Information Technology & Tourism (2019) 21:45–62
https://doi.org/10.1007/s40558-018-0129-4

ORIGINAL RESEARCH

Google Trends data for analysing tourists' online search behaviour and improving demand forecasting: the case of Åre, Sweden

Wolfram Höpken[1] · Tobias Eberle[1] · Matthias Fuchs[2] · Maria Lexhagen[2]

Received: 6 March 2018 / Revised: 20 November 2018 / Accepted: 23 November 2018 /
Published online: 30 November 2018

Abstract
Accurate forecasting of tourism demand is of utmost relevance for the success of tourism businesses. This paper presents a novel approach that extends autoregressive forecasting models by considering travellers' web search behaviour as additional input for predicting tourist arrivals. More precisely, the study presents a method with the capacity to identify relevant search terms and time lags (i.e. time difference between web search activities and tourist arrivals), and to aggregate these time series into an overall web search index with maximal forecasting power on tourism arrivals. The proposed approach enables a thorough analysis of temporal relationships between search terms and tourist arrivals, thus, identifying patterns that reflect online planning behaviour of travellers before visiting a destination. The study is conducted at the leading Swedish mountain destination, Åre, using arrival data and Google web search data for the period 2005–2012. Findings demonstrate the ability of the proposed approach to outperform traditional autoregressive approaches, by increasing the predictive power in forecasting tourism demand.

Keywords Google Trends data · Search word analysis · Online search pattern · Tourist arrival prediction · Autoregressive time series forecasting · Big data

This is an extended version of a conference paper entitled "Search engine traffic as input for predicting tourist arrivals" previously published in the proceedings of Information and Communication Technologies in Tourism 2018 Conference (ENTER 2018) held in Jönköping, Sweden, January 24–26, 2018.

✉ Wolfram Höpken
wolfram.hoepken@hs-weingarten.de

[1] Business Informatics Group, University of Applied Sciences Ravensburg-Weingarten, Doggenriedstr., 88250 Weingarten, Germany

[2] European Tourism Research Institute (ETOUR), Mid-Sweden University, Kunskapens Väg 1, 83125 Östersund, Sweden

1 Introduction

With a worldwide turnover of more than 7 trillion US dollars in 2015 and a total share of around a tenth of global GDP, travel and tourism significantly contributes to the global economy. On a global scale, nearly every tenth job relates directly or indirectly to the travel and tourism industry (WTTC 2016). However, the success of tourism-related businesses, such as airlines or hotels, largely depends on the capacity to accurately predict tourism demand. Due to the perishable nature of tourism services (i.e. the fact that services ‘perish’ in case of non-use), accurate forecasts of tourism demand are of utmost relevance (Frechtling 2002). More precisely, for tourism businesses it is pivotal to respond promptly to upcoming demand, thus, making limited resources available and ready for co-creative service production processes (Fitzsimmons and Fitzsimmons 2001; Grönroos 2008; Chekalina et al. 2018). Hence, knowledge about long-term trends, imminent changes and short-term intra-period fluctuations of customer demand is essential for tourism management in planning resource capacities. Furthermore, predictions of tourist arrivals help governments in shaping medium and long-term strategies for local and regional tourism development and planning (Fuchs et al. 2000; Edgell et al. 2008; Pike et al. 2017).

Accordingly, in the travel and tourism domain, the accuracy of demand forecasts can hardly be overestimated for businesses and policy makers, likewise (Frechtling 2002). However, predicting future tourism demand is a difficult and non-trivial task, due to the lack of historical data, seasonal fluctuations, influences of unexpected events, the variety of input factors and the complexity of visitors’ travel decision-making process (Song et al. 2010). Therefore, demand modelling and prediction has attracted great attention by academics and practitioners and ranks among the most relevant domains within tourism research.

As potential travellers extensively search the web before visiting a specific destination (Fesenmaier et al. 2010), the aim of this paper is to extend the autoregressive time series forecasting approach (i.e. prediction based on past arrivals alone) by including travellers’ web search behaviour as additional input for the prediction of tourist arrivals. More precisely, firstly, the study evaluates whether the inclusion of time series data on web search behaviour can increase the performance when forecasting tourist arrivals compared to the purely autoregressive approach. Second, the study more deeply examines behavioural aspects of travellers related to the concrete *search terms* used in online search for trip planning in different sending countries. More concretely, by considering patterns that reflect the online planning behaviour of travellers before visiting a specific destination, temporal relationships between search terms used and tourist arrivals are analysed. The study is conducted for the leading Swedish mountain destination Åre, using arrival data and Google Trend-based web search data of major sending countries (i.e. Denmark, Finland, Norway, and the United Kingdom) for the period 2005–2012.

The paper is structured as follows: Sect. 2 describes related work tackling the task of tourism demand prediction when additionally considering travellers’ online search behaviour. Section 3 discusses methodological issues and related techniques of data collection and preparation, respectively. Section 4 describes the process of model building, while major findings are discussed in Sect. 5. Finally, Sect. 6 summarizes the gained insights and provides an outlook on future research activities.

2 Related work

Being one of the important areas in tourism research, demand modelling and forecasting has attracted much attention of both academics and practitioners (Weiermair and Fuchs 1998; Song and Li 2008). Literature on *quantitative* demand modelling and prediction is dominated by two sub-categories: non-causal time series models and causal econometric approaches.

A time-series model explains a variable with regard to its own past and a random disturbance term (Höpken et al. 2017, p. 189). In the past four decades, integrated autoregressive moving average (ARIMA) models proposed by Box and Jenkins (1970) dominated the tourism literature (Song and Li 2008, p. 210). Similarly, exponential smoothing models have appeared in the literature. One of the major advantages of econometric approaches over time-series models lies in their ability to analyse causal relationships between the tourism demand (dependent) variable and its influencing factors (explanatory variables) (Peng et al. 2014). Recent econometric forecasting studies have shown strong relationships between tourism demand and the following leading economic indicators: Consumer price index, gross domestic product (as proxy for tourists' income), exchange rates, interest and unemployment rate, money supply (M3), and export/import rates (Song and Li 2008, p. 211; Cho 2001). In addition, man-made events (i.e. especially mega-events), advertising investments (Divisekera and Kulendran 2006; Kronenberg et al. 2016), but also crises (e.g. financial crises, terrorist attacks) and natural disasters (SARS, foot and mouth disease, etc.), i.e. external shocks, significantly influence tourism demand (Höpken et al. 2017, p. 189).

