HELLENIC OBSERVATORY
European Institute

# Arrivals of Tourists in Cyprus:

# Mind the Web Search Intensity

**Theologos Dergiades, Eleni Mavragani, Bing Pan**

**GreeSE Paper No.107**
**Hellenic Observatory Papers on Greece and Southeast Europe**

**February 2017**

All views expressed in this paper are those of the authors and do not necessarily represent the views of the Hellenic Observatory or the LSE
© Theologos Dergiades, Eleni Mavragani, Bing Pan

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

# TABLE OF CONTENTS

ABSTRACT ____ iii

**1. Introduction** ____ 1

**2. A brief review of the literature** ____ 7

**3. Methodology** ____ 10

3.1 Standard Granger Non-Causality Testing ____ 10

3.2 Frequency Domain Non-Causality Testing ____ 11

**4. Data and preliminary econometric analysis** ____ 13

4.1 Data sources ____ 13

4.2 Preliminary econometric analysis ____ 20

**5. Empirical results** ____ 22

5.1 Predictive Power of the Web Search Intensity per Country ____ 22

5.2 Aggregate Predictive Power of the Web Search Intensity ____ 28

**6. Discussion of findings and policy implications** ____ 33

**7. Conclusions** ____ 36

**References** ____ 39

**Appendix** ____ 42

# Arrivals of Tourists in Cyprus: Mind the Web Search Intensity

Theologos Dergiades[#], Eleni Mavragani[*], Bing Pan[†]

## ABSTRACT

This paper validates the raison d'être of the effortlessly recovered web Search Intensity Indices (SII) for predicting the arrivals of tourists in Cyprus. By using monthly data (2004-2015) and two causality testing procedures we find, for properly selected key-phrases, that web search intensity (adjusted for different languages and different search engines) turns out to convey a useful predictive content for the arrivals of tourists in Cyprus. Additionally, we show that whenever the prevailing shares of visitors come from countries in different languages, then the identification of the aggregate SII becomes complex. Hence, we argue that blindly using key-phrases to identify an aggregate SII is like an immersion into the unknown, since two sources of bias (the language bias and the search engine bias) are fully neglected. Given the importance of the tourism sector in the total economy activity of Cyprus, our findings might prove to be quite useful to governmental agencies, policy makers and other stakeholders of the sector when their purpose is to allocate effectively the existing limited resources, and to plan short- and long-run promotion and investment strategies.

**Keywords:** Cyprus tourism product; web search intensity; predictive content

[#]**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.mavragani@ihu.edu.gr;
**†Bing Pan**, Department of Recreation, Park & Tourism Management, Pennsylvania State University, e-mail: bingpan@psu.edu

# Arrivals of Tourists in Cyprus: Mind the Web Search Intensity

## 1. Introduction

Over recent years the availability of freely delivered data from copious web sources (social media, search engines, etc.), sparked a new strand in the empirical literature, the so-called real-time economics.[1] In one of the earliest studies in economics,[2] that actually inaugurated the field, authored by Hal Ronald Varian (Chief economist at Google) and Hyunyoung Choi (senior economist at Google), there are vigorous signs that properly selected query indices (provided by Google) are useful in prognosticating the activity in different economic sectors, such as the automobile industry and the tourism market.[3] This corner stone study has triggered a flurry of scientific publications that use web-related data which aim to explain upcoming trends in various markets. Among others, empirical applications have been conducted for several foreign exchange markets, stock markets, sovereign bond markets, labor markets or even real estate markets. In all the above markets, there is credible evidence that web-related data offer added value when it comes to predicting upcoming events.

[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 at later time it has been published as Choi and Varian (2012).

Paying attention on tourism markets, an essential desideratum for practitioners and policy makers, for several reasons, is the accurate prediction of the demand related to tourism products of interest. It is widely accepted that truthful forecasts provide valuable aid for: a) the development of long-run marketing strategies, b) the formation of competent pricing policies, c) the appropriate scheduling of investments into the sector and d) the effective allocation of the limited resources. Hence, the need for new leading indicators that may contribute to predicting, both effectively and timely, consumer preferences, is persistent and more than justified. Given that nowadays, web search engines constitute the major workhorse in scheduling vacations, these can be seen as a new source of information that may help us to improve our understanding with respect to the consumption of the tourism product. However, it is well recognized that a fresh source of information may not be a competent leading indicator, while a competent leading indicator may not be new. Therefore, common practice dictates that extensive empirical testing is more than imperative before the adoption of such sources of information as leading indicators. This study, concentrating on Cyprus, evaluates the impact of the relevant web search intensity, captured by Google, on the consumption of the tourism product. Accurate forecasts of tourism demand in the case of Cyprus are of major importance since the total economic activity of the island heavily relies on the tourism industry. According to the latest KPMG report (April, 2016), the overall contribution of the tourism industry to the economy, for the year 2014, is more than €3 billion, which corresponds to 21.3% of the GDP. Projections for the next 10 years show that the absolute contribution of the tourism sector is

expected to experience a steady annual growth with its magnitude to be somewhat below 5%. By 2025, the relative contribution of the tourism sector to overall economic activity is anticipated to be 25.5%.[4] Additionally, we concentrate on the search engine of Google for two major reasons. As stated in Yang et al. (2015), Google is the most popular search engine globally, with a market share equal to 66.7%, and at the same time Google provides historical information on the volume of the conducted queries.

Another aspect that makes Cyprus an ideal candidate for study is the observed arrival shares by country. While most of the studies focus on destinations where the dominant market share of the arrivals corresponds to English-speaking countries, this is not the case for Cyprus.[5] More than 70% of the arrivals in Cyprus come from the United Kingdom (UK, hereafter), Russian Federation (Russia, hereafter), Greece, Germany and Sweden. Such composition in the origin of the arrivals undeniably complicates the process we need to follow in order to identify the related aggregate web search intensity from the Google search engine. A natural difficulty in tracking the aggregate web search intensity for a specific travel destination is the selection of the appropriate language. Indisputably, English is the prevailing language in the Internet (873 million users) followed by the Chinese language (705 million users).[6]

---

[4] The 2016 tourism market report of KPMG for Cyprus, is available at: https://www.kpmg.com/cy/

[5] 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).

[6] Numbers refer to November 2015 (http://www.internetworldstats.com/stats7.htm); accessed June 2016.

Given the dominance of the English language, an apparent question is whether the aggregate web search intensity based on the usage of keywords from the English language, would be adequate to reveal the aggregate interest for a specific travel destination. The answer is yes, if and only if all the arrivals to the destination of interest come solely from English-speaking countries (or to be more precise, if all the visitors perform their web searches in English). In any other case, the aggregate constructed index will be biased.[7] In particular, as the share of the total arrivals from English speaking countries decreases progressively relatively to non-English-speaking countries, the quality of the identified aggregate web search intensity index (based only on English keywords) is expected to deteriorate in an analogous manner. Hence, failure to take into account, for our entire sample period, all the languages that correspond to the respective source markets of the destination under investigation, will give rise to the first source of bias, let's call it language bias.

To this point, we need to stress that for most of the times the construction of the aggregate web search intensity index (based on Google) for the tourist product of a country, especially when it is about a popular destination, cannot be to an absolute degree free of the language bias. The presence of the language bias, in such cases, is attributed to an inherent feature of the Google trends facility. In particular, the facility does not deliver data if the search volume for the keyword of interest is relatively small. Immediately, it becomes apparent

[7] 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.

for the source markets with small shares in the arrivals (implying small search volume), that the construction of the corresponding web search intensity index (SII, hereafter) is a non-feasible task. Consequently, even if we wish, we cannot take into account the search volume from all the languages in order to construct a unified aggregate index. As the cumulative market share in the tourism product, for the source markets that there is not enough volume to construct an index, increases, the quality of the aggregate index is expected to fade. Overall, clearly the language bias is not a question of presence or absence, but rather it is a question about its various degrees.

Even if at some point of our sample (e.g. in the beginning) all the major source markets use the same language we continue to run the risk of encountering this so-called language bias, since there is no guaranty that this will be the case at any other point of time. New source markets, using different languages, progressively may earn a greater arrival shares thus reducing or displacing the share of existing source markets. In other words, misleading web search intensity may be received once we fail to take into account source markets that gradually earn larger shares in the arrivals. For example, let's assume the following: a) over a long-period German-speaking countries are consistently the dominant source markets for a destination but with a declining share over-time and b) Russian-speaking countries initially had a small share in the arrivals (small enough in order not to have enough search volume) but with an increasing trend over time. In the above example, if we extract the web search intensity solely based on German keywords, then the aggregate web search intensity for the destination of interest is misleading. Therefore, we need to examine the dynamic evolution of the shares that

each source market has. Overall, it becomes apparent; that accurate identification of the aggregate web search intensity necessitates knowledge of all those source markets that contribute to the total arrivals for the entire sample of investigation.

