Tourism Management 82 (2021) 104208

Contents lists available at ScienceDirect

Tourism Management

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

Forecasting Chinese cruise tourism demand with big data: An optimized machine learning approach

Gang Xie[a,*], Yatong Qian[a,b], Shouyang Wang[a]

[a] CFS, MDIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, China
[b] University of Chinese Academy of Sciences, Beijing, 100049, China

ARTICLE INFO

Keywords:
Cruise
Big data
Gravitational search algorithm
Tourism demand forecasting

ABSTRACT

After more than ten years of exponential development, the growth rate of cruise tourist in China is slowing down. There is increasingly financial risk of investing in homeports, cruise ships and promotional activities. Therefore, forecasting Chinese cruise tourism demand is a prerequisite for investment decision-making and planning. In order to enhance forecasting performance, a least squares support vector regression model with gravitational search algorithm (LSSVR-GSA) is proposed for forecasting cruise tourism demand with big data, which are search query data (SQD) from Baidu and economic indexes. In the proposed model, hyper-parameters of the LSSVR model are optimized with GSA. By comparing these models with various settings, we find that LSSVR-GSA with selected mobile keywords and economic indexes can achieve the highest forecasting performance. The results indicate the proposed framework of the methodology is effective and big data can be helpful predictors for forecasting Chinese cruise tourism demand.

1. Introduction

Cruise tourism is one of the most dynamic and profitable sectors in the global tourism industry (Sun et al., 2018). With increasing incomes and leisure time, people are now more able to pursue new types of tourism. As one of the fastest growing areas in leisure tourism, the cruise industry is also growing rapidly (Wang et al., 2014; Wondirad, 2019). Cruise shipping provides services for tourists who value the experiences offered on board a ship. These services generally include catering, accommodation, transportation, and recreational activities, to name a few (Perea-Medina et al., 2019). To some extent, the cruise tourism industry is a barometer for the global political climate and an indicator of national economic health (Gibson & Parkman, 2019). In China, the high speed economic growth over the last few decades has brought about the emergence of a newly rich middle class and a large number of outbound tourists, providing a great development opportunity for the cruise tourism industry (Sun et al., 2014). According to the Cruise Lines International Association (CLIA) (2018), the Caribbean remains the most lucrative region for cruise line deployment (35.4%), followed by the Mediterranean (15.8%), Europe without the Mediterranean (11.3%) and China (6%). Incrementally cruise demand in the USA remains the largest (11.5%), followed by China (2.1%) (CLIA, 2018).

The cruise economy has received increased attention by Chinese government. Since 2012, the State Tourism Administration has approved the establishment of five *China Cruise Tourism Experimental Development Zones* in cities such as Shanghai, Tianjin, Shenzhen, Qingdao and Fuzhou. Additionally, the State Council issued three successive documents between 2015 and 2016. These reports clearly stipulate the need to promote the cruise tourism industry, to encourage entrepreneurialism in the field and to build cruise ships. Additionally, the State Council specifically stipulated to need to provide 15-day visa-free support to international tourists entering through cruise ships. In March 2017, the Ministry of Transport and the State Tourism Administration jointly issued *Some Opinions on Promoting the Integration of Transport and Tourism* to promote infrastructure construction of cruise homeports and to open new cruise routes. Consequently, there has been increasing investments in homeports, cruise ships and related promotional activities.

Due to vast interests in cruise tourism, large tourist base and enormous investment in the development of cruise tourism, China has emerged as a key player in the Asian cruise market (Hung et al., 2019). As the main driver of cruise tourism growth in Asia, China accounted for almost half of all regional tourist volume in 2015 (Hung, 2018). However, after more than ten years of exponential development, the growth rate of Chinese cruise line tourism is beginning to stagnate. Hence, for

[*] Corresponding author. 55 Zhongguancun East Road, Haidian District, Beijing, 100190, China.
E-mail addresses: gxie@amss.ac.cn (G. Xie), 473372032@qq.com (Y. Qian), sywang@amss.ac.cn (S. Wang).

https://doi.org/10.1016/j.tourman.2020.104208
Received 12 October 2019; Received in revised form 15 July 2020; Accepted 5 August 2020
Available online 19 August 2020

cruise tourism in China, cruise tourism demand forecasting is becoming a prerequisite for decision-making on investments and planning.

The capability to accurately plan incorporates the ability to analyze the destinations market. Forecasting cruise tourism demand is therefore the basis from which to generate efficient preparation (Cuhadar et al., 2014). Accurate predictions are particularly crucial because unoccupied cruise rooms cannot save costs. With accurate predictions of tourism, the government can formulate more appropriate cruise tourism policy, and cruise entrepreneurs and managers can develop more suitable marketing strategies to increase revenue and profitability.

The tourist volume data are usually available long after cruise tourism takes place. This retrospective data is of course very accurate, however, it is untimely for modeling and prediction. Yang et al. (2015) advocated that big data provide a new way to understand tourist behavior and enhance predictive accuracy. Therefore, identifying high quality big data resources is important for improving predictive accuracy of cruise tourism demand.

The development of search engines has opened new areas for the prediction of economics, management and other disciplines. People use search engines daily to find information, thus generating search query traffic, which represents potential tourist behavior preferences. The generated search query data (SQD) are real-time, which overcomes the lag of traditional forecasting methods. Consequently researchers have postulated that SQD can be helpful for tourism demand forecasting (Pan & Yang, 2017). As for cruise tourism, economic development has a significant impact on tourism demand (Sun, Feng; Gauri, 2014). Some economic indexes, which reflect the reality of business activities, the strength of consumer confidence, and relative value of the domestic currency, have been shown to correlate with tourism demand (Castillo-Manzano, Lopez-Valpuesta; Alanís, 2015; Chatziantoniou et al., 2016).

In this study, in order to enhance forecasting performance, including predictive accuracy and generalization ability, we propose a least squares support vector regression model with gravitational search algorithm (LSSVR-GSA), where SQD from Baidu as well as economic indexes are used as explanatory variables. In the determination of variables, keywords were determined by cointegration and Granger causality tests, and economic indexes are selected according to correlation coefficients. Additionally, an empirical study is conducted to illustrate the proposed methodology.

The remainder of this study proceeds as follows: Section 2 provides a review of related research. Section 3 describes the proposed framework and model in detail. Section 4 illustrates the proposed methodology by empirical study. Section 5 provides a discussion of a number of related issues and provides insights into cruise tourism policy. Section 6 provides some conclusions and the recommendations for future research.