In order to avoid spurious regression results, typically accruing with ordinary least square (OLS) techniques if applied to time series data, autoregressive distributed lag models (ADML), the error correction model (EDM), the vector autoregressive (VAR) model and the time varying parameter (TVP) model emerged as the main econometric models (Peng et al. 2014). In addition, also the linear structural equation model (SEM) has been used for tourism demand modelling (Turner and Witt 2001) (see Höpken et al. 2017, p. 190).

New, web-based data sources, like search engine traffic, web traffic, or customer feedback on online review platforms, typically have a natural relationship with tourism demand. Since the availability of such 'big data' sources has increased, they have also been used for tourism demand prediction (Höpken et al. 2017, p. 190, 2018; Fuchs et al. 2018). Thus, an increasing number of tourism researchers are demonstrating that, in particular, Google search engine traffic has the potential to greatly increase forecasting accuracy (Bangwayo-Skeete and Skeete 2015; Önder and Gunter 2016; Höpken et al. 2017). For instance, Bangwayo-Skeete and Skeete (2015) highlighted that time series data obtained by *Google Trends* show the capacity to improve the accuracy in tourism demand forecasting, both for long- and short-term predictions when using autoregressive mixed-data sampling (AR-MIDAS) models. Similarly, Önder and Gunter (2016) demonstrate that Google search engine traffic for web and image search increases accuracy of tourism demand prediction, compared to a purely autoregressive model or an exponential (i.e. Holt–Winters) smoothing time-series model. A recent study by Yang et al. (2015) uses web search volume to predict tourist arrivals for a popular tourist destination in China and demonstrates that search engine data helps to improve forecasting accuracy significantly compared to auto-regressive moving average (ARMA) models. While Pan et al. (2012) utilize search engine data to enhance the forecast accuracy of hotel (i.e. room) demand, other studies are primarily focussing on the prediction of tourist arrivals (Li et al. 2016; Höpken et al. 2017). Finally, the study by Yang et al. (2014) confirms the value of web traffic data from local destination marketing organizations (DMOs) in predicting the demand for hotel rooms in a tourist destination.

Despite that there is no standard methodology for pre-processing web search data, three main necessary tasks in pre-processing search engine queries for prediction purposes are found in the literature:

1. *Keyword selection* First, researchers start selecting domain specific keyword candidates either by using domain specific knowledge or web scraping and text mining approaches to capture domain specific grammar, or by the help of keyword recommendations from search engine providers (Liu et al. 2012).
2. *Dimensionality reduction* Second, recent studies typically have calculated temporal relationships between candidate queries and dependent (i.e. time series) variables (e.g. tourist arrivals) to identify most significant time differences between arrivals and respective search queries.
3. *Index construction* Since multi-collinearity and overfitting problems may occur when fitting linear models with a large number of high-dimensional time series data, *dimensionality reduction* is crucial when specifying input data (Varian 2014). Therefore, the third typical task when applying search engine data for forecasting purposes is the construction of an appropriate data set consisting of input variables with significant predictive power. For example, Liu et al. (2012) prevent possible collinearity by aggregating highly correlated search query series with the target series into one single index variable. More precisely, in their study the authors employed search engine queries for the prediction of the Chinese stock market. They demonstrate that using lagged search query data leads to significantly higher forecasting performance. Recently, Yang et al. (2015) adopted this approach when forecasting tourism demand by considering online search data.

The importance of identifying significant time lags between predictors and the target data series is highlighted in the literature, thus, several measures for estimating lag relationships exist. While Liu et al. (2012) use mixed metrics for measuring the similarity between lagged predictors and the target time series variable (i.e. *Pearson correlation* and *Kullback–Leibler divergences*), other studies exclusively rely on *Pearson correlation* to identify significant lags (Yang et al. 2015; Li et al. 2016; Pan et al. 2017). However, the reliability of the *Pearson correlation* coefficient is limited as it depends on statistical assumptions. Thus, it can only consider linear relations in data, wherefore it cannot capture issues such as non-stationary time series. Kristoufek (2014) has shown that the use of *Pearson correlation* coefficients is "practically useless for non-stationary time series" (ibid 2014, p. 293). This conclusion is supported by Zebende (2011) and Podobnik et al. (2011), suggesting the 'de-trended cross-correlation analysis' coefficient (DCCA) as the most appropriate method to calculate an unbiased correlation coefficient between potentially non-stationary time series.

As search query data can also contain useless information, Li et al. (2016) proposed a method for noise reduction of search query series which builds on the methodology by Yang et al. (2015). The authors conclude that noise processing is an essential step in forecasting with *Google Trends* data. Typically, 'Hilbert–Huang-Transformation' (HHT) is applied for noise processing, which reduces prediction errors significantly (Li et al. 2016). Peng et al. (2017) recently made another progress. The authors have shown that amending search queries with *Hurst exponent* values different to the target series yields higher prediction accuracy. Finally, a common drawback when using aggregated variables for forecasting purposes is that relevant information could be lost. Therefore, according to Pan et al. (2017), *principal component analysis* (PCA) or *generalized dynamic factor models* (Forni et al. 2000) are recommended.

3 Data collection and preparation

3.1 Data set specification

The employed initial data set consists of monthly aggregated tourist arrivals (i.e. December 2005–April 2012) for the leading Swedish mountain destination Åre, specified separately for its major sending countries (i.e. Denmark, Finland, Norway and the United Kingdom). Overall, the data set contains past tourist arrivals for 77 months separated by the four sending countries, which results in a total of 308 data entries. Besides past arrival data, aggregated web search traffic information is included. The latter attribute is extracted for each sending country, separately. As in previous studies, *Google Trends* was selected as an appropriate data source for web search traffic, since Internet users from most sending countries mainly use Google for searching the web. *Google Trends* is a service provided by Google which represents the relative search volume of popular search terms over time and, thus, reflects peoples' interest by specific search terms across different geographic regions and topical domains. When it comes to the sending countries analysed in this study, Google's market share is higher than 90% for Denmark, Finland, Norway and the United Kingdom, respectively (Pearson CMG 2017).

