Google Trends and tourists' arrivals: Emerging biases and proposed corrections

Theologos Dergiades [a, *], Eleni Mavragani [b], Bing Pan [c]

[a] *Department of International & European Studies, University of Macedonia, Greece*
[b] *School of Economics, Business Administration and Legal Studies, International Hellenic University, Greece*
[c] *Department of Recreation, Park & Tourism Management, Pennsylvania State University, United States*

HIGHLIGHTS

- Web based search intensity indices (SII) are used to predict tourists' arrivals.
- We define two neglected sources of bias (language bias and platform bias).
- We propose corrections to improve the predictive power of web based SII.
- We find that the bias corrected SII is preferable in forecasting arrivals to Cyprus.

ARTICLE INFO

ABSTRACT

As search engines constitute a leading tool in planning vacations, researchers have adopted search engine query data to predict the consumption of tourism products. However, when the prevailing shares of visitors come from countries in different languages and with different dominating search engine platforms, the identification of the aggregate search intensity index to forecast overall international arrivals, becomes challenging since two critical sources of bias are involved. After defining the *language bias* and the *platform bias*, this study focuses on a destination with a multilingual set of source markets along with different dominating search engine platforms. We analyze monthly data (2004–2015) for Cyprus with two non-causality testing procedures. We find that the corrected aggregate search engine volume index, adjusted for different search languages and different search platforms, is preferable in forecasting international visitor volumes compared to the non-adjusted index.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Over the recent years, the availability of data gleaned from copious web sources (social media, search engines, etc.) sparked a new interest in the area named real-time economics. In one of the earliest studies, Choi and Varian (2009) demonstrated that properly selected query indices provided by Google are useful in forecasting activities in different economic sectors, such as the automobile industry and the tourism market. Their study has triggered a flurry of scientific publications that use web-related data, which aim to explain upcoming trends in various markets, including foreign

exchange markets, stock markets, sovereign bond markets, labor markets or even real estate markets (see among others Joseph, Wintoki, & Zhang, 2011; Smith, 2012; Beracha & Wintoki, 2013; Dergiades, Milas, & Panagiotidis, 2015). Credible evidence shows that web-related data offer added value when it comes to predicting upcoming economic activities.

Forecasting tourism demand is essential for practitioners and policymakers. Accurate forecasts provide valuable aid for a) the development of medium-to long-run marketing and tourism strategies, b) the formation of pricing policies, c) the appropriate scheduling of investments (Clerides & Adamou, 2010), and d) the effective allocation of the limited resources (Song, Witt, & Jensen, 2006; Yang, Pan, Evans, & Lv, 2015). Nowadays, web search engines constitute one major tool in planning vacations and can help to improve demand forecasting for the tourism product. In this study, we argue that the failure to account for two sources of bias (*language bias* and *platform bias*) frequently encountered

* Corresponding author. Department of International & European Studies, University of Macedonia, PC 546 36, Egnantia Str., Thessaloniki, Greece.
E-mail addresses: dergiades@uom.edu.gr (T. Dergiades), e.mavragani@ihu.edu.gr (E. Mavragani), bingpan@psu.edu (B. Pan).

during the construction of *Search Intensity Indices (SII)* from search engines, deteriorates the quality of the delivered index as a predictor.

We argue that a *SII* based on search engines in one language is unbiased, only if all the visitors perform their web searches in one language. In more details, as we use the search query volume in one language (e.g., English), the web search intensity is attributed just to a set of countries (U.S.A., U.K., etc.), while neglecting entirely the web search intensity formed in other languages. Thus, failure to account for all the languages that correspond to the respective source markets will give rise to the first source of bias, *language bias*. In addition, to protect the privacy of search engine users, the dominating search platform Google does not deliver data if the search volume for certain keywords is relatively small. Consequently, one cannot construct an entirely accurate aggregate index if some international tourists who searched on Google speak a rare language. One can imagine most countries will have a small number of international arrivals speak rare languages. Hence, this *language bias* is not a question of presence or absence, but rather an existing problem in various degrees. Even if at some point in our sample, all primary source markets use the same language, there is no guarantee that this will be the case in the future.

A second bias may exist if the search engine used to collect data is not the only platform in the source market of interest - thus, the *platform bias*. In such cases, the measured volume of queries underestimates the actual volume of relevant queries (the search volume from other engines is ignored), failing to convey the precise interest of users and its evolution over time.

This study concentrates on Cyprus and evaluates the impact of the relevant web *SII*, captured by search platforms, on the consumption of the tourism product. Cyprus is an ideal candidate country since the composition of international arrivals makes both sources of bias coexist. It allows us to examine how we can deal with the effects of the *language bias* and the *platform bias*, with a purpose to construct an effective predictor for international arrivals. We concentrate on the search engine of Google for two main reasons: Google is the most popular search engine globally, with a market share amounting to 66.7% (Yang et al., 2015); Google provides the historical intensity of the conducted queries through a platform called Google Trends (https://www.google.gr/trends).

