generally used only for website quality control and to enhance website user experience.

Some examples of previous studies that have used Google Analytics include measuring the website performance of a cultural tourism website (Plaza, 2011), measuring food composition website visitor statistics (Pakkala, Presser, & Christensen, 2012), as well as understanding Google Analytics data and using it as a communications tool (Kent, Carr, Husted, & Pop, 2011). The contribution of the present study is to investigate whether Google Analytics data for a DMO's website possess predictive information that helps to improve the accuracy of tourism demand forecasts in terms of actual tourist arrivals to a destination.

There is one further study employing Google Analytics for forecasting tourism demand (Yang, Pan, & Song, 2014). However, that study only uses numbers of website visitors and website visits as website traffic indicators, whereas the present study employs 10 Google Analytics indicators in a more comprehensive forecast modeling exercise. Other than that, there are some instructions and codes on the Internet outside the tourism discipline on how to retrieve Google Analytics data through a website's application programming interface (API) and how to forecast the retrieved Google Analytics indicators as measures of website traffic. Typically, univariate forecast models are employed using statistical software such as R (see e.g. Breña Moral, 2007; Granowitz, 2014; Vadera, 2013; Zwitch, 2013).

To this end, the website of the DMO of Vienna, the Vienna Tourist Board, has been selected for scrutiny: www.wien.info. Since Google Analytics data are not open to the public, the authors received permission from the Vienna Tourist Board to use their Google Analytics data for the time period from 2008M08 until 2014M10 for this research. Apart from issues of data availability, Vienna is one of the top-10 city destinations in Europe, with more than 13 million bednights in 2013, followed by Munich, Hamburg and Amsterdam (European Cities Marketing (ECM) & MODUL University Vienna (MU), 2014), and therefore of particular relevance for such a study in the thriving strand of literature on city tourism demand forecasting.

There exist more than 20 Google Analytics indicators, 10 of which were used in the present study based on *data availability* and *potential predictive power*. Given the high dimensionality of the sample (relatively few observations in combination with relatively many variables: total tourist arrivals to Vienna plus the 10 Google Analytics website traffic indicators), methods of big data shrinking become necessary to preclude inaccurate forecasts. Big data in this sense are defined as "*data sets and analytical techniques in applications that are so large (from terabytes to exabytes) and complex (from sensor to social media data) that they require advanced and unique data storage, management, analysis, and visualization technologies*" (Chen, Chiang, & Storey, 2012, p. 1166); Google Analytics data clearly fulfill this definition.

A number of articles addressing the predictive performance of different types of big data for tourism demand forecasting have been published, typically underlining the usefulness of the employed types of big data for improvements in forecast accuracy. Gawlik, Kabaria, and Kaur (2011), for instance, use web search volume histories to predict visitor numbers. Jackman and Natiram (2015) examine the usefulness of Google Trends for predicting tourist flows to Barbados using support vector regressions (SVRs). Bangwayo-Skeete and Skeete (2015) apply a mixed-frequency approach using Google Trends to predict tourist arrivals to Caribbean islands. In a study by Önder and Gunter (2016), in which Google Trends web and image search indices are used for English and native language searches, the usefulness of Google Trends for forecasting is confirmed, particularly with respect to native language searches.

Yang, Pan, Evans, and Lv (2015) compare search engine queries from Google and Baidu to predict visitor numbers to Hainan, China, with the result that search queries from Baidu outperform those from Google. Moreover, the Baidu index has been used in a recent article that investigates the tourist flows to the Forbidden City in Beijing, China, concluding that there is a positive impact of employing the Baidu index as an explanatory variable on the accuracy of visitor numbers forecasts (Huang, Zhang, & Ding, 2016).

Since Block Granger causality tests suggest a mutual causality structure between the 11 variables, vector autoregression (VAR, Lim & McAleer, 2001; Oh, 2005; Sims, 1980; Shan & Wilson, 2001; Song & Witt, 2006) and big data shrinking tech-