Similarities and correlation between resident tourist overnights and Google Trends information in Portugal and its tourism regions

Similaridades e correlação entre as dormidas dos residentes e a informação do Google Trends para Portugal e suas regiões de turismo

Gorete Dinis
Instituto Politécnico de Portalegre, Praça da República, nº23-25, 7300 – 109 Portalegre, Portugal, gdinis@esep.pt

Carlos Costa
Universidade de Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal, ccosta@ua.pt

Osvaldo Pacheco
Universidade de Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal, orp@ua.pt

Abstract

Over the last years, we observed an exponential growth in the number of tourism consumers that use the Internet as a source of information during a destination selection process. Google Trends is a tool that displays data, almost in real time, on the interest of people in a particular topic based on search trends. This paper demonstrates that Google Trends is a tool that can provide useful and relevant information about the interests of individuals in relation to domestic tourism destinations at national and regional levels. Our findings indicate that overnights spent in hotel establishments by the residents in Portugal are strongly correlated with the Google index, mainly in mainland Portugal, Alentejo and Algarve regions, and that the results improve when more municipalities names and the national or the regional tourism brands are included as search terms.

Keywords: Google Trends, Domestic tourism, Search data, Portugal, Big Data.

Resumo

Ao longo dos últimos anos, tem-se verificado um crescimento exponencial no número de consumidores de turismo que utiliza a Internet como fonte de informação no processo de seleção do destino turístico. O Google Trends é uma ferramenta que disponibilizada dados, quase em tempo real, sobre o interesse das pessoas num determinado tema com base nas tendências de pesquisa. Este artigo demonstra que o Google Trends é uma ferramenta que disponibiliza informação útil e relevante sobre os interesses dos indivíduos em relação aos destinos de turismo doméstico ao nível nacional e regional. Os resultados indicam que as dormidas dos residentes em Portugal nos estabelecimentos hoteleiros estão fortemente correlacionadas com o índice do Google, principalmente em Portugal continental, na região do Alentejo e do Algarve, e que os resultados melhoram quando são incluídos como termos de pesquisa mais nomes de municípios e a marca de turismo nacional ou das regiões de turismo.

Palavras-chave: Google Trends, Turismo doméstico, Dados de pesquisa, Portugal, Big Data.

1. Introduction

The Internet, over the last decades, changed the way how individuals access to tourism information, select and purchase tourism products/destinations. Nowadays, individuals increasingly use the Internet in all phases of the travel cycle and usually they start the decision-making process through the use of a search engine. The most used worldwide search engine is Google and the Google-based search data is available to the public via the Google Trends [GT] tool. According to StatCounter (2014), Google was the search engine that led the market since July 2008 until July 2014, dominating almost 91% of the searches worldwide. The large amount of data about the searches performed by the potential tourist consumers is stored by Google and delivered by GT.

The GT data presents a great potential for tourism management in terms of supporting the decision making process as well as for tourism policy making, since the data is available almost in real time and several studies (see for example Choi & Varian, 2009; Chamberlin, 2010; Smith & White, 2011) concluded that the online pattern behaviour of the consumer is very much related with what happens in reality. GT data may reflect the interests and desires of the tourist consumer with the advantage of being available on time to all tourism agents.

Given the specificities of the tourism sector, the tourism information provided by the national statistical authority in Portugal (Statistics Portugal), especially at regional destinations level, is mostly insufficient and released quite late, due to the rapid changes occurring in society and the new tendencies in tourism demand and consumer behaviour. Portugal is ranked 87th regarding coverage of statistical information available to the tourism sector in the "Travel & Tourism Competitiveness Report 2015" of the World Economic Forum (WEF, 2015).

The aim of this paper is to determine whether the information provided by GT regarding tourism in Portugal and its regions has similar distribution patterns and therefore is correlated to conventional tourism statistics. Moreover, we intend to examine

whether the correlation applies in the same way to different combination of search terms.

Our results contribute to the knowledge on Big Data analytics in tourism, because it identifies data relationship patterns that increase the awareness about the Portuguese tourist interests and behaviour. Furthermore, it allows the understanding of Portuguese domestic tourism activity and its destinations at national and regional level. Additionally, our study adds to the available knowledge in the literature because we improved the selection method that can be used for searching, by identifying the word-combinations that lead to the highest correlations with overnights data in various Portuguese destination regions, that can be replicated in other studies.

This paper starts by displaying a literature review followed by research questions and a description of the used methodology. Next the results and the analysis performed are presented and ends with the final conclusions.

2. Literature review

Since last decades of last century, Information and Communications Technology (ICT) are globally revolutionizing the tourism sector (Buhalis & Law, 2008). As a result, the ICT have changed the way tourism organisations provide and tourists access to information and how they interact with each other (Azevedo et al., 2010).

The consumer behaviour in tourism has been addressed by several authors such as Mathieson and Wall (1982). They stated the decision making process develops in five steps, namely: identification of needs, information search, alternatives evaluation, choice and purchase and post-purchase evaluation. In all steps the ICT plays an important role (Buhalis & Law, 2008).

According to Xiang and Fesenmaier (2006), the search engine is an essential tool for consumers to plan their trips and it is often referred as the "first step" in a trip decision-making process. The online search process, according to Pan et al. (2009), begins with the query formulation where the user entered the search term.

