Tourism Management 70 (2019) 1–10

Contents lists available at ScienceDirect

Tourism Management

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

Forecasting tourist arrivals with machine learning and internet search index

Shaolong Sun[a,b,c], Yunjie Wei[a,d], Kwok-Leung Tsui[c], Shouyang Wang[a,b,d,*]

[a] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, China
[b] School of Economics and Management, University of Chinese Academy of Sciences, Beijing, 100190, China
[c] Department of Systems Engineering and Engineering Management, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
[d] Center for Forecasting Science, Chinese Academy of Sciences, Beijing, 100190, China

ARTICLE INFO

Keywords:
Tourism demand forecasting
Kernel extreme learning machine
Search query data
Big data analytics
Composite search index

ABSTRACT

Previous studies have shown that online data, such as search engine queries, is a new source of data that can be used to forecast tourism demand. In this study, we propose a forecasting framework that uses machine learning and internet search indexes to forecast tourist arrivals for popular destinations in China and compared its forecasting performance to the search results generated by Google and Baidu, respectively. This study verifies the Granger causality and co-integration relationship between internet search index and tourist arrivals of Beijing. Our experimental results suggest that compared with benchmark models, the proposed kernel extreme learning machine (KELM) models, which integrate tourist volume series with Baidu Index and Google Index, can improve the forecasting performance significantly in terms of both forecasting accuracy and robustness analysis.

1. Introduction

All over the world, the tourism industry contributes significantly to economic growth (Gunter & Onder, 2015; Song, Li, Witt, & Athanasopoulos, 2011). According to the China National Tourism Administration, in 2016 the tourism income of China reached 4.69 trillion RMB, increasing by 13.6% compared to the previous year, and accounted for 6.3% of China's GDP. Thus, forecasting tourist volume is becoming increasingly important for predicting future economic development. Tourism demand forecasting may provide basic information for subsequent planning and policy making (Chu, 2008; Witt & Song, 2002). Methods used in tourism modeling and forecasting fall into four groups: time series models, econometrics models, artificial intelligence techniques and qualitative methods (Goh & Law, 2011; Song & Li, 2008). In addition to simple tourist data announced by the State Statistics Bureau, Internet search queries, which reflect the behavior and intentions of tourists, have increasingly been used in tourism forecasting models (Croce, 2017; Goodwin, 2008). However, the search index has created big opportunities in the modeling process of tourism forecasting (Li, Pan, Raw & Huang, 2017).

Internet search data has been applied to many aspects, such as hotel registrations (Pan & Yang, 2017; Rivera, 2016), tourist numbers (Bangwayo-Skeete & Skeete, 2015; Yang, Pan, Evans, & Lv, 2015), economic indicators (Choi & Varian, 2012), unemployment rates (Askitas & Zimmermann, 2009), private consumption (Vosen & Schmidt, 2011), and stock returns (Zhu & Bao, 2014). When introducing the Baidu Index or Google Index into forecasting models, keywords and the composition of indexes must be selected carefully. Keywords can be selected according to the correlation coefficient, the tendency chart or the crowd-squared method (Brynjolfsson, Geva, & Reichman, 2016). Additionally, the composition of indexes can be achieved by the HE-TDC method (Peng, Liu, Wang, & Gu, 2017) or the principal component analysis (PCA). Obviously, efforts should be made to avoid problems related to multi-collinearity and over-fitting to the greatest extent possible.

In this study, we proposed a new framework integrating machine learning and Internet search index to forecast tourist volume. The forecasting power of the framework is attributable to two features: first, relevant Internet search queries greatly contribute to the goodness of fit; second, Kernel-based extreme learning machines have short computing time and good generalization ability. However, as far as we know, few studies have adopted extreme learning machine to forecast tourism demand. The proposed framework is utilized to forecasting Beijing tourist arrivals. Relevant Internet search keywords cover the various aspects of tourism including dining, lodging, recreation, shopping, tour and traffic. Different from previous studies, this paper considers both Baidu Index and Google Index, which reflect the current situation of domestic tourists and foreign travelers. The experimental

* Corresponding author. Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Zhongguancun East Road, #55, Haidian District, Beijing, 100190. China.
E-mail addresses: sunshl@amss.ac.cn (S. Sun), weiyunjie@amss.ac.cn (Y. Wei), kltsui@cityu.edu.hk (K.-L. Tsui), sywang@amss.ac.cn (S. Wang).

https://doi.org/10.1016/j.tourman.2018.07.010
Received 8 February 2018; Received in revised form 29 May 2018; Accepted 13 July 2018
0261-5177/ © 2018 Elsevier Ltd. All rights reserved.

Table 1
An overview of selected tourism forecasting studies.

References	Region focused	Research objects	Data frequency	Methodologies	Performance measure	Variables
Athanasopoulos and Hyndman (2008)	Australian	Inbound tourism	Quarterly	SSME, ES	RMSE, ME, MAE, MAPE	Tourist arrivals and economic variables
Bangwayo-Skeete & Skeete, 2015	Caribbean	Tourism demand	Monthly	AR-MIDAS	MAPE, RMSE, DM	Tourist arrivals, Google trend data
Chen, Lai, and Yeh (2012)	Taiwan	Inbound tourism	Monthly	EMD, BPNN	MAD, MAPE, RMSE	Tourist arrivals
Chu (2008)	Asian-Pacific	Tourism demand	Monthly, quarterly	ARAR model	MAPE, RMSE	Tourist arrivals
Fildes, Wei, and Ismail (2011)	UK	Air travel demand	Annual	ADLM, TVP, VAR	MAE, RMSE	Air passengers and economic variable
Gunter and Onder (2015)	Paris	Inbound tourism	Monthly	EC-ADLM, VAR, Bayesian VAR, TVP, ARMA, ETS	RMSE, MAE	Tourist arrivals in hotels and economic variable
Jungmittag (2016)	German	Travel demand	Monthly	Combination forecasts, SARIMA	MAE, RMSE, MAPE	Air passengers
Li, Pan, Law, and Huang (2017)	Beijing	Tourism demand	Monthly	GDFM, PCA	MAE, MAPE	Tourist arrivals and Baidu index
Liang (2014)	Taiwan	Inbound tourism	Monthly	SARIMA, GARCH	MAD, RMSE, MAPE	Tourist arrivals
Pan and Yang (2017)	Charleston county	Hotel demand	Weekly	ARIMAX	MAPE, RMSE	Hotel occupancy, search engine queries, website traffic, weather information
Du Preez and Witt (2003)	Seychelles	Inbound tourism	Monthly	ARIMA, SSM	MAE, RMSE, MAPE	Tourist arrivals and economic variables
Rivera (2016)	Puerto Rico	Hotel nonresident registrations	Monthly	DLM	MAE, MAPE, RMSE	Google trend data, NHNR
Shahrabi, Hadavandi, and Asadi (2013)	Japan	Inbound tourism	Monthly	MGFFS	RMSE, MAPE	Tourist arrivals
Song et al. (2011)	Hong Kong	Inbound tourism	Monthly	STSM, TVP	MAPE, RMSE	Tourist arrivals and economic variables
Wong, Song, Witt, and Wu (2007)	Hong Kong	Inbound tourism	Quarterly	ARIMA, ADLM, ECM, VAR, combining forecast	MAPE, RMSE, MAE	Tourist arrivals and economic variables
Wu and Cao (2016)	Mainland China	Inbound tourism	Monthly	SVR, FOA, SIA	MAPE, RMSE, R	Inbound tourist flow

results illustrate that the proposed forecasting framework is significantly superior to the traditional time series models and some other machine learning models. Meanwhile, the forecasting power of the models with Baidu Index and Google Index is stronger than that without one index or both indexes, which may provide solid evidence that Internet search queries are of great significance to tourism demand forecasting.

