To cite: Dergiades, T, Mavragani, E. & Pan, B. (2018). Google trends and tourists' arrivals: Emerging biases and proposed corrections. Tourism Management, 66: 411-421.

Predicting International Tourist Arrivals with Corrected Search Engine Query Data

Theologos Dergiades‡
Department of International & European Studies
University of Macedonia, Greece
e-mail: dergiades@uom.edu.gr

Eleni Mavragani
School of Economics, Business Administration and Legal Studies
International Hellenic University
e-mail: e.mavragani@ihu.edu.gr

and

Bing Pan
Department of Recreation, Park & Tourism Management
Pennsylvania State University
e-mail: bingpan@psu.edu

May 22, 2017

Abstract

As search engines constitute a leading tool in scheduling 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 to Cyprus compared to the non-adjusted index.

JEL Classification: Z31, Z38, C32, C53

Keywords: Web search intensity; forecasting tourism demand; big data search.

---

‡ Corresponding author; names are randomly ordered.

1. Introduction

Over recent years, the availability of data gleaned from copious web sources (social media, search engines, etc.) sparked a new interests in the area named real-time economics.[1] In one of the earliest studies,[2] Choi and Varian (2009) demonstrated that properly selected query indices provided by Google are useful in forecasting the activity in different economic sectors, such as the automobile industry and the tourism market.[3] 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 *et al.*, 2011; Smith, 2012; Beracha and Wintoki 2013; Dergiades *et al.*, 2014). 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 policy makers. 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 in the sector (Clerides and Adamou, 2010), and d) the effective allocation of the limited resources (Song *et al.*, 2006; Yang *et al.*, 2015). Nowadays, web search engines constitute one major tool in planning vacations and can help improve demand forecasting for the tourism product. In this study, we argue that the failure to account two sources of bias (*language bias* and *platform bias*) frequently encountered in 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 any other case, the constructed *SII* solely based on one language and platform will be biased.[4] Particularly, as the share of the total arrivals from countries with one language decreases progressively, the aggregate web *SII*, underestimates the true web search intensity. Hence, failure to account all the languages that correspond to the respective source markets will give rise to the first source

[1] The usefulness of the web search intensity data in predicting events was firstly recognized by researchers conducting studies in the field of medicine (see for example: Cooper *et al.*, 2005; Polgreen *et al.*, 2008).

[2] Ettredge *et al.* (2005) is the first study that uses web search intensity data resulted from employment related searches as a significant leading indicator for the U.S. unemployment level.

[3] See the Choi and Varian (2009) technical report, which has been published as Choi and Varian (2012).

[4] In more detail, as we use only one language (e.g. English) we reveal correctly the web search intensity that it is attributed only a set of countries (the countries that make use the English language, US, UK etc.), while at the same time we neglect entirely the web search intensity that is formed in other countries using other languages.

of bias, *language bias*. In addition, in order 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[5]. Consequently, one cannot construct a completely accurate aggregate index if some international tourists who searched on Google speak a rare language. Though without actual data, 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 it is an existing question in various degrees. Even if at some point of our sample, all major source markets use the same language, there is no guarantee that this will be the case in the future.

In addition, a second bias may exist if the search engine used to collect data is not the dominant platform in the source market of interest – thus, the *platform bias*. In such cases, the measured volume of queries underestimates the true volume of relevant queries, failing to convey the precise interest of the users and its evolution over time.

This study concentrates on Cyprus, and evaluates the impact of the relevant web *SII*, captured by various 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 purpose to receive an effective predictor for international arrivals. Finally, we concentrate on the search engine of Google for two major reasons: Google is the most popular search engine globally, with a market share amounting to 66.7% (Yang *et al.*, 2015). Google provides historical intensity of the conducted queries through a platform called Google Trends.[6]

Accurate predictions of the international arrivals in Cyprus is crucial, since the overall contribution of the tourism industry in 2014 is more than €3 billion, an 21.3% of the GDP (KPMG, April, 2016). Projections for the next 10 years show that the absolute contribution of the tourism sector 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%.[7] Cyprus is an ideal case study since only around 40% of international arrivals are from English speaking countries in 2015.[8] Around 30% of visitors speaks Russian, Greek,

[5] See: https://support.google.com/trends
[6] See: https://www.google.gr/trends
[7] The 2016 tourism market report of KPMG for Cyprus, is available at: https://www.kpmg.com/cy/
[8] To the best of our knowledge the only study that deals with a destination that receives visitors from countries with different countries is that of Choi and Varian (2012). Choi and Varian (2012) act at a disaggregated level only and they do not provide much information about the construction of the search intensity index (e.g. keywords used).

German, and Swedish as their native languages. Thus, English keyword searches might not represent a majority of searches for the country. Additionally, 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%.[9]

This study adopts two non-causality testing techniques, in the time domain and in the frequency domain. It introduces a simple 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 predictor.

In the following of the paper, Section 2 briefly reviews the literature devoted to the broad field of econometric forecasting through web-related data, paying special attention to the tourism market. Section 3 illustrates the adopted methodological framework. Section 4 presents the data and the preliminary econometric analysis, while Section 5 discusses our main findings. Section 6 concludes this study.

2. Literature Review

Researchers try to provide accurate forecasts for the arrivals of tourists implementing a wide range of techniques. Peng *et al.* (2014) summarize two broad categories of techniques: time-series econometrics and artificial intelligence methods. The former category includes econometric models ranging from very simplistic univariate specifications (Geurts and Ibrahim, 1975; Martin and Witt, 1989) to more advanced multivariate specifications (Halicioglu, 2010; or Bangwayo-Skeete and Skeete, 2015). The latter category comprises models ranging from artificial neural networks (Burger *et al.*, 2001) to genetic algorithms (see among others, Chen and Wang, 2007).[10]

The increasing availability of data capturing consumers' online activities has led to many studies with a purpose of forecasting upcoming events in the respective markets. Yang *et al.*, (2014) recognize that the major advantages of such data lie in: a) reveal preferences in real-time, b) provide data in relatively high frequency (e.g. daily or weekly) and, c) depict changes in consumers' preferences. The latter advantage, consists a solution to the inherent specification problem encountered often in traditional univariate time-

[9] See www.liveinternet.ru.
[10] For detailed review of the topic please see Peng *et al.* (2014) as well as Song *et al.* (2003).

series models (e.g. ARMA models).[11] For instance, Smith (2012) shows that the online search intensity, as captured by Google, significantly explains movements in the currency markets. Joseph *et al.*, (2011), using data from the Google Trends, forecast abnormal stock returns and the respective trading volume with search volume data for the respective stock tickers. Based on a sample of 3000 stocks, Da *et al.*, (2011) argue that a higher search volume index for the relevant stock ticker forecasts higher stock prices in the short-run. Beracha and Wintoki (2013) find that the abnormal search intensity in the real estate market of a city predicts the abnormal housing prices. Finally, Dergiades *et al.*, (2015) show that the web search intensity for the keyword *Grexit* explains future price movements of 10-year government Greek bonds.

A substantial number of web users seek information through established search engines before taking a trip (Fesenmaier *et al.*, 2011). Despite the large volume of studies dedicated to forecast the demand of the tourism product, there is relatively a small number of studies adopting web search intensity data. Xiang and Pan (2011) analyzed search queries of U.S. cities, diagnosed that "the ratio of travel queries among all queries about a specific city seems to associate with the touristic level of that city" (p. 88). Choi and Varian (2012) validate that search intensity data provided by Google for nine source markets-countries are indeed useful predictors of tourists' arrivals to Hong Kong from each respective market.

Yang *et al.* (2015) implemented an ARMA -Autoregressive Moving Average- specification and the standard Granger non-causality test, and affirmed that query volume data from two search engines - Google and Baidu - contribute significantly in decreasing forecasting errors when predicting the number of visitors to Hainan (a Chinese province). Bangwayo-Skeete and Skeete (2015) direct their interest to international visitors to five Caribbean destinations (Jamaica, Bahamas, Dominican Republic, Cayman and St. Lucia). They conducted their analysis by implementing a simple AR-MIDAS[12] model, a SARIMA model (Seasonal Autoregressive Integrated Moving Average), and a benchmark AR model (Autoregressive). The former model appeared to perform better in most of the conducted pseudo-forecasting experiments.[13] Overall, the authors argued that after the appropriate construction of the Google search intensity indicator, significant gains are achieved in forecasting tourist arrivals.

[11] Univariate time-series fail to provide robust forecasts, once sudden one-off events take place and alter the pattern of the series.
[12] The MIDAS estimation approach refers to the case where data with mixed frequencies are involved.
[13] For studies than implement web data aiming to predict the demand for hotels see Yang *et al.* (2014).