During the last years, several authors developed studies with the aim of understanding the nature and characteristics of query formulation in the tourism area, such as Sanderson and Kohler (2004); Pan et al. (2007); Xiang et al. (2009); Jansen et al (2008); Pan et al. (2006); Xiang and Pan (2011). By analysing these studies, it is possible to conclude that the search terms used in the query formulation are frequently, short, undiversified and prevailing terms related to geography, usually the name of the country, city or state, that is often combined with other specific terms in that location, particularly accommodation and transport. The search terms usually reflect the travellers' information needs, the knowledge and the image the user has about the tourism destination.

The queries formulated in search engines are recorded in databases leading to lots of data stored regarding the consumer searches. According to Martínez et al. (2016) the rapid growth of information generated by tourists is part of the Big Data revolution. Song and Liu (2017) refer that analytics are needed in order to make sense of the information within the large amount of data stored.

Currently, access to this competitive intelligence information is easier because suppliers of search engines are increasingly providing their own tools for data mining (Kaushik, 2010).

In 2012, Google Inc. launched a tool named GT, available at http://www.google.com/trends/, aiming to provide to the general public data based on search patterns performed on the Google's search engine. This tool provides relative search volume statistics, for selected search terms, over specific time ranges and geographic regions, on a daily or weekly basis. GT data is presented in relative values because they are normalized and then scaled, which means that, Google divides data sets by a common variable (the highest value of searches) to cancel the effect of the variable in the data, and then the data is multiplied by 100. This data can be downloaded in .CSV format.

GT allows comparing search volume patterns by search terms, geographic location and time ranges. Currently, GT classified the search terms in 25 categories and 288 subcategories. The search terms related to travel and tourism are classified under the travel category.

Over the last years, researchers published papers using the GT tool data addressing various areas and countries. Most publications fall under the healthcare (see for example Ginsberg et al., 2009; Yang et al., 2011; Dehkordy et al., 2014) and economy (see for example Choi & Varian, 2009ª; Schmidt & Vosen, 2009; Baker & Fradkin, 2011; Bughin, 2015) fields, but work was also performed in finance (Smith, 2012), communication and marketing (Granka, 2010); religion (Scheitle, 2011); education (Vaughan-Frias et al., 2013), and cinema (Judge & Hand, 2010). The hospitality and tourism sector was analysed in the studies developed by Chamberlin, 2010; Choi and Varian, 2009b; Shimshoni et al., 2009; Suhoy, 2009; Smith and White, 2011; Artola and Galán, 2012; Gawlik et al., 2011; Saidi et al., 2010; Pan et al., 2012; Concha et al., 2015; Kallasidis, 2015; Bangwayo-Skeete and Skeete (2015); Dinis et al. 2013, 2015, 2016ab); Jackman and Naitram, 2015); Yang et al. (2015); Rivera (2016); and Li et al., 2017).

In our literature review we found that the way the researchers selected the queries or search terms differs and therefore there is no unique and common criterion. For instance, Dehkordy et al. (2014) selected the terms using the Google keyword tool (Google Adwords), Pan et al. (2012) used the five keywords they consider "the most relevant and unique when tourists search for a destination city in the USA" (p. 200), Saidi et al. (2010) and Artola and Galán (2012) used only one keyword "Dubai" and "Spain holiday", respectively. A limitation of the study appointed by Pan et al. (2012) is precisely the low number of tourism-related queries and Varian (2014) refers that given there are billions of queries the big challenge is to "determine exactly which queries are the most predictive for a particular purpose".

When the GT data is analysed it is necessary to take into account some considerations, such as, the fact that the data is referent only to the public that uses the Internet, namely the Google search engine, as a source of information for planning and organizing a trip, identified by the Internet Protocol address of the user. Furthermore, the query results are limited to the language and expressions that people use as search terms.

3. Research questions

Within the context described in the previous sections, this paper aims to empirically verify if GT data can be used in the analysis of the Portuguese interests for what concerns domestic tourism at national and regional destination levels. This research study is novel because it focuses on domestic tourism destinations at

national and regional levels that were not addressed in previous studies. In particular, the present analysis focuses on mainland Portugal and its regions, and to our knowledge, there are no previous studies using GT data applied to these territories in the field of hospitality and tourism, a strategic sector for the Portuguese economy.

Furthermore, this study makes sense because the data indicates that, in 2013, only about 38% of the Portuguese made a trip with a duration of at least one night, involving a total of 17,9 million trips, 92% of them had as destination Portugal and 83% of those made without a prior appointment. In addition, 52% of the overnights are spent with relatives or friends and only about 15% of them are in hotel establishments or similar (INE, 2014). Moreover, in this research we contributed to the establishment of a method for obtaining Google-based search data for several search terms and it was our purpose to verify if the correlation with the official data depends on the combination of search terms.

Following such evidences, we proposed the following questions for this research:

RQ 1: Does the Google-based search data correlate with the residents overnights in Portugal at national and regional level?

RQ 2: The level of correlation between Google-based search data and the residents' overnights in Portugal at national and regional level is related with the used search-word combinations?

Analyzing our results, the researchers can evaluate whether the GT data is useful and helps to understand tourists' behaviours and desires in the context of a short scale country, medium sized geographical areas and related to a type of accommodation not often used by the Portuguese on domestic trips.

