Tourism Management 46 (2015) 386–397

Contents lists available at ScienceDirect

Tourism Management

journal homepage: www.elsevier.com/locate/tourman

Forecasting Chinese tourist volume with search engine data

Xin Yang [a, b, *], Bing Pan [c, 1], James A. Evans [b, 2], Benfu Lv [d, 3]

[a] Management School, University of Chinese Academy of Sciences, NO.80 Zhongguancun East Road Haidian District, Beijing 100190, China
[b] Department of Sociology, University of Chicago, 1126 East 59th Street, Chicago, IL 60637, USA
[c] Department of Hospitality and Tourism Management, School of Business, College of Charleston, 66 George Street, Charleston, SC 29424, USA
[d] Management School, University of Chinese Academy of Sciences, NO.80 Zhongguancun East Road Haidian District, Beijing 100190, China

HIGHLIGHTS

- Web search data help to improve visitor volume forecasting model accuracy.
- Co-integration relationship between search data and visitor volume is verified.
- Baidu data performs better than Google for predicting tourist activities in China.
- Process to select key search queries for visitor volume prediction is proposed.

ARTICLE INFO

Article history:
Received 23 March 2014
Accepted 30 July 2014
Available online 20 August 2014

Keywords:
Search engine data
Google Trends
Baidu Index
Chinese tourism market
Visitor prediction
Tourist volume forecast

ABSTRACT

The queries entered into search engines register hundreds of millions of different searches by tourists, not only reflecting the trends of the searchers' preferences for travel products, but also offering a prediction of their future travel behavior. This study used web search query volume to predict visitor numbers for a popular tourist destination in China, and compared the predictive power of the search data of two different search engines, Google and Baidu. The study verified the co-integration relationship between search engine query data and visitor volumes to Hainan Province. Compared to the corresponding auto-regression moving average (ARMA) models, both types of search engine data helped to significantly decrease forecasting errors. However, Baidu data performed better due to its larger market share in China. The study demonstrated the value of search engine data, proposed a method for selecting predictive queries, and showed the locality of the data for forecasting tourism demand.

1. Introduction

Tourism practitioners need accurate forecasts of tourist volume in order to effectively allocate resources and formulate pricing strategies (Song & Li, 2008). This has become especially important in China in recent years, due to the tremendous growth in tourism demand which has accompanied its economic growth. In China, the tourism industry is touted as one of the main sources of non-trade foreign exchange earnings. In 2013, the tourism income reached 42.9 billion RMB (approximately $7.15 billion USD), an increase of 13% compared to the previous year (China National Tourism Administration, 2014). This growth can be attributed to both the increased disposable income of Chinese citizens, as well as the government's policy of encouraging the consumption of travel products. However, the crowdedness during certain holidays has also become a major problem. For example, in one of China's National Scenic Areas, Jiuzhaigou, during one of the major holiday periods, the National Day on October 1, 2013, 4000 tourists were stuck at the entrance for five hours due to overcrowding. The administration of the scenic area had to issue a formal apology on news media (Qiu, 2013). This demonstrates the urgency for accurate forecasting, especially near-term forecasting, for anticipating and managing influxes of tourists.

Current tourist volume forecasting techniques include various statistical, econometric, and artificial intelligence methods (Song & Li, 2008). However, most of them rely on historical data to predict

* Corresponding author. Management School, University of Chinese Academy of Sciences, NO.80 Zhongguancun East Road Haidian District, Beijing 100190, China. Tel.: +86 010 8268 2865.

E-mail addresses: qwzhxyangxin@gmail.com, qwzhxyangxin@sina.com (X. Yang), bingpan@gmail.com (B. Pan), jevans@uchicago.edu (J.A. Evans), lubf@ucas.ac.cn (B. Lv).

[1] Tel.: +1 843 953 2025; fax: +1 843 953 5697.

[2] Tel.: +1 773 834 3612; fax: +1 773 702 4849.

[3] Tel.: +86 010 8268 0670.

http://dx.doi.org/10.1016/j.tourman.2014.07.019
0261-5177/© 2014 Elsevier Ltd. All rights reserved.

future tourist activities, assuming a consistent pattern and stable economic structure. During certain periods of dramatic change, or at certain one-off events, however, these methods may not provide accurate predictions. In addition, these forecasting techniques mainly focus on a long-term scale, such as annually or quarterly, instead of monthly or weekly, which limits their application in short-term forecasting.

The development of information technology, especially the Internet, has generated another type of data for forecasting. Every time a tourist interacts with the Internet, be it through a search engine, a website, a mobile phone, or a social media platform, the traces of the interaction can be captured, stored, and analyzed. As a result, a new area of forecasting with online data has blossomed. Researchers have used online data such as search engine query volumes, amount and types of tweets, website traffic, and social media posts to forecast contagious disease outbreaks (Carneiro & Mylonakis, 2009), consumer consumption (Vosen & Schmidt, 2011), popularity of songs and movies (Goel, Hofman, Lahaie, Pennock, & Watts, 2010), unemployment rates (Askitas & Zimmermann, 2009), and hotel room demand (Pan, Wu, & Song, 2012; Yang, Pan, & Song, 2014). Online data are useful as predictors not only in developed countries, but also in developing countries where the Internet adoption rate is low (Carrière-Swallow & Labbé, 2011).

Travelers use search engines to find relevant information for all aspects of a trip, including accommodations, attractions, activities, and dining (Pan, Litvin, & Goldman, 2006). In this study, we used the search query volume data provided by both Google and Baidu, two search engines used in China, to predict tourist volumes to a specific destination. According to China Internet Network Information Center (CNNIC), the number of Chinese citizens using the Internet has reached 591 million, and 80% of them queried search engines in 2013. Travel planning is one major search activity (China Internet Network Information Center, 2013). The main questions we wanted to address were: are search engine data a valid predictor for tourist volumes in China? Which search engine data are more powerful predictor, those from Google or Baidu? How can a researcher select the candidate queries for forecasting tourist volumes among possibly hundreds or thousands of queries?

In this study, we proposed a conceptual model on the role of search engines in the travel planning and travel process of tourists. We then compared the forecasting power of models with actual Google and Baidu data, with their equivalent time series counterparts. In this process, we also proposed a systematic way to obtain relevant search engine queries from search engine query volume tools. Thus, this work contributes to two areas of the literature. First, previous researchers only focused on one type of search query data, and did not discuss a search engine's locality when examining the relationship between search data and visitor numbers. Our study adds to the literature by comparing the goodness of fit and forecasting power of data from two types of search engines, Google and Baidu. Second, relative to the unclear search query selection process in previous studies, we proposed a systemic mechanism to better pick search queries for predicting visitor volumes to aid in improving the reproduction of our study.

2. Literature review

This section reviews relevant studies in the areas of tourism demand and socioeconomic activity forecasting with search engine query data. Specifically, a few recent studies on forecasting tourism demand with search engine data are also discussed and research gaps are identified.

2.1. Forecasting tourism demand

Researchers have adopted two main types of methods for forecasting tourism demand and tourist volumes. The first is based on time series or statistical techniques, such as linear regression, exponential smoothing, and autoregressive models (Song & Witt, 2000). The other type consists of artificial intelligence methods, such as artificial neural network, grey theory, rough set theory, fuzzy theory, genetic algorithm, Monte Carlo simulation, and expert systems (Abratt, Nel, & Nezer, 1995; Andrew, Cranage, & Lee, 1990; Law & Au, 1999; Weatherford & Kimes, 2003). However, recent studies have also demonstrated that no single method outperforms others in forecasting accuracy, and a combination of methods can produce better forecasting results (Chan, Witt, Lee, & Song, 2010; Song & Li, 2008).