These studies validated the value of search engine intensity data in predicting tourist arrivals. However, in most of these studies, the dominant visitor source market are English-speaking countries; almost all the source markets use Google as the dominant search engine. For instance, in the study of Bangwayo-Skeete and Skeete (2015) the three source markets for the five investigated destinations are United States, United Kingdom and Canada with the market share of Google's search engine to be 68.8%, 92.7% and 92.9%, respectively (Kennedy and Hauksson, 2012). However, the source market for many countries may use a variety of languages and Google might not be the dominant search engine. For example, Canada, among several other countries, is officially a bilingual country or in numerous other countries large linguistic minorities exist. Moreover, Google is not the dominant search engine in some large markets such as Russia and China, where Yandex (for Russia) and Baidu (for China) have a market share approximately equal to 60% and 52%, respectively (Kennedy and Hauksson, 2012). In addition, if we use only one language in search engine index or if we focus on one search engine with small market share (when another search engine is the market leader), then we will extract the web search intensity that it is attributed only to a fraction of searches. We are neglecting the web search intensity formed in other languages and other search engines.

This study adopted Cyprus, where among its source markets, English is not the dominant language; some markets do not use Google as the major search engine. We are interested in finding out how to acquire search engine query data in forecasting tourist arrivals in this country.

3. Methodology

3.1. Standard Granger Non-Causality Testing

Within a bivariate VAR framework, as in *eq.* (1), the null hypothesis of no predictive content is examined by testing whether lagged values of one variable may significantly contribute in predicting current values of another variable.

$$\mathbf{Z}_t = \mathbf{\Theta}(L)\mathbf{Z}_t + \boldsymbol{\varepsilon}_t = \begin{pmatrix} \Theta_{11}(L) & \Theta_{12}(L) \\ \Theta_{21}(L) & \Theta_{22}(L) \end{pmatrix} \begin{pmatrix} A_t \\ G_t \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} \tag{1}$$

where, $\mathbf{Z}_t = \left(A_t\ G_t\right)^T$ is a 2×1 vector of stationary variables, $\mathbf{\Theta}(L)$ is a 2×2 matrix of lag polynomials and finally, $\boldsymbol{\varepsilon}_t$ is a 2×1 vector of error terms assuming the usual

properties. The null hypothesis of non-causality running from G_t to A_t (or from A_t to G_t) is rejected if at least one coefficient of the lag polynomial $\Theta_{12}(L)$ (or $\Theta_{21}(L)$) is significantly different from zero in explaining current values of A_t (G_t).

3.2. Frequency Domain Non-Causality Testing

Based on the standard structural representation of a VAR model, implementing the well know identification process of Cholesky, the spectral density of A_t (defined as in sub-section 3.1) at frequency ω can be expressed by *eq.* (2) as follows:

$$f_x(\omega) = (1/2\pi)\left\{\left|\Psi_{11}(e^{-i\omega})\right|^2 + \left|\Psi_{12}(e^{-i\omega})\right|^2\right\} \tag{2}$$

The non-causality hypothesis within the framework of Geweke (1982) is tested from the following Fourier transformation of the moving average coefficients:

$$\begin{aligned} M_{G\to A}(\omega) &= \log\left[\frac{2\pi f_x(\omega)}{\left|\Psi_{11}(e^{-i\omega})\right|^2}\right] = \log\left[\frac{\left|\Psi_{11}(e^{-i\omega})\right|^2}{\left|\Psi_{11}(e^{-i\omega})\right|^2} + \frac{\left|\Psi_{12}(e^{-i\omega})\right|^2}{\left|\Psi_{11}(e^{-i\omega})\right|^2}\right] \\ &= \log\left[1 + \frac{\left|\Psi_{12}(e^{-i\omega})\right|^2}{\left|\Psi_{11}(e^{-i\omega})\right|^2}\right] \end{aligned} \tag{3}$$

If G_t does not cause A_t at frequency ω, $\left|\Psi_{12}(e^{-i\omega})\right|^2$ has to be equal to zero.

Provided that term $\left|\Psi_{12}(e^{-i\omega})\right|^2$ is a complicated non-linear function, Breitung and Candelon, (2006) (B&C, hereafter), propose a solution by introducing a set of linear restrictions imposed on the estimated VAR coefficients. Focusing on the $\Psi_{12}(L)$ element of the $\Psi(L)$ matrix, B&C introduce the appropriate to the case null hypothesis of no causality. The $\Psi_{12}(L)$ element is equal to:

$$\Psi_{12}(L) = -\frac{1}{c_{22}}\frac{\Theta_{12}(L)}{\left|\Theta(L)\right|} \tag{4}$$

where, $1/c_{22}$ is the positive[14] lower diagonal element of the C^{-1} matrix (this is the inverse of the lower triangular C matrix used in the Cholesky identification process) and $\left|\Theta(L)\right|$ is the determinant of $\Theta(L)$. Therefore, the non-causality hypothesis at frequency ω from G_t towards A_t is not rejected whenever the following holds:

[14] We assume that the variance-covariance matrix Σ is a positive definite.

$$\left|\Theta_{12}(e^{-i\omega})\right| = \left|\sum_{k=1}^{p}\theta_{12,k}\cos(k\omega) - \sum_{k=1}^{p}\theta_{12,k}\sin(k\omega)i\right| = 0 \quad (5)$$

where, $\theta_{12,k}$ is the upper right element of the Θ_k matrix. Subsequently, the set of restrictions that should be imposed are:[15]

$$\sum_{k=1}^{p}\theta_{12,k}\cos(k\omega) = 0 \quad \text{and} \quad \sum_{k=1}^{p}\theta_{12,k}\sin(k\omega) = 0 \quad (6)$$

The empirical procedure of the B&C approach lies on the validity of the above presented linear restrictions. For brevity if we denote $\alpha_j = \theta_{11,j}$ and $\beta_j = \theta_{12,j}$, then the VAR equation that corresponds to the A_t variable may be rewritten as:

$$A_t = \alpha_1 A_{t-1} + ... + \alpha_p A_{t-p} + \beta_1 G_{t-1} + ... + \beta_p G_{t-p} + \varepsilon_{1t} \quad (7)$$

Thus, the hypothesis of no causality $M_{G\to A}(\omega) = 0$, is equivalent to the following set of linear restrictions:

$$R(\omega)\beta = 0, \quad \text{where } \beta = \left(\beta_1,...,\beta_p\right)' \text{ and } R(\omega) = \begin{pmatrix} \cos(\omega) & ... & \cos(p\omega) \\ \sin(\omega) & ... & \sin(p\omega) \end{pmatrix} \quad (8)$$

B&C investigate the validity of the linear restrictions illustrated in *eq.* (8), for frequencies ω that receive values within the interval of $(0,\pi)$, by comparing the obtained Statistic with the 0.05 critical value of the χ^2 distribution with 2 degrees of freedom.

4. Data and Preliminary Econometric Analysis

4.1. Data Sources

This study employs monthly time-series data on tourist arrivals in Cyprus along with web search intensity data for properly selected keywords. The range of our sample is January 2004 to April 2016 (148 observations) due to data availability from Google Trends. The data for the total arrivals of tourists in Cyprus (see Fig. 1) as well as the origin per country of these arrivals come from the Statistical Service of Cyprus (Fig. 2 shows the arrivals per country as a market share). The arrivals per country are available until December 2015, 144

[15] Given that $\sin(k\omega) = 0$ in the cases where $\omega = 0$ and $\omega = \pi$, then it comes that the second restriction in *eq.* (6) is simply disregarded.

observations).[16] For the selected sample, to extract the *SII* related to the tourist product of Cyprus, we use the Google Trends facility.[17]