Our research is important because if it confirms that Google-based search data is correlated with the official data, namely the residents' overnights of Portugal in different geographical areas, the GT data potentialities would be confirmed as well as the prospect of using the data - with the best combination of search terms - in more Business Intelligence methods. Consequently, the tourism organisations, mainly the destination management organisations, could use this data to monitor the tourists' interests in a specific destination region or country, almost in real time, and even identifying interest spikes. In addition, the tourism decision makers can use GT data to help estimate the level of tourism demand in such territories or to define appropriate marketing strategies, mainly in terms of producing adequate messages to the different channels, the type of consumer and the stage of travel. Moreover, it would allow for comparing the destination performance with its competitors, clearly improving the knowledge of these organisations on the Portuguese domestic tourism market.

4. Methodology

Our purpose is to test GT data efficiency when data refers to the interests of Portuguese regarding the tourism of their own country at national and regional level and identify which combination of search queries are most used when Portuguese tourists search for a tourism destination at national and regional level in Portugal. That is to say which queries lead to results highly correlated to recognised official tourism statistics information. For that purpose, we compared a variable that resulted from the official survey-based data released by the National Statistical Institute (Statistics Portugal) with data from GT. GT data has the advantages of being available in useful time, without cost, has a large sample base and, besides that, individuals search are spontaneous and without the influence of others, which means the data is objective.

To represent the domestic tourism demand we selected as variable the overnights spent by residents in hotel establishments by month, published by Statistics Portugal. We opted to consider, as the studied population, the residents in Portugal and the hotel establishments located in mainland Portugal and its regions. For statistical purposes mainland Portugal is divided into the following five units comprising the Level II of the Nomenclature of Territorial Units (NUTS II): North (86 municipalities); Centre (100 municipalities); Lisbon (18 municipalities); Alentejo (58 municipalities); and Algarve (16 municipalities). The data was obtained, by year, since 2004 until 2012.

For what concerns GT data, we adopt a method to find search word-combinations which is described below. First, we obtained the GT data from the category Travel and subcategory Hotels & Accommodations, per year, from January 2004 to December 2012, for a specific tourism destination specified by the search terms. Furthermore, we decided to extract the GT data for different geographical areas in mainland Portugal based on the division of the country used for statistical purposes that coincide, in territorial terms, with the five tourism regional areas defined in mainland Portugal, that are: Tourism Porto and North, Tourism Centre of Portugal, Tourism of Lisboa, Tourism of Alentejo, and Tourism of Algarve.

Based in the literature review, we selected as general criterion for search, the word-combinations associated with the geographic area, mainly the municipalities' name that integrate each tourism region area. Therefore, the GT data analysed in this study covers the searches done by the Portuguese in Google search engine related to accommodation in those municipalities.

Since mainland Portugal and each tourism regional area are constituted by a different and considerable number of municipalities, we considered in this study, at first, a combination of search terms with the name of the municipalities that, according to Statistics Portugal, registered the biggest number of overnights in the accommodation establishments in mainland Portugal and the regional areas in 2011. When the municipalities' names are dubious, we changed the search terms to names of localities or tourist attractions which are well known in that specific region, for instance, "Lagoa" by "Carvoeiro", a village of Lagoa where is located the nearest beach, accommodations and golf courses. In Lisbon and Algarve regions since the number of municipalities is lower than the maximum of search terms allowed by GT (30 search terms), we opted to include as search terms the localities with tourism interest, according to the list of official accommodations provided by the Turismo de Portugal, I.P and the highest search volume of the Google Adwords, since the municipality has already been considered by the previous criteria, for instance, we added the localities of "Ericeira", "Estoril", "Guincho" and "Caparica" in the combination of search terms to represent the Lisbon region.

We grouped the search terms in a single entry using the plus sign between the search terms. Furthermore, we used the minus sign to exclude search terms that can negatively influence the results, such as other forms of accommodation, and the quotation marks when we want to detect the searches that match exactly that

expression. The tourism region brand was also used as a search term (for example Alentejo).

In the second combination of search terms, for mainland Portugal, we removed from the search terms list the name of the country (Portugal) and the brand of regional tourism areas, namely Centre, North, Alentejo and Algarve. From the regions of Portugal, we removed the search terms which represent the respective tourism region brand. Therefore, the second combination of search terms is restricted to the top 10 of the Portugal City Brand municipalities ranking by country and regions, developed by Bloom Consulting, where the purpose was to understand the attractiveness of the municipalities in Portugal as a tourism destination, for investment and for living. The performance of the municipality for what concerns tourism, investment and living, was evaluated through three dimensions: socioeconomic performance; digital demand; and online performance. The index is a composite considering these three dimensions for the three components. The socio-economic performance in the tourism component was obtained by indicators of accommodation from the demand side. The data for the digital demand was obtained through its own technological tool named "*Digital Demand*", that enables each municipality to know exactly what each Internet user searches about the municipality in the major search engines in Portugal (Bloom Consulting, 2015), and refers to the total online search volume on tourism for 16 *brandtags* by municipality. Regarding the online performance, it was obtained from the municipality website web analytics and social networking (e.g. Facebook and Twitter).

We considered this index because it was developed by Bloom Consulting, a recognized company that collaborates with the World Economic Forum to measure the nation brand appeal of every country from a tourism perspective, used in the elaboration of the Travel & Tourism Competitiveness Index. In addition, to our knowledge, this is the unique index made for Portugal and regions at municipality level that reveals the attractiveness of the municipality in the tourism category and that considers the volume of online searches related with tourism - a similarity with GT.