2. Literature review

The literature about cruise tourism has historically focused on sustainability (Diedrich, 2010; Dawson et al., 2014; Carić & Mackelworth, 2014; Paoli et al., 2017; Chiappa et al., 2018), operations management (Mak, 2008) and tourist behaviors (Henthorne, 2000; Seidl et al., 2007; Wu et al., 2018). Generally, sustainability refers to economic, social, and environmental impacts of cruise tourism on a local community (Klein, 2011; MacNeill & Wozniak, 2018). Operations management includes the selection of a cruise homeport (Wang et al., 2014), tourism strategy (Chen, 2016), price composition of the cruise product (Niavis & Tsiotas, 2018) and efficiency analysis (Chang et al., 2017; Dai et al., 2019). Tourist behavior includes expenditure behavior (Larsen & Wolff, 2016), demand determinants (Gabe et al., 2006; Chen et al., 2016), cruising experience (Hung, 2018), and satisfaction (Castillo-Manzano & López-Valpuesta, 2018; Sanz-Blas et al., 2019).

Relatively few studies exist around cruise tourism forecasting, including forecasting bookings for the cruises (Sun et al., 2011), cruise ship arrivals (Tsamboulas et al., 2013) and cruise demand forecasting (Cuhadar; Cogurcu; Kukrer, 2014; Kollwitz & Papathanassis, 2011; Pavlić, 2013) as follows.

2.1. *Cruise tourism forecasting*

Using a Delphi approach, Kollwitz and Papathanassis (2011) evaluated cruise demand forecasting practices. They investigated the scenario that under-capacity and a near-100% capacity utilisation will persist in the foreseeable future, and examined the impact and validity of this scenario. With regard to the development prospects of the European cruise industry over the next decade, experts in the cruise industry were surveyed in order to explore views on published forecasts and the corresponding impact on the market. Results showed that the prediction was reasonably effective. Due to the evident increasing demand, new cruise liners were ordered through financing, and lower cruise travel prices were set by cruise liner operators to stimulate demand, resulting in under-capacity of cruise and supporting the existing forecasting practices.

Sun et al. (2011) used the data of a major cruise company in North America, and adopted various forecasting models to obtain predictions of final bookings for the cruises that have not yet departed at a specific reading point. The results showed that autoregressive integrated moving average (ARIMA), linear regression and moving average models could achieve the most accurate forecasts. To measure the attractiveness of the port as a port of call, Tsamboulas et al. (2013) designed a cruise attractiveness index, which was used to estimate the future cruise arrivals of the port. An alternative approach was implemented by Pavlić (2013) who used a seasonal autoregressive integrated moving average (SARIMA) model to forecast cruise tourism demand to Dubrovnik over the next five years. The investigation suggested that a different management policy should be implemented to satisfy the requirements of both passengers and stationary tourists, and to improve the living standards of the local community.

Similarly, Cuhadar et al. (2014) adopted a multi-layer perceptron (MLP), radial basis function (RBF) and generalized regression neural network (GRNN) to predict cruise tourism demand, and compared the predictive accuracy of these three neural networks. The data of foreign tourist arrivals from January 2005 to December 2013 were used as a measure of inbound cruise tourism demand and monthly cruise tourist arrivals at Izmir cruise port. The experimental results showed that the RBF neural network could achieve higher predictive accuracy than both MLP and GRNN. Thus, the monthly inbound cruise demand of Izmir was estimated by using RBF.

However, the aforementioned studies only used univariate time series models with historical data of cruise tourist volume to forecast cruise tourism demand. These methods are usually based on the assumption that consistent patterns and stable economic structure will be maintained during the prediction period. Once dramatic changes and one-off irregular events occur, these methods may not provide accurate predictions. These episodic events may cause structural breaks of tourism time series. Some detection methods, such as Bai-Perron test, can find the breakpoints in time series, but they cannot evaluate the impact of events that have not yet occurred. Therefore, there is the need for an alternative forecasting model. Big data may be leading indicators, which can reflect the impact of these events in advance to a certain extent (Pan & Yang, 2017; Yang et al., 2015). As a consequence, we investigate how to find high quality explanatory variables from big data for cruise demand forecasting.

2.2. *Tourism demand forecasting with big data*

Existing studies on tourism demand forecasting with big data are presented in Table 1, where one can see most researchers utilized SQD as big data. Baidu and Google launched Baidu Index and Google search analysis functions, respectively. Through these functions, we can get the attention frequency data of a keyword in a certain period of time from

Table 1
Typical works on tourism demand forecasting with big data.

Author(s)	Data source(s)	Variable(s)	Forecasting model(s)
Bangwayo-Skeete and Skeete (2015)	SQD	Keywords	AR-MIDAS
Choi and Varian (2012)	SQD	Keywords	Linear regression
Huang et al. (2017)	SQD	Keywords	Linear regression
Li et al. (2017)	SQD	GDFM and PCA-based indexes	Linear regression
Li et al. (2018)	SQD	PCA-based indexes	BPNN
Lv et al. (2018)	SQD	Keywords	SAEN
Park et al. (2017)	SQD	Averaging keywords	ARIMAX
Rivera (2016)	SQD	Summation-based index	DLM
Sun et al. (2019)	SQD	SAS-based index	KELM
Yang et al. (2015)	SQD	SAS-based index	Linear regression
Pan and Yang (2017)	SQD, website traffic, and weather information	Keywords, website traffic volume, and snowy data	ARIMAX
Raun et al. (2016)	Mobile tracking data	Space-time tracking data	Logistic regression
Yang et al. (2014)	Web traffic	Web traffic volume	ARIMAX

Note: autoregressive integrated moving average with exogenous variables (ARIMAX); autoregressive mixed-data sampling (AR-MIDAS); back propagation neural network (BPNN); dynamic linear model (DLM); generalized dynamic factor model (GDFM); kernel extreme learning machine (KELM); principal component analysis (PCA); shift and summation (SAS); stacked autoencoder with echo-state regression (SAEN).

Baidu and Google, namely SQD. Therefore, these SQD of keywords can objectively reflect the hot spots of tourism and the interests and needs of users in a specific period (Huang et al., 2017). When query keywords with SQD are used as explanatory variables in linear regression models, the predictive accuracy is generally improved (Choi & Varian, 2012; Huang et al., 2017).