3.2 Collection of web search data

In contrast to Yang et al. (2015), the selection of appropriate keywords was limited to search engine-based keyword recommendations. Accordingly, keywords suggested by *Google*'*s Keyword Planner* tool for 'Å/åre' were obtained to generate appropriate seed queries that Åre visitors from different sending countries are likely to use. Next, queries were filtered by region and language for each sending country to reflect sending country specific search behaviour. Subsequently, the suggested keywords were filtered according to two matching rules in order to prevent potential noise caused by irrelevant search queries. Therefore, only keywords strongly related to Åre (i.e. keywords containing 'Åre', 'åre', 'are' or 'ore') as well as keywords containing 'ski' and either 'sweden' or 'sverige' were chosen. For query extraction, an algorithm for automatically crawling *Google Trends* with keyword suggestions has been developed. The algorithm iterates over the suggested keywords and extracts corresponding query series. In case a query series was found for a specific keyword, the algorithm further tries to resolve related queries (i.e. specified alternatives for the given keyword). In case no related queries were found for a given keyword, the keyword is skipped. The algorithm for retrieving queries has been implemented with *Spyder*® ('Scientific Python Development Environment'), thus, heavily depending on the 'pyTrends' framework for the Python programming language.

It should be noted that *Google Trends* data used in this study are provided in a normalized format. First, the search intensity for a given search term or topic is provided as a proportion of all searches on all topics on Google at that time and location. Second, *Google Trends* data is normalized between zero and 100 for the selected time-period and location.

3.3 Normalization of search terms

Search terms were further examined for close similarity based on linguistic variations, synonyms or misspellings (Liu 2008). More precisely, similar search terms were merged for two reasons. Due to low search intensity, some query series were found to contain an over proportional amount of zero values and, therefore, might lack data quality. It is further assumed that normalization of semantically identical search terms can improve predictive power (ibid 2008). Additionally, the intention was to capture travellers' interests more naturally. As there is no difference whether users search for "skiing in sweden" or "sweden skiing" or "ski sweden", these queries are likely to point at the same topical subject: Skiing in Sweden. To sum up, the normalization procedure aims at generating most meaningful search queries, which reflect travellers' intentions accurately. Therefore, the search terms of the query series were, first, transformed using text processing techniques in order to achieve similarity matches between search terms. More precisely, tokenization, character substitution, stemming and elimination of stop-words has been performed (Liu 2008). After these pre-processing steps, the search terms for each query were transformed into a word vector in order to calculate similarity matches based on cosine similarity. The cosine similarity cos(θ) between two vectors A and B is defined as follows:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}_2\| \|\mathbf{B}_2\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

with A_i and B_i being character occurrences of word vectors A and B, respectively. Search queries with cosine similarity equal to one (i.e. containing the same words and having the same word occurrences) have been merged. Overall, four types of irregularities concerning query names were handled. First, queries were found to contain the same terms but are arranged differently (e.g. 'åre ski' vs. 'ski åre'). Second, keywords which differed from others only by the presence of stop-words (e.g. 'skiing åre' vs. 'skiing in åre'). Third, nearly similar query names were identified by travellers searching for 'åre ski' rather than 'åre skiing'. As a consequence, queries were transformed in correspondence of their word stem (e.g. 'skiing' was transformed to 'ski'). Furthermore, some travellers prefer searching for travel-related information in their native language, while others feel comfortable to search for similar information in English or bilingually as well. Thus, semantically identical keywords occurred in different languages (e.g. 'åre sää' and 'åre weather'). Other acquired queries were found to be spelled either with or without special characters (e.g. 'åre', 'are' and 'ore'). Therefore, Nordic special characters {Å, å, Ä, ä, Æ, æ} and {Ø, ø Ö, ö} were substituted by {a} and {o}, respectively. Finally, search queries were analysed to find additional semantical conformances within the datasets. By doing so, it was found that some queries pointed to the same topic, although they were formulated differently. Accordingly, semantically related queries pointing to the same topic, like 'åre' and 'åre sweden' or 'copperhill åre' and 'copperhill mountain lodge åre' were merged. As cosine similarity is not sufficient for identifying topic-based similarities, merging those queries was performed manually.

4 Construction of web search indices with high predictive power

4.1 Construction of aggregated search indices

As query series reflect certain behavioural aspects of tourism demand, the entirety of all queries may represent drivers behind future tourism demand thereby showing the capacity to discover important trends for the development of a specific destination (Liu et al. 2012). According to the methodology proposed by Yang et al. (2015), search query series for each sending country were aggregated to compound search indices by shifting single search query series by the most appropriate time lag. Therefore, before constructing search indices, temporal relationships between search query series and tourism arrival data need to be identified, since the use of time-lagged predictors can raise forecasting performance significantly (Liu et al. 2012). As a common method to verify temporal relationships of different economic indicators with certain target variables, *cross-correlations* were calculated to identify time lags with maximum correlation between search queries and arrival series. However, as suggested in the literature, instead of the *Pearson correlation* coefficient, the *de-trended cross-correlation analysis* coefficient (DCCA) was used to reliably handle the correlation between possibly non-stationary time series (Zebende 2011; Podobnik et al. 2011).