Since the GT data is available by week and the official data by month, we transformed the GT data to the same scale by arithmetic average, similarly to Schmidt and Vosen (2009) and Willard and Nguyen (2011) In the figure 1, we can see the first (highlighted in bold) and the second combinations of search terms used in the empirical study.

Fig 1. Search queries used in the empirical study

MAINLAND PORTUGAL	**portugal+centro+ norte +lisboa+alentejo+algarve + "porto"+ albufeira + vilamoura+ portimao+ gaia+ coimbra+ cascais+ braga+ evora+ matosinhos+ ourem + fatima+ covilha+ viseu+ oeiras+ tavira + setubal+ faro+ figueira+aveiro + carvoeiro-rural-campismo-**
	lisboa-macau+"porto"-seguro+albufeira+portimao+cascais+lagos+faro+matosinhos+evora+coimbra-rural-campismo-juventude-hostel
NORTH	**norte+douro+"porto"- seguro+gaia+braga+matosinhos+guimaraes+ bragança+povoa+chaves+ viana+ tirso+ mirandela+ bouro +lamego + maia+ "vila real"+ regua+esposende+ feira+valença+penafiel+ miranda+ caminha+conde+cerveira-rural-campismo-**
	"porto"-seguro+ matosinhos+braga+guimaraes+"viana do castelo"+gaia+varzim+esposende+"ponte de lima"+bragança-rural-campismo-juventude-hostel
CENTRE	**centro+coimbra+ourem+fatima+covilha+viseu+ figueira+ aveiro+leiria+ vedras+guarda+ branco+peniche+ "são pedro do sul"+anadia+ovar+mealhada+tomar+nazare+obidos+"marinha grande"+"caldas da rainha"+ nelas+seia + estrela+alcobaça-rural-campismo-juventude-hostel**
	coimbra+aveiro+nazare+peniche+covilha+obidos+leiria+figueira+ovar+viseu-rural-campismo-juventude-hostel
LISBON	**lisboa+loures+xira+alcochete+montijo+cascais+oeiras+setubal+almada+sesimbra+sintra+amadora+mafra+ericeira+palmela+odivelas+barreiro+moita+seixal-macau+estoril+guincho+caparica+belem-campismo-rural-juventude-hostel**
	lisboa-macau+sintra+cascais+oeiras+setubal+almada+seixal+mafra+loures+xira-campismo-rural-juventude-hostel
ALENTEJO	**alentejo+evora+grandola+beja+elvas+estremoz+sines+santarem+ "santiago do cacem" +alcacer+moura+odemira+viçosa+marvao+serpa+reguengos+ferreira+barrancos+alqueva+covo+milfontes+troia+zambujeira+com porta-rural -campismo-hostel -juventude**
	evora+beja+sines+odemira+grandola+santarem+elvas+moura+"santiago do cacem"+estremoz-rural -campismo-hostel -juventude
ALGARVE	**algarve+albufeira+loule+portimao+montegordo+tavira+faro+carvoeiro+lagos+sagres+silves+"castro marim"+monchique+olhao+aljezur+vilamoura+alvor+quarteira+eulalia+almancil+altura+armaçao-rural -campismo-hostel-juventude**
	albufeira+portimao+tavira+lagos+faro+loule+silves+olhao+lagoa+bispo-rural-campismo-hostel-juventude

Source: Authors.

We have performed a univariate and bivariate analysis. The series were evaluated in relation to the distribution using the normality test, subject to logarithm transformations to achieve the normality of the distribution when necessary, i.e., when the variable presented a distribution positively skewed (Corrar et al., 2007), and we applied correlation coefficients to assess the relation between the time series.

The bivariate correlation analysis or the behaviour of a set of two variables can be measured by the correlation coefficients, which are defined as a function of the measurement scale of the variables considered (Maroco, 2007). The coefficients most used are the Pearson and the Spearman correlation coefficients.

In continuous variables it is normally applied the Pearson coefficient, however the variables need to meet other conditions, such as normal distribution, the existence of a linear relationship between the variables and absence of significant outliers in the data. The more the Pearson coefficient approaches the +1 value, stronger the relationship between the two variables, which means that changes in one variable are strongly correlated with changes in the other variable. On the other hand, the more the Pearson coefficient approaches 0, the weaker is the relation between the variables. If the Pearson correlation is -1 this mean that the variables are identical but in opposing directions (Pallant, 2001). According to Franzblau (1958) the Pearson correlation coefficient is interpret as shown in Table 1.