Choi and Varian (2012) used the keyword "Hong Kong" from Google Trends as an exogenous variable to predict monthly visitor arrivals at Hong Kong from nine countries. Huang et al. (2017) used the keywords "The Forbidden City", "The Forbidden City in Beijing", and "Tickets of The Forbidden City" from Baidu for predicting tourism flows to the Forbidden City. Likewise, Bangwayo-Skeete and Skeete (2015) used SQD from Google Trends to construct a new indicator, which was introduced into AR-MIDAS models for tourism demand forecasting. This indicator was based on the keyword "hotels and flights", which was searched comprehensively from three major source countries to five popular tourist destinations in the Caribbean. The results indicated that AR-MIDAS outperformed other methods in most out-of-sample prediction experiments.

Besides SQD, there are other types of big data used for tourism demand forecasting. Using mobile positioning data of foreign tourists in Estonia from 2011 to 2013, Raun et al. (2016) analyzed spatial, temporal and compositional dimensions of a tourist destination, and demonstrated applications of big data in destination management. Results showed that smaller destinations could be distinguished within a country by geographical, temporal and compositional parameters, indicating that mobile positioning data could be helpful for forecasting tourism demand.

Yang et al. (2014) examined the application of web traffic data from a destination marketing organization (DMO) in predicting demand for hotel rooms. The results suggest that local DMOs, as well as convention and visitor bureaus, can provide local hotel markets with their web traffic data, which can help hotels obtain more accurate forecasts. To accurately predict weekly hotel occupancy for a destination, Pan and Yang (2017) developed a time series model incorporating several tourism big data sources, which were SQD, website traffic, and weekly weather information. The results showed the superiority of ARIMAX models with both SQD and website traffic data in accurate forecasts.

As for forecasting models, besides econometric models such as linear regression models (Choi & Varian, 2012; Huang et al., 2017; Li et al., 2017; Yang et al., 2015), ARIMAX (Park et al., 2017; Pan & Yang, 2017; Yang et al., 2014), DLM (Rivera, 2016), logistic regression model (Raun et al., 2016) and AR-MIDAS (Bangwayo-Skeete & Skeete, 2015), there are machine learning models such as BPNN (Li et al., 2018), SAEN (Lv et al., 2018) and KELM (Sun et al., 2019).

As mentioned above, many models have been developed for forecasting tourism demand with big data. The uniqueness of this study is that big data are firstly used for forecasting Chinese cruise tourism demand to improve the predictive accuracy. In the prediction, using both SQD and economic indexes, we develop an optimized machine learning model. The importance of this study is that accurate forecasts can effectively support relevant decisions such as appropriate cruise tourism policy and marketing strategies.

3. Methodology

In this section, we first present the framework of the methodology. Then, we describe the proposed model for forecasting cruise tourism demand.

3.1. *Framework of the methodology*

As previously mentioned, cruise tourism demand is closely related with tourist behavior and economic development. Therefore, we utilized search query volume and economic indexes as explanatory variables. Additionally, to enhance predictive accuracy, we adopted the gravitational search algorithm (GSA) for parameter optimization in the LSSVR model. The framework of the methodology is described in Fig. 1, which includes three modules: data preparation, GSA for parameter optimization and LSSVR for cruise tourism demand forecasting.

3.2. *LSSVR-GSA model*

The support vector machine (SVM) is based on statistical learning theory and the principle of structural risk minimization (Vapnik, 2013). However, the support vector regression (SVR) model is time-consuming when large-scale samples are processed. In order to improve the speed of calculations, Suykens and Vandewalle (1999) proposed the LSSVR model, which first maps the original data, y_t, into a high-dimensional feature space by using a nonlinear kernel function $\varphi(y_t)$, and then performs a linear regression analysis using the following formula:

$$\hat{y}_t = w^T\varphi(y_t) + b \tag{1}$$

where parameters w and b are the weight and bias, respectively. The two parameters are estimated according to the function of structural risk minimization, as follows:

$$\text{Min } \left(w^T w\right)\Big/ 2 + \left(\gamma \sum_{t=1}^{T} e_t^2\right)\Big/ 2 \tag{2}$$

$$s.t\ \hat{y}_t = w^T\varphi(y_t) + b + e_t\ ,\quad (\mathrm{t}=1,\ 2, \ldots, \mathrm{T})$$

where γ is the penalty parameter and e_t is the estimation error at time t.

After solving the function in Eq. (2), we can obtain the solution to the optimization problem using the following formula:

$$\hat{y} = \sum_{t=1}^{T} \alpha_t K(y, y_t) + b, \tag{3}$$

where $K(y, y_t) = \exp(-\|y - y_t\|^2/\sigma^2)$ is a kernel function with Gaussian

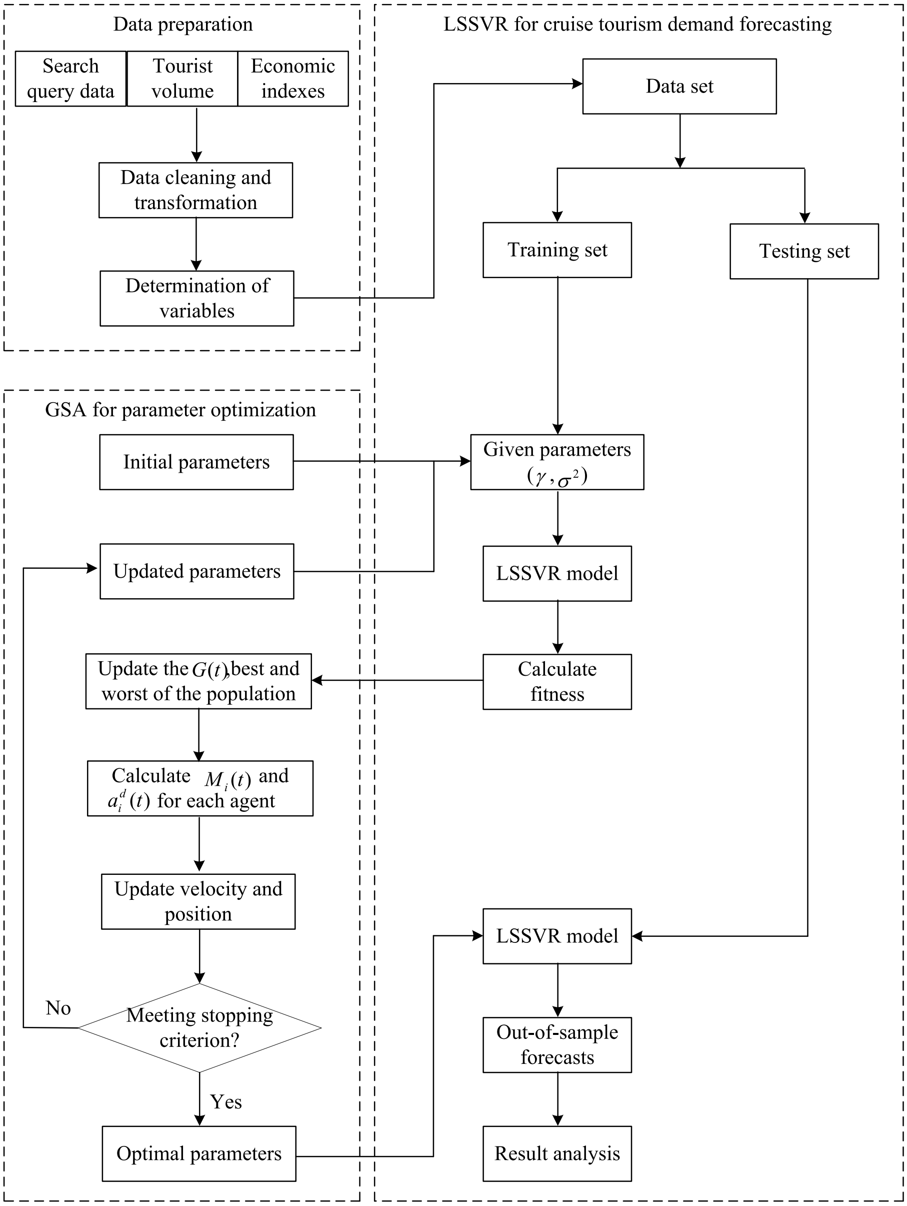
Data preparation
Search query data
Tourist volume
Economic indexes
Data cleaning and transformation
Determination of variables
LSSVR for cruise tourism demand forecasting
Data set
Training set
Testing set
GSA for parameter optimization
Initial parameters
Given parameters (γ, σ^2)
Updated parameters
LSSVR model
Update the $G(t)$,best and worst of the population
Calculate fitness
Calculate $M_i(t)$ and $a_i^d(t)$ for each agent
Update velocity and position
LSSVR model
No
Meeting stopping criterion?
Yes
Out-of-sample forecasts
Optimal parameters
Result analysis