As it is assumed, that $\mathrm{y} = \{y_1, y_2, \ldots, y_n\}$ is the target variable (i.e. time series of tourist arrivals) and $\mathrm{x} = \{x_1, x_2, \ldots, x_n\}$ is an indicator time series (i.e. search query from *Google Trends*), then the *de-trended cross-correlation analysis* coefficient r for time lag l can be calculated for up to L time lags as follows:

$$r_l = \rho_{DCCA}(s)_l = \frac{F^2_{DCCA}(s)}{F_{DFA,x}(s)F_{DFA,y}(s)}, \quad l = 0, 1, 2, \ldots, L,$$

where, $F^2_{\mathrm{DCCA}(s)}$ is the *de-trended co-variance* between partial sums $\{y_t\}$ and $\{x_{\mathrm{t}}\}$ for a window size s, while $F_{\mathrm{DFA},x}(s)$ and $F_{\mathrm{DFA},y}(s)$ are de-trended variances of partial sums $\{y_t\}$ and $\{x_t\}$ for a window size s, respectively. Window size was specified as $s = 25$ to capture long-term dependencies between both series. The largest cross-correlation of r^n_i is denoted by R^n_i and represents the most significant lag for x when estimating values of y. From a statistical point of view, the maximum correlation coefficient indicates that keyword i was most likely queried n periods prior to the specific arrivals, and is denoted as:

$$R^n_i = \max(r^l_i, r^{l+1}_i, \ldots, r^L_l).$$

All search queries were lagged by up to 6 months in order to capture travellers' short- and mid-term online travel planning behaviour (Fesenmaier et al. 2010). Accordingly, *de-trended cross-correlations* were calculated between the arrival series and each of the search queries series at lag {0, 1, 2, 3, 4, 5, 6), respectively. Thus, in total, 7 correlation coefficients were calculated for each search query and corresponding tourist arrivals, 0–6 months ahead, respectively. Based on the results of the *de-trended cross-correlation analysis* (DCCA), the queries were weighted by their maximum *cross correlation* coefficient by multiplying each query i by R^n_i. Next, each query $R^{n<0}_i$ was shifted according to n time lags towards the arrivals series, while queries with R^0_i were excluded from the data set, since as suggested by Yang et al. (2015), only web search activities executed at least 1 month prior to departure typically show any predictive power in forecast models for tourism demand.

Furthermore, according to Peng et al. (2017), the queries were filtered by *Hurst exponent* in order to assure the search indices to be constructed following the same auto-correlative patterns as its corresponding tourist arrival series. Finally, for each sending country, the number of queries to be included in each of the search indices was limited by a backward-stepwise regression, assuring that only significant predictor series are selected for index aggregation. Hence, the limitation of search query series to be included in the index aggregation procedure is based on a trade-off between forecasting performance and model parsimony and generalizability, respectively.

4.2 Evaluation of search indices

Although high correlation between input and target series suggests that most fluctuations of the target series can be explained well, it does not necessarily imply predictive power. Thus, as mentioned, in addition to squared correlation and *de-trended cross correlation* coefficients describing the structural similarity between time series, the *fluctuation memorability* was analysed by calculating *Hurst exponents* for both series (Hurst et al. 1965). Moreover, as suggested in the literature (Song et al. 2010), Granger causality has been chosen as an additional criterion for index evaluation. According to Granger (1969, 1988), a variable x_t is causally related to y_t, if the forecasting performance of y_t marginally improves by the inclusion of x_t. Results in Table 1 show that the structural similarity between indices and corresponding target series is high (i.e. all squared correlation coefficients greater than 0.5; all *de-trended cross-correlation* coefficients greater than 0.75). As expected, results also show that both the predictors and target series follow long-term positive auto-correlation patterns, i.e. following the *Hurst exponent*, all arrival and corresponding search index data sets are in the range 0.5–1. This means, that high values will likely be followed by high values. Moreover, values of both series tend to increase over time. Finally, the statistically significant results of the *Granger-Causality* test empirically show that prediction accuracy can be improved when autoregressive forecasting models are extended by an additional predictor variable in terms of search query indices. For the four sending countries, Table 1 summarizes the squared correlation and de-trended cross correlation between the search index and the arrival data series (for testing structural similarity), the Hurst exponent (for testing fluctuation memorability) for both series, and the t statistic with corresponding F values (for testing Granger causality between the two series).

5 Model building and evaluation

As a traditional forecasting approach, linear regression has been chosen as the statistical technique for predicting tourist arrivals (Frechtling 2002; Song et al. 2010). Hence, this study puts a clear emphasis on adding search traffic data as an additional input to predicting tourist arrivals, and not on comparing different forecasting

Table 1 Index evaluation metrics

Data set:	Structural similarity		Fluctuation memorability		Granger causality	
Sending country	Squared correlation	DCCA	Hurst exponent (arrivals/index)		*t* statistic	*F* value
Denmark	0.606	0.891	0.572	0.635	3.663	0.061
Finland	0.504	0.825	0.541	0.613	5.638	0.023
Norway	0.530	0.794	0.590	0.569	6.699	0.011
United Kingdom	0.610	0.834	0.564	0.603	5.684	0.025

approaches like linear regression and *machine learning* methods, like artificial neural networks or even non-parametric approaches from the area of deep learning. While predictions of future tourist arrivals using univariate approaches typically rely on past arrivals exclusively, multivariate forecasts make use of additional 'exogenous' variables. In the study at hand, travel-related search engine traffic over time is used as additional input attribute. The aim of this study is to evaluate how the accuracy of the prediction is increased by the inclusion of search traffic data. Thus, a purely autoregressive approach, using a window of *p* past arrival values as input data (i.e. an autoregressive model AR[p] of order p), is used as a baseline, and is compared to the extended approach, adding search engine traffic. Forecast accuracy is usually defined as a reduction of prediction errors (Frechtling 2002). In this study, prediction accuracy is operationalized by the root mean square error (RMSE), which according to Frechtling (2002) and Kim and Kim (2016), is among the most commonly used metrics when evaluating the performance of time series forecasting.

Besides choosing the right measure to evaluate forecasting accuracy, another important consideration is to choose the appropriate *validation* method for time series data (Frechtling 2002). In order to avoid overfitting, the evaluation of the forecasting performance in this study, is based on a *sliding window* approach (Song et al. 2010). Accordingly, a fraction of data entries is used as *training* data, while a consecutive fraction of data entries is used as *test* data (Liu 2008). Both fractions are successively shifted along the complete data set to compute specific forecasting performance measures for each fraction. Prediction accuracy was calculated by averaging forecasting errors of each fraction. To validate whether the regression models captured the relationship between input and output attributes well, the residuals of the forecasting models were tested for normal distribution by applying the *Shapiro–Wilk* test (Hill et al. 2011).