The remainder of this paper is organized as follows: Literature review is provided in Section 2. Kernel extreme learning machine is introduced in Section 3. Forecasting framework is shown in Section 4. The empirical study is given in Section 5. Finally, Section 6 offers concluding work and implications for further research.

2. Literature review

This section reviews relevant literature about tourist arrivals forecasting and tourism forecasting with search engine query data. A list of these literature is provided in Table 1.

NOTE: state space models with exogenous variables (SSME) model; exponential smoothing (ES) model; Autoregressive Mixed-Data Sampling (AR-MIDAS) models; empirical mode decomposition (EMD); back propagation neural network (BPNN); autoregressive distributed lag model (ADLM); time-varying parameter (TVP); vector autoregressive (VAR) model; error correction autoregressive distributed lag model (EC-ADLM); Bayesian vector autoregressive (BVAR); autoregressive moving averaging (ARMA); seasonal autoregressive integrated moving average (SARIMA); generalized dynamic factor model (GDFM); principal component analysis (PCA); generalized autoregressive conditional heteroskedasticity (GARCH); autoregressive integrated moving average with exogenous variables (ARIMAX); state space models (SSM); Dynamic linear model (DLM); modular genetic-fuzzy forecasting system (MGFFS); structural time series model (STSM); support vector regression (SVR); fruit fly optimization algorithm (FOA); seasonal index adjustment (SIA); root mean square error (RMSE); mean error (ME); mean absolute error (MAE); mean absolute percentage error (MAPE); Diebold-Mariano (DM) statistic; mean absolute deviation (MAD); the number of hotel nonresident registrations (NHNR).

2.1. Tourist volume forecasting

Autoregressive integrated moving average (ARIMA) is the most widely used time series forecasting model. This model has also been widely applied to tourism forecasting and performed well (Athanasopoulos, Hyndman, Song, & Wu, 2011; Brida & Risso, 2011; Chang & Liao, 2010; Chen et al., 2012; Du Preez & Witt, 2003; Jungmittag, 2016; Li & Sheng, 2016; Liang, 2014; Lim & McAleer, 2002; Shahrabi et al., 2013). However, ARIMA models do not always outperform others. These models perform well in the traditional econometric models, but they are sometimes inferior to intelligence methods. Song and Witt (2000) used a variety of techniques to predict tourism demand for a specified region and found that a neural network approach outperformed ARIMA model. Similarly, previous research on tourist arrivals in China demonstrated that support vector regression (SVR) outperformed back propagation neural network (BPNN) and ARIMA models.

Exponential smoothing (ES) has been widely used in tourism forecasting and many scholars use the model as a benchmark (Fildes et al., 2011; Park, Rilett, & Han, 1999; Witt & Witt, 1995). Other time series models include but are not limited to state space models (Athanasopoulos & Hyndman, 2008; Beneki, Eeckels, & Leon, 2012; Du Preez & Witt, 2003), error correction models (ECM) (Lee, 2011; Shen, Li, & Song, 2009; Vanegas, 2013; Wong et al., 2007), and generalized autoregressive conditional heteroskedastic (GARCH) models (Chan, Lim, & Mcaleer, 2005; Liang, 2014).

In recent years, artificial intelligence (AI) techniques have emerged in the tourism study, such as fuzzy logic theory, artificial neural

networks (ANNs), support vector machines (SVM) and genetic algorithms (GA). The key advantage of AI is that they do not need any assumptions such as stationarity or distribution. Hence, these AI techniques have been widely applied to tourism demand forecasting.

For example, ANNs are soft computational techniques used in computer science and other research disciplines. The unique features of ANNs, such as the adaptability, nonlinearity, make this technique a useful alternative to the classical regression forecasting models (Song & Li, 2008). The first computational model for ANNs was proposed by McCulloch and Pitts (1943) by means of threshold logic algorithms and mathematics. After this, ANN had a rapid development and continuous improvement. For tourism forecasting, the overall performance of ANNs has been shown to be better than traditional time series models and econometric models. ANN applied to predict Japanese tourist arrivals in Hong Kong outperformed naïve, multiple regressions, exponent smoothing, and moving average (Law & Au, 1999).

Generally, SVMs are used to solve classification, regression estimation and forecasting problems. SVM was firstly introduced to tourism forecasting in the late 1990s and improved versions of SVM have continued to appear post-2000. Sencheong and Turner (2005) given the literature review of the applications of SVMs to tourism forecasting. The empirical results demonstrate that SVMs commonly outperform the time series models and multiple regression models in tourism forecasting. For instance, Claveria, Monte, and Torra (2016) showed that the SVM improved the forecasting performance with respect to the benchmark model. Similar findings were also obtained by Pai, Hung, and Lin (2014), who concluded that the performance of SVM outperformed ARIMA model in tourism demand forecasting at Hong Kong and Taiwan.

2.2. Tourism forecasting with search engine data

In recent years, a number of authors have paid more attention to the application of web search data for tourism forecasting. A common web search using Google Trends or Baidu Index, which are essentially search volumes of keywords, can be used to identify potential tourists and as indicators of tourist behaviors, including where and how tourists travel.

The use of web search data can significantly improve the precision of tourist volume forecasting. For instance, Pan, Wu, and Song (2012) showed that forecasting accuracy improved significantly when Google search data were included in the ARMA models, which provide strong support for the use of search engine data in demand forecasting of hotel rooms. Similar conclusions were reached by Artola, Pinto, and Garcia (2015), who improved forecasts of tourism inflows into Spain using Google Indexes on internet searches measuring the relative popularity of keywords associated with travelling to Spain. In addition, Baidu Index has been used to predict tourism demand in China. Yang et al. (2015) used search engine query data to forecast tourist arrivals to Hainan Province. Their empirical results showed that search engine query data from both Google and Baidu helped significantly improve forecasting performance; furthermore, Baidu index data performed better more, likely due to its larger market share in China. Similarly, Li et al. (2017) employed search engine query data to create a composite search index and used a generalized dynamic factor model (GDFM) to forecast tourist arrivals of Beijing.

3. Kernel extreme learning machine

Extreme learning machine (ELM) is a type of single-hidden layer feed-forward neural networks (SLFNs) (Huang, Zhu, & Siew, 2006). ELM models have been widely used to many fields by means of its fast learning speed and generalization ability. The key highlight of the ELM models is that the input weights and biases are randomly generated and the hidden layer parameters need not be tuned. The output weights are obtained by simple matrix computations, so the computing time is very short.