Table 1. Interpretation of Pearson correlation coefficient

Pearson coefficient (r)	Correlation
|r| < 0,20	No correlation
0,20 < |r| < 0,40	Weak
0,40 < |r| < 0,60	Moderate
0,60 < |r| < 0,80	Strong
|r| > 0,80	Very strong

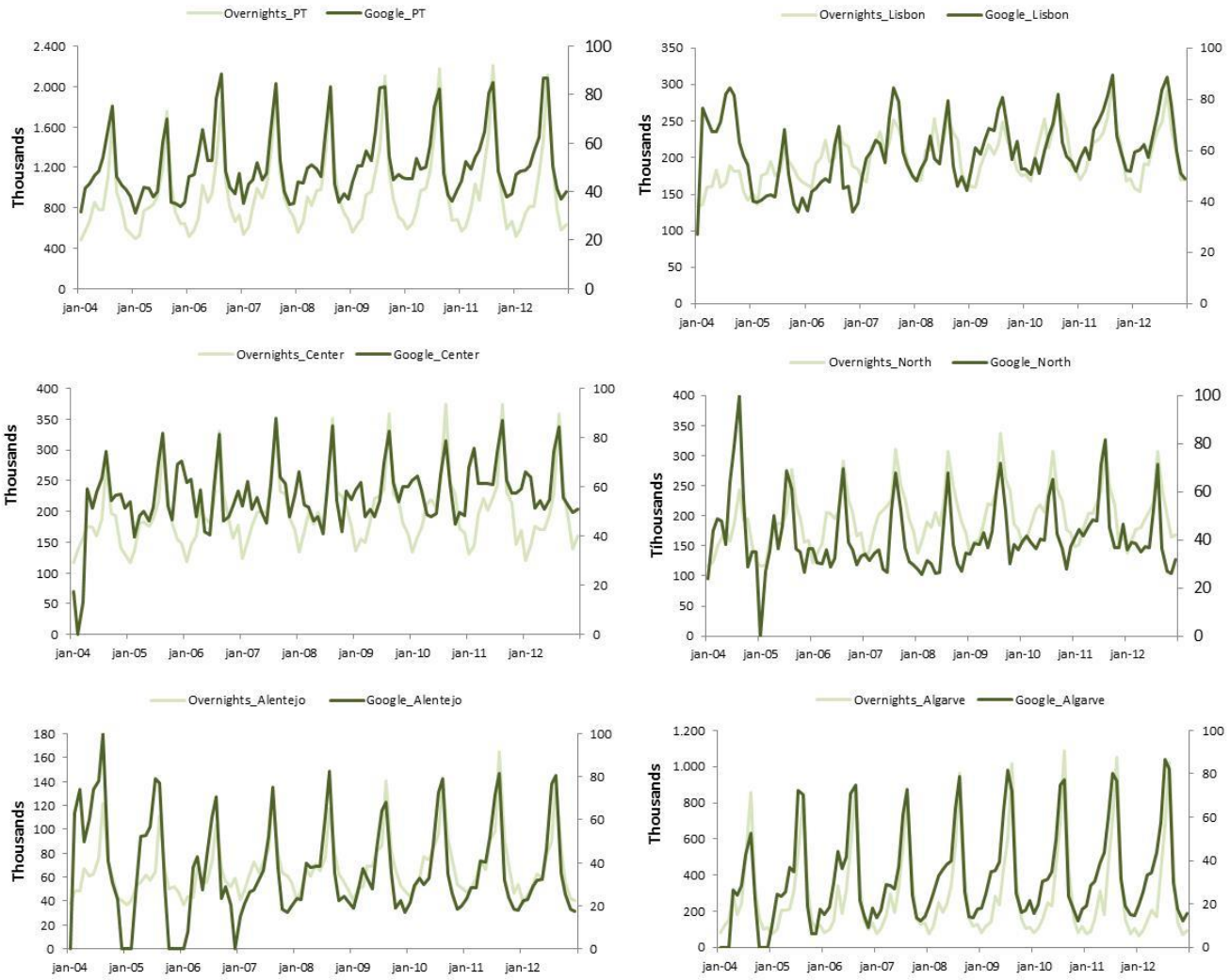
Source: Authors.

The Pearson coefficient was applied to variables that presented a bivariate normal distribution, i.e., to the Lisbon and Alentejo variables, in the first combination of search terms, and to the North and Lisbon variables in the second combination of search terms. To the other variables we decided to apply the Spearman coefficient. The data was analysed using the SPSS tool, version 20.0.

5. Findings

The comparison between the time series related to the variable "overnights spent in hotel establishments" and the GT index for Portugal and its regions, obtained for the first combination of search terms is represented in Figure 2. Analysing this figure, we found that the annual behaviour of the variables is very similar both in Portugal and its regions, reaching maximum values in the summer and minimum values during the winter months. The series, during the time interval in analysis, show a seasonal pattern, which is repeated year after year.

Fig.2. Overnights spent by residents in Portugal in hotel establishments (Portugal and regions) vs. the GT index



Source: Authors.

Observing the bivariate correlations represented in Table 2, it is noted that the variables "Google_Centre"/ "Overnights_Centre" are those that have a lower correlation coefficient (0.267) in the first combination of search terms (0.371), which means a low correlation between these two variables at *p*=0.01. Furthermore, the second combination of search terms, whose municipalities

represents about 43% of the total overnights spent by the Portuguese in hotel establishments in that region compared to the 73% of the first combination of terms, shows a higher correlation coefficient than the first one, which means that results did not benefit with the inclusion of the search queries made by the potential travellers that included the tourism region or other municipalities names.

Regarding the sets of variables "Overnights_Lisboa"/ "Google_Lisboa";"Google_North"/"Overnights_North" it was found that there is a moderate correlation (*r* is approximately 0.5) but also with these variables the correlation is superior in the second combination of search terms, although the difference is more significant in the North region.