The traditional time series model and many of its derivatives are well-established and widely adopted in forecasting tourism demand, and are superior to other methods (Song, Witt, & Li, 2008). In the Asian-Pacific area, Lim and McAleer (2002) compared several exponential smoothing models in order to estimate quarterly tourist volumes from Hong Kong, Malaysia and Singapore to Australia during the period of 1975–1999. The results showed that the Hoyt-Winter additive model and multi-variant seasonal model performed better than the secondary Hoyt-Winter non-seasonal exponential smoothing model. Liu (2008) demonstrated that the seasonal product model was more accurate than the autoregressive and exponential smoothing models in predicting the visitors to Guilin, China. Feng (2008) built an ARMA model that predicted visitors to Yue Temple in China from 1980 to 2007 and achieved a better result than other models. Feng (2008) used Hong Kong travel demands as an example and compared three typical visitor forecast techniques: exponential smoothing, univariate ARIMA, and Artificial Neural Networks. He concluded that Artificial Neural Networks performed better than the other two time series prediction methods. Chaitip and Chaiboonsri (2009) explored two different statistical models, one named X-12-ARIMA seasonal adjustment, and an autoregressive fractionally integrated moving average (ARFIMA), in the application of travel amount prediction and conducted an empirical study on India from 2007 to 2010. They argued that in order to achieve the most optimal forecast performance, one needs to diligently test parameters when using different statistical models.

The development of computer technology has facilitated the adoption of artificial intelligence methods in predicting tourism demand. The neural network method is especially useful when the relationship between predictors and predicted variables is nonlinear. For example, Chen, Chen, Xing, and Fu (2005) built an effective back propagate neural network model of tourism demand to Yunnan Province. Rough set, grey theory, fuzzy theory and other artificial intelligence methods have also been adopted. For instance, Goh and Law (2003) used rough set theory to predict visitor volumes from 10 source markets to Hong Kong with an accuracy of 87.2%. Weng, Zhen, Liu, and Zhang (2008) used the GM-Markov model to analyze and predict the amount of inbound tourists to China and achieved superior accuracy. Kan, Lee, and Chen (2010) adopted six grey models, each with different parameter settings, to predict Taiwan travel demands from 2009 to 2013. The simulation results showed that the highest growth rate would reach about 6% while the lowest growth rate would be around 5%. Alvarez-Diaz, Mateu-Sbert, and Rossello-Nadal (2009) employed a Genetic Program (GP) to forecast monthly visitors from UK and Germany to the Balearic Islands of Spain. Their empirical results showed the GP was more robust, easy-to-use, and allowed for simple ad hoc interpretation compared to other non-parametric methods.

However, all of these forecasting methods have their limitations. As stated before, time series and statistical analyses rely on a consistent historical pattern and a stable economic structure. Any dramatic structural changes in the economy or large scale one-off events may decrease their forecasting accuracy. In addition, artificial intelligence methods are both complicated and labor-intensive, and also require a large amount of training data.

2.2. *Forecasting of socioeconomic behavior with search engine data*

One way to increase forecasting accuracy is by incorporating more powerful predictors. The deluge of online data, in the form of traces of users' online behavior, provides much potential for increasing forecasting accuracy.

With the wide adoption of the Internet, search engines have become one primary way for finding information online. For example, in 2012, 85% of Americans used the Internet, and 91% of those used search engines to find information, which was the number one online activity (Pew Internet & American Life Project, 2013). Similarly, in China in 2012, 40% of people used the Internet, and 80% used search engines to find information, which was the second most popular activity online, exceeded only by instant messaging (China Internet Network Information Center, 2013). The traces the users left behind, in the form of search engine query types and volume, became indicators of users' interests, behavior, and attitude. Furthermore, major search engines, such as Google and Baidu, publish their scaled search engine volume data publicly online (Zhang & Li, 2011). These types of data have become an invaluable source for monitoring and predicting socioeconomic activities, which has led to a new, burgeoning field of research: predicting economic activities based on search engine data.

Search engine data has been widely used in forecasting diseases (Ginsberg et al., 2009), ranking universities (Vaughan & Romero-Frías, 2013), and gathering public opinions (Baram-Tsabari & Segev, 2011; Ripberger, 2011). Search engine data can be used to make valuable predictions in developing countries, even where the Internet adoption rates are low. Carrière-Swallow and Labbé (2011) constructed an Automotive Index in Chile with Google Trends, and their model outperformed benchmark models. Althouse, Ng, and Cummings (2011) used search engine query volume to predict Dengue disease occurrences in Singapore and Thailand, and the model performed better than any other models.

Search engine data are also useful in forecasting general economic indicators such as unemployment rates (Askitas & Zimmermann, 2009; Choi & Varian, 2012; D'Amuri, 2009; Marcucci, 2009) and general consumer consumptions (Dzielinski, 2012; Kholodilin, Podstawski, Silverstovs, and Bürgi, 2009; McLaren & Shanbhogue, 2011; Vosen & Schmidt, 2011). Furthermore, search engine data have also been applied to other specific consumption categories, like the housing market (Wu & Brynjolfsson, 2009), box-office revenue (Hand & Judge, 2012; Wu & Brynjolfsson, 2009), as well as gun sales (Scott & Varian, 2013).

2.3. *Forecasting of tourism demand with search engine data*

Several researchers have examined search engine data when trying to understand tourist behavior and predict tourism demand. Xiang and Pan (2011) explored the nature of search engine marketing for destinations by analyzing the relationship between travelers' search behavior and the popularity of a specific city. They concluded that search volume could be a direct indicator of the size of that city's tourism industry. They examined the user queries specifically related to tourist cities, and claimed that a relatively small size of queries dominated the travel and hotel related searches. Still, the presence of a “long tail” to the data, consisting of many search queries with small amounts, reflected the unique and heterogeneous traveling experience. Gawlik, Kabaria, and Kaur (2011) used search volume histories to predict tourist rates in Hong Kong. Similarly, Choi and Varian (2012) considered the application of Google Trends index for Hong Kong, and forecasted the actual visitation by country of origin, including tourists from the top nine source countries. They found high R^2 values for the forecasting models. Pan and his colleagues also employed search engine queries (Pan et al., 2012) and a tourism bureau's website traffic (Yang et al., 2014) to forecast hotel room demand. They found that both types of data significantly contributed to the increase in forecasting accuracy of hotel demand for one destination.

Overall, there were two main limitations in the previous studies, related to forecasting tourist activities with search engine queries. First, most of these studies derived data from Google Trends. Although Google is the largest search engine in the world today, many countries, such as China, may have different dominant search engines. Will a global or local search engine have more predictive power? It is reasonable to assume that query data from a local search engine may provide better forecasting accuracy. We therefore propose to compare the fitness and prediction power of both Google and Baidu search data in predicting tourist volumes to a Chinese destination. Second, previous studies selected query predictors on an ad-hoc basis, without proposing a systematic selection method to distinguish among millions of potential query candidates. Most of these studies manually and randomly selected 10 queries on average to serve as the predictors. In this study, we propose a systematic way to better select queries for forecasting.

3. Conceptual framework

During a travel planning process, tourists need to make many different decisions regarding the various facets of a trip, such as the selection for a destination, dining, transportation, attractions, etc. (Jeng & Fesenmaier, 2002). Prior to their arrival, each tourist will make those different decisions during their own time frame, which will vary from person to person. In all of the decisions, one could employ search engines to look for information. As a result, the different queries, reflecting different types of information need, could be captured online in different time periods on those search engines. Thus, to predict the tourist volumes, different lag periods would need to be adopted in order to maximize the prediction accuracy (shown in Fig. 1).

4. Empirical testing

Hainan Province was selected as the destination for empirical testing of the conceptual model. Hainan is the biggest island, and the southernmost province, in China. Its tourism industry has increased rapidly in recent years. In 2013, Hainan Province received over 3.6 million overnight tourists, with 756,400 (about 21.0%) of them being foreign visitors (Hainan Tourism Bureau, 2013).

4.1. *Data source*

To compare search data from different search engines for modeling and predicting visitor volumes, we collected search query volume data from two major search engines, Google and Baidu. Google is the most popular search engine in the world, with a 66.7% market share, but with only a 2.1% market share in China in 2013; Baidu has the biggest market share in China (69%) (Sterling, 2013). Both of them provide free services of offering historic search engine query volume data.