In our effort to measure web search intensity, another source of bias may result from the usage of the Google trends facility itself, if Google is not the dominant search engine in the source market of interest; let's call it search engine bias. In such cases, the measured volume of queries by the Google trends facility underestimates the true volume of relevant queries, failing this way to convey the precise interest of the users and its evolution over time. Obviously, the bias of the SII delivered by the Google trends facility will be zero if the share of Google for the total number of web searches in the source market is 100%, and increases as the above share of Google decline.

By using two alternative causality testing techniques (the first test takes place in the time domain while the second one in the frequency domain) and introducing a simple way to select appropriate key-words, we investigate the predictive power of Google's SII towards the arrivals of tourists in Cyprus at an aggregate and disaggregate level. The findings from our analysis are the following: a) All the country-specific SII are highly significant in predicting arrivals from the respective source market, b) the presence of both sources of bias, the language bias and the search engine bias, render as ineffective the aggregate SII to predict the total number of tourist arrivals and finally, c) once we consider the two sources of bias, the corrected aggregate SII now turns out to convey a precious predictive content in relation to the arrivals that come from

the respective market sources. In practice, these findings validate the usage of the web search intensity as an important leading indicator for the demand of the tourism product. At the same time, we make clear that when it comes to predicting the demand of the tourism product, then it is preferable that this task is conducted at a disaggregated level. Of course, we do not support to ostracize approaches that attempt to predict arrivals at an aggregate level. Instead, we argue that aggregate SII are exposed to two significant sources of bias and hence special handling is needed.

Our study has the following structure: Section 2 briefly reviews the literature devoted to the broad field of market predictability 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. Finally, Section 6 concludes.

## 2. A brief review of the literature

As already mentioned, the increasing availability of data revealing consumers' online activities, has led to several projects that take advantage of these data with purpose to forecast upcoming events in the respective markets. Yang et al. (2014) recognize that the major advantages of such data lie behind in the fact that: a) can reveal preferences in real time, b) can be provided in relatively high frequency (e.g. daily or weekly) and most importantly, c) can depict changes in

consumers' preferences, providing this way a solution to the inherent problem which is encountered often in traditional univariate time-series models (e.g. ARMA models).[8] Empirical applications using web-related data can be found for several markets. For instance, Smith (2012) shows that the online search intensity, as captured by Google, explains significantly movements in the currency markets. Joseph et al. (2011) using again data from the Google trends facility (former Google insights) support that searching for stock tickers helps to forecast abnormal stock returns as well as the respective trading volume. Da et al. (2011) by using a sample of 3000 stocks 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 key-word Grexit explains future price movements for the 10-year government Greek bonds.

Turning now to the tourism market, there are long-lasting efforts from researchers to provide accurate forecasts for the arrivals of tourists implementing a wide range of techniques. According to Peng et al. (2014), who provide a very comprehensive review of the relevant literature, these techniques can be classified into two broad categories. The first includes studies that use time-series econometrics, while the second one embraces studies that implement various artificial intelligence methods. Within the former category, numerous econometric methods are implemented ranging from very simplistic

[8] Univariate time-series fail to provide robust forecasts, once sudden one-off events take place and alter the pattern of the series.

univariate specifications (mainly over the early literature; see among others Geurts and Ibrahim, 1975; or Martin and Witt, 1989), to relatively more advanced multivariate specifications (mainly over the most-recent literature; see among others Halicioglu, 2010; or Bangwayo-Skeete and Skeete, 2015). Similarly, for the latter broad category, several alternative approaches can be identified throughout the literature. These approaches range from the very popular artificial neural networks (see among others, Burger et al., 2001) to the more recent genetic algorithms (see among others, Chen and Wang, 2007).[9]

Given the substantial number of web users who seek information through established search engines before taking a trip (Fesenmaier et al. 2011), a natural way for the researchers to proceed is to take advance of these valuable signals. Despite the large volume of studies dedicated to forecast the demand of the tourism product, so far there is relatively a small number of studies that make use of web search intensity data. Xiang and Pan (2011) adopting a qualitative approach and focusing on U.S. cities, diagnose that "the ratio of travel queries among all queries about a specific city seems to associate with the touristic level of that city". Choi and Varian (2012), focusing on Hong-Kong, confirm, based on a standard dynamic regression specification, that search intensity data (provided by Google) at a disaggregated level (for nine source markets-countries) are indeed useful predictors of tourists' arrivals from each respective market.

Yang et al. (2015) based on query volume data from two search engines widely used in China (Google and Baidu), affirm (implementing an ARMA

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

-Autoregressive Moving Average- specification and the standard Granger non-causality test) that when it comes to predicting the number of visitors in Hainan (a Chinese province), both sources contribute significantly in decreasing forecasting errors. Bangwayo-Skeete and Skeete (2015) by directing their interest to five Caribbean destinations (Jamaica, Bahamas, Dominican Republic, Cayman and St. Lucia), they support that after the appropriate construction of the Google search volume indicator then significant gains in forecasting tourists arrivals can be observed. Bangwayo-Skeete and Skeete (2015) conduct their analysis by implementing a simple AR-MIDAS[10] model, a SARIMA model (Seasonal Autoregressive Integrated Moving Average) and finally; a benchmark AR (Autoregressive) model. The former model appeared to perform better in the majority of the conducted pseudo-forecasting experiments.[11]

# 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.

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

[10] The MIDAS estimation approach refers to the case where data with mixed frequencies are involved.
[11] For studies than implement web data aiming to predict the demand for hotels see Yang *et al*. (2014).

where, $\mathbf{Z}_t = (A_t\ G_t)^T$ is a $2\times1$ vector of stationary variables, $\mathbf{\Theta}(L)$ is a $2\times2$ matrix of lag polynomials and finally, $\boldsymbol{\varepsilon}_t$ is a $2\times1$ 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 $\omega$ 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 $\omega$, $\left|\Psi_{12}(e^{-i\omega})\right|^2$ 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:

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

where, $1/c_{22}$ is the positive[12] 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 $|\Theta(L)|$ is the determinant of $\Theta(L)$. Therefore, the non-causality hypothesis at frequency $\omega$ from $G_t$ towards $A_t$ is not rejected whenever the following holds:

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

where, $\theta_{12,k}$ is the upper right element of the $\Theta_k$ matrix. Subsequently, the set of restrictions that should be imposed are:[13]

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

The empirical procedure of the B&C approach lies on the validity of the above presented linear restrictions. For brevity reasons 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:

$$A_t = \alpha_1 A_{t-1} + \ldots + \alpha_p A_{t-p} + \beta_1 G_{t-1} + \ldots + \beta_p G_{t-p} + \varepsilon_{1t} \tag{7}$$

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

[13] 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.

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

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

B&C investigate the validity of the linear restrictions illustrated in *eq*. (8), for frequencies $\omega$ that receive values within the interval of $(0,\pi)$, by comparing the obtained Statistic with the 0.05 critical value of the $\chi^2$ 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 data that capture the web search intensity for properly selected keywords. The range of our sample is January, 2004 to April, 2016 (148 obs.) and is dictated solely by the data availability. 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 obs.).[14] For the selected sample, in order to extract the *SII* related to the tourist product of Cyprus, we use the Google trends facility.[15]

[14] See: http://www.mof.gov.cy/mof/cystat/statistics.nsf; accessed June 2016.
[15] See: http://www.google.com/trends/; accessed June 2016.