When modelling time series with statistical approaches it is common to ensure that the time series is within probabilistic limits of *stationarity*. Time series are stationary when their mean and variance are constant and auto-correlations between two values only depend on the time lag but not the point in time within the series (Frechtling 2002). Thus, "when regressing over non-stationary time series, traditional statistical approaches fail to generate reliable results" (Mukherjee et al. 1998, p. 335). Therefore, before building the regression model for evaluating the predictive power of the indices, the series were checked for stationarity by applying the *Augmented Dickey–Fuller* (ADF) test, which tests an autoregressive model for the existence of *unit-roots* as an indicator for non-stationarity (Baddeley and Barrowclough 2009). Additionally, the *Kwiatkowski–Phillips–Schmidt–Shin* (KPSS) test has been applied (Hill et al. 2011). In contrast to the ADF test having non-stationarity as the null hypothesis, the KPSS test's null hypothesis is stationarity. Additionally, *co-integration* relationships were tested by applying the *Johansen* test to check whether any further data transformations were necessary in case a time series was found to be non-stationary (ibid 2011).

Results in Table 2 confirm stationarity for the data sets Denmark (DK), Finland (FI) and the United Kingdom (UK) as no unit roots could be found with statistical significance. KPSS results of Arrivals$_{(DK)}$ can be considered as borderline, as test statistics range slightly above the outlined *p* value. However, the null hypothesis

Table 2 Tests for stationarity and co-integration for arrival data sets and corresponding indices

Data set	Stationarity-tests				Cointegration-test	
	ADF-test		KPSS-test		Johansen-test (max. eigenvalue $r=1$)	
	Test statistic	p value	Test statistic	p value	Test statistic	c value $^{1\%}$
Arrivals$_{(DK)}$	4.827	0.01	0.107	0.100	20.410	11.650
Index$_{(DK)}$	5.703	0.01	0.050	0.100		
Arrivals$_{(FI)}$	5.130	0.01	0.027	0.100	15.842	11.650
Index$_{(FI)}$	5.645	0.01	0.057	0.100		
Arrivals$_{(NO)}$	4.217	0.01	0.690	0.014	18.146	11.650
Index$_{(NO)}$	4.910	0.01	0.502	0.041		
Arrivals$_{(UK)}$	5.425	0.01	0.027	0.100	17.548	11.650
Index$_{(UK)}$	6.301	0.01	0.046	0.100		

of stationarity was rejected for both data sets corresponding to Norway, suggesting that both data sets are non-stationary. Nevertheless, the results of the *Johansen* test clearly show the existence of co-integration relationships between the constructed search indices and their corresponding arrival series, as the null hypothesis (i.e. no co-integration) could be rejected for all the data sets at a significance level of 1% (i.e. all values greater than corresponding *c* values). Therefore, no further data transformation was necessary to prevent spurious regression caused by non-stationary data sets.

R-Statistics® has been used for all statistical computations in this study. Moreover, the process for search query-based tourism demand prediction has been implemented with *Rapid Miner Studio*®, a data mining tool-set with integration capabilities for software modules written in 'R' and 'Python', respectively.

6 Results

6.1 Comparison of the forecasting performance

The prediction of tourist arrivals has been executed *autoregressively* (i.e. based on past tourist arrivals alone) and based on search engine data as additional model input (Frechtling 2002). The prediction models (i.e. 'autoregressive only' and 'autoregressive with search query indices') have been learned and evaluated for all four sending countries, separately. For evaluating different forecasting characteristics of the specific search indices, the prediction task was executed with forecasting horizons of 3, 6 and 12 months, respectively. Table 3 shows the *root mean squared error* (RMSE) of the different autoregressive models, the error reduction by adding *Google Trends* data to the pure autoregressive approaches. Finally, a *Shapiro–Wilk* test has been executed for models with *Google Trends* data in order to check if residuals (i.e. the difference between predicted and actual values) are normally distributed. The latter

Table 3 Comparison of prediction accuracy at different forecasting horizons

Forecasting horizon	Sending country	Autoregressive only (RMSE)	With Google Trends (RMSE)	Difference in (%)	Shapiro–Wilk
3 months	Denmark	1035.40	966.25	– 6.68	0.00
	Finland	782.79	737.92	– 5.73	0.00
	Norway	1439.54	949.48	– 34.04	0.01
	UK	327.64	326.55	– 0.33	0.00
6 months	Denmark	833.44	809.09	– 2.92	0.00
	Finland	935.69	740.03	– 20.91	0.04
	Norway	1336.86	1080.63	– 19.17	0.03
	UK	328.83	322.32	– 1.98	0.00
12 months	Denmark	892.15	726.95	– 18.52	0.05
	Finland	1052.90	719.11	– 31.70	0.06
	Norway	1335.23	1307.61	– 2.07	0.01
	UK	426.23	381.13	– 10.58	0.00

condition is considered as an indicator that the regression model has reliably captured the relationship between input and output attributes (Baddeley and Barrowclough 2009; Hill et al. 2011).

Results in Table 3 clearly show that utilizing online search traffic in forecasting tourism demand raises the performance significantly, as the RMSE is reduced for all sending countries and at any forecasting horizon, if autoregressive models were extended by search query series. Finally, results of the *Shapiro–Wilk* test clearly coincide with the reduction of the RMSE by adding Google Trends data. In all cases with a significant reduction of the RMSE, the likelihood of the residuals being normally distributed (i.e. the *p* value of the *Shapiro–Wilk* test) is higher than in cases without a significant RMSE reduction. At the same time, the *Shapiro–Wilk* test shows that most models still offer room for improvement. As mentioned before, more flexible and often more powerful machine learning approaches, like artificial neural networks, constitute promising approaches to further increase prediction accuracy. It is important to note that the results of the Shapiro–Wilk test do by no means affect or question the reliability of the tests on stationarity or co-integration (cf. Table 2) or the validity of the overall approach.