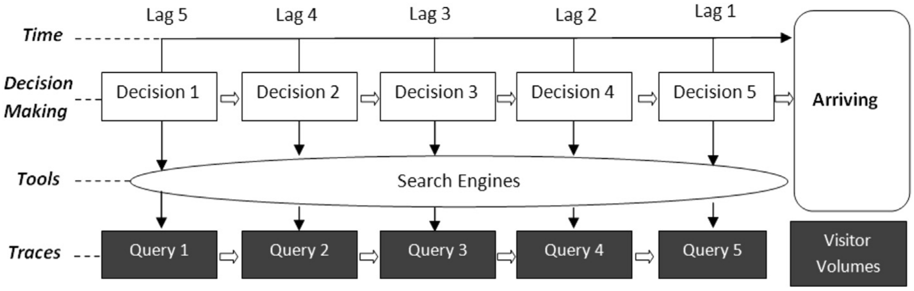


Fig. 1. A conceptual model of search behavior and forecasting.

Google Trends (http://www.google.com.hk/trends/?hl=en) provides Google query data, from January 2004 to the present, on a weekly or monthly basis. It does not report the raw volumes for a given search query; rather, it reports a query index, which displays how frequently a search query has been searched relative to the total search volume from different areas and different languages. Baidu Index (http://index.baidu.com/), a similar service by Baidu Inc., provides Baidu query volume data from June 2006 to the present on a daily basis, and the search data are in absolute numbers. Even though Google Trends and Baidu Index have calculated their indices with different methods, both of them reflect the popularity of a particular query and users' interests at a given moment in time. However, due to the dispute between the Chinese government and Google Inc., Google exited from the mainland China market in 2010 (Levy, 2011). As a result, the traffic of Google dramatically decreased to a minimal level in 2011. Thus, in order to accurately compare the two search engines, we chose the overlapping period from June, 2006 to December, 2010 for modeling. In addition, in order to increase the relevancy of the forecasting model to current time, we further explored the prediction power of Baidu search query data after 2010 by comparing the fitness performance and prediction error between models with visitor history data and models with Baidu search query data from June, 2006 to September, 2013. The forecasted variable, Hainan's monthly visitor volume data, were obtained from Wind database, a popular financial data provider in China (Zhang, 2011).

Fig. 2 shows the correlation between Hainan's monthly visitor volumes and two query volumes of “Hainan Airlines” from Google and Baidu. A strong concordance between visitor volume and query volume was present. However, there are potentially thousands of relevant queries related to Hainan's tourism industry. The following section details a systematic way to select queries for forecasting purposes.

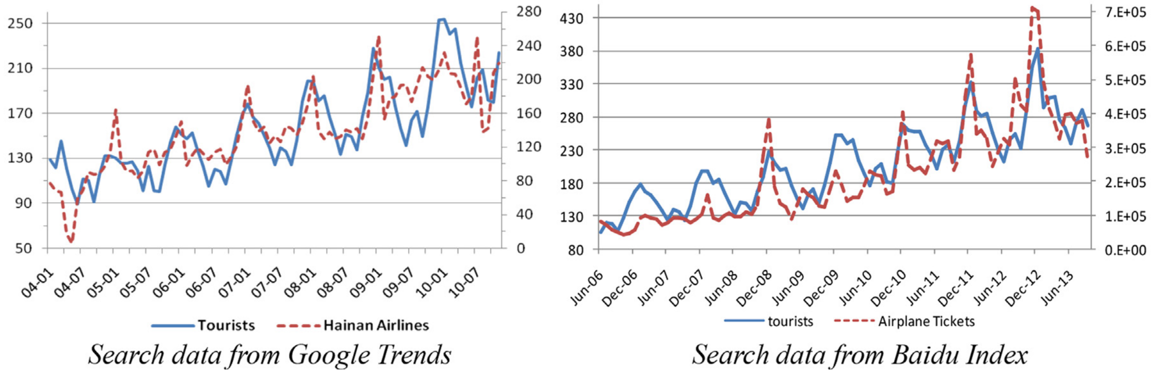
4.2. Selection of search queries

We followed a four stages process in selecting the candidate queries with the most predictive power, by relying on the “related searches” function on Google Trends and Baidu Index:

(1) We initially chose 20 basic search queries (in Chinese) based on various aspects of trip planning, including transportation, dining, lodging, shopping, etc. related queries. The translated queries are listed in Table 1 with their corresponding categories.

(2) We entered the 20 queries in Google Trends and Baidu Index as seed queries and retrieved the related queries. We then iteratively obtained the related queries for the second round of queries. We repeated this process for a few rounds. The number of queries converged to a total of 201. Only 164 queries remained in the Search Query Library (SQL) after duplications were removed.

(3) We calculated the Pearson correlation coefficient between Hainan monthly visitor volumes and each of the search queries with different lag periods. In total, eight correlation coefficients were calculated for each search query, including the correlations between visitor volumes of the current period and search query volumes of 0–7 months ahead, respectively. Furthermore, we chose the queries with the highest correlation values in the modeling process. A total of 10 search queries from Google Trends, and 25 queries from Baidu Index, were selected (shown in Table 2 and Table 3). In order to obtain the appropriate numbers of queries, we used 0.76 as the threshold for the correlation between Google Trends data and visitor volumes, and 0.8 as the threshold for Baidu Index data.



Search data from Google Trends

Search data from Baidu Index

Fig. 2. Trend of Hainan visitor volume and search query volumes.

Table 1
Basic search queries related to tourists.

Tourism	Hainan tourism	Lodging	Hainan accommodation
	Travel agents		Hainan hotel
	Travel site		Hainan hotel reservation
	Hainan travel site	Clothing	Hainan weather
Traffic	Hainan Airlines		Climate of Hainan
	Air China		Hainan weather forecast query
	China Eastern Airlines	Eating	Hainan cuisine
	Aircraft	Shopping	Hainan specialty
	Train tickets		Hainan characteristical fruit
	Hainan map	Tour	Hainan spots

We chose different thresholds due to the following reasons. First, if we intended to adopt the same numbers of keywords, say 25, then the threshold for Google Trends data would be lower than 0.76. As a result, the contribution of those keywords to forecasting accuracy would be non- significant. Second, if we set the same threshold, say 0.80, only four keywords from Google would have met that standard, and such few keywords would result in a low level of forecasting accuracy. Third, if we set a threshold of 0.76, the result of 41 keywords from Baidu might reduce the parsimony and generalizability of our models. In conclusion, the selection threshold is the result of a trade-off between prediction accuracy and model parsimony. In consideration of these two factors, we set 0.76 as the threshold for Google Trends and 0.80 for Baidu Index to obtain the appropriate number of predictors.

(4) For the purpose of predicting future visitor volumes, we only chose the queries with at least one lag prior to the arrival month, since Google and Baidu only publish the data at the end of each month. Finally, 5 queries with 1 or 2 lags were picked as Google Trends predictors and 17 queries were selected as Baidu Index predictors.

The data traces revealed interesting travel planning behavior. Tables 2 and 3 show that the lags of the maximum correlation coefficient of the search queries varied from 0 to 6, with most being 0 and 1. First, travel agent-related queries (e.g. Travel Agent Rank, Kanghui Travel Solution, and China Travel Solution) had the largest lag term, followed by flight information, and finally weather and shopping information. This indicated that most Chinese visitors to Hainan first searched travel agent information when they wanted to go on a trip, and that happened about six months prior to the trip. Next, they would search for flight information, about four months prior. Finally, just before their trip, during the month leading up to their departure, they looked for the destinations' latest weather and shopping information (Haikou Weather, Hainan Weather, Sanya Weather, Hainan Weather, Weather Forecast, Hainan Specialty, etc.). This validated the proposed conceptual framework: travelers make different aspects of travel decisions, and search for corresponding travel-related information, at different stages along the travel planning process. Search queries and their lag order reflected this behavior.