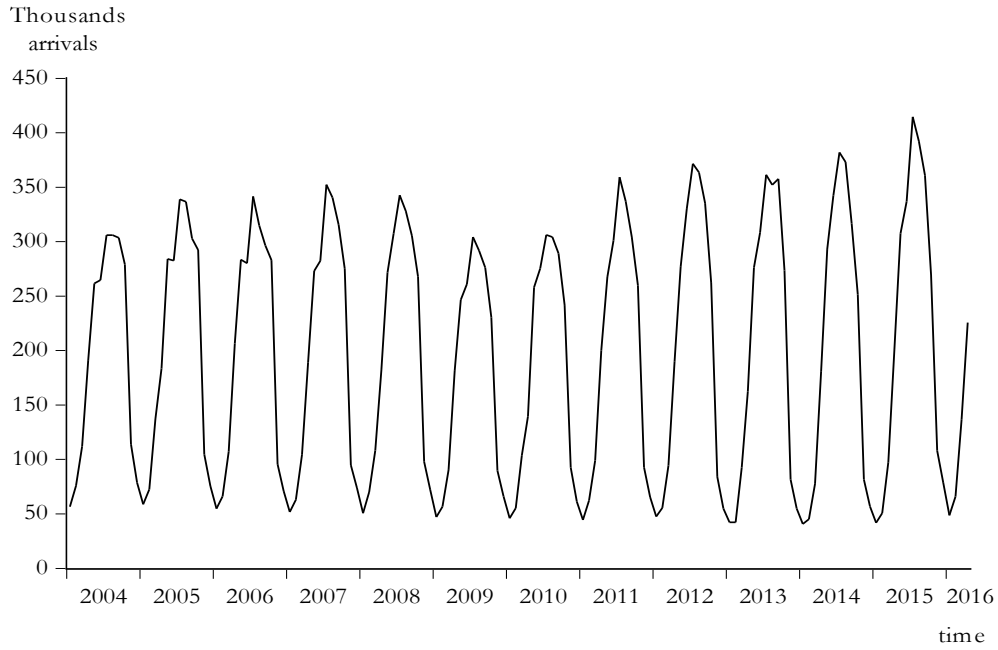

Figure 1. Monthly arrivals of tourists in Cyprus

To capture the entire web search intensity for a destination, we first consider possible existence of the *language bias* and the *platform bias*. To overcome the first problem, we disentangle the aggregate number of tourist arrivals in Cyprus by country of origin to specify the corresponding languages. The market share in the total arrivals per country is illustrated in Fig. 2. Visual inspection suggests that five countries are the main source markets, representing jointly 74.1%[18] of the market share, while the respective share for all the other countries is 25.9%. The major source markets countries are the following: UK (45.4%), Russia (10.1%), Greece (7.7%), Germany (7.2%) and Sweden (3.7%). Hence, we concentrate on the respective languages (English, Russian, Greek, German and Swedish).

[16] See: http://www.mof.gov.cy/mof/cystat/statistics.nsf; accessed June 2016.
[17] See: http://www.google.com/trends/; accessed June 2016.
[18] The reported value is the average share of the total monthly arrivals for the period of study (2004-2015).

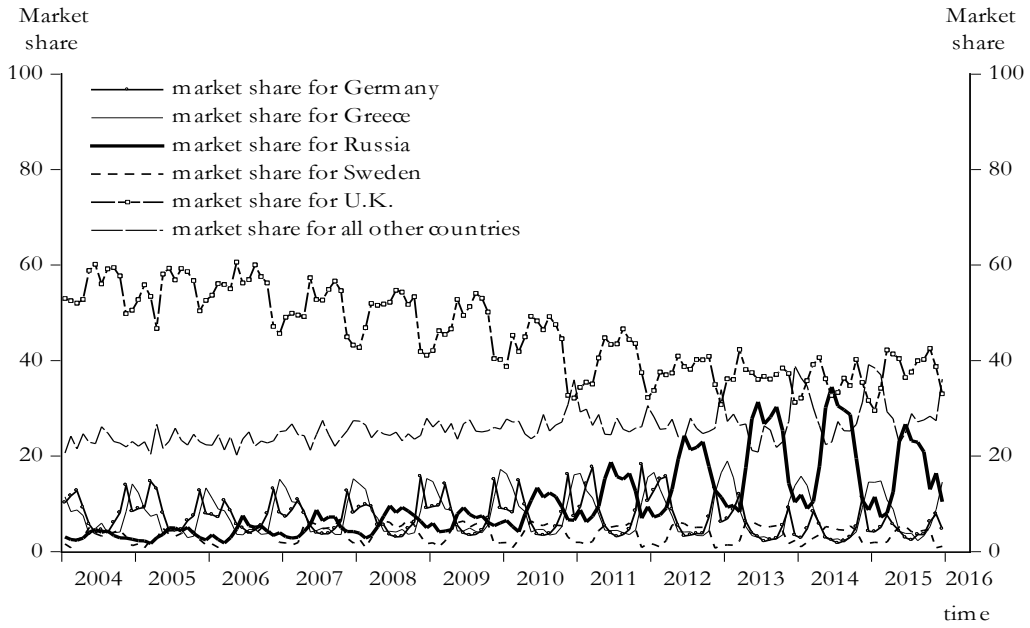

Figure 2. Markets shares in tourists arrivals for Cyprus per country

Fig. 2, shows that the market shares for each country evolve differently. The market share of the U.K. shrinks from 55.2% in 2004 to 38.0% in.[19] The German share in 2004 was 7.2% reaching the value of 4.9% in 2015. The Russian market share increased from 3.3% in 2004 to 16.3% at the end of the sample. The total market share of the rest of countries increased from 23.2% in 2004 to 29.6% in 2015. Finally, Greece and Sweden has stayed constant.

The next step is to identify the appropriate keywords that are directly related to a potential visit in Cyprus for each language. It is reasonable to assume that the performed searches contain the term *Cyprus*. The identification of appropriate keywords involves the following steps: 1) by first selecting the source market of interest, we type the term *Cyprus* (in the language of the source market) in the Google Correlate tool[20] to attain other queries that present similar patterns (the similarity is ascertained through a simple correlation coefficient). From the delivered queries, ranked in terms of correlation, we select the query that presents the highest correlation to our search term and its meaning refers explicitly to a visit in Cyprus (e.g. *flights to Cyprus*). 2) From Google Trends facility, we extract a monthly frequency series for the keyword identified in the previous step.[21] We verify the validity of our chosen keyword by examining the top related queries as suggested by Google Trends. If the vast majority of the related queries imply interest for visiting Cyprus, we may argue

[19] A significant time trend is verified by running a simple regression of the market share on a standard time trend. The significance level is 0.01. The results are available upon request.
[20] See: https://www.google.com/trends/correlate
[21] The search term is not enclosed in quotation marks.

in favor of our keyword. 3) For those cases where in step 1 our initial key term (e.g. *Cyprus*) does not deliver keywords that convey direct interest for a trip to the destination, we type our key-term (*Cyprus*) to the Google Trends facility and from the delivered related queries we select the one that expresses explicit interest to visit the destination.

For instance, in the case of U.K., the Google Correlate facility suggests that the first most highly correlated term to *Cyprus* (which implies explicit intention to visit Cyprus) is the keyword *hotel Cyprus*. At the second stage, we type *hotel Cyprus* in the Google Trends facility and we examine the relevant queries. All the relevant queries verify the validity of our selected keyword since they imply direct interest to visit Cyprus.[22] The finally extracted index in monthly frequency is presented in Fig. 3a below.[23] Implementing the same strategy for the remaining source markets, we end up with the following keywords. For the Russian market, the identified keyword is туры кипр (tours Cyprus) and the respective index is illustrated in Fig. 3b. For the German market, the keyword is *hotel zypern* (hotel Cyprus) and depicted in Fig. 3c, and for the Swedish market the keyword is *cypern resor* (Cyprus travel) shown in Fig. 3d. Finally, the strategy failed to deliver a keyword that expresses an intention to visit Cyprus for the case of Greece. We tried keywords that are similar to those identified for the other countries, as for example *ξενοδοχεία Κύπρος* (hotels Cyprus) or *διακοπές Κύπρος* (holidays Cyprus), and the Google Trends facility indicated that there is not enough search volume to deliver results. Therefore, we are unable to construct a web *SII* for Greece, and we proceed with the remaining markets.[24]

[22] The relevant queries in order are: hotel in Cyprus, Paphos Cyprus, Paphos, hotels Cyprus, Cyprus holidays, Portaras Cyprus, Portaras.

[23] Figure 3, along with the web *SII*, also presents the arrivals for each country.

[24] For the rest major markets (U.K., Russia, Germany and Sweden), the average share of total monthly arrivals in Cyprus, for the period of study (2004-2015), is 66.4%. Ridderstaat and Croes (2016), investigate the effect that money supply cycles in three major source markets may have on tourism arrivals for the case of Aruba and Barbados. The market shares of these three markets for the two destinations are 68.1 and 70.9, respectively. These shares are of similar magnitude to the market share covered by our study.