For N arbitrary samples (x_i, y_i), $x_i \in \Re^N$, $y_i \in \Re^N$, $i = 1,2,\cdots,N$, if the activation function of hidden layer is $h(x)$ and the output matrix is Y, then the typical SLFNs can be defined as

$$Y = \begin{bmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{mj} \end{bmatrix}_{m\times N} = \begin{bmatrix} \sum_{i=1}^{l} \beta_{i1} h(w_i x_j + b_i) \\ \sum_{i=1}^{l} \beta_{i2} h(w_i x_j + b_i) \\ \vdots \\ \sum_{i=1}^{l} \beta_{im} h(w_i x_j + b_i) \end{bmatrix}_{m\times N}, (j = 1,2, \cdots N) \quad (1)$$

where β represents the network output weights between the hidden layer and the output layer, $w_i = [\omega_{i1}, \omega_{i2}, ..., \omega_{iN}]^T$ is the input weight between the ith hidden layers and input layers the, l is the number of hidden nodes, and b is the threshold of the hidden layer. The above equations can also be written as:

$$H\beta = Y, \quad Y \in \Re^{N\times m}, \beta \in \Re^{N\times m} \quad , \quad H = H(\omega, b) = h(\omega x + b) \quad (2)$$

where H is the output matrix of the hidden layer. The weights and biases of input layer are randomly produced instead of being tuned. The only unknown parameter is the output weight β which can be solved by the ordinary least square (OLS) method. The solution of the above equation is given by

$$\hat{\beta} = H^{\dagger}Y, \quad H^{\dagger} = H^T(HH^T)^{-1} \quad (3)$$

where $H^{\dagger}$ denotes the Moore-Penrose generalized inverse of H (Huang et al., 2006). According to Ridge regression theory and the orthogonal projection method, β can be calculated by adding a positive penalty factor $1/C$ as follows:

$$\hat{\beta} = H^T(1/C + HH^T)^{-1}Y \quad (4)$$

Then, the output function of ELM can be expressed as follows,

$$f(x) = H\hat{\beta} = HH^T(1/C + HH^T)^{-1}Y \quad (5)$$

This method overcomes some shortcomings of the typical gradient-based learning algorithms, such as over-fitting, local minima and long computation time. The topology structure of ELM is shown in Fig. 1.

A Kernel-based ELM was proposed by Huang (2014). In Huang's proposal, the activation function $h(x)$ of the hidden layer is replaced by a kernel function in terms of Mercer's conditions. The output function of KELM can be formulated as follows,

$$f(x) = h(x)\hat{\beta} = \begin{bmatrix} k(x, x_1) \\ k(x, x_2) \\ \vdots \\ k(x, x_n) \end{bmatrix}^T (1/C + HH^T)^{-1}Y \quad (6)$$

in this formula, the feature mapping $h(x)$ need not be known to users;

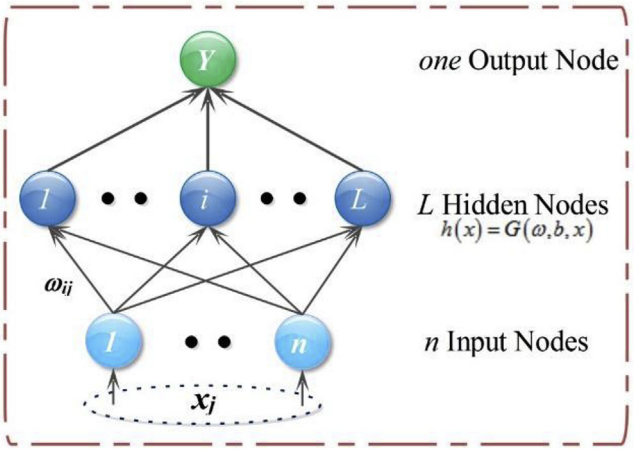

Fig. 1. The topology structure of ELM.

instead one may use its corresponding kernel $k(x, x_i)$. This situation means that a kernel function can replace the random mapping of the ELM, and the output weights are more stable. Therefore, the KELM achieves better generalization ability than the ELM. In this paper, four different kernel functions are employed as follows:

(1) Linear kernel function:

$$K(x, x_i) = x^T x_i \tag{7}$$

(2) Polynomial kernel function:

$$K(x, x_i) = (\gamma x^T x_i + r)^p, \quad \gamma > 0 \tag{8}$$

(3) RBF kernel function:

$$K(x, x_i) = \exp(-\gamma \|x - x_i\|^2), \quad \gamma > 0 \tag{9}$$

(4) Wavelet kernel function:

$$K(x, x_i) = \cos(\alpha * (x - x_i)) \exp(-\gamma (x - x_i)^2), \quad \alpha, \gamma > 0 \tag{10}$$

4. Forecasting framework

During the travel planning process, visitors need to make a large number of decisions about all aspects of travel, such as selecting a destination, traffic, lodging, and dining. Prior to arriving, visitors make these decisions based on their own time, and vary from person to person. To aid in decision-making, visitors often employ search engines. Therefore, the different types of information required by visitors reflected by the search query may be captured at different times on these search engines. Hence, we propose a forecasting framework that uses machine learning and internet search indexes to forecast tourist arrivals (Fig. 2). The framework describes the modeling process starting with data extraction, data fusion and data computing.

5. Experimental study

Beijing was selected as the location of this empirical study to evaluate the effectiveness of the proposed forecasting framework. The experimental design is introduced in Section 5.1 and the experimental results are provided in Section 5.2. Finally, these results are summarized and discussed in **Section 5.3**.

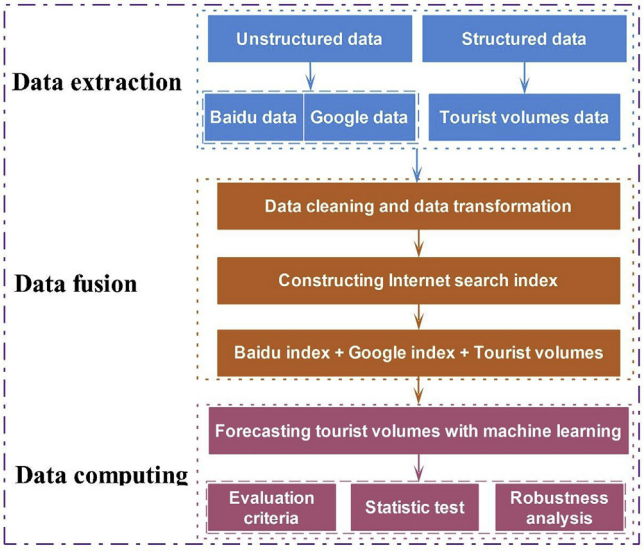

Fig. 2. Forecasting framework with machine learning and Internet search index.

5.1. Experimental design

5.1.1. Data collection

Data for monthly Beijing tourist arrivals, which is the sum of mainland tourist arrivals and overseas tourist arrivals, during the period of January 2011 to April 2017 were obtained from the Wind Database (http://www.wind.com.cn/). The data were divided into in-sample subsets and out-of-sample subsets. In-sample subsets were used for model training with data from January 2011 to April 2016, as shown in Fig. 3, whereas out-of-sample subsets were used for empirical testing with data from May 2016 to April 2017. The detailed data can be obtained from the Wind Database or the authors upon request.

Furthermore, in order to verify the internet search results from different search engines for forecasting tourist arrivals, this study collected search query data from two major search engines, Baidu Index (http://index.baidu.com/) and Google Trends (https://trends.google.com/trends/). Baidu occupies the biggest market share in China accounting for 80.5% approximately and is used to reflect search behaviors of domestic tourists in this paper; Google is the world's most popular search engine, accounting for 92.5% market share, and is used to reflect the search behaviors of overseas tourists in this paper. The search query history generated by the two search engines is available to the public. Although Google Trends and Baidu Index calculate their indexes by different methods, they both reflect the popularity of specific queries and user interests at specific times (Yang et al., 2015). Thus, to compare the two search engines, we chose monthly search query data of the two search engines respectively, from January 2011 to April 2016. The following sections detail a systematic way to choose search keywords and to construct internet search indexes for forecasting Beijing tourist arrivals.