Table 2
Search query maximum correlation coefficient of Google Trends.[a]

Search query	Lag order	Search query	Lag order
Hainan Airline tickets search	2	*Air China*	0
Eastern miles	1	*Haikou Weather*	0
Hainan Sanya Spot	1	*Hainan tourism strategy*	0
Hainan Sanya weather	1	*Sanya tourist attractions*	0
Hainan tourist attractions	1	*Hainan Airlines*	0

[a] Queries with lag order of 0 were discarded in the modeling process.

Table 3
Maximum correlation coefficient of search queries from Baidu Index.[a]

Search query	Lag order	Search query	Lag order
Kanghui travel solution	6	China Eastern Airlines	1
Travel agent rank	6	China Eastern Airlines website	1
China travel solution	6	Hainan Island	1
China Eastern Miles	4	Air China membership	1
China Eastern Airline	4	Hainan weather	1
Aircraft	4	*Airplane tickets*	0
Hainan Sanya weather	2	*AC*	0
Haikou Weather	1	*Hainan Airlines website*	0
SanYa weather	1	*HA phone*	0
Hainan tourism strategy	1	*Air China website*	0
Hainan produces	1	*Airplane tickets search*	0
Hainan weather	1	*Train tickets search*	0
China Eastern Airlines	1	*Hainan Airlines*	0

[a] Queries with lag order of 0 were discarded in the modeling process.

4.3. Search data composite

We further aggregated search data into one index using a shift and sum method. All the selected queries were shifted according to lag order (they ranged from 0 to 7) of the maximum Pearson correlation coefficient; and all of the shifted search queries in the same model were summed up to form a new time series. The following analyses were based on the search index composite, for both Google and Baidu data.

4.4. Co-integration analysis of search index and Hainan visitors

We constructed two time series models in this section, based on Google Trends data and Baidu Index data.

In Eqs. (1)–(3), T_{t1} denotes the predicted variable, Hainan monthly visitor volume, from June, 2006 to December, 2010. Eq. (1) was the baseline model, denoted L_{b1}, using historical visitor volume data to predict future volumes. Due to the seasonality of Hainan tourism, we used the 12 periods T_t as the dependent variables in the baseline. Eq. (2) represented the first model $L1$, with search index derived from Google (GI_{t1}); Eq. (3) represented $L2$ with search index from Baidu (BI_{t1}).

$$\text{Log}\ T_{t1} = c_0 + \beta_1\ \text{Log}\ T_{t1}(-12) + u_t \tag{1}$$

$$\text{Log}\ T_{t1} = c_0 + \beta_1\ \text{Log}GI_{t1} + u_t \tag{2}$$

$$\text{Log}\ T_{t1} = c_0 + \beta_1\ \text{Log}BI_{t1} + u_t \tag{3}$$

To reduce the impact of outliers, the three variables were converted to logarithm form (LogT_{t1}, LogGI_{t1} and LogBI_{t1}), and u_t donated the residual series. Stationary tests (unit root tests) with extended Dickey–Fuller test method (ADF) were performed for all three variables. All of the three original time series were not stable, but the first difference of LogT_{t1}, LogGI_{t1} and LogBI_{t1} was. Two pairs of time series from model $L1$ and model $L2$ were co-integration series with the same order. This supported Granger causality analysis and ARMAX (Autorgressive Moving Average with External Variables) models. Thus, three regression models were constructed based on Eqs. (1)–(3). In all the models, we used the sample data from June, 2006 to June, 2010 and left the last 6 months, from July, 2010 to December, 2010, for out-sample prediction (shown in Table 4).

The coefficients of all explanatory variables in the three models were significant at the 0.05 level. Both LogGI_{t1} and LogBI_{t1} were statistically significant at the 0.01 level, which indicated a pronounced correlation between web search data and tourist volumes. Model $L1$ with Google Trends actually performed worse than the baseline model L_{b1}; Model $L2$ with Baidu Index data performed more or less as well as the baseline model L_{b1}, depending on

Table 4
Regression comparison of model $L1$, $L2$ with baseline model L_{b1} (2006.6–2010.6).

Data source		Visitor history data		Google Trends		Baidu Index	
Model		Baseline L_{b1}		Model $L1$		Model $L2$	
Independent variables		Log$T_{t1}(-12)$	1.022***	LogGI_{t1}	0.446***	LogBI_{t1}	0.176***
			(492.004)		(5.873)		(8.695)
		$MA(1)$	0.510***	C	2.404***	C	3.033***
			(4.525)		(4.928)		(12.135)
				$MA(1)$	0.997***	$AR(1)$	0.450***
					(14.521)		(4.001)
				$AR(12)$	0.220**	$AR(6)$	−0.565***
					(2.228)		(−4.577)
						$MA(12)$	0.881***
							21.929)
R^2		0.943		0.856		0.949	
Log likelihood		61.408		41.348		78.712	
AIC		−2.914		−2.334		−3.196	
SC		−3.057		−2.151		−3.118	
DW		1.952		1.834		2.053	
Residual stationary	ADF	−2.638***		−5.229***		−4.519***	
	1% Crit.	−2.616		−2.642		−2.635	
	5% Crit.	−1.948		−1.952		−1.951	
	10% Crit.	−1.612		−1.610		−1.611	
	Conclude	Stationary		Stationary		Stationary	
Conclusion		co-integration		co-integration		co-integration	
Adjusted observations		48		32		36	

Note: *, ** and *** represent significance at the 10%, 5% and 1% level.

different measurements. Based on various tests like R^2, Log likelihood, AIC, SC, and DW, model $L2$ with Baidu Index performed much better than model $L1$ with Google Trends. Model $L1$ and model $L2$ can be expressed as Eqs. (4) and (5) respectively:

$$\begin{cases} \text{Log } T_{t1} = 2.404 + 0.446 \text{ Log}GI_{t1} + u_t \\ u_t = 0.220u_{t-12} + \varepsilon_t + 0.997\varepsilon_{t-12} \end{cases} \quad (4)$$

$$\begin{cases} \text{Log } T_{t1} = 3.033 + 0.176 \text{ Log}BI_{t1} + u_t \\ u_t = 0.450u_{t-1} - 0565\varepsilon_{t-6} + \varepsilon_t + 0.881\varepsilon_{t-12} \end{cases} \quad (5)$$

Unit root test indicated that the residuals of the models $L1$ and $L2$ were stationary at the 0.01 level. Therefore, the co-integration relationship existed between search volume data and visitor volume in Hainan. The coefficient of the search data in Model $L2$ was 0.176, which indicated an increase of 1% of Baidu search volume data, corresponding to a visitor volume rise of 0.176%; for the model with Google Trends, it was 0.446%. The coefficient of u_t in both models $L1$ and $L2$ were statistically significant, which indicated that the independent variable of search index LogBI_{t1} and LogGI_{t1} could not explain all the fluctuation seen in the number of visitors. Other factors also affected visitors' short-term fluctuation.

4.5. Granger causality analysis

Since Model $L2$ presented better overall fitness, the Granger causality relationship between LogT_{t1} and LogBI_{t1} was tested. The Granger causality test examined whether the variable had predictive power for other variables. Due to the test's great sensitivity to the lag order, we considered five test criteria: AIC (Akaike Information Criterion), SC (Schwarz Information Criterion), HQ (Hannan-Quinn Information Criterion), LR (Likelihood Ratio Test) and FPE (Final Prediction Error Criterion Minimum) to determine the optimal lag length (for more details about these five information criteria, please refer to Appendix A). The optimal lag order of the majority test methods, a lag order of 5, was selected as the lag order for the test. Table 5 shows the results. LogBI_{t1} and LogT_{t1} Granger caused each other, that is, Baidu Index data can predict Hainan visitor volumes and vice versa.

4.6. Forecasting with web search data

To further examine the predictive accuracy and power of search indices for the numbers of Hainan visitors, the data were segmented into two parts: a training set and a testing set. For models $L1$ and $L2$, we used the last 6 months for testing, and the rest were included in the training set. We compared the actual results to the fitted results for both models in Fig. 3. Model $L2$ seemed to outperform model $L1$. The mean absolute error of LogT_{bt} with Google Trends data was 9.31%, while the MAE using the Baidu Index data was 3.64%, an improvement of almost 6%.