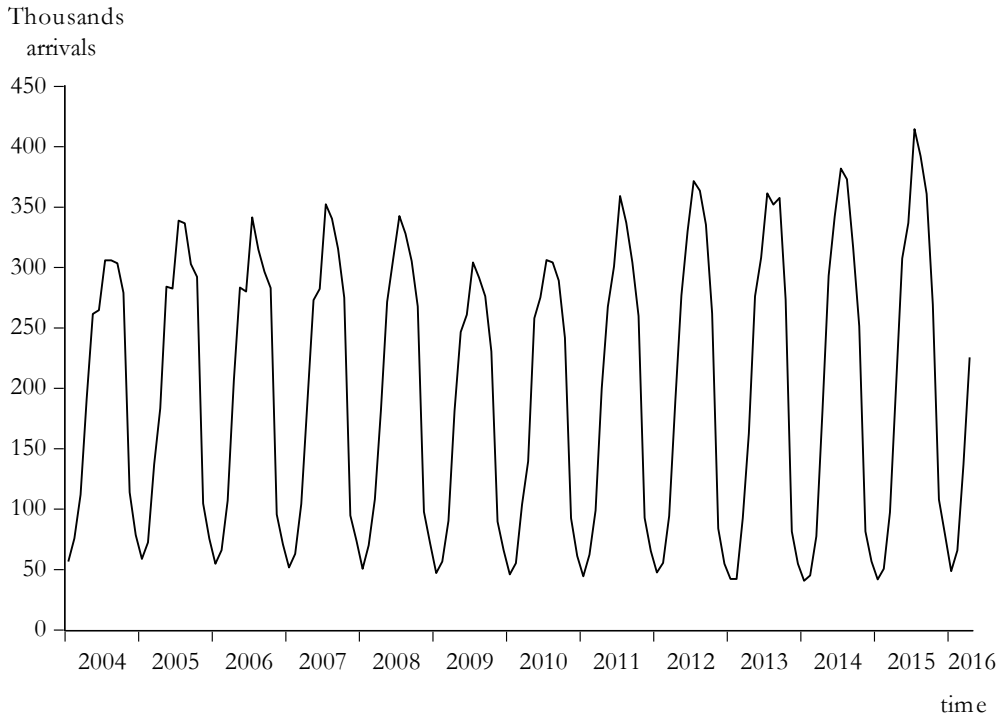


**Figure 1.** Monthly arrivals of tourists in Cyprus

From the discussion in the introduction section, it becomes apparent that the formation of a bias free aggregate index, intended to capture the entire web search intensity for a destination, is a complex and challenging task and in several cases almost impossible to be constructed. Therefore, before we proceed, it is more than imperative to take into account the existing sources of bias. In order to overcome the first source of bias, we need to disentangle the aggregate number of tourist arrivals in Cyprus by country of origin. Acting this way, we will be able to identify the key source markets and therefore to specify the corresponding languages. The market share in the total arrivals per country is illustrated in Fig. 2. Visual inspection of Fig. 2 suggests that five countries are the main source markets, representing jointly 74.08%[16] of the market share, while the respective share for all the other

[16] The reported value is the average share of the total monthly arrivals for the period of study (2004-2015).

countries is 25.92%. The major source markets countries are the following: UK (45.44%), Russia (10.13%), Greece (7.68%), Germany (7.18%) and Sweden (3.65%). Hence, if we wish to track the web activity related to the arrivals in Cyprus, our attention has to be concentrated on the respective languages (English, Russian, Greek, German and Swedish).

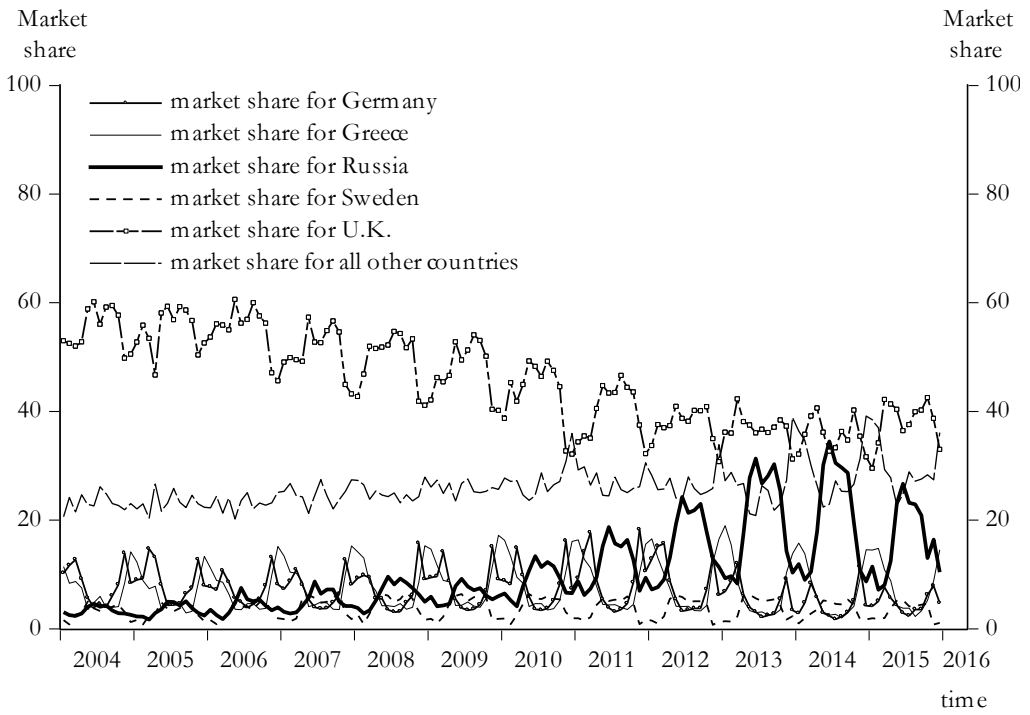


**Figure 2.** Markets shares in tourists arrivals for Cyprus per country

Another interesting feature of the data in Fig. 2, is that the market shares for each country evolves quite differently. In more detail, there is a distinct negative and significant trend in the market share of the U.K., with the average monthly market share to shrink from 55.17% in 2004 to 38.01% in 2015 (annual compound growth: -3.3%).[17] A similar negative and significant trend is observed for Germany. The German share in 2004 was 7.17% reaching the value of 4.85% in 2015 (annual compound

[17] 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.

growth rate: -3.5%). On the other hand, a highly significant positive trend is observed for the Russian market share. The average market share increased from 3.29% in 2004 to 16.32% at the end of the sample (annual compound growth rate: 15.7%). Positive and significant trend also experiences the market share that is formed by all the other countries. The average market share increased from 23.18% in 2004 to 29.63% in 2015 (annual compound growth rate: 2.3%). Finally, Greece and Sweden indicate no significant time trend.

Having established the languages of interest, it is essential to identify for every language the appropriate key-phrases that are directly related to a potential visit in Cyprus. Incontestably, we are powerless to know the infinite number of thinkable key-phrases that someone may use in order to schedule a visit to Cyprus. However, what we know with relative certainty is that the majority of the performed searches are expected to contain the term *Cyprus*. Taking as a starting point the destination name (Cyprus), the strategy that we pursue to identify appropriate key-phrases involves the following steps: 1) by first selecting the source market of interest (e.g. U.K., we actually determine the geographical location for the conducted searches), we type the term *Cyprus* in the Google correlate facility[18] to attain other queries that illustrate similar patterns (the similarity is ascertained through a simple correlation coefficient). From the delivered queries which are 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) Shifting from the Google correlate to the Google

[18] See: https://www.google.com/trends/correlate

trends facility, we extract on a monthly frequency the *SII* for the key-phrase identified in the previous step.[19] We verify the validity of our chosen key-phrase by examining the top related queries as these are suggested by Google trends. If the vast majority of the related queries imply interest for visiting Cyprus, we may argue in favor of our key-phrase. 3) For those cases where in step 1 our initial key term (e.g. *Cyprus*) does not deliver key-phrases 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.

By applying the above strategy for all the languages of the major source markets, we can extract the web *SII* for each source market. Starting from the 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 key-phrase *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 key-phrase since they imply direct interest to visit Cyprus.[20] The finally extracted index in monthly frequency is presented in Fig. 3a below.[21] Implementing the same strategy for the remaining source markets, we end up with the following key-phrases. For the Russian market, the identified key-phrase is *туры кипр* (tours Cyprus) and the respective index is illustrated in Fig. 3b. For the German market, the key-phrase is *hotel zypern* (hotel Cyprus) and depicted in Fig. 3c, and for the Swedish

[19] The search term is not enclosed in quotation marks.
[20] The relevant queries in order are: hotel in Cyprus, Paphos Cyprus, Paphos, hotels Cyprus, Cyprus holidays, Portaras Cyprus, Portaras.
[21] Figure 3, along with the web *SII*, also presents the arrivals for each country.

market the key-phrase is *cypern resor* (Cyprus travel) shown in Fig. 3d. Finally, the strategy failed to deliver a key-phrase that expresses an intention to visit Cyprus for the case of Greece. We tried key-phrases 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.[22]