5.1.2. Evaluation criteria

To verify the forecasting accuracy of different models, we adopted two main evaluation criteria to compare the in-sample and out-of-sample forecasting performance: normalized root mean squared error (NRMSE) and mean absolute percentage error (MAPE).

$$NRMSE = \frac{100}{\bar{x}} \sqrt{\frac{1}{N} \sum_{t=1}^{N} (x_t - \hat{x}_t)^2}, \; MAPE = \frac{100}{N} \sum_{t=1}^{N} \left| \frac{x_t - \hat{x}_t}{x_t} \right|, \tag{11}$$

Where N is the number of observations, x_t denotes the actual tourist volume, and $\hat{x}_t$ indicates the forecast value of tourist volume.

Additionally, in order to evaluate forecasting performance from a statistical perspective, the Diebold-Mariano (DM) statistic was employed to test the statistical significance of all models (Diebold & Mariano, 2002). The DM statistic was used to test the null hypothesis of equality of expected forecast accuracy against the alternative of different forecasting abilities across models. In this study, mean square prediction error (MSPE) was used as the loss function. Thus, the null hypothesis of the DM test was that the MSPE of the tested model *te* is not smaller than that of the benchmark model *be*. For the tested model *te* and the benchmark model *be*, the DM statistic can be defined as:

$$S_{DM} = \frac{\bar{g}}{\left(\widehat{V}_{\bar{g}}/N\right)^{1/2}} \tag{12}$$

where $\bar{g} = (\sum_{t=1}^{N} g_t)/N$ $(g_t = \sum_{t=1}^{N} (x_t - \hat{x}_{te,t})^2 - \sum_{t=1}^{N} (x_t - \hat{x}_{be,t})^2$) and $\widehat{V}_{\bar{g}} = \gamma_0 + 2\sum_{l=1}^{\infty} \gamma_l$ ($\gamma_l = \mathrm{cov}(g_t, g_{t-l})$). $\hat{x}_{te,t}$ and $\hat{x}_{be,t}$ are forecasting values of x_t calculated by the tested model *te* and the benchmark model *be*, respectively, for a period *t*.

5.2. Experimental results

In this section, we first construct both Baidu and Google internet search indexes by composite leading search index. Secondly, we test the co-integration and Granger causality between the tourist arrivals and

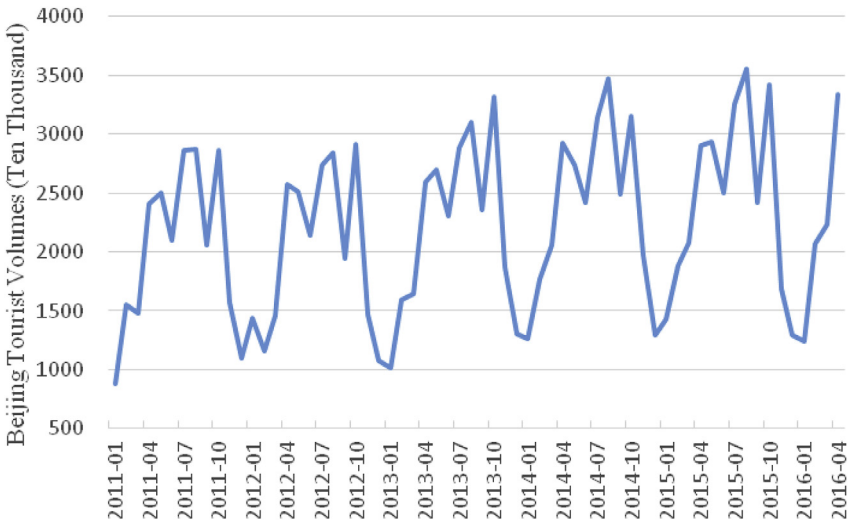

Fig. 3. Trend of monthly Beijing tourist arrivals.

these two indexes. Thirdly, we design different forecasting models in terms of tourist arrivals series and these two indexes, and evaluate the in-sample and out-of-sample forecasting performance of each model. Finally, the robustness of all forecasting models is analyzed, and conclusions are drawn from the empirical analysis.

5.2.1. Internet search index

This subsection followed a five-stage process to select candidate queries and construct an Internet search index using the related searches function of Baidu Index and Google Trends (Yang et al., 2015):

(1) We initially chose 24 basic search queries based on all aspects of tourism planning, including travel, traffic, lodging, dining, recreation and shopping (Li et al., 2017). The queries are listed in Table 2 with the corresponding category.
(2) We first searched the 24 keywords in Baidu Index and Google Trends as seed keywords and retrieved related keywords. We then iteratively got the recommended keywords as the second round of keywords. This process was repeated for several rounds, during which keywords with unavailable or extremely low volume data were eliminated. The number of keywords converged to a total of 154 for Baidu Index, and 69 for Google Trends.
(3) We calculated the Pearson correlation coefficient between Beijing tourist arrivals and each of the search keywords with different lag periods. A total of four correlation coefficients were calculated for each search keyword, including the correlations between tourist arrivals of the current period and search keyword volumes of 0–3 months previous, respectively. In addition, we chose the keywords with the highest correlation coefficient when building our internet search index. A total of 24 keywords from Baidu Index and 16 keywords from Google Trends were selected (shown in Table 3 and Table 4). To obtain the right number of keywords, we used a threshold correlation coefficient between tourist arrivals and the two engine indexes: 0.75 for Baidu and 0.7 for Google.
(4) To forecast future tourist arrivals, we chose only the keywords that have at least one lag period prior to the arrival month, because Baidu and Google only release the data at the end of each month. Finally, a total of 22 keywords with one or two lag periods were selected as Baidu Index predictors, and 13 keywords as Google Trends predictors.

Table 2
Search queries related to Beijing tourism.

No	Search queries	No.	Search queries	No.	Search queries
	Tourism		Traffic		Lodging
1	Beijing tourism	5	Beijing airlines	9	Beijing hotels
2	Beijing weather	6	Beijing shuttle bus	10	Beijing accommodation
3	Beijing maps	7	Beijing railway tickets	11	Beijing farmhouses
4	Beijing travel agency	8	Beijing bus schedules	12	Beijing resorts
	Dining		Recreation		Shopping
13	Beijing food	17	Beijing nightlife	21	Beijing shopping
14	Peking duck	18	Beijing recreation	22	Beijing specialties
15	Beijing food websites	19	Beijing bars	23	Dashilan Street
16	Beijing snacks	20	Beijing shows	24	Panjiayuan Center

Table 3
Maximum correlation coefficients of search queries from Baidu.

No.	Search queries	Lag order	No.	Search queries	Lag order
1	Beijing travel agency	2	13	National Aquatics Center	1
2	Beijing travel solution	2	14	National Stadium	1
3	Beijing snacks	1	15	Central TV Tower	1
4	Beijing hotels	1	16	Badaling Great Wall	1
5	Beijing farmhouses	1	17	Hotel booking	1
6	Beijing accommodation guides	1	18	Beijing airports	1
7	Beijing tourism	1	19	Beijing flights	1
8	Beijing travel guides	1	20	Beijing amusement parks	1
9	Beijing travel sites	1	21	Panjiayuan Center	1
10	The Palace Museum	1	22	Dashilan Street	1
11	Xidan district	1	23	Beijing weather	0
12	Ming Tombs	1	24	Beijing maps	0

Table 4
Maximum correlation coefficient of search queries from Google.