Fig. 1. The flow chart of the methodology.

RBF.

Generally, in traditional LSSVR model, hyper-parameters γ and σ^2 are given before training. Obviously, these parameters are not optimal. To optimize the hyper-parameters of LSSVR model, we implemented GSA, which is a heuristic optimization algorithm proposed by Rashedi et al. (2009). Originating from swarm intelligence optimization algorithm, GSA simulates the universal gravitation in physics. The principle of the GSA is that the search particles are regarded as a group of agents attracted by the interaction of universal gravitation in space. The movement of agents follows the law of dynamics which is that, gravitation between two agents is positively correlated with the masses of the two agents, but negatively correlated with the distance between the two agents.

For a system with N_a agents, the position of the ith agent is defined, as follows:

$$X_i = \left(x_i^1, \ldots, x_i^d, \ldots x_i^n\right) \tag{4}$$

where n is the number of the dimensions of the ith agent and x_i^d is the position of the ith agent in the dth dimension, $i = 1, 2, \ldots, N_a$. The force between agents i and j in the dth dimension at time t is defined as:

$$F_{ij}^d(t) = G(t)\frac{M_{pi}(t) \times M_{aj}(t)}{R_{ij}(t) + \varepsilon}\left(x_j^d(t) - x_i^d(t)\right) \tag{5}$$

where, at time t, $G(t)$ is the constant of gravitation, $M_{pi}(t)$ is the passive gravitational mass of the ith agent, $M_{aj}(t)$ is the active gravitational mass of the jth agent, ε is a small constant, and $R_{ij}(t)$ is the Euclidian distance between two agents i and j, i.e., $R_{ij}(t) = \left\|X_i(t), X_j(t)\right\|_2$. The total force that the ith agent has in dimension d is defined as:

$$F_i^d(t) = \sum_{j=1, j\neq i}^{N_a} rand_j F_{ij}^d(t) \tag{6}$$

where $rand_j$ is a random variable and uniformly distributed in the interval [0,1]. Then, we can get the acceleration, which is derived from Newton's second law of motion as follows:

$$a_i^d(t) = \frac{F_i^d(t)}{M_{ii}(t)} \tag{7}$$

where $M_{ii}(t)$ is the inertial mass of the ith agent.

Let $M_{ai}(t) = M_{pi}(t) = M_{ii}(t) = M_i(t)$. The gravitational and inertial masses of the ith agent at time t are

$$m_i(t) = \frac{fit_i(t) - worst(t)}{best(t) - worst(t)} \tag{8}$$

$$M_i(t) = \frac{m_i(t)}{\sum_{j=1}^{N_a} m_j(t)} \tag{9}$$

where $fit_i(t)$ is the fitness value of the ith agent at time t.

In the LSSVR-GSA model, GSA is used to optimize the hyper-parameters γ and σ^2 of LSSVR model. For more information about GSA, please refer to Rashedi et al. (2009). In order to compare with LSSVR-GSA, we also use LSSVR-CV model, in which the values of the hyper-parameters are determined via a cross-validation grid search method.

4. Empirical study

In this section, using the time series of cruise tourist volume in China and corresponding big data, we conduct an empirical study for the purposes of validating and verifying the proposed analytical model.

4.1. Data description and experiment design

In recent years, the volume of cruise tourism in China has reached a new high. However, the growth rate of tourism demand has dropped substantial, as shown in Fig. 2. When making travel plans, tourists generally enter basic query keywords in search engines such as Baidu in order to acquire relevant information, which may include destination details, ticket prices and past customer feedback. The contents of the retrieved webpages related to the keywords assist potential tourists when deciding travel plans.

Cruise tourism potentially has a significant impact on local economies (Chen et al., 2019; Dwyer & Forsyth, 1998) which may be small but beautiful inlets or harbors around the world, somewhat cut off from standard economic development. This phenomenon also has a knock-on effect on economic development because this influences the cruise tourism industry (Gibson & Parkman, 2019; Sun, Feng; Gauri, 2014). In order to explore this, our study is designed to incorporate three economic indexes related to cruise tourism demand and data are collected, as follows:

Purchasing manager index (PMI) is a monthly survey index of purchasing managers, which can reflect the changing trend of economic situation. In the survey, the purchasing managers make qualitative judgments on relevant questions and choose one of three answers (increase, unchanged or decrease compared with the last month). The PMI is constructed by counting the percentage of three answers. 50% of PMI value is the dividing line of prosperity and withering, and the dynamic change of the index reflects the cyclical state of economic activities. The PMI business reports of manufacturing and non-manufacturing industries are released on the 1st and 3rd of each month, far exceeding the statistical reports of other government departments. Therefore, PMI has become a timely and reliable leading indicator for monitoring economic prospects (Müller, 2013).