Table 6 lists the actual and fitted data for baseline model L_{b1} and the two models $L1$ and $L2$ with search data. Model $L1$ predicted 6 months of visitor volumes more accurately than the baseline model with an improvement of 6%. As indicated before, compared to baseline L_{b1}, model $L2$ had almost the same level of fitness. However, the prediction function with search data performed fairly well, with an improvement of over 12%. When we looked at the forecast results of models $L1$ and $L2$, the MAPE went from 9.86% using the Google Trends forecast to 3.11% using the Baidu data, which was a 6.75% reduction in prediction mean error.

4.7. Search queries' prediction power analysis after 2010

In order to increase the relevancy of the forecasting model to current time, we further examined the predictive power of search data after 2010. Considering that Google exited the Chinese market at the end of 2010, we were only able to build a model with Baidu data, and compared it with a model containing visitor history data. We used all available Baidu data from June, 2006 to September, 2013.

Table 5
Granger causality test results for search data and visitor volume.

Null hypothesis	Observations (Adjusted)	F- statistics	P-value
LogBI_{t1} does not Granger Cause LogT_{t1}	43	4.65283	0.0026
LogT_{t1} does not Granger Cause LogBI_{t1}		2.96244	0.0262

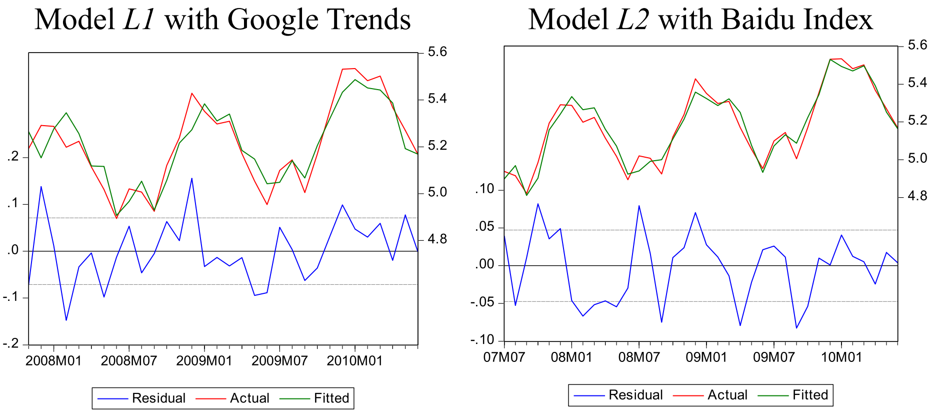


Fig. 3. Residual, actual and fitted value of model $L1$ and model $L2$.

In Eq. (6), T_{t2} denotes the predicted variable, Hainan monthly visitor volume from June, 2006 to September, 2013. BI_{t2} is the monthly search index derived from Baidu Index. We built one baseline model, denoted by L_{b2}, which used historical data to predict future volumes. L_{b2} is different from L_{b1} in previous section. L_{b1} only included samples before 2011, while L_{b2} contained data from 2006 to 2013. As indicated before, we included the 12 periods T_{t2} into the baseline model L_1 considering the seasonality of Hainan tourism. Eq. (7) represents the forecasting model $L3$ with search index BI_{t2}.

$$\text{Log } T_{t2} = c_0 + \beta_1 \text{ Log } T_{t2}(-12) + u_t \tag{6}$$

$$\text{Log } T_{t2} = c_0 + \beta_1 \text{ Log}BI_{t2} + u_t \tag{7}$$

Similarly, we first converted all-time series to logarithm form (LogT_{t2}, LogBI_{t2}), in which u_t donated the residual series. Then ADF tests implied that LogT_{t2} and LogBI_{t2} followed co-integration with the same order. The regression results are summarized in Table 7. The predictive power of these two models is displayed in Table 8.

Similar to the results in Table 4, the regression performance of model $L3$ was not significantly improved compared to baseline model L_{b2}. That is, the goodness of fit of the model with Baidu Index was roughly the same as the model with visitor history data. However, in terms of the predictive power of these two models, Table 8 shows that model $L3$ with Baidu Index data performed better than baseline, and the MAPE of $L3$ decreased almost 2%, compared with the baseline 6 month forecast. These results showed that Baidu Index data were also valuable in forecasting Hainan tourist volume in current time.

4.8. Robustness testing

In order to ensure the robustness of the new forecasting model using web search data, we separated the time series, a total of 88 data points from June 2006 to September 2013, into 20 different pairs, with different lengths of training and testing sets. We constructed 20 models for both Baseline L_{b2} and model $L3$, and compared the MAPE of the two models, where the testing sets varied from 4 months to 24 months. Appendix B summarizes the regression results of robustness test. For example, in the first pair comparison, model $C1$ contained the last 4 months as the testing set, and the remaining 84 months were the training set. All of the 20 models with Baidu Index data performed better than their corresponding baseline models at the same prediction period. The minimum MAPE of the forecasting models with Baidu Index was only 2.97% when we predicted 5 months, while the maximum MAPE was 7.14% when we predicted 12 months in advance. The average MAPE for the 21 models was 5.48%, while the average MAPE for the 21 baseline autoregressive models was 7.19%. This indicates the robustness of using Baidu Index data to forecast visitor volumes on any a forecasting horizon from 4 months to 24 months.

Table 6
Forecast comparison of models $L1$ and $L2$ with baseline model L_{b1} (2010.7–2010.12).

Data source		Visitor history data		Google Trends		Baidu Index	
Model		Baseline L_{b1}		Model $L1$		Model $L2$	
Month	Actual	Fitted	Error%	Fitted	Error%	Fitted	Error%
2010M07	201.19	214.74	0.12%	170.79	15.11%	189.16	5.98%
2010M08	208.73	210.42	1.90%	210.11	4.43%	209.06	0.16%
2010M09	181.8	169.14	21.14%	164.80	18.09%	175.39	3.53%
2010M10	179.61	220.81	2.94%	172.68	14.17%	184.76	2.87%
2010M11	224.03	258.59	20.56%	191.40	4.87%	213.63	4.64%
2010M12	268.34	257.79	20.19%	196.11	2.52%	264.36	1.48%
MAPE		15.61%		9.86%		3.11%	

5. Conclusions

This paper reviewed studies on the forecasting of socio-economic behavior and visitor volumes based on web searches. The correlation relationship between the search index data and actual visitor numbers to Hainan Province, China was analyzed, and a conceptual framework was proposed. The comparison of the

Table 7
Regression comparison of $L3$ with baseline model L_{b2} (2006.7–2013.3).

Data source		Visitor history data		Baidu Index	
Model		Baseline L_{b2}		Model $L3$	
Independent variables		LogT_{t2}(−12)	1.022^{***} (450.622)	LogBI_{t2}	0.411^{***} (74.483)
		MA(1)	0.618^{***} (5.525)	AR(1)	0.842^{***} (14.191)
				MA(6)	-0.415^{***} (−6.202)
				MA(12)	0.875^{***} (30.881)
R^2		0.941		0.942	
Log likelihood		78.937		98.512	
AIC		−2.959		−2.554	
SC		−2.884		−2.430	
DW		1.834		1.934	
Residual stationary	ADF	-4.343^{***}		-8.228^{***}	
	1% Crit.	−2.597		−2.597	
	5% Crit.	−1.945		−1.945	
	10% Crit.	−1.614		−1.614	
	Conclude	Stationary		Stationary	
Conclusion		co-integration		co-integration	

Table 8
Forecast comparison of *L*3 with baseline model L_{b2} (2013.4–2013.9).