No.	Search query	Lag order	No.	Search query	Lag order
1	China travel	2	9	Beijing travel	1
2	Beijing weather	2	10	Great Wall	1
3	Peking duck	1	11	Beijing flights	1
4	Duck recipes	1	12	Beijing airports	1
5	Beijing hotels	1	13	Beijing railways	1
6	Beijing restaurants	1	14	Beijing maps	0
7	Beijing shopping	1	15	Beijing bars	0
8	Zhongguancun	1	16	Beijing shows	0

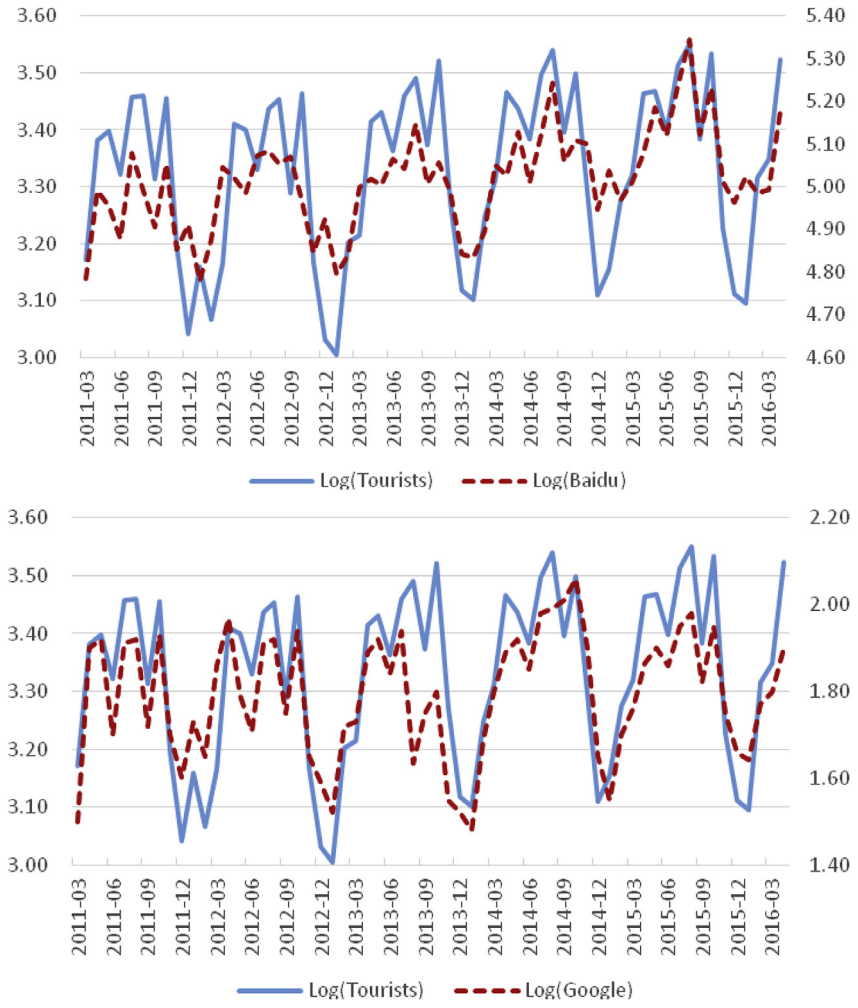

Fig. 4. Trend of Beijing tourist arrivals and two Internet search indexes.

(5) We further aggregated the search data into a composite index by means of shift and summation. All selected keywords were moved through the lag of the maximum Pearson correlation coefficient, and all of the shifted search keywords in the same model were summed to form a new time series. Fig. 4 shows the correlation between the log of Beijing monthly tourist arrivals and the two Internet search indexes. The following analyses were based on the internet composite search index, for both Baidu and Google data.

5.2.2. Co-integration and granger causality analysis

In order to reduce the impact of outliers, these three variables were converted to logarithmic form (Log*T*, Log*BI* and Log*GI*). Table 5 provides the stability test and the Johansen co-integration test among Log*T*, Log*BI* and Log*GI*. These three time series are stable validated by augmented Dickey-Fuller test. The co-integration results demonstrate that Log*T* and Log*BI* are co-integrated. Similarly, Log*T* and Log*GI* are also co-integrated. Therefore, a long-term co-integration relationship exists between Internet search indexes and Beijing tourist arrivals. These findings suggest that adopting Internet search indexes to predict tourist arrivals from an econometric perspective is feasible.

The purpose of the Granger causality tests is to verify whether these two Internet search indexes are predictors of Beijing tourist arrivals. As shown in Table 6, Log*BI* and Log*GI* are the Granger cause of Log*T*, indicating a causal relationship between the data of these two Internet search indexes and actual Beijing tourist arrivals.

Table 5
Co-integration test results.

Augmented Dickey-Fuller tests		
t statistics		p value
Log*T*	−3.8056	0.0027
Log*BI*	−3.7143	0.0041
Log*GI*	−3.4034	0.0115

Cointegration between Log*T* and Log*BI*				
	Eigenvalue	Trace statistic	Critical value	Prob[b]
None[a]	0.07	27.31	15.98	0.00
At most 1[a]	0.05	14.11	4.02	0.00

Cointegration between Log*T* and Log*GI*				
	Eigenvalue	Trace statistic	Critical value	Prob[b]
None[a]	0.06	17.58	15.43	0.01
At most 1[a]	0.03	1.19	3.42	0.22

[a] Denotes rejection of the null hypothesis at the 0.05 confidence level.
[b] MacKinnon, Haug, and Michelis (1999) p-value.

5.2.3. Forecasting with machine learning and internet search index

Machine learning techniques were used to further examine the forecasting power of Internet search indexes for Beijing tourist arrivals forecasting. The independent variables of forecasting models were

Table 6
Granger causality tests between the Internet search indexes and tourist arrivals.

Null hypothesis	F-statistics	Prob.
Log*BI* does not Granger cause Log*T*	27.86	0.00[a]
Log*T* does not Granger cause Log*BI*	0.43	0.53
Log*GI* does not Granger cause Log*T*	26.17	0.00[a]
Log*T* does not Granger cause Log*GI*	0.02	0.94

[a] Indicates the significance level of 1%.

classified into four types: "time series", "time series + Baidu index", "time series + Google index", and "time series + Baidu index + Google index". In this study, the inputs of the machine learning models and the numbers of hidden neurons of ANN and KELM models were determined by trial-and-error testing for minimizing in-sample forecasting errors. The Gaussian kernel function was applied in LSSVR and SVR models. The optimal form of ARIMA models was estimated by minimizing the Schwarz Criterion (SC) and Akaike Information Criterion (AIC).

The forecasting performance of the four different independent variables and eight models as mentioned above is provided in this section. Table 7 shows the comparison results of MAPE and NRMSE evaluation criteria.