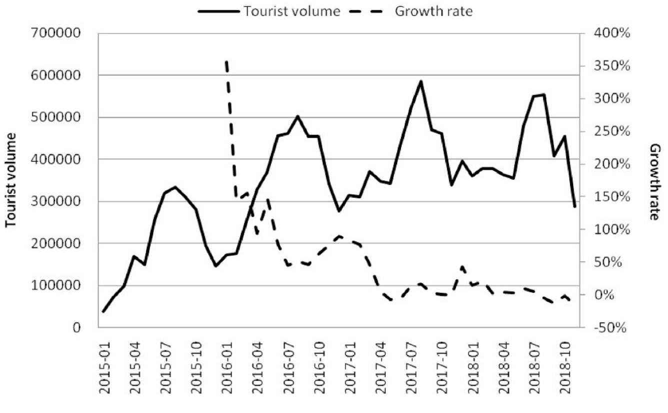

Fig. 2. Monthly Chinese cruise tourist volume.

Consumer confidence index (CCI) is a leading indicator to reflect the strength of consumer confidence. It comprehensively reflects and quantifies consumers' subjective feelings on the current economic situation, economic prospects, income level, income expectation and consumption psychological state, and predicts the economic and consumption trend. CCI is composed of consumer satisfaction index and consumer expectation index. The former refers to consumers' evaluation of current economy and life, including satisfaction with income, quality of life, macro-economy, consumption expenditure, employment status, purchase of durable consumer goods and savings; the latter refers to consumers' expectation of future changes, including the expectations of the above indicators in the next year, purchasing houses, decoration and automobiles in the next two years, and changes in the stock market in the next six months. Therefore, this index is a useful source for estimating economic and prospective consumption trends (Gounopoulos et al., 2012). The Chinese CCI data takes the end of 1997 as the baseline, and the consumer confidence increases in CCI.

The real effective exchange rate index (REERI) is the nominal effective exchange rate adjusted by the relative price level or cost index between a country and the selected country. According to the statistics of China Customs, in 2018, China's total import and export value ranked first in the world, reaching 4.62 trillion US dollars, with a trade surplus of US $351.76 billion. The USA, Japan, South Korea, Taiwan, Germany, Australia, Vietnam, Brazil, Malaysia and Russia were the top 10 major trading partners with China. REERI not only takes into account exchange rate fluctuations for countries which are major trading partners, but also excludes the inflation factor. An increase of REERI represents an increase in the relative value of the domestic currency, while a decrease indicates a devaluation of the local currency (Chang et al., 2013; Pavlic; Svilokos; Tolic, 2015).

Monthly data of Chinese cruise tourist volumes (Fig. 2) and economic indexes (PMI, CCI and REERI in Fig. 3) are collected from the Wind Database (http://www.wind.com.cn/). In addition, weekly SQD from Baidu (https://index.baidu.com/) are collected.

The sample data cover the period from January 2015 to November 2018. Here, the keywords of SQD are generated through two main sources: personal computer (PC) and mobile terminals. To match monthly data with weekly SQD, the monthly time series of tourist

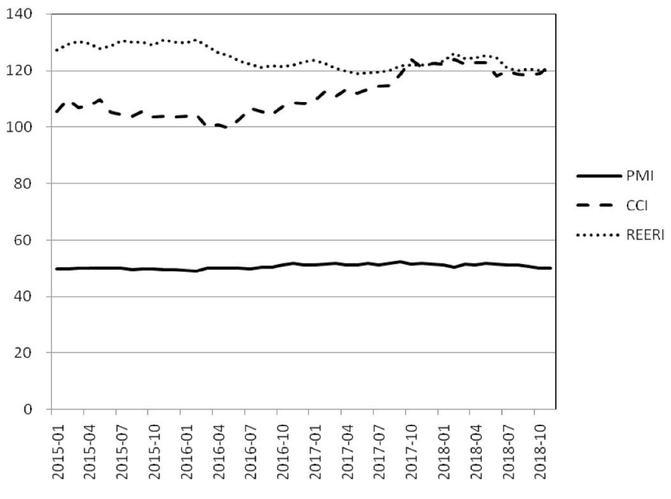

Fig. 3. Monthly economic indexes: PMI, CCI and REERI.

volume and economic indexes were converted into weekly series using the “quadratic-match average” method (Vogelvang, 2005). There are 205 observations for each time series analyzed. The first 193 and the last 12 observations are used as training and testing sets, respectively.

To predict Chinese cruise tourist volume, the following three modules are incorporated into the proposed LSSVR- GSA model.

Data preparation. After referring to related literature (Li et al., 2017; Park et al., 2017; Yang et al., 2015), we selected the possible seed keywords, which are divided into three categories: cruise, cruise tourism, and cruise sites. During the search, the keywords recommended by Baidu were used as search keywords for the next round. 23 keywords were garnered through Baidu, as is shown in Table 2. These keywords were generated through two electronic data sources: PC and mobile terminals.

The Pearson correlation coefficient between SQD series of keyword x with lag order l (l=0,1 …, L) and tourist volume series y is calculated as:

$$r_l = \frac{\sum_{t=1}^{T-l}(x_t - \overline{x})(y_{t+l} - \overline{y})}{\sqrt{\sum_{t=1}^{T-l}(x_t - \overline{x})^2 \sum_{t=1}^{T-l}(y_{t+l} - \overline{y})^2}}, \tag{10}$$

where $l < T$, $\overline{x} = \sum_{t=1}^{T-l} x_t/(T-l)$, and $\overline{y} = \sum_{t=1}^{T-l} y_{t+l}/(T-l)$. We set the threshold value for the maximum correlation coefficients of keywords, and then keywords having correlation coefficients higher than the threshold value were retained. After data cleaning and transformation, we conducted stationarity, cointegration and Granger causality tests to select keywords as explanatory variables. In addition to SQD, economic indexes, namely PMI, CCI and REERI, were selected according to correlation coefficients.

GSA for parameter optimization. This predefined algorithm searches for optimal solutions by moving agent positions across the population. With the circulation of the algorithm, agents move continuously within search parameters using interspatial gravitation. When agents move to the best positions, optimal solutions for the parameters can be identified. The main steps of GSA are as follows:

Step 1. Initialization. Set population size N_a, initial value of gravitational constant G_0, agent positions, and maximum number T of iterations.

Step 2. Calculate fitness.

Step 3. Calculate the inertial mass of agents, the gravitation and acceleration of each agent in each direction.