The highest correlation coefficients are found, in the first set of search terms on the variables "Overnights_Alentejo"/"Google_Alentejo" (0,725); "Overnights_PT"/"Google_PT_PT" (0,712); and "Google_Algarve"/ "Overnights _Algarve" (0,804). This means that there is a strong relationship between the variables and that changes in the "Google" variables are strongly correlated with the variations in the "Overnights" variables. Moreover, once the coefficients are all positive values, when the value of the "Google" variables increases or decreases, the value of the "Overnights" variables also increases or decreases, respectively.

The relationship between these variables decreases in the cases of the second combination of search terms, mainly in the case of the Alentejo region, which indicates - in opposite to what happens in the North and Centre regions that have a higher number of municipalities -, to meet the interests of Portuguese potential travellers for this region, it is important to include the other municipalities which also have a significant weight in terms of overnights in hotel establishments, finer geographic areas (for instance Comporta, Alqueva and Tróia), and the Alentejo brand.

The Lisbon and Algarve regions are the smallest regional areas because of that and in both analyses, almost all the municipalities are addressed, which represent more than 90% of the overnights in case of the Lisbon region and 80% in the Algarve region. Analysing the coefficients results, we can conclude that the search terms selected are more helpful to define the interests of Portuguese potential travellers by the Algarve rather than the Lisbon region, this could mean that they do not search in Google equally for this regions or, in the case of Lisbon, they can use other search terms which might include a different geographic level (e.g. quarter), hotel brand or tourism attractions.

Additionally, it can be seen that the correlation coefficients are all positive, which means that variables tend to relocate together, in other words, large values of the variables "overnights" tend to be associated with large values of the "Google" variables.

Table 2. Correlations between overnights spent by residents in Portugal in hotel establishments (Portugal and regions) and GT index

First combination of search terms					Second combination of search terms		
Date	**Overnights|Google**	**Variable**	**p**	**Correlation**	**Variable**	**p**	**Correlation**
2004-2012	Portugal	O	0,000	0,712**	O	0,000	0,657**
	North	O	0,000	0,479**	O	0,000	0,578**
	Centre	O	0,000	0,267**	O	0,000	0,371**
	Lisbon	O	0,000	0,560**	O	0,000	0,562**
	Alentejo	T	0,000	0,725**	O	0,000	0,569**
	Algarve	O	0,000	0,804**	O	0,000	0,799**
O: Original Variable; T: transformed variable							
** Correlation is significant at the 0.01 level (2-tailed)							

Source: Authors.

Through the quantitative analyses performed in this research, we were able to answer the research questions proposed in this study. Therefore, we can conclude that the Google-based search data is correlated with overnights stays of domestic tourists in Portugal both at national and regional level, although this correlation is greater in the case of national level and in some regions such as Algarve, Alentejo and Lisbon. Regarding the research question "The level of correlation between Google-based search data and the residents' overnights in Portugal at national and regional level is related with the used search-word combinations?", from the correlation analyses, we verified that the selection of search terms is very important and the correlation varies accordingly. We used several search terms related to geography to define Portugal and its regions, and in the case of Algarve and Alentejo regions, the correlation results are improved when we used as search queries the regional tourism brand and extended the geographical scope, i.e. include the names of more municipalities.

6. Conclusions

GT is a great tool for analysing people's interest and intentions in relation to a specific tourist destination, and can be used to anticipate tourist's desires and movements because the GT data is available almost in real time, well before the release of official tourism statistics.

Several researchers have shown the correlation between the GT data and official data and the usefulness of GT data for nowcasting and forecasting of different areas' phenomena, but apart from the

few studies in the field of hospitality and tourism, none has focused on domestic tourism at national or regional level, or applied the studies to the Portuguese people. As a result of this study, we concluded that GT data can help the tourism organisations to understand the domestic tourist behaviour and interests in Portugal at national and regional level, and since one of the major challenges of this data is to determine which search terms are more adequate to a particular purpose, we showed which search terms structure is more adequate for defining and understanding domestic tourism demand for Portugal and its regions.

Our results show that there are similar movements and positive correlations between official statistics on the number of overnights spent in hotel establishments by the residents of Portugal in their own country and tourism regions and the GT index for the search terms designated by us. Furthermore, we proved that in the same destinations, such as Portugal, Alentejo or Algarve, the correlation results are improved by including in the analysis more search terms related to the geographical location, mostly the municipalities' names which present the highest number of overnight stays of residents in Portugal by country and its regions. On the other hand, we demonstrated that in some regions like the Centre or North regions of Portugal, although larger geographical areas, the correlation is improved when the analysis is restricted to less municipalities and tourism regions brands as search terms. These results indicate that there are tourism regions and municipalities with more appeal for the Portuguese as tourism destinations than others, likewise, we think it is important that tourism destination organisations ponder on the attractiveness of their brand image and align their strategic decisions towards the necessities and interests of potential travellers to their destinations.

Moreover, this study presents meaningful results concerning the similarities and correlation in interests between residents in Portugal and the individuals searching on Google for tourism destinations in Portugal, but also presents limitations, as is the case when there exists only a low number of Portuguese tourists that use Google to plan their trips, mainly when the destination is in their own country. In North, Centre and Alentejo the study did not include all the municipalities as search terms, and even if we tried to eliminate searches that may not be related with hotels and accommodation, it is possible that some online searches might be made with another intention or the search terms is referent to destinations in other countries.