As Table 7 shows, the proposed KELM-rbf model with independent variables of "time series + Baidu index + Google index" has the lowest MAPE and NRMSE in in-sample forecasting. Furthermore, in the out-of-sample forecasting, "time series + Baidu index + Google index" constantly outperforms other independent variables in forecasting Beijing tourist arrivals in respect to MAPE and NRMSE, followed by "time series + Baidu index" and "time series + Google index", whereas "time series" ranks last. Moreover, the proposed KELM models produced 3.41–8.53% smaller MAPE and 4.15–9.32% smaller NRMSE than ARIMAX models respectively, reaching an accuracy rate of 0.643% and 0.702% in out-of-sample respectively.

Table 7
Forecasting performance evaluation.

Models	In-sample		Out-of-sample	
	MAPE (%)	NRMSE (%)	MAPE (%)	NRMSE (%)
Time series				
ARIMA	8.142	9.061	9.168	10.021
ANN	3.071	3.689	4.067	4.967
SVR	2.953	3.267	3.591	4.301
LSSVR	2.916	3.261	3.407	4.016
KELM-lin	2.637	3.014	3.056	3.986
KELM-poly	1.784	2.569*	1.921	2.914
KELM-rbf	1.627*	2.571	1.709*	2.726*
KELM-wav	1.709	2.602	1.846	2.843
Time series + Baidu Index				
ARIMAX	5.516	5.763	5.962	6.167
ANN	2.571	2.698	2.423	2.564
SVR	1.914	2.069	2.196	2.306
LSSVR	1.706	1.817	1.834	1.902
KELM-lin	1.047	1.156	1.314	1.397
KELM-poly	0.972	1.098	1.196	1.264
KELM-rbf	0.958*	1.006*	1.026*	1.127*
KELM-wav	0.969	1.037	1.088	1.191
Time series + Google Index				
ARIMAX	5.967	6.129	6.118	6.237
ANN	2.261	2.342	2.367	2.468
SVR	2.174	2.228	2.306	2.325
LSSVR	1.918	2.106	2.118	2.267
KELM-lin	1.446	1.609	1.546	1.674
KELM-poly	1.369	1.438	1.438	1.598
KELM-rbf	1.297	1.392	1.357	1.416
KELM-wav	1.011*	1.126*	1.348*	1.425*
Time Series + Baidu Index + Google Index				
ARIMAX	4.593	4.672	4.054	4.856
ANN	1.698	1.783	1.967	2.016
SVR	1.732	1.914	1.933	2.065
LSSVR	1.426	1.678	1.704	1.816
KELM-lin	0.814	0.973	0.896	1.013
KELM-poly	0.674	0.784	0.792	0.804
KELM-rbf	0.492*	0.622*	0.643*	0.702*
KELM-wav	0.571	0.713	0.725	0.891

The asterisk numbers indicate the lowest error rate (MAPE and NRMSE).

5.2.4. DM test of out-of-sample forecasting

To assess forecasting accuracy of different models from a statistical perspective, we applied the DM test to eight models with four different independent variables. The results of the DM test are shown in Table 8. When time series, Baidu index, and Google index were integrated, DM statistics and p-values for the KELM-rbf model were less than −6.1125 and almost zero, respectively. This suggests that the KELM-rbf model significantly outperforms other benchmark models under the 100% confidence level.

These analyses have revealed some interesting findings: (1) when KELM models were considered as the test goal, all *p*-values were less than 0.00, suggesting that the KELM models are significantly superior to all other benchmark models at an almost 100% confidence level; (2) the forecasting performance of SVR and ANN were quite similar and neither of them statistically outperformed the other; (3) the ARIMAX model had the lowest forecasting performance with four different independent variables.

5.2.5. Robustness analysis

The robustness of the eight forecasting models with four different independent variables are assessed in this subsection. As ARIMAX, ANN, SVR, LSSVR and KELM models tend to produce different forecasting results with different initial settings, we ran all the forecasting models twenty times, and analyzed their robustness according to standard deviation of MAPE and NRMSE. These analyses are provided in Table 9, and provide evidence that (1) KELM is the most stable among all forecasting models, since its standard deviations from NRMSE and MAPE are far smaller than other benchmark models; (2) All forecasting models based on "time series + Baidu index + Google index" are the most robust approaches, and their standard deviations from MAPE and NRMSE are far smaller than all the corresponding models; (3) ARIMAX is the most unstable among all the forecasting models with different independent variables.

5.3. Summary

To summarize:

(1) The forecasting performance of "time series + Baidu index + Google index" is superior to other independent variables, followed by "time series + Baidu index" and "time series + Google index", whereas "time series" ranks the last.
(2) Due to the intrinsic complexity of the data of tourist arrivals, AI techniques are much more appropriate than the ARIMAX model in forecasting tourist arrivals.
(3) The proposed KELM models with different kernel functions outperform all other benchmark models in both forecasting accuracy and robustness.
(4) The forecasting power of the proposed KELM models is most stable and effective according to accuracy and robustness analysis, followed by LSSVR, SVR, ANN and ARIMAX.

6. Conclusions

In this paper, we proposed a forecasting framework that uses machine learning and internet search indexes to forecast tourist arrivals for popular destinations in China and compared its forecasting performance to the search data generated by Google and Baidu, respectively. This study verified the co-integration and Granger causality

Table 8
DM test results of out-of-sample datasets.

Tested model	Reference model			
	LSSVR	SVR	ANN	ARIMAX
Time series				
KELM-rbf	−6.2687 (0.0000)	−8.6782 (0.0000)	−11.6592 (0.0000)	−15.2679 (0.0000)
LSSVR		−0.6874 (0.2459)	−1.5789 (0.0572)	−8.2698 (0.0000)
SVR			−0.9876 (0.1617)	−8.0694 (0.0000)
ANN				−7.1627 (0.0000)
Time series + Baidu Index				
KELM-rbf	−5.9372 (0.0000)	−8.1695 (0.0000)	−10.0128 (0.0000)	−14.6872 (0.0000)
LSSVR		−1.7929 (0.0365)	−2.6197 (0.0044)	−6.0158 (0.0000)
SVR			−0.1185 (0.4528)	−5.0246 (0.0000)
ANN				−4.7691 (0.0000)
Time series + Google Index				
KELM-wav	−5.8564 (0.0000)	−7.9854 (0.0000)	−9.8756 (0.0000)	−14.6872 (0.0000)
LSSVR		−1.7543 (0.0397)	−2.5876 (0.0048)	−6.1229 (0.0000)
SVR			−0.2069 (0.4180)	−5.1267 (0.0000)
ANN				−4.5968 (0.0000)
Time Series + Baidu Index + Google Index				
KELM-rbf	−6.1125 (0.0000)	−8.5692 (0.0000)	−9.6485 (0.0000)	−14.1167 (0.0000)
LSSVR		−1.8467 (0.0324)	−2.5691 (0.0051)	−5.0168 (0.0000)
SVR			−0.2182 (0.4136)	−4.9613 (0.0000)
ANN				−4.6916 (0.0000)

relationship between internet search index and the volume of tourists in Beijing. The experimental results suggest that the proposed KELM models with integrated tourist volume series of Baidu index and Google Index can significantly improve forecasting performance. Compared to other popular benchmark forecasting methods, our KELM model, which integrates “tourist volume series + Baidu index + Google index”, is more accurate and more robust. Consequently, our KELM model is a promising approach towards resolving difficulties in forecasting tourist volume flows.