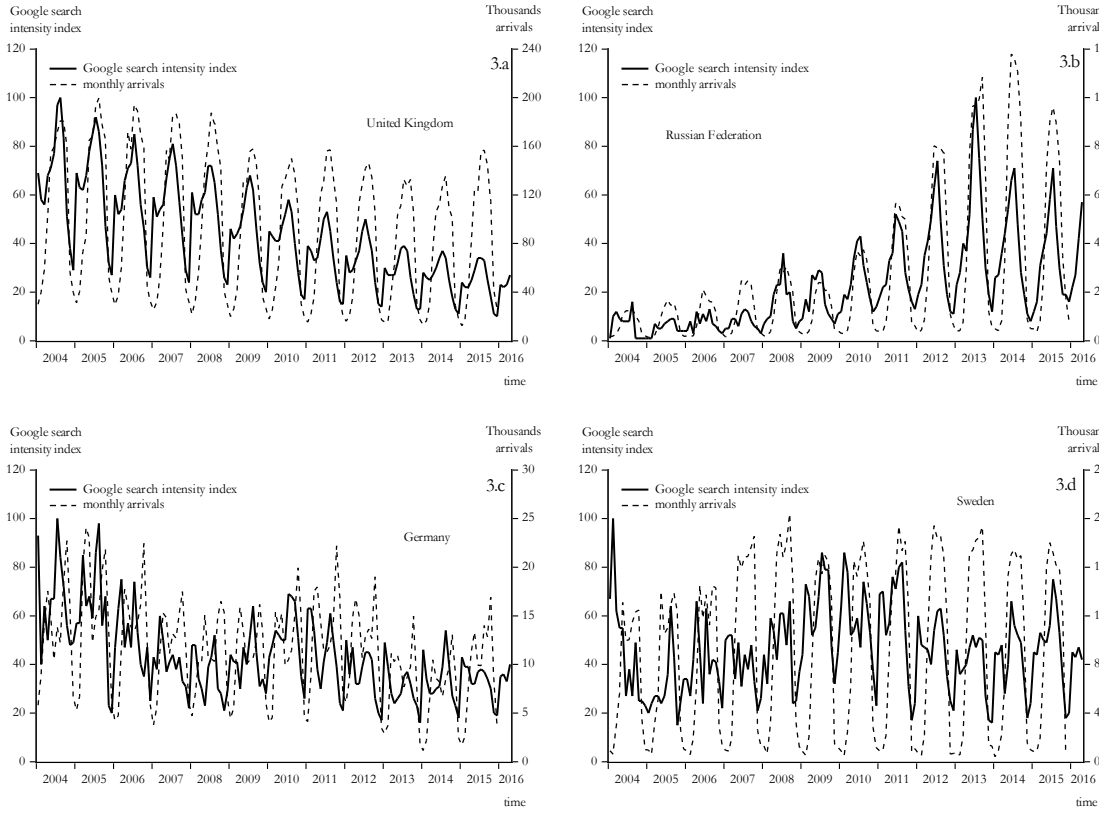

Figure 3. Google *SII* and tourists arrivals per country.

In the case of Russia, the *platform bias* is also essential. Since Google is not the dominant search engine, we run the risk to misidentify the precise interest of the users and its evolution over-time. To cross-check the validity of our selected keyword (туры кипр) from Google, we execute the same identification strategy by using a similar facility offered by Yandex (relevant phrases). The delivered keyword is туры на кипр, which is almost identical to the keyword identified by Google Trends (туры кипр). To check the keywords' evolution overtime, we take advantage of another feature offered by Yandex, which delivers the absolute number of searches for a keyword of interest. The common sample correlation coefficient between the *SII* of Google Trends (туры кипр) and the number of searches in Yandex (туры на кипр) is 0.97.[25] Hence, we may argue that the *SII* obtained from the Google, despite its' relatively small share in the market, reveals the true pattern over-time. Nevertheless, what is still an issue with the case of Russia is the fact that in Google the true intensity of searches is underestimated.

[25] The absolute number of a search in Yandex is available, on monthly basis, for the past two years.

To construct the aggregate uncorrected for the two sources of bias *SII*, we combine the volumes of all the previously identified keywords (one for each country) to a single search.[26] The constructed index is presented in Fig. 4.

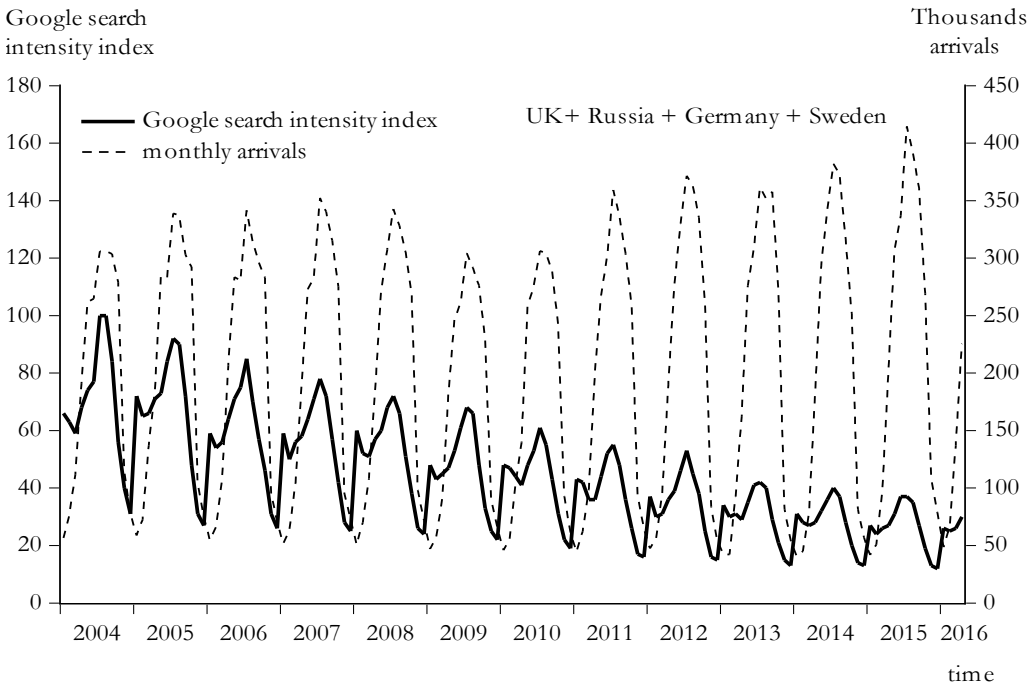

Figure 4. Aggregate Google *SII* and total arrivals of tourists.

4.2. Preliminary Econometric Analysis

The data of country specific and aggregate of the arrivals and the web *SII* (see Figs. 3 and 4), show a clear seasonal variation. The well-known adverse effects of seasonality in statistical inference dictate a need for a seasonal adjustment procedure. In our case, we remove the deterministic seasonal parts of the series by implementing the TRAMO/SEATS approach as part of the X-13ARIMA-SEATS program. Fig. 5 illustrates the seasonally adjusted series.

In addition, the stationarity properties of all the de-seasonalized series are examined by conducting the Phillips and Perron (1988) test, with and without the presence of a deterministic linear trend (Table 1). Hence, we reject the null hypothesis of a unit root, at the 0.01 significance level, for the aggregate arrivals and the arrivals that origin from Germany and Sweden, while the opposite is true (failing to reject) for the arrivals that come from the U.K. and Russia. However, once we allow for the presence of a linear trend, the

[26] The conducted single search is: hotel Cyprus + туры кипр + hotel zypern + cypern resor.

arrivals from the U.K. and Russia prove to be trend stationary. In a similar fashion, the null hypothesis is rejected, at the 0.01 significance level, when the test is conducted to the Google *SII* of two countries, Germany and Sweden, while this is not the case for the remaining indices. The inference for the remaining indices is reversed in the presence of a linear trend. Overall, we may treat all the involved variables as stationary or trend stationary. In the case where a variable is trend stationary, it is incorporated into our analysis after removing the linear time-trend.