6.2 Analysis of relevant search queries

As highlighted above, analysing the correlation between tourist arrivals and query series with different time lags enables conclusions about consumers' online search behaviour (Fesenmaier et al. 2010). While travellers from Denmark start to search for inspiration first by activity-related topics (i.e. queries related to skiing) 3 months prior to departure without mentioning any destination, search queries are formulated more precisely 2 months ahead of departure (i.e. by mentioning Sweden as a possible destination). Queries executed 1 month prior to departure are formulated even

Table 4 Significant query lags (sig. level 0.001) for sending country Denmark

Lag	DCCA	Query	Topic	Category
−1	0.599	åre ski	Skiing in Åre, Sweden	Activities
−1	0.793	åre	Åre, Sweden	Location
−1	0.670	åre sverige + åre sweden	Åre, Sweden	Location
−1	0.841	ski sverige + ski i sverige + skisport sverige	Skiing in Sweden	Activities
−1	0.515	skistar åre	Skistar Åre, Sweden	Brand
−1	0.740	skisteder sverige + skisteder i sverige + skisportssteder sverige	Ski resorts in Sweden	Activities
−2	0.783	skiferie + ski ferie	Skiing holiday	Activities
−2	0.694	skiweekend sverige	Skiing holiday in Sweden	Activities
−2	0.737	val thorens skiferie	Skiing holiday in Val Thorens, France	Activities
−3	0.805	billig skiferie	Cheap skiing holiday	Activities
−3	0.807	skiferie sverige + skiferie i sverige	Skiing holiday in Sweden	Activities
−3	0.563	skihytte sverige	Chalet in Sweden	Activities

more precisely, as Åre is mentioned more frequently. Interestingly, in contrast to visitors from other examined sending countries, travellers from Denmark perform web searches for trip planning at least 1 month prior to the trip, as none of the shifted queries pointed to lag zero. Table 4 lists search queries with their most relevant time lag, the corresponding DCCA value, as well as topic and category the query deals with for the sending country Denmark.

For Finnish travellers, a rather high correlation coefficient of $r = 0.7$ can be observed for the queries 'Levi' and 'Ruka'. Since Levi and Ruka are popular skiing resorts in Finland, those queries indicate that Finish travellers more critically evaluate various ski resorts before they finally choose to visit Åre. However, in contrast to Danish travellers, visitors from Finland start searching more specifically for Åre as a destination. For instance, searches for cottages in Åre are performed already

Table 5 Significant query lags (sig. level 0.001) for sending country Finland

Lag	DCCA	Query	Topic	Category
−1	0.611	åre ski + åre laskettelu	Skiing in Åre, Sweden	Activities
−1	0.713	ruka	Ruka, Finland	Location
−1	0.742	skistar	Skistar	Brand
−1	0.682	skistar åre	Skistar in Åre, Sweden	Brand
−1	0.625	åre + are	Åre, Sweden	Location
−2	0.531	holiday club åre	Hotel in Åre, Sweden	Lodging
−2	0.704	levi	Levi, Finland	Location
−3	0.650	åre majoitus	Hotels in Åre, Sweden	Lodging
−3	0.671	åre matkat	Travel to Åre, Sweden	Location
−4	0.148	åre mökit	Cottages in Åre, Sweden	Lodging

4 months prior to arrival, followed by queries for the destination Åre as a whole 3 months ahead of the trip. Finally, 1 month prior to arrival, especially skiing-related queries are executed. Table 5 lists search queries with their most relevant time lag, the corresponding DCCA value, as well as the topic and category the query deals with for the sending country Finland.

Interestingly enough, the search queries of Norwegian tourists are characterized by high diversity, thus, show much more detail concerning travellers' demand. More precisely, Norwegian travellers search for lodging and activity-related information before visiting Åre. The elicited queries suggest that Åre is a very popular skiing destination for travellers from Norway, as, instead of comparing different ski resorts, like travellers from Denmark and Finland are doing, almost all the queries are pointing specifically to Åre. These findings are in line with results gained by Kronenberg et al. (2016), who identified customers from Denmark and Finland as being more price-elastic and more aware of competing destinations than customers from Norway (as well as UK and Russia). Table 6 lists search queries with their most relevant time lag, the corresponding DCCA value, as well as the topic and category the query deals with for the sending country Norway.

Finally, the UK data set primarily contains search queries related to the topic 'Skiing in Are' (i.e. 'åre ski', 'åre sweden ski', 'åre ski resort', 'skistar åre'), queries with the topic 'Skiing in Sweden' (i.e. 'sweden ski resorts' and 'sweden skiing') as well as location-based queries that point to Åre and its neighbouring destination Östersund (i.e. 'ostersund' and 'ostersund sweden'). Table 7 lists search queries with their most relevant time lag, the corresponding DCCA value, as well as the topic and category the query deals with for the sending country United Kingdom.

Table 6 Significant query lags (sig. level 0.001) for sending country Norway

Lag	DCCA	Query	Topic	Category
− 1	0.41	copperhill åre + copperhill mountain lodge i åre + ⋯ + copperhill	Hotel in Åre, Sweden	Lodging
− 1	0.63	hytte åre	Cottages in Åre, Sweden	Lodging
− 1	0.42	åre continental inn	Hotel in Åre, Sweden	Lodging
− 1	0.37	holiday club åre	Hotel in Åre, Sweden	Lodging
− 1	0.46	hotell åre	Hotels in Åre, Sweden	Lodging
− 1	0.52	overnatting åre	Accommodation in Åre, Sweden	Lodging
− 1	0.45	skistar åre	Skistar in Åre, Sweden	Brand
− 1	0.64	åre	Åre, Sweden	Location
− 1	0.01	åre sverige + åre sweden	Åre, Sweden	Location
− 3	0.07	fjellgården åre	Hotel in Åre, Sweden	Lodging
− 4	0.36	åre skipass	Skiing in Åre, Sweden	Activities
− 5	0.09	åreskutan	Åreskutan, Sweden	Location
− 5	0.34	åre bike park	Bike park in Åre, Sweden	Activities
− 6	0.19	holiday club	Hotel in Åre, Sweden	Lodging
− 6	0.51	åre camping + camping åre	Camping in Åre, Sweden	Lodging
− 6	0.02	åre skianlegg	Skiing in Åre, Sweden	Activities