Our study can provide some inspiration. First, forecasting the volume of tourist accurately can help the tourism practitioners to optimize resources allocate and formulate pricing strategies rationally. Furthermore, forecasting tourist volume accurately may contribute to various industries that directly or indirectly depend on tourism. Second, it will provide solid evidence for policy makers and foresee the trends of tourist volume, which can help the government to adjust policy decisions, design infrastructure for tourism residential planning and transportation system. Third, the search engine query data is a user generated data and can be acquired freely from the main search engine. Many studies have verified that it can significantly improve the forecasting accuracy of tourist arrivals. Therefore, it can be used as an alternative data sources for the tourism resource management. Hence, tourism managers can employ our proposed forecasting framework to forecast tourism demand by collecting a variety of internet search data for the strategic decision-making in practical application.

In addition to tourist volume flow forecasting, the proposed forecasting framework with machine learning and internet search index can be applied to solving other complex and difficult forecasting problems, including stock trend forecasting, crude oil price forecasting, and exchange rates forecasting.

However, this study does have some limitations, mainly because we only used the Beijing travel market as a test case. The ability to summarize keyword selection way, as well as the concepts that keywords will converge no matter what search engine people choose, is limited. More research explores the use of Internet search data in other destinations as well as empirical research with a larger sample, are necessary to address these limitations. In addition, due to constant changes in web user information needs, establishing a comprehensive and dynamic keyword selection way that can effectively cope with changing market competition should be the research direction in the future.

Conflicts of interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

Table 9
Robustness analysis.

Std.[a]	Forecasting models							
	ARIMAX	ANN	SVR	LSSVR	KELM-lin	KELM-poly	KELM-rbf	KELM-wav
	Time series							
Std. of MAPE	0.0029	0.0032	0.0028	0.0033	0.0021	0.0003	0.0001	0.0005
Std. of NRMSE	0.0081	0.1012	0.0084	0.0088	0.0067	0.0045	0.0039	0.0042
	Time series + Baidu Index							
Std. of MAPE	0.0031	0.0029	0.0030	0.0032	0.0014	0.0001	0.0000	0.0001
Std. of NRMSE	0.0078	0.0094	0.0079	0.0083	0.0052	0.0041	0.0023	0.0038
	Time series + Google Index							
Std. of MAPE	0.0027	0.0026	0.0023	0.0033	0.0018	0.0000	0.0000	0.0001
Std. of NRMSE	0.0072	0.0095	0.0073	0.0072	0.0041	0.0036	0.0023	0.0033
	Time Series + Baidu Index + Google Index							
Std. of MAPE	0.0025	0.0021	0.0018	0.0019	0.0016	0.0001	0.0001	0.0003
Std. of NRMSE	0.0071	0.0091	0.0062	0.0067	0.0035	0.0029	0.0018	0.0035

Note: Std.[a] refers to the standard deviation.

Authors' contributions list

Shaolong Sun and **Yunjie Wei** conceived of the presented idea. **Shaolong Sun** developed the forecasting framework and performed the computations. **Shaolong Sun** and **Yunjie Wei** contributed to the interpretation of the results. **Shouyang Wang** and **Kwok-Leung Tsui** encouraged **Shaolong Sun** and **Yunjie Wei** to investigate the Internet search data and supervised the findings of this work. **Shaolong Sun** took the lead in writing the manuscript. **All authors** provided critical feedback and helped shape the research, analysis and manuscript. **All authors** read and approved the manuscript.

Acknowledgement

This work was partly supported by the National Natural Science Foundation of China (Project No. 51505307 and Project No. 11471275), General Research Fund (Project No. CityU 11216014) and a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. T32-101/15-R). Authors would like to express their sincere appreciation to the editor and the three independent referees in making valuable comments and suggestions to this paper. Their comments and suggestions have improved the quality of the paper immensely.