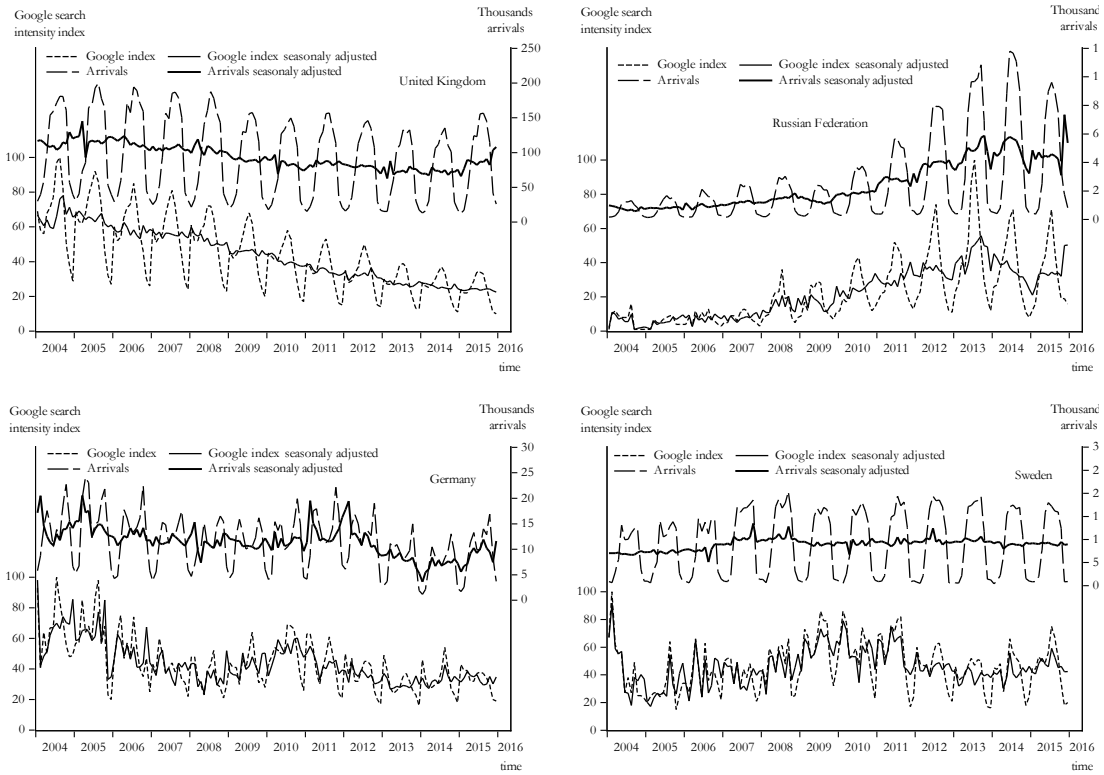

Figure 5. Seasonally adjusted series for the *SII* and the arrivals per country.

Table 1. Phillips-Perron unit-root tests for the de-seasonalized series.

Country	Arrivals		Inference	Google *SII*		Inference
	no trend	trend		no trend	trend	
UK	-1.75	-4.03***	I(0) /	-0.74	-3.83**	I(0) /
Russia	-0.44	-3.15*	I(0) /	-1.08	-3.17*	I(0) /
Germany	-3.58***	-4.68***	I(0)	-5.22***	-7.49***	I(0)
Sweden	-5.33***	-6.48***	I(0)	-6.21***	-6.28***	I(0)
Aggregate	-4.87***	-5.16***	I(0)	-0.58	-5.63***	I(0) /

Notes: the symbols * and *** denote the rejection of the null hypothesis at the 0.1 and 0.01 significance level, respectively. I(0) that implies that the series is stationary, while I(0) / implies that the series is stationary under a linear time-trend. Finally, the bandwidth for the Phillips-Perron test was chosen based on the Newey-West selection procedure, while the spectral estimation method used is the Bartlett kernel.

5. Empirical Results

5.1 Predictive Power of the Web Search Intensity per Individual Country

To evaluate the predictive content of the constructed Google *SII* for every individual country towards the respective arrivals in Cyprus, we implement two alternative causality tests: the standard linear Granger non-causality test in the time domain and the B&C non-causality test in the frequency domain. The B&C allows us to identify whether a verified causal relationship is short-run or long-run, and is capable of revealing potential non-linear causal relationships.

Table 2 shows that the hypothesis of no predictability running from the *SII* to the arrivals is consistently rejected for all countries of interest. In particular, predictability is verified at the 0.05 significance level for the U.K. and Sweden, while the same inference is drawn for Russia and Germany at the 0.01 significance level. Additionally, we fail to reject the hypothesis of no predictability that runs from the arrivals to the *SII*. The only exception is Sweden where bidirectional causality is established. Overall, our findings based on the standard Granger test suggest that arrivals in Cyprus from the four major source markets can be predicted by the respective *SII*.

Table 2. Standard Granger non-causality test results (per country).

Country	Google *SII* → Arrivals		Arrivals → Google *SII*	
	F-statistic	(lag length)	*F*-statistic	(lag length)
UK	3.67**	(3)	1.51	(3)
Russia	9.96***	(3)	1.52	(3)
Germany	4.54***	(4)	0.73	(4)
Sweden	2.31**	(5)	3.41***	(5)

Notes: the symbols ** and *** denote the rejection of the null hypothesis of non-causality at the 0.05 and 0.01 significance level, respectively. The numbers within the parentheses indicate the lag length of the underlying bivariate VAR specification. Finally, the arrow signifies the direction of causality.

The B&C test results for the U.K., in Fig. 6.a, show that the null hypothesis of no predictability running from *SII* to tourist arrivals, is rejected at the 0.05 significance level, when $\omega \in [0, 1.24]$. This finding suggests that low and medium cyclical components of the *SII*, with wavelengths of more than five months, are those that contribute significantly in predicting arrivals. The opposite hypothesis is clearly rejected for the whole range of frequencies. The results for Russia are shown in Fig. 6.b. In particular, the predictability of the arrivals through *SII* is verified for the entire set of frequencies ($\omega \in [0, \pi]$). Again, the opposite hypothesis is rejected for the complete set of frequencies. For Germany (Fig. 6.c), predictability is not verified for the medium cyclical components but rather for the low and the high cyclical components of the series ($\omega \in [0, 0.75] \cup [1.88, \pi]$). Therefore, significant predictability is confirmed for wavelengths of less than 3.3 months and more

than 8.4 months. Again, arrivals appear not to predict the *SII*. Finally, our findings for Sweden (see Fig. 6.d) show that only the high-frequency components of the *SII* series are significant in predicting arrivals ($\omega \in [1.85, \pi]$). Hence, predictive power exists for wavelengths of less than 3.4 months. As was the case with the linear Granger non-causality test, we reject the non-predictability for the opposite hypothesis in high frequencies ($\omega \in [1.97, \pi]$), implying predictability for wavelengths of less than 3.2 months. In other words, for the case of Sweden short-run bidirectional predictability is established. Overall, our findings from the B&C test are qualitatively similar to those of the linear Granger non-causality test.

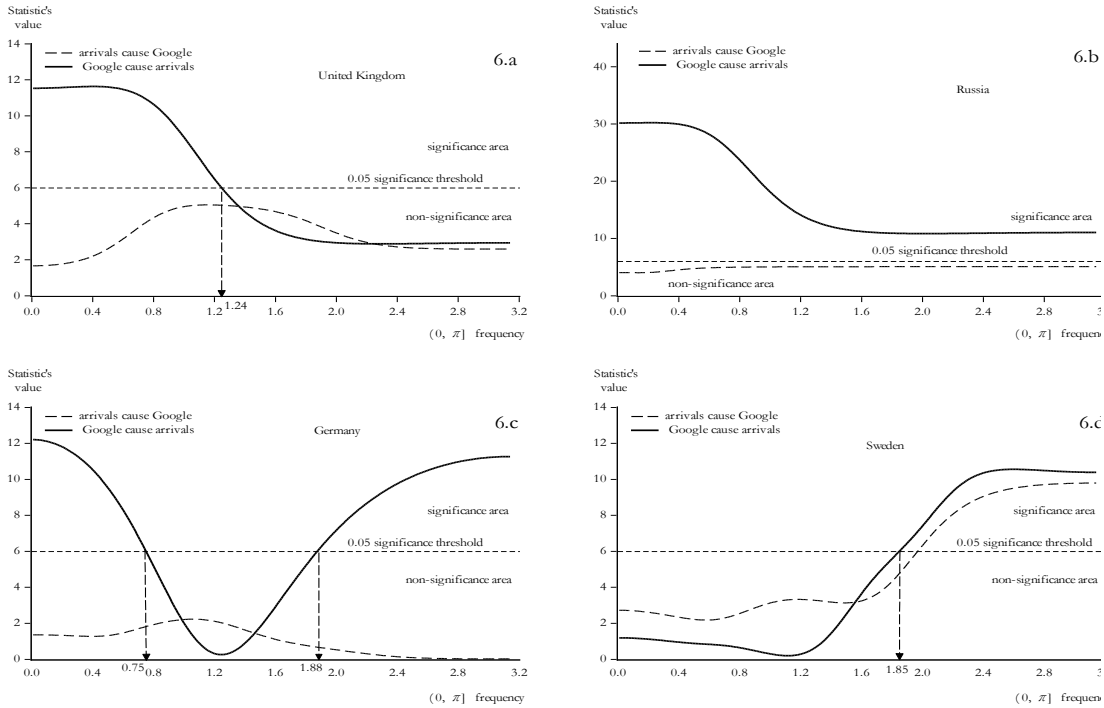

Figure 6. B&C Granger non-causality test per country

The results show that the predictive content of the constructed Google *SII* (with respect to the arrivals) is dissimilar among the examined countries. According to Mc Cabe *et al.*, (2016), national cultures integrate idiosyncratic features which affect the search information behavior. Similar in nature are the findings of Gursoy and Umbreit (2004), who verify for a set of European countries that national cultures influence traveller's search behavior resulting this way to clearly distinct consuming patterns.