Table 7 Significant query lags (sig. level 0.001) for sending country United Kingdom

Lag	DCCA	Query	Topic	Category
−1	0.79	åre	Åre, Sweden	Location
−1	0.66	are ski + ski are + are sweden ski	Skiing in Åre, Sweden	Activities
−1	0.47	are sweden + are in sweden	Åre, Sweden	Location
−1	0.39	skistar åre	Skistar in Åre, Sweden	Brand
−1	0.45	are webcam	Webcam for Åre, Sweden	Location
−2	0.36	ostersund + östersund + ostersund sweden	Östersund, Sweden	Location
−2	0.12	åre ski resort	Ski resorts in Åre, Sweden	Lodging
−3	0.45	skiing in sweden + ski sweden + sweden skiing	Skiing in Sweden	Activities
−6	0.19	are hotel	Hotels in Åre, Sweden	Lodging
−6	0.33	sweden weather forecast	Weather forecast for Sweden	Environment

7 Conclusion and outlook

The present study compared an autoregressive approach to forecast tourist arrivals by using only past arrivals as input attributes with an extended model that includes big data-based information sources as additional input. More concretely, *web search traffic* (i.e. obtained via *Google Trends*) has been added as additional input for predicting tourist arrivals. As a prediction method, the study used traditional linear regression (Frechtling 2002; Song et al. 2010). In addition, *Granger causality* tests were performed to examine the evidence for predictive power between constructed search indices and tourist arrival series. The proposed approach has been executed and evaluated for the leading Swedish mountain destination Åre by using arrival data and Google search data for the time period 2005 to 2012.

As a theoretical contribution, the study presented a novel approach to construct tourism-related search indices from Google Trends data as additional input to predict tourism demand. In contrast to Yang et al. (2015), the keyword selection is solely based on search engine-based keyword recommendations by *Google*'*s Keyword Planner* tool. Following the work of Liu (2008), a domain-specific mechanism for iteratively extracting, filtering and normalizing search terms has been developed. Compared with existing literature, cosine similarity has been used to improve query normalization. Adapting findings from Liu (2008), Yang et al. (2015) and Zebende (2011), time-lagged search queries for relevant search terms are finally aggregated into a compound search index. Instead of the *Pearson correlation* coefficient, used by previous studies (Liu et al. 2012; Yang et al. 2015; Li et al. 2016; Pan et al. 2017), the *de-trended cross-correlation analysis* coefficient (DCCA) has been used to identify relevant search queries. To the best of the authors' knowledge, there is no study utilizing DCCA analysis, when forecasting tourism demand with Google Trends data, so far. When using resulting search indices and past arrivals as input to demand prediction and in contrast to existing (autoregressive) approaches, an automatic selection of the most appropriate time lags is performed by a backward feature selection mechanism. Findings clearly revealed that tourism-related search queries show the capacity to *significantly increase accuracy levels in predicting tourist arrivals* compared to using past arrivals alone (i.e. pure autoregressive forecasting approach, as discussed by Song and Li 2008).

From a managerial perspective, the study demonstrates that *analysing search queries can reveal meaningful and managerially valuable insights* from sending country specific search behaviour. Therefore, the results reveal important implications for tourism managers and policy makers: *Google Trends* data can be effectively used as a tool for forecasting short- and mid-term tourism demand as well as for the detection of future (i.e. long-term) trends and demand fluctuations. Additionally, search engine data can be used by local tourism suppliers for marketing purposes to better understand the decision-making process of travellers when choosing a specific destination, e.g. which tourism services and attractions are most heavily searched and, thus, is of particular relevance for travellers from various sending countries (Fesenmaier et al. 2010).

When it comes to study limitations, the matching capabilities of cosine similarity for merging semantically identical search queries are likely restricted. Thus, in order to detect further cases of semantically identical search queries automatically, for information extraction we recommend the application of text analytics tools with the potential to identify entities and related topics without the need for human intervention (Schmunk et al. 2014; Menner et al. 2016; Höpken et al. 2016). Additionally, in future research, causal chain patterns between lagged queries could be explored by *Granger causality* analysis in order to analyse in greater detail how travellers behave when planning a trip to a specific destination and with respect to specific motivations for certain tourism and travel activities. Finally, the current study was limited to using the statistical approach of linear regression to estimate future tourist arrivals. In future research studies, more flexible machine learning approaches can be used, like artificial neural networks or even non-parametric approaches from the area of deep learning. The latter methods offer the advantage of being more robust against violations of input data requirements and noisy data.

References

Baddeley MC, Barrowclough D (2009) Running regressions—a practical guide to quantitative research in economics, finance and development studies. University Press, Cambridge

Bangwayo-Skeete PF, Skeete RW (2015) Can Google data improve the forecasting performance of tourist arrivals? A mixed-data sampling approach. Tour Manag 46:454–464

Box GE, Jenkins GM (1970) Time series analysis, forecasting and control. Holden Day, San Francisco

Carrière-Swallow Y, Labbé F (2013) Nowcasting with Google Trends in an emerging market. J Forecast 32(4):289–298

Chekalina T, Fuchs M, Lexhagen M (2018) Customer-based destination brand equity modelling—the role of destination resources, value-for money and value-in-use. J Travel Res 57(1):31–51

Cho V (2001) Tourism forecasting and its relationship with leading economic indicators. J Hosp Tour Res 25:399–420

Divisekera S, Kulendran N (2006) Economic effects of advertising on tourism demand. Tour Econ 12:187–205

Edgell DL Sr, Del Mastro Allen M, Smith G, Swanson JR (2008) Tourism policy and planning—yesterday, today and tomorrow. Routledge, New York

Fesenmaier DR, Xiang Z, Pan B, Law R (2010) An analysis of search engine use for travel planning. In: Gretzel U, Law R, Fuchs M (eds) Information and communication technologies in tourism. Springer, New York, pp 381–392

Fitzsimmons JA, Fitzsimmons MJ (2001) Service management—operations, strategy & technology, 3rd edn. McGraw Hill, New York