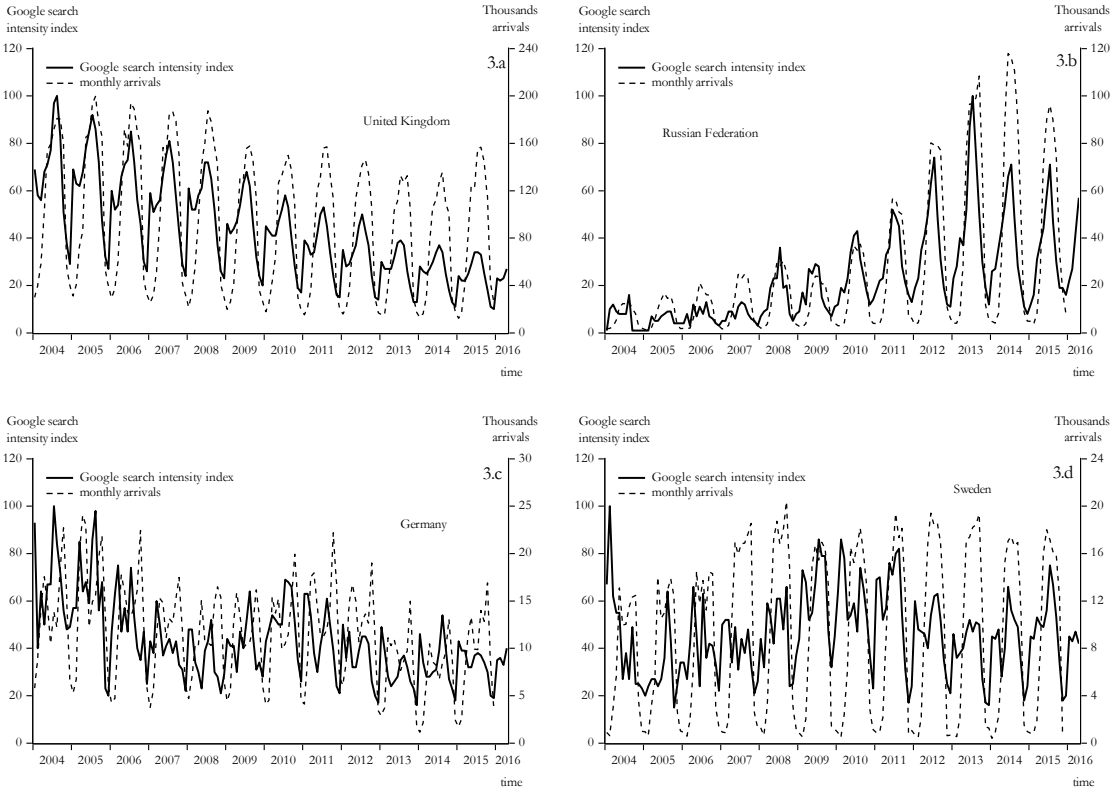


**Figure 3.** Google *SII* and tourists arrivals per country.

[22] 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.

To this point, it is worth mentioning that in Russia the second source of bias may be essential. Google is not the dominant search engine in the market. The search engine called Yandex operates, on average over the period of our study, approximately 60% of the market, while Google's respective share is about 25%.[23] Therefore, 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 key-phrase (туры кипр) from Google, we execute the same identification strategy by using a similar facility offered by Yandex (relevant phrases). The delivered key-phrase is *туры на кипр*, which is almost identical to the key-phrase identified by Google Trends (туры кипр). Although both search engines deliver almost identical key-phrases, their evolution overtime may be dissimilar across the two engines. To assess this possibility we take advantage of another feature offered by Yandex, which delivers the absolute number of searches. 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.[24] Therefore, we may argue that the *SII* obtained from the Google, despite its' relatively small share in the market, reveals quite accurately the true pattern over time. Nevertheless what is still an issue with the case of Russia is the fact that in Google the true volume of searches is underestimated.

Since we have discussed the main reasons for the occurrence of biased search intensity measures, we may now extract the aggregate *SII* based on the four major source markets.

[23] See www.liveinternet.ru.
[24] The absolute number of a search in Yandex is available, on monthly basis, for the past two years.

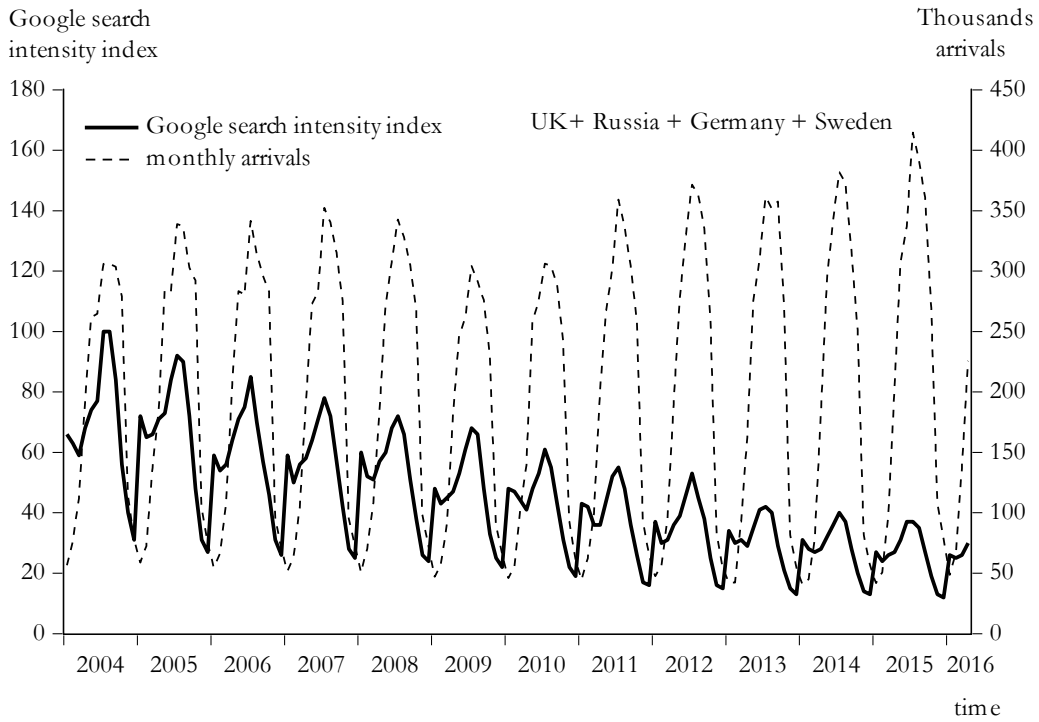


**Figure 4.** Aggregate Google *SII* and total arrivals of tourists.

Thus, to construct aggregate *SII* we combine all the previously identified key-phrases to a single search.[25] The constructed index is presented in Fig. 4.

## 4.2 Preliminary econometric analysis

By simply observing the data (country specific and aggregate) related to the arrivals as well as to the web *SII* (see Figs. 3 and 4), the existence of a seasonal variation is evident. The well-known adverse effects of seasonality in statistical inference dictate that a seasonal adjustment procedure needs to be performed. 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. The

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

seasonally adjusted series to be used in our analysis are illustrated in Fig. 5.

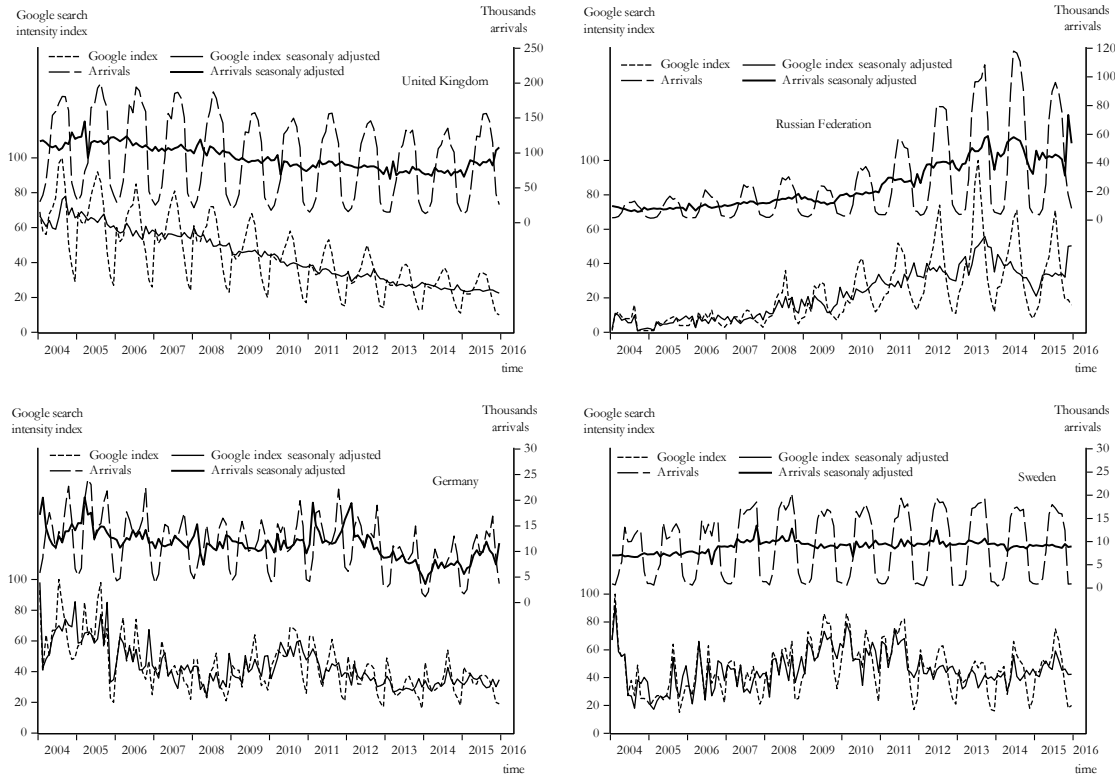


**Figure 5.** Seasonally adjusted series for the *SII* and the arrivals per country.

Another issue that may lead to biased causal inferences is the presence of unit-roots in the data. As a result, the stationarity properties of all the de-seasonalized series are examined by conducting the well-know Phillips and Perron (1988) test, with and without the presence of a deterministic linear trend. The tests results are displayed in Table 1. From these results, 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 (fail 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 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 characterized as trend stationary, it is incorporated into our analysis after removing the linear time trend.