References

Artola, C., Pinto, F., & Garcia, P. D. (2015). Can internet searches forecast tourism inflows? *International Journal of Manpower, 36*, 103–116.
Askitas, N., & Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. *Applied Economics Quarterly, 55*, 107–120.
Athanasopoulos, G., & Hyndman, R. J. (2008). Modelling and forecasting Australian domestic tourism. *Tourism Management, 29*, 19–31.
Athanasopoulos, G., Hyndman, R. J., Song, H. Y., & Wu, D. C. (2011). The tourism forecasting competition. *International Journal of Forecasting, 27*(3), 822–844.
Bangwayo-Skeete, P. F., & Skeete, R. W. (2015). Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. *Tourism Management, 46*, 454–464.
Beneki, C., Eeckels, B., & Leon, C. (2012). Signal extraction and forecasting of the UK tourism income time series: A singular spectrum analysis approach. *Journal of Forecasting, 31*, 391–400.
Brida, J. G., & Risso, W. A. (2011). Research note: Tourism demand forecasting with SARIMA models - the case of South Tyrol. *Tourism Economics, 17*, 209–221.
Brynjolfsson, E., Geva, T., & Reichman, S. (2016). Crowd-squared: Amplifying the predictive power of search trend data. *MIS Quarterly, 40*(4), 941–961.
Chang, Y. W., & Liao, M. Y. (2010). A seasonal ARIMA model of tourism forecasting: The case of Taiwan. *Asia Pacific Journal of Tourism Research, 15*, 215–221.
Chan, F., Lim, C., & Mcaleer, M. (2005). Modelling multivariate international tourism demand and volatility. *Tourism Management, 26*, 459–471.
Chen, C. F., Lai, M. C., & Yeh, C. C. (2012). Forecasting tourism demand based on empirical mode decomposition and neural network. *Knowledge-Based Systems, 26*, 281–287.
Choi, H. Y., & Varian, H. (2012). Predicting the present with Google trends. *The Economic Record, 88*, 2–9.
Chu, F. L. (2008). Forecasting tourism demand with ARMA-based methods. *Tourism Management, 30*, 740–751.
Claveria, O., Monte, E., & Torra, S. (2016). Combination forecasts of tourism demand with machine learning models. *Applied Economics Letters, 23*, 428–431.
Croce, V. (2017). Business confidence and international tourism Demand: Evidence from a global panel of experts. *Global Journal of Management and Business Research, 16*(1), 29–42.
Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics, 20*(1), 134–144.
Du Preez, J., & Witt, S. F. (2003). Univariate versus multivariate time series forecasting: An application to international tourism demand. *International Journal of Forecasting, 19*(3), 435–451.
Fildes, R., Wei, Y., & Ismail, S. (2011). Evaluating the forecasting performance of econometric models of air passenger traffic flows using multiple error measures. *International Journal of Forecasting, 27*(3), 902–922.
Goh, C., & Law, R. (2011). The methodological progress of tourism demand forecasting: A review of related literature. *Journal of Travel & Tourism Marketing, 28*(3), 296–317.
Goodwin, P. (2008). A quick tour of tourism forecasting. *Foresight, 10*, 35–37.
Gunter, U., & Onder, I. (2015). Forecasting international city tourism demand for Paris: Accuracy of uni- and multivariate models employing monthly data. *Tourism Management, 46*, 123–135.
Huang, G. B. (2014). An insight into extreme learning machines: Random neurons, random features and kernels. *Cognitive Computation, 6*(3), 376–390.
Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing, 70*(1), 489–501.
Jungmittag, A. (2016). Combination of forecasts across estimation windows: An application to air travel demand. *Journal of Forecasting, 35*(4), 373–380.
Law, R., & Au, N. (1999). A neural network model to forecast Japanese demand for travel to Hong Kong. *Tourism Management, 20*, 89–97.
Lee, K. N. (2011). Forecasting long-haul tourism demand for Hong Kong using error correction models. *Applied Economics, 43*, 527–549.
Liang, Y. H. (2014). Forecasting models for Taiwanese tourism demand after allowance for Mainland China tourists visiting Taiwan. *Computers & Industrial Engineering, 74*(1), 111–119.
Lim, C., & McAleer, M. (2002). Time series forecasts of international travel demand for Australia. *Tourism Management, 23*, 389–396.
Li, X., Pan, B., Law, R., & Huang, X. K. (2017). Forecasting tourism demand with composite search index. *Tourism Management, 59*, 57–66.
Li, Z. C., & Sheng, D. (2016). Forecasting passenger travel demand for air and high-speed rail integration service: A case study of Beijing-guangzhou corridor, China. *Transportation Research Part A: Policy and Practice, 94*(1), 397–410.
MacKinnon, J. G., Haug, A. A., & Michelis, L. (1999). Numerical distribution functions of likelihood ratio tests for cointegration. *Journal of Applied Econometrics, 14*, 563–577.
McCulloch, W. S., & Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*, 115–133.
Pai, P. F., Hung, K. C., & Lin, K. P. (2014). Tourism demand forecasting using novel hybrid system. *Expert Systems with Applications, 41*(8), 3691–3702.
Pan, B., Wu, C. D., & Song, H. (2012). Forecasting hotel room demand using search engine data. *Journal of Hospitality and Tourism Technology, 3*(3), 196–210.
Pan, B., & Yang, Y. (2017). Forecasting destination weekly hotel occupancy with big data. *Journal of Travel Research, 56*(7), 957–970.
Park, D., Rilett, L. R., & Han, G. (1999). Spectral basis neural networks for real-time travel time forecasting. *Journal of Transportation Engineering, 125*(6), 515–523.
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. *Journal of Systems Science and Systems Engineering, 26*(2), 163–182.
Rivera, R. (2016). A dynamic linear model to forecast hotel registrations in Puerto Rico using Google trends data. *Tourism Management, 57*, 12–20.
Sencheong, K., & Turner, L. W. (2005). Neural network forecasting of tourism demand. *Tourism Economics, 11*, 301–328.
Shahrabi, J., Hadavandi, E., & Asadi, S. (2013). Developing a hybrid intelligent model for forecasting problems: Case study of tourism demand time series. *Knowledge-Based Systems, 43*, 112–122.
Shen, S., Li, G., & Song, H. (2009). Effect of seasonality treatment on the forecasting performance of tourism demand models. *Tourism Economics, 15*(4), 693–708.
Song, H., & Li, G. (2008). Tourism demand modelling and forecasting-A review of recent research. *Tourism Management, 29*(2), 203–220.
Song, H. Y., Li, G., Witt, S. F., & Athanasopoulos, G. (2011). Forecasting tourist arrivals using time-varying parameter structural time series models. *International Journal of Forecasting, 27*, 855–869.
Song, H., & Witt, S. F. (2000). Tourism demand modelling and forecasting: Modern econometric approaches. *Journal of Retailing and Consumer Services, 9*(1), 54–55.
Vanegas, M. (2013). Co-integration and error correction estimation to forecast tourism in El Salvador. *Journal of Travel & Tourism Marketing, 30*, 523–537.
Vosen, S., & Schmidt, T. (2011). Forecasting private consumption: Survey-based indicators vs. Google trends. *Journal of Forecasting, 30*(6), 565–578.
Witt, S., & Song, H. (2002). Forecasting future tourism flows. In S. Medlik, & A. Lockwood (Eds.). *Tourism and hospitality in the 21st century* (pp. 106–118). Oxford: Butterworth-Heinemann.
Witt, S. F., & Witt, C. A. (1995). Forecasting tourism demand: A review of empirical research. *International Journal of Forecasting, 11*(3), 447–475.
Wong, K. K., Song, H., Witt, S. F., & Wu, D. C. (2007). Tourism forecasting: To combine or not to combine? *Tourism Management, 28*(4), 1068–1078.
Wu, L. J., & Cao, G. H. (2016). Seasonal SVR with FOA algorithm for single-step and multi-step ahead forecasting in monthly inbound tourist flow. *Knowledge-Based Systems, 110*, 157–166.
Yang, X., Pan, B., Evans, J. A., & Lv, B. F. (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management, 46*, 386–397.
Zhu, Y., & Bao, W. B. (2014). The impact of investors' attention on stock returns -study based on Baidu index. *Service systems and service management (ICSSSM), 2014 11th international conference on* (pp. 1–5). IEEE.

Shaolong Sun is a PhD candidate majoring in Management Science and Engineering at the Institute of Systems Science, Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences, China. He is currently a research assistant at the Department of Systems Engineering and Engineering Management, City University of Hong Kong. His research interests include artificial intelligence, big data mining, machine learning, social networks analysis, knowledge management, economic and financial forecasting. He has published over 10 papers in journals including *Applied Energy*, and *Journal of Environmental Management*.

Yunjie Wei received her PhD degree in Management Science and Engineering at Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China in 2017 and also received a PhD degree in Management Science at City University of Hong Kong in 2018. She is currently an assistant professor at Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China. Her research interests include economic modeling, analysis and forecasting. She has published over 10 papers in journals including *Applied Energy*, and *Journal of Systems Science and Complexity*.

Kwok-Leung Tsui received his PhD degree in statistics from University of Wisconsin at Madison. He is currently the Head and a Chair Professor of Industrial Engineering of the Department of Systems Engineering and Engineering Management at the City University of Hong Kong, and the founder and Director of Center for Systems Informatics Engineering. Prior to joining City University of Hong Kong, he was a professor at the School of Industrial & Systems Engineering at the Georgia Institute of Technology. Professor Tsui was a recipient of the National Science Foundation's Young Investigator Award. He is a fellow of the American Statistical Association, American Society for Quality, and International Society of Engineering Asset Management; a U.S. representative to the ISO Technical Committee on Statistical Methods. Professor Tsui was a Chair of the INFORMS Section on Quality, Statistics, and Reliability and the Founding Chair of the INFORMS Section on Data Mining. Professor Tsui's current research interests include data mining, surveillance in healthcare and public health, prognostics and systems health management, calibration and validation of computer models, process control and monitoring, and robust design and Taguchi methods.

Shouyang Wang received his PhD degree in Operations Research from the Institute of Systems Science, Chinese Academy of Sciences, China, in 1986. He is currently a Bairen Distinguished Professor of Management Science at the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China and a Changjiang Chair Professor of Management Science and Engineering at the University of Chinese Academy of Sciences. He was the President of the International Society of Knowledge and Systems Sciences and is currently the President of the China Society of Systems Engineering. He is a fellow of the World Academy of Sciences, an academician of International Academy of Systems and Cybernetics Sciences, and a fellow of Asia Pacific Industrial Engineering and Management Society. He is the author or coauthor of 30 monographs, of which 15 published by Springer-Verlag, and more than 300 papers in leading journals. He is/was a co-editor of 16 journals including *Information & Management and Energy Economics*. He was a guest editor of special issues/volumes of more than 15 journals including *Decision Support Systems*, *Annals of Operations Research* and *European Journals of Operational Research*. His research interests include decision analysis, risk management, economic analysis and forecasting.