Data source		Visitor history data		Baidu Index	
Model		Baseline L_{b2}		Model *L*3	
Month	Actual	Fitted	Error%	Fitted	Error%
2013M04	276.83	301.29	8.84%	296.50	7.10%
2013M05	264.34	275.81	4.34%	258.40	2.25%
2013M06	240.18	241.64	0.61%	244.99	2.00%
2013M07	273.40	276.37	1.09%	275.30	0.70%
2013M08	290.76	269.48	7.32%	272.67	6.22%
2013M09	267.70	242.00	9.60%	277.52	3.67%
MAPE		5.30%		3.66%	

baseline autoregressive model with alternative models using Google Trends and Baidu Index data showed that the forecasting models with Baidu data performed better than the baseline model as well as the models with Google Trends.

Compared to previous studies on using web search query to forecast visitor volumes, our study makes two contributions. First, we compared the forecasting power of the data from two different search engine indices, and validated the locality of search engine data to maximize forecasting power. Second, we proposed a systemic search query selection mechanism to better fit and predict visitor volumes.

Our results have several implications both for theory and managerial application. Theoretically, first, we proposed a conceptual framework connecting visitors' travel planning with the forecasting model selection. It showed that different search keywords may possess different lag structures when compared with the actual tourist activities, since the travelers make a variety of travel decisions in different stages. Specifically, this research showed that Chinese visitors mostly likely search for travel agents first, followed by information on lodgings, flights, and last, shopping and weather. This study also exhibited that the keywords chosen were similar in Baidu Index and Google Trends. Even though travelers could use different search engines, the information needs were similar.

Second, we demonstrated the significant co-integration relationship and long-term stability between the web search index data and visitors. Even though adding search data into baseline models only increased in-sample fitness slightly, both models significantly reduced the out-of-sample forecast MAPE. Compared with traditional prediction methods, the models with search index data performed better, no matter whether the data were from Google Trends or Baidu Index. However, in comparing the fitness and predictive power of Baidu Index with Google Trends on forecasting tourism volumes in China, Baidu Index performed better. It indicated that search engine data were localized and specific to a certain country or area, since Baidu has a much higher market share in China than Google does.

In terms of application, our results provide policy makers and managers in tourism and hospitality services a new way to track and monitor visitor volumes and behavior in the near-term. Compared to traditional large-scale surveys, the advantages of applying web search data on tourism demand analysis lie in its timeliness as well as low cost. In addition, compared to traditional methods of monitoring visitor numbers, the predictive power of the search data models in this study was much higher.

Specifically, policy makers could track the search volumes of specific tourism related keywords, and release a forecasted visitor volumes index as a surrogate benchmark for local tourism and hospitality services to gauge their performance. For example, if the visitor volumes' index showed an increase of 20% for a certain area next month, an increase of 10% in reservation volume for a specific hotel at the same time period should not be considered good performance.

More significantly, the results of this study encourage future research on various types of big data sets other than search engine query data, like tweets, blogs, and other social media, and the impact they have on tourism demand prediction and visitor behavior analysis. The large scale of big data could make up for the limitation of sample size issue faced by survey data users, as well as provide us a new way to understand visitors' information demand. The combined power of a variety of online traces of tourist behavior could further increase the accuracy and timeliness of forecasting and monitoring tourist activities, and help to better manage tourism on a small or large scale (Yang et al., 2014). In addition, as shown in this paper, the analysis of big data on tourism can also help build and validate tourist behavioral models, and thus, informs theory development.

However, this study had a number of limitations, the primary one being that we only focused on the Hainan tourism market as a test case. The ability to generalize the query selection method, as well as the notion that queries will converge no matter what search engine people choose, is limited. In order to address these limitations, more research exploring the application of web search data to other destinations, as well as empirical studies with larger sample sizes, would be required. In addition, web users' information needs are constantly changing. Therefore, the establishment of a comprehensive, dynamic, query selection method that would be able to effectively respond to changing market competition would also be a useful future research direction.

Acknowledgments

This research is supported by National Natural Science Foundation of China (Grant No. 71202115, 71172199 and 71203218), Beijing Natural Science Foundation (Grant No. 9143021), and Postdoctoral Science Foundation (Grant No. 2013T60158). The first author would like to thank China Scholarship Council (201304910217) for the financial support.

Appendix A. Granger causality test

Appendix A summarizes five lag order selection criteria and LR statistic for model *L*2. Model L2 is based on Baidu Index during the period of June, 2006 to June, 2010. *LR* is sequential modified *LR* test statistic (each test at 5% level); *FPE* is Final Prediction Error; *AIC* is Akaike Information Criterion; *SC* is Schwarz Information Criterion; *HQ* is Hannan-Quinn Information Criterion.

Appendix A
Lag order selection criteria

Lag	LogL	LR	FPE	AIC	SC	HQ
0	−15.418	NA	7.71E-03	0.810	0.892	0.840
1	74.525	167.336	1.42E-04	−3.187	−2.941[a]	−3.097
2	80.420	10.418	1.30E-04	−3.275	−2.866	−3.124
3	81.342	1.544	1.50E-04	−3.132	−2.559	−2.921
4	87.763	10.155	1.35E-04	−3.245	−2.508	−2.973
5	98.567	16.079[a]	9.97e-05[a]	−3.561[a]	−2.660	−3.229[a]

[a] Indicates optimal lag order selected by the criterion.

Appendix B. Robustness testing

Appendix B.1 and B.2 show the MAPE of 20 pairs of models in order to test the robustness of employing web search data to predict visitor volumes. We separated the time series, a total of 88 data points from June 2007 to September 2013, into 20 different pairs, with different lengths of training and testing sets. We constructed 20 models for both Baseline L_{b2} (Appendix B.2) and model *L*3 (Appendix B.1), and compared the MAPE of the two models, where the testing sets varied from 4 months to 24 months. For example, model *C*1 in both appendix had the last 4 months as the testing set, and the remaining 84 months as the training set. Model *C*1 in Appendix B.1 contains Baidu Index, but C1 in Appendix B.2 does not.

Appendix B.1
Results of comparison models with Baidu Index

C1	C2	STD	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
																				6.09%
																			13.54%	14.10%
																		5.08%	5.65%	6.37%
																	10.68%	11.08%	12.58%	13.47%
																0.20%	0.54%	0.66%	0.49%	0.03%
															0.72%	1.05%	0.11%	1.32%	2.89%	2.83%
														6.12%	6.04%	5.95%	6.22%	6.14%	5.90%	4.64%
													0.98%	0.71%	0.73%	0.87%	0.27%	0.57%	3.72%	5.37%
												7.79%	8.07%	8.18%	8.09%	8.00%	8.20%	7.34%	6.36%	5.13%
											9.00%	8.90%	8.67%	8.62%	8.70%	8.72%	7.71%	7.70%	7.91%	8.25%
										2.61%	2.56%	2.70%	2.90%	2.97%	2.90%	2.97%	3.09%	3.08%	2.61%	1.87%
									2.79%	2.97%	3.27%	3.34%	3.57%	3.58%	3.52%	3.46%	3.66%	3.76%	4.70%	5.03%
								14.85%	14.56%	14.43%	14.53%	14.49%	14.35%	13.96%	14.03%	14.07%	13.98%	13.88%	14.29%	13.76%
							6.91%	3.64%	3.15%	3.03%	3.42%	3.29%	3.17%	3.00%	3.05%	3.02%	3.15%	3.24%	1.50%	2.28%
						4.98%	4.51%	6.65%	6.64%	6.61%	7.25%	6.63%	6.46%	6.08%	6.19%	6.33%	5.80%	3.29%	2.39%	1.90%
					16.84%	16.89%	16.50%	21.60%	22.24%	22.53%	23.25%	22.83%	22.77%	22.56%	22.62%	22.80%	17.68%	17.19%	14.75%	13.27%
				5.97%	5.91%	5.89%	5.73%	6.00%	5.96%	5.74%	6.17%	5.78%	5.52%	5.34%	5.44%	6.23%	6.08%	5.92%	6.21%	6.79%
			7.46%	7.56%	6.95%	6.89%	6.73%	5.75%	4.72%	4.46%	3.85%	4.09%	4.02%	4.19%	4.20%	4.06%	4.60%	5.28%	6.30%	6.48%
		7.10%	6.95%	6.79%	6.22%	6.23%	5.92%	8.64%	9.32%	9.60%	10.20%	10.01%	10.08%	7.04%	7.06%	7.26%	6.57%	6.12%	4.15%	4.10%
	2.32%	2.25%	2.37%	2.48%	3.22%	3.18%	4.37%	3.94%	4.27%	4.27%	5.41%	4.64%	4.79%	4.41%	4.53%	4.73%	4.10%	3.45%	0.95%	0.70%
2.02%	1.91%	2.00%	1.92%	1.82%	2.44%	1.11%	1.32%	1.44%	1.59%	1.75%	1.75%	0.43%	0.39%	0.30%	0.34%	0.31%	0.39%	0.90%	1.13%	1.40%
0.71%	0.56%	0.70%	0.51%	0.37%	2.61%	2.64%	2.60%	0.30%	0.91%	0.98%	0.30%	0.86%	1.21%	1.79%	1.62%	1.26%	0.65%	1.39%	1.85%	1.40%
6.16%	6.24%	6.22%	6.33%	6.87%	7.16%	7.14%	7.01%	6.40%	6.23%	6.47%	6.59%	6.48%	6.38%	6.28%	6.32%	6.53%	6.62%	6.69%	7.42%	8.01%
3.64%	3.83%	3.67%	3.04%	2.84%	0.56%	0.60%	0.63%	6.49%	5.27%	5.38%	5.02%	5.16%	5.19%	5.32%	5.28%	5.34%	5.09%	4.76%	2.55%	1.23%
3.13%	**2.97%**	**3.66%**	**4.08%**	**4.34%**	**5.77%**	**5.56%**	**5.66%**	**7.14%**	**6.74%**	**6.49%**	**6.84%**	**6.71%**	**6.38%**	**6.14%**	**5.86%**	**5.66%**	**5.49%**	**5.40%**	**5.65%**	**5.60%**