For future research, we suggest that the GT data for the different combinations of search terms defined in this study could also be tested in relation to their capacity to improve forecasting accuracy in tourist demand in Portugal and its regions using the GT data as explanatory variable in a domestic tourism forecasting model, such as the Transfer Function. In addition, the search word-combinations presenting the best correlation with overnights can be explored in the theory development of the field of regional destination brand studies.

References

Artola, C. & Galan, E. (2012). Tracking the future on the web: construction of leading indicators using Internet searches. *Banco de Espana Occasional Paper*, (1203). Retrieved 25 November 2014 from http://bit.ly/XRLfCf.

Azevedo, C., Dinis, M.G. & Breda, Z. (2010). Understanding visitors' spatio-temporal distribution through data collection using information and communication technologies, Proceedings of the 10th International Forum on Tourism Statistics.

Baker, S. & Fradkin, A. (2011). What drives job search? Evidence from Google Search Data. SIEPR Discussion Paper No. 10-020. Stanford Institute For Economic Policy Research. Retrieved 17 July 2014 from http://stanford.io/1oKwdIJ.

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

Bloom consulting. (2015). Bloom Consulting Portugal City Brand Ranking 2015. Retrieved 5 September 2015 from http://bit.ly/1KX6Tcs.

Bughin, J. (2015). Google searches and twitter mood: nowcasting telecom sales performance. *Netnomics*, 16(1-2), 87-105.

Buhalis, D. & Law, R. (2008). Progress in information technology and tourism management: 20 years on and 10 years after the Internet-The state of eTourism research. *Tourism Management*, 29(4), 609–623.

Chamberlin, G. (2010). Googling the present. *Economic* & *Labour Market Review*, 4(12), 56.

Choi, H.C. & Varian, H. (2009a). Predicting Initial Claims for Unemployment Benefits. Retrieved 12 November 2015 from http://bit.ly/1snBiam.

Choi, H.C. & Varian, H. (2009b). Predicting the Present with Google Trends. Retrieved 12 November 2015 from http://bit.ly/1snBiam.

Concha, A., Pinto, F. & Pedraza García, P. (2015). Can internet searches forecast tourism inflows?. *International Journal of Manpower*, 36 (1), 103 – 116.

Corrar, L., Paulo, E. & Filho, J. (2007) *Análise multivariada para os cursos de administração, ciências contábeis e economia*. São Paulo: Atlas.

Dehkordy, S. Carlos, R., Hall, K. & Dalton, V. (2014). Novel Data Sources for Women's Health Research: Mapping Breast Screening Online Information Seeking Through Google Trends. *Academic Radiology*, (21)9, 1172–1176.

Dinis, M.G., Costa, C. & Pacheco, O. (2013). Using Google Trends to obtain information about tourism. *Innovation and Technology in Tourism and Hospitality applied research - Proceedings of the ISITH 2012*, Colecção Politécnico da Guarda, Guarda, 91–104.

Dinis, G., Costa, C. & Pacheco, O. (2015). Nós Googlamos! Utilização da ferramenta Google Trends para compreender o interesse do público pelo Turismo no Algarve. *Dos Algarves: A Multidisciplinary e-Journal*, 26(1), 64-84.

Dinis, G., Costa, C. & Pacheco, O. (2016a). Tendências e interesse de pesquisa do público por museus, locais e edifícios históricos e festivais de música: A ferramenta Google Trends. *Revista de Turismo Contemporâneo*, 4(2), 177-195.

Dinis, G., Costa, C. & Pacheco, O. (2016b).The Use of Google Trends Data as Proxy of Foreign Tourist Inflows to Portugal. *International Journal of Cultural and Digital Tourism*, (3)1, 66 - 75.

Franzblau, A. (1958). *A Primer of Statistics for Non-Statisticians*. New York: Harcourt, Brace & World, Inc.

Gawlik, E., Kabaria, H. & Kaur, S. (2011). Predicting tourism trends with Google Insights. Retrieved 1 December 2015 from http://stanford.io/V1lAWl.

Ginsberg, J., Mohebbi, MH., Patel, RS., Brammer, L., Smolinski, MS. & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457, 1012-1014.

Granka, L. (2010). Measuring agenda setting with online search traffic: influences of online and traditional media. Prepared for delivery at the

2010 Annual Meeting of the American Political Science Association, September 2-5, 2010. Retrieved from http://bit.ly/2Il6fZi. Accessed 28 th November 2012.

INE (2014). *Estatísticas do Turismo 2013*. Lisboa: Instituto Nacional de Estatística.

Jackman, M. & Naitram, S. (2015). Nowcasting tourist arrivals to Barbados. Just Google It! *Tourism Economics* 21, 1309-1313.

Jansen, B.J., Ciamacca, C.C. & Spink, A. (2008). An analysis of travel information searching on the web. *Information Technology* & *Tourism*, 10(2), 101-118.

Judge. G., Hand, C. (2010). Searching for the picture: forecasting UK cinema admissions making use of Google Trends data. Department of Economics Discussion Paper Nº 162. University of Portsmouth Business School. Retrieved 20 November 2014 from http://bit.ly/1o12y0j.