Forni M, Hallin M, Lippi M, Reichlin L (2000) The generalized dynamic-factor model: identification and estimation. Rev Econ Stat 82(4):540–554

Frechtling DC (2002) Forecasting tourism demand. Butherworth-Heinemann, Oxford

Fuchs M, Rijken L, Peters M, Weiermair K (2000) Modelling Asian incoming tourism—a shift-share approach. Asia Pac J Tour Res 5(2):1–10

Fuchs M, Höpken W, Lexhagen M (2018) Business Intelligence for Destinations: Creating Knowledge from Social Media. In: Sigala M, Gretzel U (eds) Advances in social media for travel, tourism and hospitality: new perspectives, practice and cases. Routledge, New York, pp 290–310

Granger CW (1969) Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37(3):424–438

Granger CW (1988) Some recent developments in a concept of causality. J Econom 39(1–2):199–211

Grönroos C (2008) Service logic revisited—who creates value? And who co-creates? Eur Bus Rev 20(4):298–314

Hill RC, Griffith WE, Lim GC (2011) Principles of econometrics, 4th edn. Wiley, New York

Höpken W, Fuchs M, Menner Th, Lexhagen M (2016) Sensing the online social sphere—the sentiment analytical approach. In: Xiang Zh, Alzua A, Fesenmaier D (eds) Analytics in smart tourism design—concepts and methods. Springer, Berlin, pp 129–146

Höpken W, Ernesti D, Fuchs M, Kronenberg K, Lexhagen M (2017) Big data as input for predicting tourist arrivals. In: Schegg R, Stangl B (eds) Information and communication technologies in tourism, Springer, Cham, pp 187–199

Höpken W, Eberle Th, Fuchs M, Lexhagen M (2018) Search engine traffic as input for predicting tourist arrivals. In: Stangl B, Pesonen J (eds) Information and communication technologies in tourism 2018. Springer, New York, pp 381–393

Hurst HE, Black RP, Simaika YM (1965) Long-term storage: an experimental study. Constable, London

Kim S, Kim A (2016) A new metric of absolute percentage error for intermittent demand forecasts. Int J Forecast 32(3):669–679

Kristoufek L (2014) Measuring correlations between non-stationary series with DCCA coefficient. Phys A 402:291–298

Kronenberg K, Fuchs M, Salman K, Lexhagen M, Höpken W (2016) Economic effects of advertising expenditures—a Swedish destination study of international tourists. Scand J Hosp Tour Res 16(4):352–374

Li X, Wu Q, Peng G, Lv B (2016) Tourism forecasting by search engine data with noise processing. Afr J Bus Manag 10(6):114–130

Liu B (2008) Web data mining—exploring hyperlinks, contents, and usage data. Springer, Heidelberg

Liu Y, Lv B, Peng G, Yuan Q (2012) A pre-processing method of Internet search data for prediction improvement. In: Proceedings of the data mining and intelligent knowledge management workshop, New York, ACM 2012:3:1–3:7

Menner Th, Höpken, Fuchs M, Lexhagen M (2016) Topic detection – Identifying relevant topics in tourism reviews. In: Inversini A, Schegg R (eds) Information and communication technologies in tourism 2016. Springer, New York, pp 411–423

Mukherjee C, White H, Wuyts M (1998) Econometrics and data analysis for developing countries. Routledge, New York

Önder I, Gunter U (2016) Forecasting tourism demand with Google Trends for a major European city destination. Tour Anal 21:203–220

Pan B, Wu C, Song H (2012) Forecasting hotel room demand using search engine data. J Hosp Tour Technol 3(3):196–210

Pan B, Li X, Law R, Huang X (2017) Forecasting tourism demand with composite search index. Tour Manag 59(1):57–66

Pearson CMG (2017) Internet and search engine use by country: global search engine marketing. http://ptgmedia.pearsoncmg.com/images/9780789747884/supplements/9780789747884_appC.pdf. Accessed 20 Feb 2018

Peng B, Song H, Crouch G (2014) A meta-analysis of international tourism demand forecasting and implications for practice. Tour Manag 45:181–193

Peng G, Liu Y, Wang J, Gu J (2017) Analysis of the prediction capability of web search data based on the HE-TDC method—prediction of the volume of daily tourism visitors. J Syst Sci Syst Eng 26(2):163–182

Pike A, Rodríguez-Pose A, Tomaney J (2017) Local and regional development, 2nd edn. Routledge, New York

Podobnik B, Jiang Z-Q, Zhou W, Stanley HE (2011) Statistical tests for power-law cross-correlated processes. Phys Rev E 84(066118):1–8

Schmunk S, Höpken W, Fuchs M, Lexhagen M (2014) Sentiment analysis—implementation and evaluation of methods for sentiment analysis with Rapid-Miner®. In: Xiang Ph, Tussyadiah I (eds) Information and communication technologies in tourism 2014. Springer, New York, pp 253–265

Song H, Li G (2008) Tourism demand modelling and forecasting: a review of recent research. Tour Manag 29:203–220

Song H, Li G, Witt StF, Fei B (2010) Tourism demand modelling and forecasting: how should demand be measured? Tour Econ 16(1):63–81

Turner LW, Witt SF (2001) Factors influencing demand for international tourism: tourism demand analysis using structural equation modelling. Tourism Economics 16(1):63–81

Varian H (2014) Big data: new tricks for econometrics. J Econom Perspect 28(2):3–28

Weiermair K, Fuchs M (1998) On the use and usefulness of economics in tourism: a critical survey. Int J Dev Plan Lit 13(3):255–273

WTTC (2016) Travel & tourism: economic impact 2016—world. World Travel & Tourism Council, London

Yang Y, Pan B, Song H (2014) Predicting hotel demand using destination marketing organizations' web traffic data. J Travel Res 53(4):433–447

Yang X, Pan B, Evans JA, Lv B (2015) Forecasting Chinese tourist volumes with search engine data. Tour Manag 46(3):386–397

Zebende G (2011) DCCA cross-correlation coefficient: quantifying level of cross-correlation. Phys A 390:614–618