Note: The bold numbers on the bottom row are Mean Absolute Percentage Error.

Appendix B.2
Results of comparison models with auto-regression (Using historical data to predict)

C1	C2	STD	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
																				6.11%
																			1.79%	4.08%
																		0.76%	1.97%	4.30%
																	17.89%	16.90%	15.03%	11.64%
																10.69%	11.73%	10.76%	9.09%	5.90%
															4.62%	4.88%	5.85%	4.94%	3.38%	0.37%
														7.16%	7.53%	7.79%	8.78%	7.86%	6.27%	3.22%
													8.54%	9.23%	9.60%	9.87%	10.86%	9.94%	8.34%	5.28%
												2.77%	3.90%	4.55%	4.90%	5.14%	6.07%	5.21%	3.71%	0.86%
											0.82%	1.29%	2.44%	3.09%	3.44%	3.69%	4.63%	3.76%	2.25%	0.64%
										5.66%	5.01%	4.56%	3.49%	2.88%	2.54%	2.31%	1.43%	2.25%	3.66%	6.37%
									7.73%	7.29%	6.66%	6.24%	5.20%	4.61%	4.29%	4.07%	3.22%	4.01%	5.37%	7.98%
								11.68%	11.53%	11.10%	10.47%	10.05%	9.02%	8.44%	8.12%	7.90%	7.06%	7.84%	9.19%	11.78%
							7.78%	8.72%	8.56%	8.10%	7.42%	6.97%	5.86%	5.23%	4.89%	4.64%	3.74%	4.58%	6.04%	8.83%
						4.60%	4.78%	5.77%	5.60%	5.11%	4.40%	3.92%	2.76%	2.09%	1.73%	1.48%	0.52%	1.41%	2.95%	5.88%
					26.69%	26.54%	26.25%	24.74%	24.99%	25.73%	26.82%	27.54%	29.32%	30.34%	30.91%	31.29%	32.80%	31.42%	28.86%	24.21%
				8.70%	10.79%	10.65%	10.41%	9.06%	9.28%	9.96%	10.93%	11.59%	13.20%	14.12%	14.62%	14.97%	16.29%	15.06%	12.94%	8.90%
			4.50%	3.93%	5.92%	5.79%	5.56%	4.27%	4.49%	5.13%	6.06%	6.68%	8.21%	9.09%	9.56%	9.89%	11.15%	9.98%	7.96%	4.13%
		8.84%	8.66%	8.08%	10.12%	9.98%	9.75%	8.43%	8.65%	9.31%	10.25%	10.90%	12.46%	13.36%	13.84%	14.19%	15.47%	14.27%	12.21%	8.27%
	2.98%	4.34%	4.17%	3.62%	5.55%	5.42%	5.20%	3.95%	4.16%	4.78%	5.67%	6.28%	7.76%	8.60%	9.06%	9.39%	10.60%	9.47%	7.52%	3.81%
0.28%	0.67%	0.61%	0.45%	0.07%	1.75%	1.63%	1.42%	0.25%	0.44%	1.03%	1.87%	2.44%	3.83%	4.63%	5.06%	5.36%	6.50%	5.44%	3.60%	0.11%
0.17%	0.23%	1.09%	0.92%	0.39%	2.26%	2.14%	1.92%	0.71%	0.91%	1.52%	2.38%	2.97%	4.40%	5.22%	5.67%	5.98%	7.15%	6.06%	4.17%	0.57%
8.16%	8.52%	7.32%	7.47%	7.95%	6.25%	6.36%	6.56%	7.66%	7.48%	6.93%	6.14%	5.60%	4.29%	3.54%	3.14%	2.85%	1.78%	2.78%	4.51%	7.79%
10.40%	10.75%	9.60%	9.74%	10.21%	8.58%	8.68%	8.87%	9.93%	9.75%	9.22%	8.47%	7.95%	6.71%	5.99%	5.60%	5.33%	4.31%	5.26%	6.91%	10.05%
4.75%	**4.63%**	**5.30%**	**5.13%**	**5.37%**	**8.66%**	**8.18%**	**8.05%**	**7.93%**	**7.97%**	**7.92%**	**7.56%**	**7.36%**	**7.73%**	**7.90%**	**7.85%**	**8.09%**	**8.94%**	**8.18%**	**7.29%**	**6.29%**
1.62%	***1.66%***	***1.64%***	***1.05%***	***1.03%***	***2.89%***	***2.62%***	***2.39%***	***0.79%***	***1.23%***	***1.43%***	***0.72%***	***0.65%***	***1.35%***	***1.76%***	***1.99%***	***2.43%***	***3.45%***	***2.78%***	***1.64%***	***0.69%***

Note: The bold numbers on the second to the bottom row are Mean Absolute Percentage Error. The numbers on the bottom row are MAPE gap between historic AR and new models with Baidu Index.

References

Abratt, R., Nel, D., & Nezer, C. (1995). Role of the market maven in retailing – a general marketplace influencer. *Journal of Business and Psychology, 10*(1), 31–55.

Althouse, B. M., Ng, Y. Y., & Cummings, D. A. (2011). Prediction of dengue incidence using search query surveillance. *PLoS Neglected Tropical Diseases, 5*(8), e1258.

Alvarez-Diaz, M., Mateu-Sbert, J., & Rossello-Nadal, J. (2009). Forecasting tourist arrivals to Balearic Islands using genetic programming. *International Journal of Computational Economics and Econometrics, 1*(1), 64–75.

Andrew, W. P., Cranage, D. A., & Lee, C. K. (1990). Forecasting hotel occupancy rates with time series models: an empirical analysis. *Journal of Hospitality & Tourism Research, 14*(2), 173–182.

Askitas, N., & Zimmermann, K. (2009). Google econometrics and unemployment forecasting. *Applied Economics Quarterly, 55*(2), 107–120.

Baram-Tsabari, A., & Segev, E. (2011). Exploring new web-based tools to identify public interest in science. *Public Understanding of Science, 20*(1), 130–143.

Carneiro, H. A., & Mylonakis, E. (2009). Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases, 49*(10), 1557–1564.