**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 Country

To evaluate the predictive content of the constructed Google *SII* for the countries of interest towards the respective arrivals, we implement two alternative causality tests. On the one hand, we conduct the standard

linear Granger non-causality test in the time domain and on the other hand, the B&C non-causality test in the frequency domain. The proposed methodology by B&C integrates some advantages that cannot be traced in the classic Granger non-causality test. First, the B&C test allows us to identify whether a verified causal relationship is short-run or long-run and second the same test is capable of revealing potential non-linear causal relationships. Overall, working within the frequency domain could help us to disclose causal relationships that may not be distinguishable in the time domain.

The country-specific results for the standard linear Granger test are illustrated in Table 2. Consistently, the hypothesis of no predictability running from the *SII* to the arrivals is 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 but this time at the 0.01 significance level. Turning now to the opposite direction, 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*.

Provided that the B&C test may deliver wealthier information with respect to the predictive power that one variable may carry, our attention shifts to the frequency domain. The 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 wave lengths 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. When our attention goes to Germany (see Fig. 6.c) the non-causal inference is now slightly altered. More specifically, predictability is not verified for the medium cyclical components but rather for the low and the high cyclical components of the series ($\omega \in [0, 0.75] \cup [1.88, \pi]$).

**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.

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 in any significant manner 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 wave lengths of less than 3.2 months. In other words, for the case of Sweden short-run bidirectional predictability is established. Overall, we may argue that our findings from the B&C test are qualitatively similar to those of the linear Granger non-causality test.

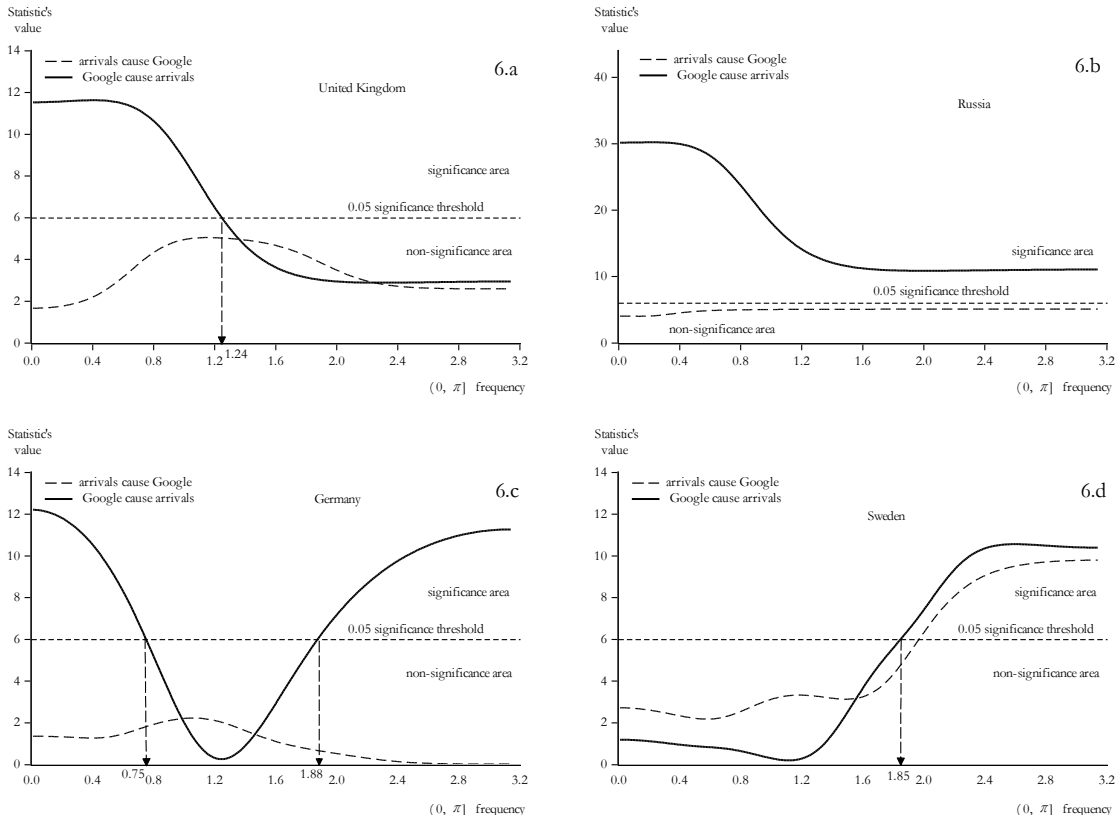


**Figure 6.** B&C Granger non-causality test per country

The fact that the predictive content of the constructed Google SII (with respect to the arrivals) is dissimilar among the examined countries comes as no surprise. According to Mc Cabe *et al*. (2016), national cultures integrate idiosyncratic features which affect the search

information behaviour. As a result, tourist's decision making process for visiting a destination follows a different path for different cultures. 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 behaviour resulting this way to clearly distinct consuming patterns.

Heterogeneous consuming patters imply variation in the decision-making lead times, and this is what our results reveal. In particular, we find for the three major source markets (UK, Russia and Germany) that 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,[26] which investigates the holiday travel patterns of German tourists, 70% of the travellers 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 not contradict our empirical findings. An exception to the observed relatively long-run trip planning behavior, are the tourists that come from Sweden, since we verify 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

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

therefore personal search for additional information may take place only few months prior their trip.[27]

An inherent weakness of the implemented Granger non-causality tests is that while both are unable to reveal whether the variables of interest are connected in a positive or negative manner. In other words, we would be very much interested in knowing the response of one variable when a positive shock takes place in another variable. Such knowledge is valuable since we gain significant insights about the nature of the identified causal relationship. A natural way to deal with the above-mentioned issue is to conduct impulse response analysis. The Cholesky defined accumulated impulse response functions of interest along with their associated $\pm 2$ standard errors confidence bands are presented in Figs. 7a to 7d. In more detail, Fig. 7.a shows, 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 (confidence bands do not include 0) 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. Again, the results are in concordance to our findings from the B&C tests. Finally, for the case of

[27] 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.

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. Yet again, 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.

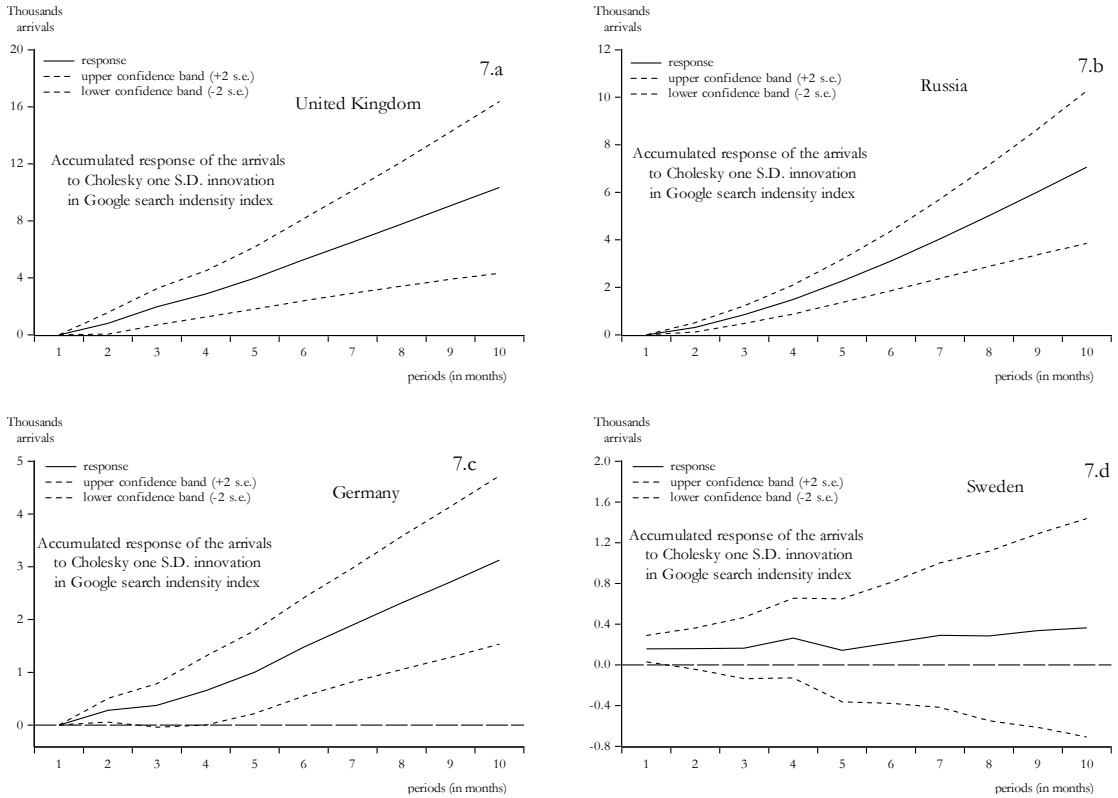


**Figure 7.** Impulse response functions per country

## 5.2 Aggregate Predictive Power of the Web Search Intensity

Having completed the country-specific analysis for the four major source markets of tourist arrivals in Cyprus our attention now shifts to the predictive content encompassed in 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* [28] (see the unit-root test results in Table 1) we conduct the standard Granger non-causality test. The linear non-causality test results for the aggregate series are illustrated in Table 3. Clearly, 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) for all the conventional levels of significance. 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 once we examine the same hypothesis 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 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,

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

while there is strong evidence of predictability at a country level, this predictability vanishes once we use the aggregate data.