Kallasidis, F. (2015). Web Search activity: A Forecast Tool for Tourist Arrivals in Cyprus. Master of Science in Management. School of Economics, Business Administration & Legal Studies. Greece. Retrieved 22 February 2017 from http://bit.ly/2mciL3H.

Kaushik, A. (2010). *Web analytics 2.0: The art of online accountability* & s*cience of customer centricity*. Indianapolis: Wiley.

Li, X., Pan, B., Law, R. & Huang, X. (2017). Forecasting tourism demand with composite search index. *Tourism Management*, 59, 57-66.

Martínez, R, Herráez, B. & Yábar, D. (2016). Actividad de búsquedas en Internet como variable para determiner la afluencia a museos. *Cuadernos de Turismo*, 38, 203-219.

Maroco, J. (2007). *Análise Estatística com utilização do SPSS* (3ª Ed.). Lisboa: Edições Silabo.

Mathieson, A. & Wall, G. (1982). *Tourism, economic, physical and social impacts*. London: Wiley.

Pallant, J. (2001). *SPSS – Survival Manual*. Philadelphia: Open University Press.

Pan, B., Litvin, S. & Goldman, H. (2006). Real users, real trips, and real queries: An analysis of destination search on a search engine. Paper presented at the Annual Conference of Travel and Tourism Research Association (TTRA 2006). Ireland, Dublin, 16 - 18 June.

Pan, B, Litvin, S.W. & O'Donnell, T.E. (2007). Understanding accommodation search query formulation: The first step in putting 'heads in beds'. *Journal of Vacation Marketing*, 13(4), 371–381.

Pan, B., Xiang Z, Fesenmaier, D. & Law, R. (2009). Destination Online Competitiveness and Search Engine Marketing. Retrieved 16 August 2016 from http://bit.ly/1LfPnDe.

Pan, B., Wu, D.C. & Song, H. (2012). Forecasting hotel room demand using search engine data, *Journal of Hospitality and Tourism Technology*, 3(3), 196–210.

Rivera, R. (2016). A dynamic linear model to forecast hotel registrations in Puerto Rico using Google Trends data. *Tourism Management* 57, 12-20.

Saidi, N., Scacciavillani, F. & Ali, F. (2010). Forecasting Tourism in Dubai. Dubai International Finance Centre: Economic Note n.º 8. Retrieved 6 August 2014 from http://www.difc.ae/publications.

Sanderson, M. & Kohler, J. (2004). Analyzing geographic queries. In the proceedings of SIGIR Workshop on Geographic Information Retrieval: The 27th Annual International ACM SIGIR Conference. Sheffield, UK, 25 - 29 July 2004.

Scheitle, C.P. (2011). Google's Insights for Search: A Note Evaluating the Use of Search Engine Data in Social Research. *Social Science Quarterly*, 92(1), 285–295.

Schmidt, T. & Vosen, S. (2009). Forecasting private consumption: survey-based indicators vs. Google Trends, Ruhr Economic Papers 155, RWI. Retrieved 24 November 2014 from http://bit.ly/1so26rQ.

Shimshoni, Y., Efron, N. & Matias, Y. (2009). On the Predictability of Search Trends (draft). Google, Israel Labs. Retrieved 15 November 2014 from http://bit.ly/1zjL573.

Smith, G.P. (2012). Google Internet search activity and volatility prediction in the market for foreign currency. *Finance Research Letters*, 9(2), 103-110.

Smith, E. & White, S. (2011). What Insights Can Google Trends Provide About Tourism in Specific Destinations? UK: ONS. Retrieved 10 June 2015 from http://bit.ly/1o13x0r.

StatCounter (2014). StatCounter Global Stats: search engine. StatCounter. Retrieved 1 July 2014 from http://bit.ly/1v0s56K.

Song, H. & Liu, H. (2017). Predicting Tourist Demand Using Big Data. In Z. Xiang, D.R. Fesenmaier (eds.), *Analytics in Smart Tourism Design: Concepts and Methods*. Switzerland: Springer

Suhoy, T. (2009). Query Indices and a 2008 Downturn: Israeli Data. Bank of Israel: Research Department. Discussion Paper No. 2009.06. Retrieved 15 January 2014 from http://bit.ly/Z29izI.

Varian, H. (2014). Big Data: New Tricks for Econometrics. Retrieved 30 November 2015 from http://bit.ly/1e2m9V4.

Vaughan, L. & Romero-Frías, E. (2013). Web search volume as a predictor of academic fame: An exploration of Google trends. *Journal of the Association for Information Science and Technology*, 65(4), 707-720.

Willard, S. D. & Nguyen, M.M. (2013). Internet search trends analysis tools can provide real-time data on kidney stone disease in the United States. *Urology*; 81 (1), 37–42.

Xiang, Z., Fesenmaier, D.R. (2006). Assessing the initial step in the persuasion process: Meta tags on destination marketing websites. *Information Technology* & *Tourism*, 8(2), 91–104.

Xiang, Z., Gretzel, U. & Fesenmaier, D.R. (2009). Semantic representation of the online tourism domain. *Journal of Travel Research*, 47(4), 440 – 453.

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

Yang, A.C., Tsai, S.J., Huang, N.E. & Peng, C.K. (2011). Association of Internet search trends with suicide death in Taipei City, Taiwan, 2004–2009. *Journal of affective disorders*, 132(1), 179–184.

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

Received: 25 July 2016
Accepted: 21 March 2017