Carrière-Swallow, Y., & Labbé, F. (2011). Nowcasting with google trends in an emerging market. *Journal of Forecasting, 32*(4), 289–298.

Chaitip, P., & Chaiboonsri, C. (2009). Forecasting with X-12-ARIMA and ARFIMA: international tourist arrivals to India. *Annals of the University of Petrosani Economics, 9*(3), 147–162.

Chan, C., Witt, S., Lee, Y., & Song, H. (2010). Tourism forecast combination using the CUSUM technique. *Tourism Management, 31*(6), 891–897.

Chen, J., Chen, Z. X., Xing, L., & Fu, X. D. (2005). Forecasting of yunnan's international tourism demand based on bp neural network. *Journal of Kunming Teachers College, 27*(4), 89–91.

China Internet Network Information Center (CNNIC). (2013). *Statistical report on Internet development in China*. Available online at http://www1.cnnic.cn/AU/MediaC/rdxw/hotnews/201307/t20130722_40723.htm.

China National Tourism Administration. (2014). *Tourism statistics*. Available online at http://www.cnta.gov.cn/html/rjy/index.html.

Choi, H., & Varian, H. (2012). Predicting present with google trends. *Economic Record, 88*(S1), 2–9.

D'Amuri, F. (2009). *Predicting unemployment in short samples with internet job search query data*. Germany: University Library of Munich. Available online at http://mpra.ub.uni-muenchen.de/18403/.

Dzielinski, M. (2012). Measuring economic uncertainty and its impact on the stock market. *Finance Research Letters, 9*(3), 167–175.

Feng, Y. (2008). An application of ARMA model in tourist prediction. *Zhejiang Statistics, 31*(10), 8–10.

Gawlik, E., Kabaria, H., & Kaur, S. (2011). *Predicting tourism trends with google insights*. Retrieved from http://cs229.stanford.edu/proj2011/GawlikKaurKabaria-PredictingTourismTrendsWithGoogleInsights.pdf.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*(7232), 1012–1014.

Goel, S., Hofman, J. M., Lahaie, S. B., Pennock, D. M., & Watts, D. J. (2010). Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences, 107*(41), 17486–17490.

Goh, C., & Law, R. (2003). Incorporating the rough sets theory into travel demand analysis. *Tourism Management, 24*(5), 511–517.

Hainan Tourism Bureau. (2013). *Tourist volume and income of Hainan province, December, 2012*. Available at http://www.visithainan.gov.cn/government/lvyoutongji/tongjihuizong/2012/201301/t20130129_37128.html Accessed 26.11.13.

Hand, C., & Judge, G. (2012). Searching for the picture: forecasting UK cinema admissions using google trends data. *Applied Economics Letters, 19*(11), 1051–1055.

Jeng, J., & Fesenmaier, D. R. (2002). Conceptualizing the travel decision-making hierarchy: a review of recent developments. *Tourism Analysis, 7*(1), 15–32.

Kan, M. L., Lee, Y. B., & Chen, W. C. (2010). Apply grey prediction in the number of Tourist. In *2010 Fourth International Conference on IEEE Genetic and Evolutionary Computing* (pp. 481–484). Shenzhen, China http://doi.ieeecomputersociety.org/10.1109/ICGEC.2010.126.

Kholodilin, K. A., Podstawski, M., Siliverstovs, B., & Bürgi, C. (2009). *Google searches as a means of improving the nowcasts of key macroeconomic variables (No. 946)*. German Institute for Economic Research. Discussion papers.

Law, R., & Au, N. (1999). A neural network model to forecast Japanese demand for travel to Hong Kong. *Tourism Management, 20*(1), 89–97.

Levy, S. (2011). Inside google's China misfortune. *Fortune*. Available online at http://tech.fortune.cnn.com/2011/04/15/googles-ordeal-in-china/.

Lim, C., & McAleer, M. (2002). Time series forecasts of international travel demand for Australia. *Tourism Management, 23*(4), 389–396.

Liu, X. (2008). The selection of econometric model on forecasting Guilin tourist number. *Journal of Guangxi University Wuzhou Branch, 18*(2), 17–21.

Marcucci, J. (2009). *"Google it!" Forecasting the US unemployment rate with a google job search index.* Germany: University Library of Munich. Available online at http://mpra.ub.uni-muenchen.de/18732/.

McLaren, N., & Shanbhogue, R. (2011). Using internet search data as economic indicators. *Bank of England Quarterly Bulletin, 51*(2), 134–140.

Pan, B., Litvin, S. W., & Goldman, H. (2006). Real users, real trips, and real queries: an analysis of destination search on a search engine. In *Annual Conference of Travel and Tourism Research Association (TTRA 2006). Dublin, Ireland*.

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.

Pew Internet & American Life Project. (2013). *Internet user demogrpahics*. Available online at http://www.pewinternet.org/data-trend/internet-use/latest-stats/.

Qiu, X. (2013). *Scenic area apologized and denied being attacked*. Available online at http://news.xinhuanet.com/fortune/2013-10/04/c_125482946.htm.

Ripberger, J. T. (2011). Capturing curiosity: using Internet search trends to measure public attentiveness. *Policy Studies Journal, 39*(2), 239–259.

Scott, S. L., & Varian, H. R. (2013). *Bayesian variable selection for nowcasting economic time series*. NBER working paper. Available online at: http://www.nber.org/papers/w19567.pdf.

Song, H., & Li, G. (2008). Tourism demand modeling and forecasting—A review of recent research. *Tourism Management, 29*(2), 203–220.

Song, H., & Witt, S. (2000). *Tourism demand modeling and forecasting: Modern econometric approaches*. Oxford: Pergamon Press.

Song, H., Witt, S. F., & Li, G. (2008). *The advanced econometrics of tourism demand*. New York: Routledge.

Sterling, G. (2013). *Shake up in Chinese search market as engines merge*. Available online at http://searchengineland.com/shake-up-in-chinese-search-market-as-secondary-engines-merge-171818.

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

Vosen, S., & Schmidt, T. (2011). Forecasting private consumption: survey-based indicators vs. Google Trends. *Journal of Forecasting, 30*(6), 565–578.

Weatherford, L. R., & Kimes, S. E. (2003). A comparison of forecasting methods for hotel revenue management. *International Journal of Forecasting, 19*(3), 401–415.

Weng, G., Zhen, Z., Liu, Y., & Zhang, X. (2008). Tourism forecasting based on GM-Markov model: a case study of inbound tourism in China. *Journal of Yanshan University: Philosophy and Social Sciences Edition, 2*, 109–112.

Wu, L., & Brynjolfsson, E. (2009). *The future of prediction: How google searches foreshadow housing prices and sales*. Working paper, Available at http://ssrn.com/abstract=2022293.

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

Zhang, Y. J. (2011). The impact of financial development on carbon emissions: an empirical analysis in China. *Energy Policy, 39*(4), 2197–2203.

Zhang, Z., & Li, Q. (2011). QuestionHolic: hot topic discovery and trend analysis in community question answering systems. *Expert Systems with Applications, 38*(6), 6848–6855.

Xin Yang, is a PhD candidate at the Management School of University of Chinese Academy of Sciences, and visiting scholar at the Sociology Department of University of Chicago. Her research focuses on the impacts of users' online behavior on offline economic decisions. She is especially interested in the application of web search and social media data in economic fields, such as tourism, finance, as well as macroeconomics market.

Bing Pan, Ph.D., is an Associate Professor and head of research in the Office of Tourism Analysis, School of Business, at the College of Charleston. His research focuses on uses of information technologies in tourism industry, information systems, online behavior, and consumer behavior in tourism.

James A. Evans, Ph.D., is an Associate Professor in Sociology Department of University of Chicago. He especially focuses on the role that social and technical institutions, such as the Internet, play in collective cognition and discovery. He uses machine learning, social and semantic network representations to explore knowledge processes.

Benfu Lv, Ph.D., is a Professor in the School of Management of University of Chinese Academy of Sciences. His research interests include online behavior and internet economics.