Heterogeneous consuming patterns imply variation in the decision-making lead time. In particular, for three major source markets (UK, Russia and Germany) there is an essential magnitude of tourists who choose Cyprus as a destination at least half year ahead from their arrival time. Characteristic example is the case of Germany. According to the

Reise Monitor survey conducted by the ADAC Verlag,[27] which investigates the holiday travel patterns of German tourists, 70% of the travelers intending to visit European destinations start planning their trip half year ahead. Similarly, the respective percentage for those who plan their trip three months ahead until the last minute is approximately 20%. Such pattern clearly does corroborate our empirical findings. The tourists that come from Sweden have a predictive content for wavelengths of less than 3.4 months. This pattern can be attributed to the idiosyncratic features of those Swedish tourists who plan to visit Cyprus. For example, their booking practices may be heavily depend on travel agencies and therefore personal search for additional information may take place only few months prior their trip.[28]

However, both non-causality tests are unable to reveal whether the variables of interest are connected in a positive or negative manner. Cholesky defined accumulated impulse response functions of interest along with their associated ± 2 standard errors confidence bands and they are presented in Figs. 7a to 7d. For the case of the U.K., the accumulated response of tourist arrivals to one standard deviation shock in the *SII* for a 10-month period. Clearly, the response of the arrivals is constantly positive and significant for the entire period. Additionally, the impulse response analysis supports further our findings in the B&C test for the existence of causality that is long-run in nature. The impulse response analysis for Russia (see Fig. 7.b) and Germany (see Fig. 7.c) provides qualitatively similar inference to that of the U.K.. Hence, we observe a constantly positive and significant response of the arrivals to one standard deviation shock in the *SII* for both countries. Finally, for the case of Sweden (see Fig. 7.d), the impulse response function is positive throughout the examined period, but it proves to be significant only in the first few months. This finding chain with the B&C test results which support causality only in the short-run. Overall, we may claim that the response of the arrivals in Cyprus to one standard deviation shock in the *SII*, as this is captured by the Google Trends, is positive and in harmony with the B&C test results.

[27] The survey is available at: http://www.pot.gov.pl/component/rubberdoc/doc/1897/raw

[28] Given that the official statistical agency of Cyprus does not provide data about the decision making lead times of tourists we conducted the Cyprus Tourism Organization (CTO). CTO officials come to verify our empirical findings related to the decision making lead times of the four major source markets.

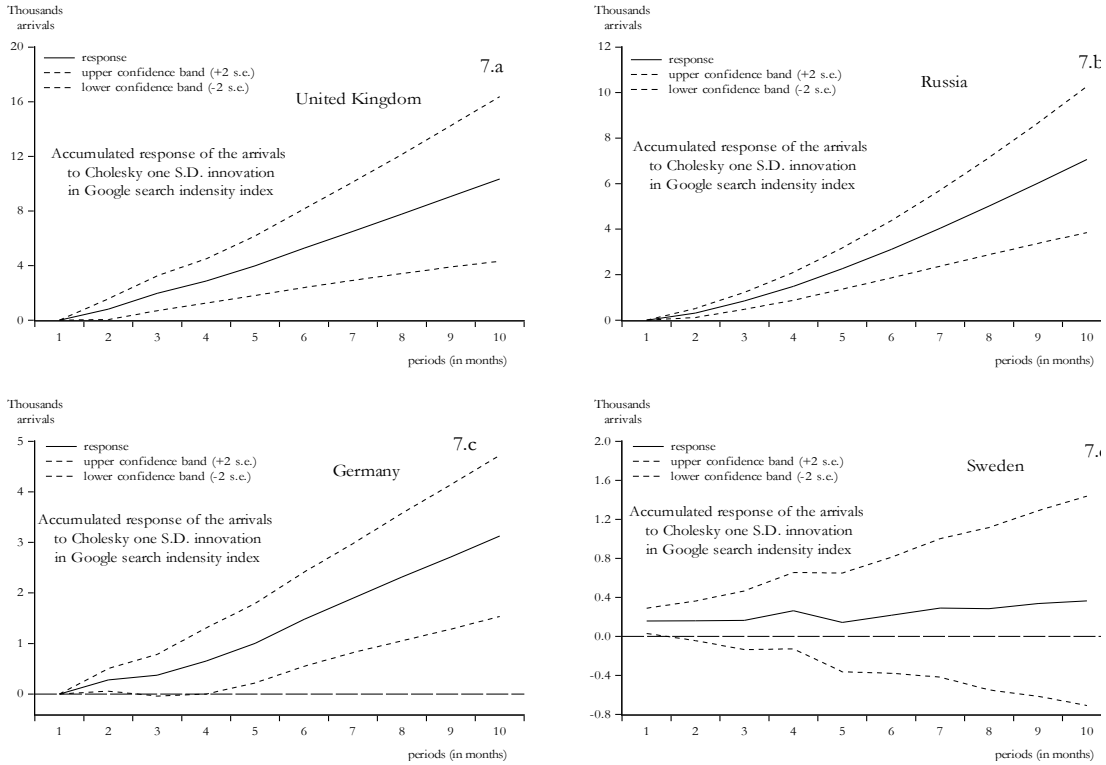

Figure 7. Impulse response functions per country

5.2 Aggregate Predictive Power of the Web Search Intensity

Next, we perform the causality tests for the aggregate web *SII* with respect to the total arrivals (see Fig. 4). After, de-seasonalizing both series and de-trending the aggregate web *SII* [29] (see the unit-root test results in Table 1) we conduct the standard Granger non-causality test (Table 3). We fail to reject the null hypothesis of no predictability that runs from the aggregate *SII* to the total arrivals (see 1st line in Table 3). Finally, the same inference holds for the opposite hypothesis.

Table 3. Standard Granger non-causality test results (aggregate).

Country	Google *SII* → Arrivals		Arrivals → Google *SII*	
	F-statistic	(lag length)	*F*-statistic	(lag length)
Aggregate	1.37	(3)	0.03	(3)
Aggregate corrected	4.09***	(3)	1.07	(3)

Notes: the symbols ** and *** denote the rejection of the null hypothesis of non-causality at the 0.05 and 0.01 significance level, respectively. The numbers within the parentheses indicate the lag length of the underlying bivariate VAR specification. Finally, the arrow signifies the direction of causality.

The hypothesis of no predictability is also certified within the framework of the B&C test. In particular, the null hypothesis of no predictability running from the *SII* to tourist arrivals is not rejected, at the conventional levels of significance, for the entire set

[29] To save space these results are not presented here. They are available upon request.

of frequencies ($\omega \in [0, \pi]$). Similarly, arrivals fail to predict in any significant manner the *SII* (see Fig. 8.a). The associated impulse responses while they prove to be consistently positive the relevant confidence bands include throughout the examined period the zero value. These findings are consistent with the B&C test results. Overall, while there is strong evidence of predictability at a country level, this predictability vanishes once we use the aggregate data.

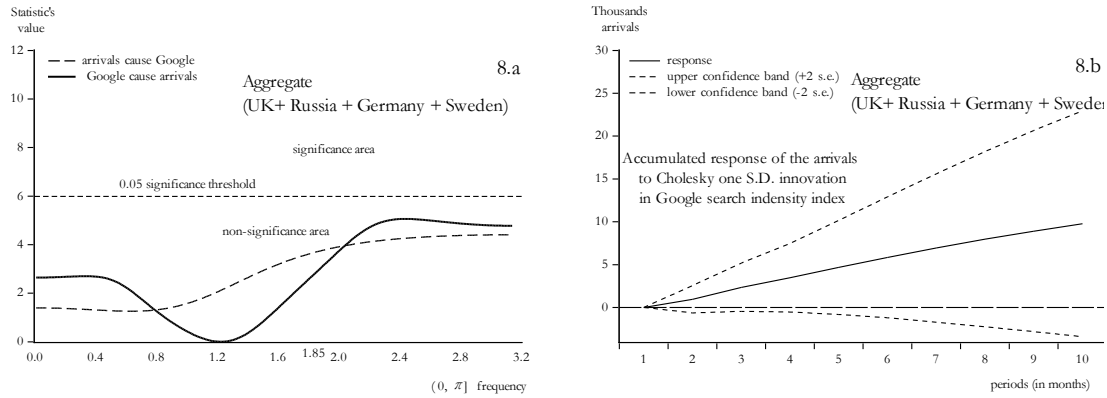

Figure 8. B&C test and impulse responses at the aggregate level.