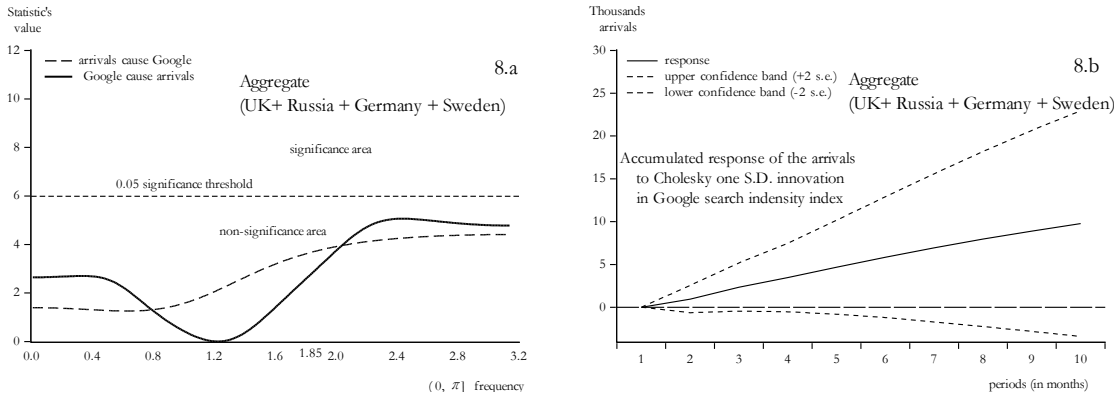


**Figure 8.** B&C test and impulse responses at the aggregate level.

The verified lack of predictability for the aggregate series it comes as no surprise given the biases discussed within the introduction section. Therefore, to reassess the predictive content of the aggregate index, we construct a corrected version which takes into account, as this is possible, the two discussed sources of bias. In particular, the steps we follow to construct the aggregate corrected index are two. First, to overcome the so-called *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). Acting this way, we ensure that the four corresponding languages, used to construct the aggregate index, reflect truly the web search intensity which linked to the arrivals from these countries.[29]

Second, to rectify our aggregate index from the so-called *search engine bias*, which is present in the case of Russia, we need to correct for the

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

low market share of Google in the Russian Internet market. To perform such a correction it is necessary to reconsider the way we followed in section 3.1, in order to extract the aggregate *SII*. Instead, of constructing a unified index by combining our four key-phrases (hotel Cyprus + туры кипр + hotel zypern + cypern resor), we extract four separate indices (one for each key-phrase) 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), underestimate the true volume of searches.

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 key-phrase туры на кипр) correlates strongly to the index delivered from Google (for the key-phrase туры кипр), we may use the ratio of the respective market shares ($S_2/S_1$) as a volume correction factor. Once we multiply Google's web *SII* that corresponds to Russia with the volume 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 volume 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.

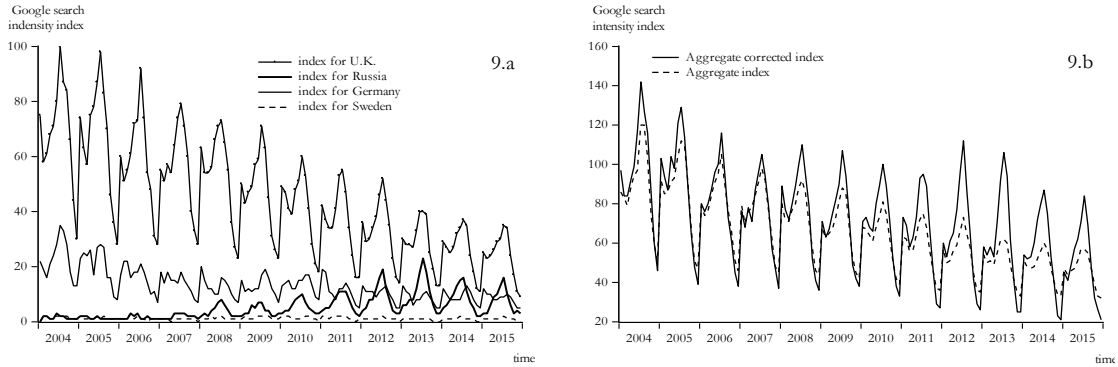


**Figure 9.** Search volume of the selected key-phrases and the aggregate corrected index.

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.[30] 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. Moving now to the B&C test, we receive qualitatively analogous inference. The results show that 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 response function is consistently positive with the confidence bands not to include the zero value (See Fig. 10.b).

[30] 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.

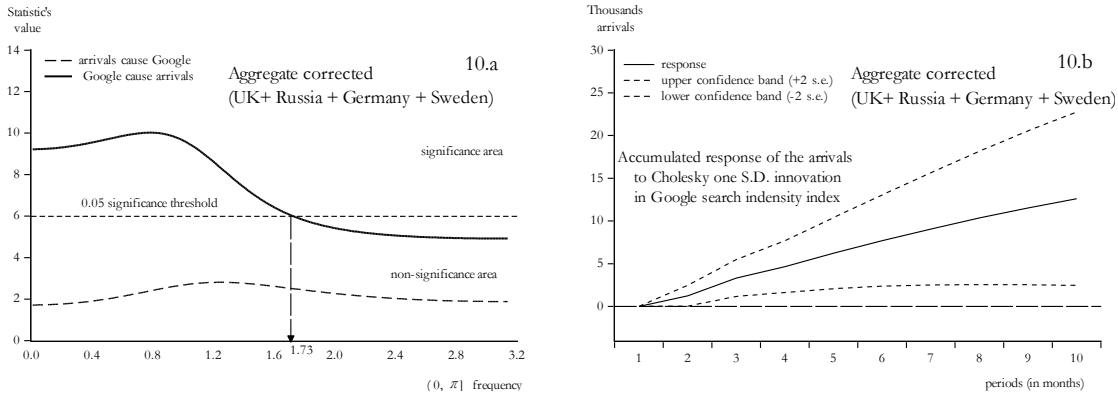


**Figure 10.** B&C test and impulse responses at the aggregate level (corrected *SII*).

# 6. Discussion of findings and policy implications

Undeniably, search engines are among the most popular online planning sources for travelers (Google Travel Study, June 2014, Ipsos MediaCT[31]). The outcome of the present study, focusing on search engines, offers valuable insights to the understanding of those travelers' behavior who plan to visit Cyprus. Overall, we argue that carefully identified web search activity indices encompass early signals that can assist significantly to the prediction of tourists' arrivals in Cyprus. The predictive content of the web search activity is verified for the four major source markets of Cyprus (UK, Russia, Germany and Sweden), which jointly add approximately 70% in the total arrivals.

As mentioned already, the decision-making process of travelers from different countries is not following a uniform pattern as national cultures influence the search information behavior (Mc Cabe *et al*., 2016; Gursoy

[31] See: https://storage.googleapis.com/think/docs/2014-travelers-road-to-decision_research_studies.pdf

and Umbreit, 2004). Our results come to confirm the above statement. British travelers search for information in the web at least five months before a visit to Cyprus, while Russians are searching in the web for holidays to Cyprus throughout the year. On the other hand, Germans illustrate also an interesting purchasing behavior. Germans search the web for information in two distinct terms; at least eight months prior their trip to Cyprus and again three months prior to their arrival at the island. Finally, Swedes search for information to visit Cyprus only during the last three months. This may be due to the existence of reliable travel agencies, which act as intermediaries in booking holidays and Swedes tourists make their personal search for information only the very last months prior their trip. Obviously, the above unveiled behavioral patterns can be valuable to Cypriot tourism authorities to plan their communication and pricing strategy accordingly.

In more detail, the findings of the present study might be proved quite useful to governmental agencies, stakeholders of the sector and Destination Management Organizations (DMO's), when their purpose is to identify the upcoming future demand. Accurate prediction of the tourism demand is vital to the tourism industry, especially when tourism constitutes a key driving force for one country's economic growth as is the case with Cyprus. The tourism sector in Cyprus, has a large share in the national income (more than 20%; see Clerides and Pashourtidou, 2007) and is a major job-creator (directly or indirectly). The Cypriot government, with improved knowledge on the total magnitude of the arrivals, can assess more accurately sectors' contribution to the economy. Therefore, projections about the country's future growth path

involve lower uncertainty (see Clerides and Adamou, 2010). Moreover, our results may help the policy makers of the Cyprus Tourism Organization to attend tourism exhibitions and to conduct advertising campaigns on the right timing for each source market. In other words, by knowing the proper time that every action needs to take place, policy makers could achieve cost savings and efficient allocation of the resources spent. Furthermore, prediction of decreased arrivals from one destination, can assist in adopting actions of promotion in other promising markets. This way, in reduced arrivals can be reversed by increasing the last-minute bookings.