Accurate prediction of the international arrivals in Cyprus is crucial since the overall contribution of the tourism industry in 2014 is more than €3 billion, a 21.3% of the GDP (KPMG, April 2016, available at: https://www.kpmg.com/cy/). Projections for the next ten years show that the absolute contribution of the tourism industry is expected to grow at a steady annual rate of around 5%. By 2025, the relative contribution of the tourism sector is anticipated to reach 25.5%. In addition, only around 40% of international arrivals are from English-speaking countries in 2015.[1] Around 30% of visitors speaks Russian, Greek, German, and Swedish as their native languages. Thus, English keyword searches might not represent a majority of searches for the country. Furthermore, Google is not the dominant search engine in the Russian market. A search engine called Yandex on average operates approximately 60% of the Russian market, while Google's respective share is about 25% (see www.liveinternet.ru).

This study adopts two non-causality testing techniques, in the

time domain and the frequency domain. It introduces an uncomplicated way to select appropriate keywords, and investigates the predictive power of Google's *SII* towards the arrivals of international tourists in Cyprus at an aggregate and disaggregate level. The findings show that the presence of the *language bias*, and the *platform bias* render the simple aggregate *SII* ineffective in predicting the total number of international arrivals. The corrected aggregate *SII* conveys a more valuable predictive content.

Our study has the following structure: Section 2 briefly reviews the literature devoted to the broad field of econometric forecasting through web-related data, paying particular attention to the tourism market. Section 3 illustrates the methodological framework and section 4 presents the data and the preliminary econometric analysis. Section 5 presents our main empirical findings while the resulting behavioral and managerial implications are discussed in Section 6. Finally, Section 7 concludes this study.

2. Literature review

Researchers try to provide accurate forecasts for the arrivals of tourists implementing a wide range of techniques. Peng, Song, and Crouch (2014) summarize two broad categories of techniques: time-series econometrics and artificial intelligence methods. The former category includes econometric models ranging from very simple univariate specifications (Geurts & Ibrahim, 1975; Martin & Witt, 1989) to more advanced multivariate specifications (Halicioglu, 2010; or; Bangwayo-Skeete & Skeete, 2015). The latter category comprises models ranging from artificial neural networks (Burger, Dohnal, Kathrada, & Law, 2001) to genetic algorithms (see among others, Chen & Wang, 2007). A detailed review on the topic is discussed in Peng et al. (2014) and Song, Witt, and Jensen (2003).

The empirical studies on tourism demand introduce an extensive set of explanatory factors to model arrivals. Using a diverse set of criteria, several researchers have grouped these factors (see Frechtling, 2001; Middleton, Fyall, & Morgan, 2009). Frechtling (2001) groups tourism demand factors into: 1) push, 2) pull, and 3) resistance factors. All groups above embrace both quantitative and qualitative factors, with the former to be those most frequently used in the empirical analysis since they are easily measurable and accessible effortless. In contrast, while qualitative attributes play a very crucial role in determining arrivals, rarely are these incorporated in demand specifications as their quantification is an arduous task.

In more detail, push factors include features related to the source markets. For example, Martins, Gan, and Ferreira-Lopes (2017) find that the per capita income is critical in explaining arrivals (based on a large panel of 218 countries) while Goh, Law, and Mok (2008) show, for the case of Hong Kong, that leisure time (in the two sources markets - U.S.A. and U.K.), influences arrivals stronger than economic factors. Additionally, Dragouni, Filis, Gavriilidis, and Santamaria (2016), focusing on the U.S. outbound tourism, support that the effect of consumers' sentiment and mood on the demand for tourism appears significant but time and event dependent.

Pull factors refer to attributes of the destination country (the quality the natural resources, Foreign Direct Investments (FDI) or social ties, etc.). For instance, Deng, King, and Bauer (2002) mention that natural resources constitute one of the leading attractions for tourism demand. A report of the World Tourism Organization (UNWTO) in 2012, evaluates that the number of travelers attracted by natural resources is predicted to rise rapidly over the upcoming decades, at a rate higher than the average of the tourism industry. Furthermore, by surveying the attitude of 2356 individuals from Italy with respect to the exploitation of natural resources, Meleddu and Pulina (2016) identify a positive propensity to pay a premium

---

[1] To the best of our knowledge the only study that deals with a destination that receives visitors from different countries is that of Choi and Varian (2012). Choi and Varian (2012) act at a disaggregate level only, and they do not provide many details about the construction of the search intensity index (e.g., keywords used).