Step 4. Update the location, fitness and global optimum of each agent.

Step 5. When meeting stopping criterion, the optimal parameters are output for the LSSVR model. Otherwise, return to Step 3 for the next iteration.

LSSVR for cruise tourism demand forecasting. The LSSVR-GSA model, together with selected big data, were used for static out-of-sample forecasting of Chinese cruise tourist volume. In the experiment, the training dataset was utilized to determine unknown parameters of the pre-defined models, and the testing dataset was used for model evaluation.

Table 2
Search queries from Baidu.

Cruise	cruise	Tourism	cruise tourism
	cruise company		cruise strategy
	cruise website		cruise travel notes
	cruise lines		cruise tourism strategy
	Costa Cruise		cruise tour price
	Costa Cruise website		ship tourism
	Costa Serena		ship travel
	Royal Caribbean	Cruise sites	cruise home port
	Royal Caribbean International		cruise terminal
	Royal Caribbean International website		Sanya cruise
	Star cruises		Shanghai cruise
	Viking cruises		

Five models, which include ARIMA, BPNN, RBF, LSSVR-CV and LSSVR-GSA, were implemented for weekly cruise tourism demand forecasting. The best ARIMA model for each training sample was determined using the minimization of the Schwarz Criterion (SC) and Akaike Information Criterion (AIC). In ANN models, NN(I–H-1) was established and the number I of the input nodes was determined by partial autocorrelation analysis. The number H of the hidden nodes of ANN models was determined through trial-and-error testing to minimize the mean square error (MSE) of training. In LSSVR-CV model, the Gaussian RBF function was employed as the kernel function, and the values of γ and σ^2 were determined via a k-fold cross-validation grid search method in the range [0.01, 10,000]. As for the LSSVR-GSA model, GSA was used to optimize the hyper-parameters of LSSVR model.

In order to compare the forecasting performance of the models, we adopted the root mean squared error (RMSE), mean absolute percentage error (MAPE), and Willmott’s index of agreement (WIA) as the evaluation criteria:

$$RMSE = \sqrt{\sum_{t=1}^{N}(y_t - \widehat{y}_t)^2 \Big/ N}, \tag{11}$$

$$MAPE = 100 \times \sum_{t=1}^{N}|1 - \widehat{y}_t / y_t| \Big/ N, \tag{12}$$

$$WIA = 1 - \sum_{t=1}^{N}(\widehat{y}_t - y_t)^2 \Big/ \sum_{t=1}^{N}(|\widehat{y}_t - \overline{y}| + |y_t - \overline{y}|)^2, \tag{13}$$

where N is the number of observations in the testing set. In the evaluation of forecasting performance, RMSE and MAPE are used to evaluate the predictive accuracy, which decreases in the value of criteria. Because WIA can measure the outside prediction ability of a model, WIA is used to evaluate the generalization ability, which increases in the value of WIA.

In this study, the calculation of Pearson correlation coefficients, BPNN, RBF, LSSVR-CV and LSSVR-GSA models were implemented in MATLAB, while ARIMA(X) models were implemented in Eviews software.

4.2. Determination of variables

Search queries on Baidu are provided in Table 2. We calculated Pearson correlation coefficients between Chinese weekly tourist volume (TV) and SQD of each keyword with 0–12 lag periods, i.e. $L = 12$. The selection threshold is the result of tradeoff between predictive accuracy and model simplicity (Yang et al., 2015). Considering these two factors, we set the threshold value as 0.6 to obtain the appropriate number of predictors. The keywords with correlation coefficients of more than 0.6 became explanatory variables for cruise tourism demand forecasting.

The maximum correlation coefficients of keywords are listed in Table 3, where VC_PC denotes keyword “Viking cruises” generated by PC terminals and SHC, CTS, CS, CT, VC, CR, CC, CHP, SC denote corresponding keywords generated by mobile terminals. Obviously, mobile terminals generate more potential explanatory variables than do PC terminals.

Table 3 shows that the lag orders of the maximum correlation coefficient of search queries range from 0 to 9. First of all, China’s earliest and most famous cruise tourism (Shanghai cruise) has the largest lag term, followed by tourism strategy (cruise tourism strategy, cruise strategy), then the specific information related to cruise tourism (cruise

Table 3
Maximum correlation coefficients of keywords.

Data source	Keyword	Lag order
PC	Viking cruises (VC_PC)	5
	Shanghai cruise (SHC)	9
	cruise tourism strategy (CTS)	8
	cruise strategy (CS)	6
	cruise tourism (CT)	5
	Viking cruises (VC)	5
Mobile	cruise (CR)	4
	cruise company (CC)	1
	cruise home port (CHP)	0
	Star cruises (SC)	0

tourism, cruise), and finally the cruise company and the place of cruise terminal (cruise company, cruise home port, Star cruises). This shows that most potential Chinese cruise tourists first search relevant information of the most famous cruise tourism in Shanghai in order to get a preliminary understanding of this new type of tourism, which happened about nine weeks before the trip. After that, they will search for travel information about the cruise experience to prepare for travel plans, about six to eight weeks in advance. Then, four to five weeks before the trip, they determine the specific itinerary plan of cruise tourism, and finally, one week before departure, they learn about the service details of the cruise company. This shows that in different stages of tourism planning process, tourists make different tourism decisions and search for relevant tourism information. The search query keywords and their lag orders reflect this behavior process.

After data cleaning and logarithmic transformation, we conducted augmented Dickey-Fuller, Johnson cointegration and Granger causality tests to check whether the keywords can be used as predictors (Huang et al., 2017). Results show that the dependent variable (Log*TV*) and independent variables (Log*VC_PC*, Log*SHC*, Log*CTS*, Log*CS*, Log*CT* and Log*VC*) are stationary in levels and there is co-integrative relationship between the independent and dependent variables. Moreover, independent variables Granger cause the dependent variable but the dependent variable does not Granger Cause independent variables. Therefore, VC_PC, SHC, CTS, CS, CT, VC are selected as explanatory variables for forecasting Chinese cruise tourism demand. As for economic indexes in Fig. 3, PMI and CCI have positive relationships with TV, while REERI has an unreasonable negative correlation with TV. Therefore, PMI and CCI are used as explanatory variables.

4.3. *Prediction results*

In univariate time series forecasting, the ARIMA (1,1,1) models was used. In BPNN and RBF models, NN(3-10-1) is established (3 input nodes, 10 hidden nodes and 1 output node). In LSSVR-CV and LSSVR-GSA models, $k = 10$, $N_a = 50$, $T = 100$, and inputs are the same as ANN models. In the prediction with big data, keywords (VC_PC, SHC, CTS, CS, CT, VC) with their lag orders and contemporaneous economic indexes (PMI and CCI) are added as inputs additionally. With different settings, the forecasting performance of models is shown in Table 4.