Overall, knowledge about the upcoming trends in the arrivals along with the unveiled behavioral patterns of the tourists from the major source markets, may help the tourism sector to improve the quality of the provided services and will allow potential investors to plan their projects (e.g. development of infrastructures) with greater certainty. The government as well as all the stakeholders of the sector would be more informed in order to allocate effectively the existing limited resources and to plan short and long-run promotion and investment strategies.

# 7. Conclusions

Over the recent years, Google Inc. provides data on the volume of queried subjects conducted in their quite renowned search engine known as Google. This policy initiated an outbreak of scientific projects aiming to explain upcoming trends in various markets based, of course, on these data. As search engines constitute a leading tool in scheduling vacations, the communicated signals from these engines can be exploited to improve predictions on the consumption of the tourism product. Under this prism, we examine the predictive power of a relevant web *SII*, as these are 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 common language (see for instance: Bangwayo-Skeete and Skeete, 2015), our work is the first that focuses on a destination with a multilingual set of source markets. However, once the arrivals come from countries in different languages then the identification of the aggregate *SII* grows into a laborious task. Hence, the big red flashing light of this study is that blindly using key-phrases to identify aggregate *SII* may give rise to two significant sources of bias, the *language bias* and the *search engine bias*, weakening this way the predictive power of the respective index.

We test our hypothesis by using monthly data (2004-2015) for Cyprus and by conducting two Granger non-causality tests. These two tests are the standard linear Granger non-causality test as well as the B&C non-causality test. We implement the B&C test since it encompasses some advantages over the standard linear Granger non-causality test. In

particular, it allows us to distinguish between short-run and long-run causality, and at the same time it helps us to disclose causal relationships that are not distinguishable in the time domain. By introducing a simple way to select appropriate key-words and working within the above framework our findings are summarized as follows: a) country-specific *SII* (for U.K., Russia, Germany and Sweden) are highly significant in predicting positively (higher intensity leads to more arrivals) the arrivals from the corresponding source markets, under both testing techniques, b) the initially constructed aggregate *SII*, without considering the *language bias* as well as the *search engine bias*, proves inadequate to predict the total number of arrivals, again under both testing techniques, and finally c) once we construct a corrected version of the aggregate *SII*, taking into account the two sources of bias, then it turns out that the corrected version predicts in a significant and positive manner the arrivals that come from the respective countries (UK, Russia, Germany and Sweden).

Overall, our study not only 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. We flag emphatically, for destinations that accept arrivals from countries in different languages, that the formation of a bias free 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 sources of bias and hence special handling is needed. In failing to account for these biases, 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.

Clerides, S. and N. Pashourtidou (2007). Tourism in Cyprus: Recent Trends and Lessons from the Tourist Satisfaction Survey. *Cyprus Economic Policy Review* **1**(2), 50-72.

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.

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 (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 key-phrases 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 key-phrases 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 introduce the volume correction factor to the series of interest, and then the four separate series are added in order to form a single index. However, let's be more precise.

Let's assume that we wish to compare four key-phrases. 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 $V_{e,t}^{q}$ 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 ($V_{e,t}^{q}$), that is:

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

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:

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

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:

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

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.

## Previous Papers in this Series

106. **Kougias, Konstantinos**, 'Real' Flexicurity Worlds in action: Evidence from Denmark and Greece, January 2017

105. **Jordaan, Jacob A. and Monastiriotis, Vassilis**, *The domestic productivity effects of FDI in Greece: loca(lisa)tion matters!*, December 2016

104. **Monokroussos, Platon; Thomakos, Dimitrios D.; Alexopoulos, Thomas A.,** *The Determinants of Loan Loss Provisions: An Analysis of the Greek Banking System in Light of the Sovereign Debt Crisis*, November 2016

103. **Halikiopoulou, Daphne, Nanou Kyriaki, Vasilopoulou Sofia,** *Changing the policy agenda? The impact of the Golden Dawn on Greek party politics*, October 2016

102. **Chalari, Athanasia; Sealey, Clive; Webb, Mike,** *A Comparison of Subjective Experiences and Responses to Austerity of UK and Greek Youth*, September 2016

101. **Monokroussos, Platon; Thomakos, Dimitrios D. and Alexopoulos, Thomas A.,** *Explaining Non-Performing Loans in Greece: A Comparative Study on the Effects of Recession and Banking Practices*, August 2016

100. **Gourinchas, Pierre-Olivier, Philippon, Thomas and Vayanos, Dimitri,** *The Analytics of the Greek Crisis*, July 2016

99. **Labrianidis, Louis and Pratsinakis, Manolis,** *Greece's new Emigration at times of Crisis*, May 2016

98. **Vasilaki, Rosa,** *Policing the Crisis in Greece: The others' side of the story*, April 2016

97. **Makrydemetres, Anthony, Zervopoulos, Panagiotis D and Pravita, Maria-Eliana,** *Reform of Public Administration in Greece; Evaluating Structural Reform of Central Government Departments in Greece: Application of the DEA Methodology*, February 2016

96. **Huliaras, Asteris and Kalantzakos, Sophia,** *Looking for an Oasis of Support: Greece and the Gulf states*, January 2016

95. **Simiti, Marilena,** *'Social Need' or 'Choice'? Greek Civil Society during the Economic Crisis*, November 2015

94. **Ifantis, Kostas, Triantaphyllou, Dimitrios and Kotelis, Andreas,** *National Role and Foreign Policy: An Exploratory Study of Greek Elites' Perceptions towards Turkey*, August 2015

93. **Tsirbas, Yannis and Sotiropoulos, Dimitri A.,** *What do Greek political elites think about Europe and the crisis? An exploratory analysis*, July

2015

92. **Tsekeris, Charalambos, Kaberis, Nikos and Pinguli, Maria,** *The Self in Crisis: The Experience of Personal and Social Suffering in Contemporary Greece*, June 2015

91. **Thomadakis, Stavros B.,** *Growth, Debt and Sovereignty: Prolegomena to the Greek Crisis*, May 2015

90. **Arapoglou, Vassilis and Gounis, Kostas,** *Poverty and Homelessness in Athens: Governance and the Rise of an Emergency Model of Social Crisis Management*, March 2015

89. **Dimelis Sophia, Giotopoulos, Ioannis and Louri, Helen,** *Can firms grow without credit? Evidence from the Euro Area, 2005-2011: A Quantile Panel Analysis*, February 2015

88. **Panagiotidis, Theodore and Printzis, Panagiotis,** *On the Macroeconomic Determinants of the Housing Market in Greece: A VECM Approach*, January 2015

87. **Monokroussos, Platon,** *The Challenge of Restoring Debt Sustainability in a Deep Economic Recession: The case of Greece*, October 2014

86. **Thomadakis, Stavros, Gounopoulos, Dimitrios, Nounis, Christos and Riginos, Michalis,** *Financial Innovation and Growth: Listings and IPOs from 1880 to World War II in the Athens Stock Exchange*, September 2014

85. **Papandreou, Nick,** *Life in the First Person and the Art of Political Storytelling: The Rhetoric of Andreas Papandreou*, May 2014

84. **Kyris, George,** *Europeanisation and 'Internalised' Conflicts: The Case of Cyprus*, April 2014

83. **Christodoulakis, Nicos,** *The Conflict Trap in the Greek Civil War 1946-1949: An economic approach*, March 2014

82. **Simiti, Marilena,** *Rage and Protest: The case of the Greek Indignant movement*, February 2014

81. **Knight, Daniel M.,** *A Critical Perspective on Economy, Modernity and Temporality in Contemporary Greece through the Prism of Energy Practice*, January 2014

80. **Monastiriotis, Vassilis and Martelli, Angelo,** *Beyond Rising Unemployment: Unemployment Risk Crisis and Regional Adjustments in Greece*, December 2013

79. **Apergis, Nicholas and Cooray, Arusha,** *New Evidence on the Remedies*

*of the Greek Sovereign Debt Problem*, November 2013

78. **Dergiades, Theologos, Milas, Costas and Panagiotidis, Theodore,** *Tweets, Google Trends and Sovereign Spreads in the GIIPS*, October 2013

77. **Marangudakis, Manussos, Rontos, Kostas and Xenitidou, Maria,** *State Crisis and Civil Consciousness in Greece*, October 2013

76. **Vlamis, Prodromos,** *Greek Fiscal Crisis and Repercussions for the Property Market*, September 2013

75. **Petralias, Athanassios, Petros, Sotirios and Prodromídis, Pródromos,** *Greece in Recession: Economic predictions, mispredictions and policy implications*, September 2013

74. **Katsourides, Yiannos,** *Political Parties and Trade Unions in Cyprus*, September 2013