To reassess the predictive content of the aggregate index, we construct a corrected index which considers the *language bias* and *platform bias*. In particular, we follow two steps in constructing the aggregate corrected index. First, to overcome the *language bias*, instead of using the total number of tourist arrivals, we restrict our exercise only to the arrivals that correspond to the four major source markets (UK, Russia, Germany and Sweden, which jointly pose almost 70% of the overall share in the arrivals). The constructed aggregate index reflects truly the web search intensity which linked to the arrivals from these countries.[30]

Second, to rectify our aggregate index from the *platform bias*, which is present in the case of Russia, we need to correct for the low market share of Google in the Russian Internet market. Instead, of constructing a unified index (as we did previously) by combining our four keywords (hotel Cyprus + туры кипр + hotel zypern + cypern resor), we extract four separate indices (one for each keyword) which are now compared jointly in terms of search volume (See Fig. 9). As Google has a low market share (aprox. 25%, S_1) in the Russian Internet market, naturally the *SII* that corresponds to Russia (see Fig. 9.a), underestimates the true search intensity.

[30] To reduce the "contamination" of the aggregate corrected index by search queries which may be irrelevant, we restrict our search to the travel category.

At the same time, as Yandex dominates the Russian Internet market (with market share aprox. 60%, S_2) and given that the volume delivered from Yandex (for the keyword туры на кипр) correlates strongly to the index delivered from Google (for the keyword туры кипр), we may use the ratio of the respective market shares (S_2/S_1) as a market share correction factor. Once we multiply Google's web *SII* that corresponds to Russia with the correction factor (S_2/S_1), then we can add the corrected index for Russia to the remaining three indices in order to form the aggregate corrected index. Consequently, the corrected aggregate index is expected to receive values above 100. This scale adjustment is attributed to the alternative scaling factor as well as to the introduced correction factor (these details are analytically discussed in the Appendix). For comparison purposes, the aggregate corrected *SII* along with the initial aggregate *SII*, both are illustrated in Fig. 9.b.

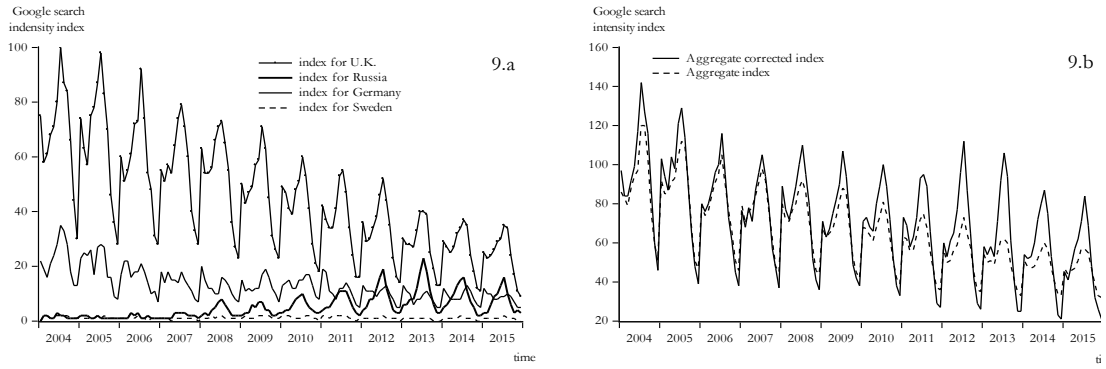

Figure 9. Search volume of the selected keywords and the aggregate corrected index.

Working within the same methodological framework, we can examine the predictive content of the aggregate corrected *SII* (See Fig. 9.b) with respect to the total arrivals from the four main source markets.[31] Starting from the standard linear non-causality test, we now fail to reject the hypothesis of no predictability that runs from the aggregate corrected index to the total arrivals from the four major source markets (see the 2nd line in Table 3) for all the conventional levels of significance. Regarding the opposite hypothesis, the testing results imply no predictability. The B&C test shows qualitatively analogous inference. The predictability running from the aggregate corrected index to the total arrivals from the four major source markets is verified at the 0.05 significance level, for wavelengths of more than 3.6 months ($\omega \in [0, 1.73]$) (See Fig. 10.a), while for the opposite hypothesis, there is no predictability at any frequency. Finally, the associated impulse

[31] Before we test for non-causality we de-seasonalize the arrivals from the four major source markets and the corrected aggregate index, while we de-trend only the later.

response function is consistently positive with the confidence bands not to include the zero value (See Fig. 10.b).

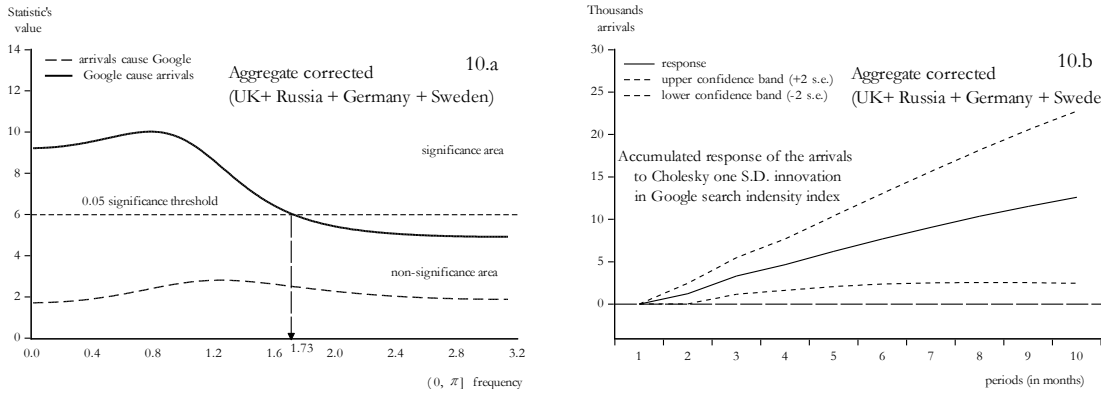

Figure 10. B&C test and impulse responses at the aggregate level (corrected *SII*).

6. Conclusions

In recent years, Google Inc. provides data on the intensity of queries conducted on their search engine. This leads to an outbreak of scientific projects aiming to explain upcoming trends in various markets based on these data. As search engines constitute a leading tool in planning vacations, the digital traces can be exploited to improve predictions on the consumption of tourism products. Under this prism, we examine the predictive power of a relevant web *SII*, as captured by Google, on the total number of arrivals at a destination of interest. While existing studies emphasize at destinations that receive arrivals from countries with one major language and Google to be the dominant web search platform (Bangwayo-Skeete and Skeete, 2015), our work is the first that focuses on a destination with a multilingual set of source markets and with different dominant search platforms. As such, we introduce an approach to correct for the *language bias* and the *platform bias*, improving this way the predictive power of the constructed index.

We test our hypothesis by using monthly tourist arrival data (2004-2015) to Cyprus and by conducting two Granger non-causality tests, the standard linear Granger non-causality test and the B&C non-causality test. By introducing a simple way to select appropriate keywords and working within the above framework, our findings show: a) country-specific *SII* (for U.K., Russia, Germany and Sweden) are highly significant in predicting the arrivals from the corresponding source markets, b) the initially constructed aggregate *SII*, without considering the *language bias* and the *platform bias*, proves inadequate to predict the total number of arrivals, and finally c) a corrected version of the aggregate

SII, taking into account the two problems, predicts in a significant and positive manner the arrivals (UK, Russia, Germany and Sweden).

Overall, our study validates the usage of the *SII* as an important leading indicator for the upcoming arrivals at a destination, but also reveals one very crucial methodological aspect. For destinations that accept arrivals from countries in different languages, that the formation of a precise aggregate *SII* (intended to capture the entire web activity) is a challenging task and in several cases almost impossible to be constructed. Therefore, we argue that when it comes to predicting the consumption of the tourist product based on the *SII*, then it is preferable that this task is conducted at a disaggregated level. In other words, every major source market has to be investigated separately. Acting such, we actually use a richer set of information allowing each country's idiosyncratic characteristics to be revealed. Clearly, we do not claim that approaches aiming to predict arrivals at an aggregate level have to be ostracized. Instead, we support that aggregate *SII* are exposed to two significant problems and hence special handling is needed. In failing to account for these problems, misleading prediction inferences may be conducted.