Table 4
The forecasting performance of models.

Model	RMSE	MAPE	WIA	RMSE	MAPE	WIA
	Time series			Time series + Economic indexes		
ARIMAX	34475.855	6.384	0.9456	31664.27	5.944	0.9542
BPNN	33836.939	6.261	0.9505	32740.62	5.86	0.9542
RBF	32448.538	6.290	0.9536	30384.029	5.842	0.9605
LSSVR	31811.728	5.858	0.9557	29842.032	5.563	0.9561
LSSVR-GSA	27710.239	5.362	0.9633	23380.277	4.305	0.9669
	Time series + PC keywords			Time series + PC keywords + Economic indexes		
ARIMAX	34023.511	6.278	0.9465	30603.692	5.712	0.9567
BPNN	30227.744	6.127	0.9535	29396.441	5.591	0.9561
RBF	28497.638	5.948	0.9591	27977.112	5.459	0.9633
LSSVR	21957.279	5.304	0.9797	20741.608	5.154	0.9801
LSSVR-GSA	20399.558	4.401	0.9810	17174.676	3.861	0.9869
	Time series + Mobile keywords			Time series + Mobile keywords + Economic indexes		
ARIMAX	33256.406	6.055	0.9487	30493.220	5.596	0.9571
BPNN	32926.785	5.898	0.9464	26661.749	4.783	0.9673
RBF	31548.110	5.866	0.9501	25437.250	4.593	0.9697
LSSVR	29275.590	4.766	0.9556	20004.678	4.376	0.9819
LSSVR-GSA	17465.026	3.861	0.9855	14958.891	3.311	0.9896

Table 5
Robustness analysis.

Model	Std. of RMSE	Std. of MAPE	Std. of WIA	Std. of RMSE	Std. of MAPE	Std. of WIA
	Time series			Time series + Economic indexes		
ARIMAX	9.3835	0.0034	0.0005	8.2117	0.0029	0.0005
BPNN	8.1815	0.0026	0.0005	7.9861	0.0022	0.0004
RBF	8.0710	0.0024	0.0004	7.9088	0.0020	0.0004
LSSVR	7.6098	0.0019	0.0003	4.7771	0.0015	0.0003
LSSVR-GSA	4.5253	0.0008	0.0002	3.6287	0.0006	0.0002
	Time series + PC keywords			Time series + PC keywords + Economic indexes		
ARIMAX	8.5242	0.0030	0.0005	8.1454	0.0027	0.0005
BPNN	8.0721	0.0024	0.0004	7.7893	0.0020	0.0004
RBF	8.0574	0.0021	0.0004	7.8867	0.0019	0.0003
LSSVR	4.9068	0.0014	0.0003	4.5344	0.0012	0.0002
LSSVR-GSA	3.7253	0.0006	0.0001	2.9706	0.0004	0.0001
	Time series + Mobile keywords			Time series + Mobile keywords + Economic indexes		
ARIMAX	8.0966	0.0028	0.0005	8.0510	0.0026	0.0005
BPNN	7.7053	0.0018	0.0003	7.6006	0.0016	0.0003
RBF	7.6699	0.0018	0.0003	7.4205	0.0016	0.0003
LSSVR	4.3236	0.0011	0.0002	4.1820	0.0010	0.0002
LSSVR-GSA	2.8348	0.0003	0.0001	2.3867	0.0001	0.0000

Note: Std. refers to the standard deviation.

4.4. *Robust analysis*

We ran five models with different initial settings, each ten times over, and analyzed the robustness of the models in terms of RMSE, MAPE and WIA. From the results provided in Table 5, we can conclude that: (1) LSSVR-GSA is the most stable forecasting model, since the corresponding standard deviations from RMSE, MAPE and WIA outputs are smaller than other comparable models. (2) All forecasting models under the setting “Time series + Mobile keywords + Economic indexes” are the most robust, and related standard deviations of the three criteria are smaller than corresponding models under alternative settings. (3) ARIMAX is the most unstable among all the forecasting models under different settings. (4) Similarly, all forecasting models with mobile keywords are more stable than corresponding models with PC keywords.

5. Discussion

Based on the prediction results, we first analyzed the forecasting performance of the proposed models. Then, some insights into the cruise tourism policy were provided according to the analysis.

5.1. *Result analysis*

Clearly, according to the evaluation criteria of predictive accuracy in Table 4, the LSSVR-GSA models with keywords produced the best forecasting performance. The ARIMA and RBF models are benchmark methods currently adopted for forecasting cruise tourism demand. Due to the significant nonlinearity of the cruise tourist volume, the BPNN, RBF, LSSVR-CV, and LSSVR-GSA models achieve higher accuracy than do the ARIMA models.

In terms of RMSE, MAPE and WIA, improvement rates (IR) of LSSVR-GSA over other four models are shown in Fig. 4. Clearly, LSSVR-GSA outperforms these models in both predictive accuracy and generalization ability. Particularly, LSSVR-GSA is better than LSSVR-CV model, the reason may be that the selection of k is rather subjective and there is over-fitting in LSSVR-CV model.

When both mobile keywords and economic indexes are used as explanatory variables, the IR of models with big data over models without big data is shown in Fig. 5. In the setting, the ARIMAX, BPNN, RBF, LSSVR-CV and LSSVR-GSA models exhibited 11.55%, 21.21%, 21.61%, 37.12% and 46.02% decreases in RMSE, 12.34%, 23.61%, 26.98%, 25.30% and 38.25% decreases in MAPE, and 1.22%, 1.77%, 1.69%, 2.74% and 2.73% increases in WIA, suggesting that the selected big data had significantly improved the forecasting performance of models.

Furthermore, Friedman's nonparametric one-way ANOVA test was employed to examine model superiority, which is widely considered standard practice (Derrac et al., 2011). The results in Table 6 show that the LSSVR-GSA model ranks the top for RMSE and MAPE, indicating that it has the highest predictive accuracy. For criterion WIA, the LSSVR-GSA model ranks the bottom, which proves that it has the best generalization ability. Moreover, the corresponding p-values for the criteria are all less than 0.01, indicating that it has the best performance at a 1% level of significance.