73. **Ifantis, Kostas,** *The US and Turkey in the fog of regional uncertainty*, August 2013

72. **Mamatzakis, Emmanuel,** *Are there any Animal Spirits behind the Scenes of the Euro-area Sovereign Debt Crisis?*, July 2013

71. **Etienne, Julien,** *Controlled negative reciprocity between the state and civil society: the Greek case*, June 2013

70. **Kosmidis, Spyros,** *Government Constraints and Economic Voting in Greece*, May 2013

69. **Venieris, Dimitris,** *Crisis Social Policy and Social Justice: the case for Greece*, April 2013

68. **Alogoskoufis, George,** *Macroeconomics and Politics in the Accumulation of Greece's Debt: An econometric investigation 1974-2009*, March 2013

67. **Knight, Daniel M.,** *Famine, Suicide and Photovoltaics: Narratives from the Greek crisis*, February 2013

66. **Chrysoloras, Nikos,** *Rebuilding Eurozone's Ground Zero - A review of the Greek economic crisis*, January 2013

65. **Exadaktylos, Theofanis** and **Zahariadis, Nikolaos,** *Policy Implementation and Political Trust: Greece in the age of austerity*, December 2012

64. **Chalari, Athanasia,** *The Causal Powers of Social Change: the Case of Modern Greek Society*, November 2012

63. **Valinakis, Yannis,** *Greece's European Policy Making*, October 2012

62. **Anagnostopoulos, Achilleas** and **Siebert, Stanley,** *The impact of Greek labour market regulation on temporary and family employment - Evidence from a new survey*, September 2012

61. **Caraveli, Helen** and **Tsionas, Efthymios G.,** *Economic Restructuring, Crises and the Regions: The Political Economy of Regional Inequalities in Greece*, August 2012

60. **Christodoulakis, Nicos,** *Currency crisis and collapse in interwar Greece: Predicament or Policy Failure?*, July 2012

59. **Monokroussos, Platon** and **Thomakos, Dimitrios D.,** *Can Greece be saved? Current Account, fiscal imbalances and competitiveness*, June 2012

58. **Kechagiaras, Yannis,** *Why did Greece block the Euro-Atlantic integration of the Former Yugoslav Republic of Macedonia? An Analysis of Greek Foreign Policy Behaviour Shifts*, May 2012

57. **Ladi, Stella**, *The Eurozone Crisis and Austerity Politics: A Trigger for Administrative Reform in Greece?*, April 2012

56. **Chardas, Anastassios,** *Multi-level governance and the application of the partnership principle in times of economic crisis in Greece*, March 2012

55. **Skouroliakou, Melina,** *The Communication Factor in Greek Foreign Policy: An Analysis*, February 2012

54. **Alogoskoufis, George,** *Greece's Sovereign Debt Crisis: Retrospect and Prospect*, January 2012

53. **Prasopoulou, Elpida,** *In quest for accountability in Greek public administration: The case of the Taxation Information System (TAXIS)*, December 2011

52. **Voskeritsian, Horen** and **Kornelakis, Andreas,** *Institutional Change in Greek Industrial Relations in an Era of Fiscal Crisis*, November 2011

51. **Heraclides, Alexis,** *The Essence of the Greek-Turkish Rivalry: National Narrative and Identity*, October 2011

50. **Christodoulaki, Olga; Cho, Haeran; Fryzlewicz, Piotr**, *A Reflection of History: Fluctuations in Greek Sovereign Risk between 1914 and 1929*, September 2011

49. **Monastiriotis, Vassilis** and **Psycharis, Yiannis,** *Without purpose and strategy? A spatio-functional analysis of the regional allocation of public investment in Greece*, August 2011

**SPECIAL ISSUE edited by Vassilis Monastiriotis**, *The Greek crisis in focus: Austerity, Recession and paths to Recovery*, July 2011

48. **Kaplanoglou, Georgia** and **Rapanos, Vassilis T.**, *The Greek Fiscal Crisis and the Role of Fiscal Government*, June 2011

47. **Skouras, Spyros** and **Christodoulakis, Nicos**, *Electoral Misgovernance Cycles: Evidence from wildfires and tax evasion in Greece and elsewhere*, May 2011

46. **Pagoulatos, George** and **Zahariadis, Nikolaos**, *Politics, Labor, Regulation, and Performance: Lessons from the Privatization of OTE*, April 2011

45. **Lyrintzis, Christos**, *Greek Politics in the Era of Economic Crisis: Reassessing Causes and Effects*, March 2011

44. **Monastiriotis, Vassilis** and **Jordaan, Jacob A.**, *Regional Distribution and Spatial Impact of FDI in Greece: evidence from firm-level data*, February 2011

43. **Apergis, Nicholas**, *Characteristics of inflation in Greece: mean spillover effects among CPI components*, January 2011

42. **Kazamias, George**, *From Pragmatism to Idealism to Failure: Britain in the Cyprus crisis of 1974*, December 2010

41. **Dimas, Christos**, *Privatization in the name of 'Europe'. Analyzing the telecoms privatization in Greece from a 'discursive institutionalist' perspective*, November 2010

40. **Katsikas, Elias and Panagiotidis, Theodore**, *Student Status and Academic Performance: an approach of the quality determinants of university studies in Greece*, October 2010

39. **Karagiannis, Stelios, Panagopoulos, Yannis,** and **Vlamis, Prodromos,** *Symmetric or Asymmetric Interest Rate Adjustments? Evidence from Greece, Bulgaria and Slovenia*, September 2010

38. **Pelagidis, Theodore,** *The Greek Paradox of Falling Competitiveness and Weak Institutions in a High GDP Growth Rate Context (1995-2008)*, August 2010

37. **Vraniali, Efi,** *Rethinking Public Financial Management and Budgeting in Greece: time to reboot?*, July 2010

36. **Lyberaki, Antigone,** *The Record of Gender Policies in Greece 1980-2010: legal form and economic substance*, June 2010

35. **Markova, Eugenia,** *Effects of Migration on Sending Countries: lessons*

*from Bulgaria*, May 2010

34. **Tinios, Platon,** *Vacillations around a Pension Reform Trajectory: time for a change?*, April 2010

33. **Bozhilova, Diana,** *When Foreign Direct Investment is Good for Development: Bulgaria's accession, industrial restructuring and regional FDI*, March 2010

32. **Karamessini, Maria,** *Transition Strategies and Labour Market Integration of Greek University Graduates,* February 2010

31. **Matsaganis, Manos** and **Flevotomou, Maria,** *Distributional implications of tax evasion in Greece,* January 2010

30. **Hugh-Jones, David, Katsanidou, Alexia** and **Riener, Gerhard,** *Political Discrimination in the Aftermath of Violence: the case of the Greek riots,* December 2009

29. **Monastiriotis, Vassilis** and **Petrakos, George** *Local sustainable development and spatial cohesion in the post-transition Balkans: policy issues and some theory,* November 2009

28. **Monastiriotis, Vassilis** and **Antoniades, Andreas** *Reform That! Greece's failing reform technology: beyond 'vested interests' and 'political exchange',* October 2009

27. **Chryssochoou, Dimitris,** *Making Citizenship Education Work: European and Greek perspectives,* September 2009

26. **Christopoulou, Rebekka** and **Kosma, Theodora,** *Skills and Wage Inequality in Greece:Evidence from Matched Employer-Employee Data, 1995-2002,* May 2009

25. **Papadimitriou, Dimitris** and **Gateva, Eli,** *Between Enlargement-led Europeanisation and Balkan Exceptionalism: an appraisal of Bulgaria's and Romania's entry into the European Union,* April 2009

24. **Bozhilova, Diana,** *EU Energy Policy and Regional Co-operation in South-East Europe: managing energy security through diversification of supply?,* March 2009

23. **Lazarou, Elena,** *Mass Media and the Europeanization of Greek-Turkish Relations: discourse transformation in the Greek press 1997-2003,* February 2009

22. **Christodoulakis, Nikos,** *Ten Years of EMU: convergence, divergence and new policy priorities*, January 2009

21. **Boussiakou, Iris** *Religious Freedom and Minority Rights in Greece: the case of the Muslim minority in western Thrace* December 2008

20. **Lyberaki, Antigone** *"Deae ex Machina": migrant women, care work and women's employment in Greece*, November 2008

19. **Ker-Lindsay, James**, *The security dimensions of a Cyprus solution*, October 2008

18. **Economides, Spyros**, *The politics of differentiated integration: the case of the Balkans*, September 2008

17. **Fokas, Effie**, *A new role for the church? Reassessing the place of religion in the Greek public sphere*, August 2008

**Online papers from the Hellenic Observatory**

All GreeSE Papers are freely available for download at

http://www.lse.ac.uk/europeanInstitute/research/hellenicObservatory/pubs/GreeSE.aspx

Papers from past series published by the Hellenic Observatory are available at http://www.lse.ac.uk/europeanInstitute/research/hellenicObservatory/pubs/DP_oldseries.aspx