References

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

Breitung, J. and B. Candelon (2006). Testing for short- and long-run causality: A frequency-domain approach. *Journal of Econometrics* **132**(2), 363-378.

Beracha, E. and M.B. Wintoki (2013). Forecasting residential real estate price changes from online search activity. *Journal of Real Estate Research* **35**(3), 283-312.

Burger, C., M. Dohnal, M. Kathrada and R. Law (2001). A practitioners guide to time-series method for tourism demand forecasting - a case study of Durban, South Africa. *Tourism Management* **22**(4), 403-409.

Chen, K.Y. and C.H. Wang (2007). Support vector regression with genetic algorithms in forecasting tourism demand. *Tourism Management* **28**(1), 215-226.

Choi, H. and H. Varian (2009). Predicting the Present with Google Trends. Technical report, Google Inc.

Choi, H. and H. Varian (2012). Predicting the Present with Google Trends. *Economic Record* **88**(281, S1), 2-9.

Clerides, S. and A. Adamou (2010). Prospects and Limits of Tourism-Led Growth: The International Evidence. *Review of Economic Analysis* **2**(3), 287-303.

Cooper, C., K. Mallon, S. Leadbetter, L. Pollack and L. Peipins (2005). Cancer Internet Search Activity on a Major Search Engine, United States 2001-2003. *Journal of Medical Internet Research* **7**(3), 1-13.

Da, Z., J. Engelberg and P. Gao (2011). In search of attention. *Journal of Finance* **66**(5), 1461-1499.

Dergiades, T., C. Milas and T. Panagiotidis (2014). Tweets, Google trends, and sovereign spreads in the GIIPS. *Oxford Economic Papers* **67**(2), 406-432.

Ettredge, M., J. Gerdes and G. Karuga (2005). Using Web-based Search Data to Predict Macroeconomic Statistics. *Communications of the ACM* **48**(11), 87-92.

Fesenmaier, D., Z. Xiang, B. Pan and R. Law (2011). A framework of search engine use for travel planning. *Journal of Travel Research* **50**(6), 587-601.

Geurts, M.D. and I.B. Ibrahim (1975). Comparing the Box-Jenkins approach with the exponentially smoothed forecasting: model application to Hawaii tourists. *Journal of Marketing Research* **12**(2), 182-188.

Gursoy, D. and T. Umbreit (2004). Tourist information search behavior: cross-cultural comparison of European Union member states. *International Journal of Hospitality Management* **23**(1), 55-70

Halicioglu, F. (2010). An econometric analysis of the aggregate outbound tourism demand of Turkey. *Tourism Economics* **16**(1), 83-97.

Joseph, K., M.B. Wintoki and Z. Zhang (2011). Forecasting abnormal stock returns and trading volume using investor sentiment: evidence from online search. *International Journal of Forecasting* **27**(4), 1116-1127.

Kennedy, A.F. and K.M. Hauksson (2012). *Global Search Engine Marketing: Fine-Tuning Your International Search Engine Results*. Indianapolis: Que Publishing.

Martin, C.A. and S.F. Witt (1989). Forecasting tourism demand: a comparison of the accuracy of several quantitative methods. *International Journal of Forecasting* **5**(1), 7-19.

Mc Cabe, S., C. Li and Z. Chen (2016). Time for Radical Reappraisal of Tourist Decision Making? Toward a New Conceptual Model. *Journal of Travel Research* **55**(1), 3-15.

Peng, B., H. Song and G. Crouch (2014). A meta-analysis of international tourism demand forecasting and implications for practice. *Tourism Management* **45**(1), 181-193.

Phillips, P. and P. Perron (1988). Testing for a Unit Root in Time Series Regression. *Biometrika* **75**(2), 335-346.

Polgreen, P.M., Y. Chen, D.M. Pennock and F.D. Nelson (2008). Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases* **47**(11), 1443-1448.

Ridderstaat, J. and R. Croes (2016). The Link between Money Supply and Tourism Demand Cycles: A Case Study of Two Caribbean Destinations. *Journal of Travel Research*, forthcoming.

Smith, G.P. (2012). Google Internet search activity and volatility prediction in the market for foreign currency. *Finance Research Letters* **9**(2), 103-10.

Song, H., S.F. Witt and T.C. Jensen (2006). Forecasting international tourist flows to Macau. *Tourism Management* **27**(2), 214-224

Song, H., S.F. Witt and T.C. Jensen (2003). Tourism forecasting: accuracy of alternative econometric models. *International Journal of Forecasting* **19**(1), 123-141.

Xiang, Z. and B. Pan (2011). Travel queries on cities in the United States: Implications for search engine marketing for tourist destinations. *Tourism Management* **32**(1), 88-97.

Yang, Y., B. Pan and H. Song (2014). Predicting hotel demand using destination marketing organization's web traffic data. *Journal of Travel Research* **53**(4), 433-447.

Yang, X., B. Pan, J. Evans and B. Lv (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management* **46**(1), 386-397.

Appendix

To construct the corrected aggregate intensity index, instead of conducting a joint search for the four keywords of interest (hotel Cyprus + туры кипр + hotel zypern + cypern resor) we act slightly in a different manner. In particular, we perform a separate search by adding sequentially all the keywords of interest (see the *compare multiple search terms* in the help function of Google trends). Acting this way we receive four separate series, which are directly comparable in terms of search volume (only one series receives the maximum value of 100). Having extracted the raw data for each search phrase, we adjust the series of interest with the market share correction factor, and then the four-separate series are added to form a single index. However, let's be more precise.

Let's assume that we wish to compare four keywords. The search volume for each one of the queries, for the period of interest (t=1,2,..,n), can be denoted as: $V_{1,t}^{q}$, $V_{2,t}^{q}$ $V_{3,t}^{q}$ and $V_{4,t}^{q}$, respectively or more compactly as $V_{i,t}^{q}$ (i=1,2,3,4). Let now $V_{\mathrm{e},t}^{q}$ to represent, at time t, the entire volume of queries, then the first step of the normalization process that Google implements is to express the search volume of each query ($V_{i,t}^{q}$ with i=1,2,3,4) as a fraction of the entire search volume of queries ($V_{\mathrm{e},t}^{q}$), that is:

$$r_{1,t},\ r_{2,t},\ r_{3,t} \text{ and } r_{4,t} \text{ or } \frac{V_{i,t}^{q}}{V_{\mathrm{e},t}^{q}} = r_{i,t} \quad (i=1,2,3,4) \tag{A.1}$$

Once the fractions have been estimated the four-normalized series can be constructed by multiplying each series with the scaling factor: $100/r^{*}$, where r^{*} is the maximum observed fraction among the fractions that come from the four-constructed series, that is:

$$\max_{r_{1,t},r_{2,t},r_{3,t} \text{ and } r_{4,t}\in R^{+}} \{r_{1,t}, r_{2,t}, r_{3,t}, r_{4,t}\} = \{r^{*}\} \tag{A.2}$$

The four normalized directly compared series can be denoted as: $S_{i,t}^{n} = (r_{i,t}/r^{*})100$, with i=1,2,3,4. Once we have at our disposal the normalized series (this is the form that the Google Trends facility deliver's the series), we may now implicate the market share correction factor for the intensity index that corresponds to Russia, say $S_{4,t}^{n} = (r_{4,t}/r^{*})100$. In particular, the volume adjusted series for Russia is now given by: $S_{4,t}^{n,\mathrm{va}} = r_{4,t}(m/r^{*})100$, where m is a scalar and represents the market share correction factor.

Given that the denominator is common, it comes that all four series can be added in order to form a unified, volume corrected, search intensity as follows:

$$S_t^f = \sum_{i=1}^{3} S_{i,t}^n + S_{4,t}^{n,\text{va}} \quad \text{or} \quad S_t^f = \frac{\frac{\sum_{i=1}^{3} V_{i,t}^q}{V_{\text{e},t}^q} + \frac{mV_{4,t}^q}{V_{\text{e},t}^q}}{r^*} 100 \tag{A.3}$$

From A.3 it is obvious that it is possible to receive series that are scaled above 100. The difference of A.3 from the standard case, where the search of multiple keywords delivers a unique *SII* with a maximum value of 100, lies on the fact that a) the scaling factor, r^*, is now different and b) the market share correction factor is introduced. Given that both factors are simple scalars, the resulted series from the two alternative approaches are expected to illustrate almost identical evolution over time and therefore, a high degree of correlation. In other words, both approaches deliver qualitatively similar results.