5.2. Insights on cruise tourism policy

From the above analysis, we can provide some insights on cruise tourism policy as follows:

When potential tourists have cruise travel intention, they will search tourism related information through search engines. In each step of the decision-making process, if everything can meet their own needs and economic affordability, the potential tourists may realize their tourism plans. Therefore, the SQD of cruise tourism related query keywords generated by search behavior reflects tourists' travel intention and tourism trend. The above research also shows that SQD of some keywords can be used to predict cruise tourism demand. In addition, cruise tourism is regarded as a luxury activity, providing services for tourists with relatively high income (Hung, 2018). When the economy is booming and the expected consumption trend is good, more potential tourists will pay for cruise tourism. Otherwise, the cruise tourism demand will decrease. Since some economic indicators can objectively reflect the economic prospect and the expected consumption trend, adding them into the forecasting model can form a powerful supplement to keyword variables. In this way, the exogenous variables surrounding cruise tourism, including query keywords and economic indicators, will have an impact, and can produce higher predictive accuracy than the univariate time series model in the existing literature.

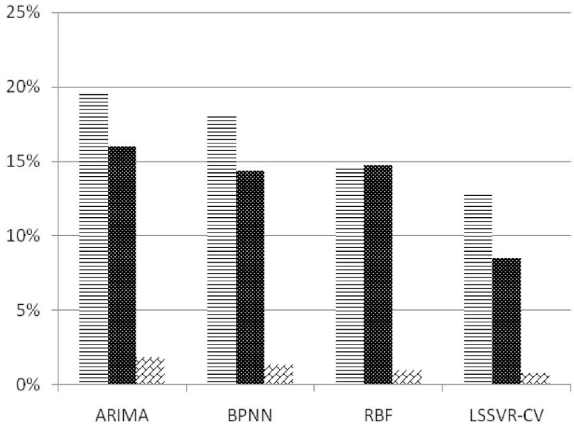

Fig. 4. IR of LSSVR-GSA over other models.

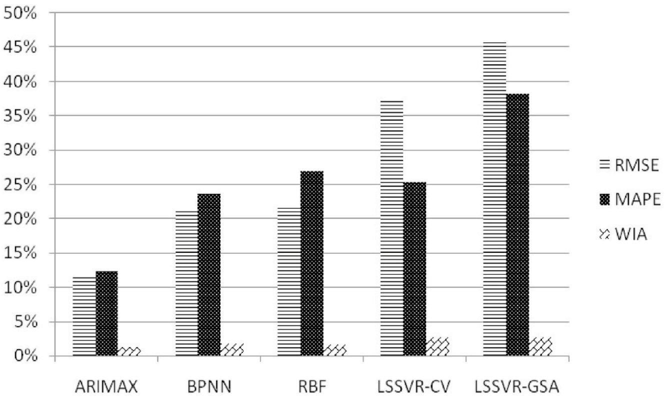

Fig. 5. IR of models with "Time series + Mobile keywords + Economic indexes" over models without big data.

Table 6
Ranks and p-values in Friedman tests.

	RMSE	MAPE	WIA
ARIMAX	4.83	5	1
BPNN	4.17	3.83	2.17
RBF	3	3.17	2.83
LSSVR-CV	2	2	4.42
LSSVR-GSA	1	1	4.58
p-value	0.0003	0.0003	0.0008

The selected keywords, namely 'Viking cruises (VC_PC)' generated from PC terminals, and 'Shanghai cruise (SHC)', 'cruise tourism strategy (CTS)', 'cruise strategy (CS)', 'cruise tourism (CT)' and 'Viking cruises (VC)' generated from mobile terminals, are effective keywords for forecasting Chinese cruise tourism demand. This indicates that cruise tourism is still a novelty for most potential Chinese tourists and searching keywords through Baidu is the preferable way to acquire experience information on cruise tourism influencing decision-making. In addition, contemporaneous PMI and CCI are effective economic indexes, which improve the forecasting performance of RMSE, MAPE, and WIA. The results suggest that both selected keywords and economic indexes are helpful predictors for cruise tourism demand forecasting. Hence, related decision-makers on cruise tourism can estimate the demand by collecting these data.

An issue deserving specific attention is that mobile terminals can generate more keywords as predictors than can PCs. The reason may be that intelligent mobile phones are more popular and readily available in China. Moreover, mobile phones are generally personal belongings, which may better reflect individual needs. In contrast, PCs are more likely to be used at workplaces. Therefore, potential tourists are more likely to search for travelling-related information on their mobile phones.

Despite the fact that the growth rate of cruise tourism demand has at the very least begun to stagnate, China's cruise tourism market still has great potential. In current cruise tourism market, the original potential demand has been met, but new market demand still needs to be developed. The cruise tourism market of the United States may be considered as a good example. To popularize the cruise tourism, network surveys can be conducted to feed back the needs of potential tourists.

In the era of the Internet, big data and the Internet provide an opportunity for tourism service providers to formulate suitable tourism policy. The tourism-related SQD can provide insight into tourists' preferences and predict their consumption behavior. Such insight can help to improve the quality of service and tourism marketing efficiency. Since Chinese cruise tourists will search related information in Baidu, cruise operators can advertise in Baidu for network marketing, which recommends specific cruises and destinations through advertising to netizens. In addition, cruise tourism-related enterprises should provide equipment access on the Internet, update their capability to acquire and analyze big data, thereby balancing cruise supply and market demand.

6. Conclusions

To enhance forecasting performance, we proposed a new LSSVR-GSA model for forecasting cruise tourism demand based on SQD from Baidu and economic indexes. In the prediction, keywords were selected as explanatory variables through cointegration and Granger cause tests, and economic indexes were selected according to correlation coefficients. An empirical study was conducted to highlight the related issues. The results indicate that the proposed model using mobile keywords and economic indexes can enhance forecasting performance.

The contribution of this study is that both SQD and economic indexes were adopted for forecasting cruise tourism demand. The proposed method enhances the forecasting performance effectively. In theory, we propose a new framework, using big data from different sources and an optimized machine learning approach, which can be used not only for forecasting cruise tourism demand, but also for other big data related prediction. Through the comparison and analysis of forecasting performances, we can summarize cruise tourism demand forecasting, as follows:

First, big data can be used to generate accurate predictions for cruise tourism demand forecasting. The results of this novel study demonstrate that modeling with big data can achieve better forecasting performance compared to those using alternative data sources.

Second, mobile keywords and economic indexes are better predictors than other variables used in the past. The models with setting “Time series + Mobile keywords + Economic indexes” are the most robust, and they can achieve the highest predictive accuracy and generalization ability.

Due to the nonlinearity and complexity of cruise tourist volumes, machine learning models can achieve higher predictive accuracy than the traditional linear models. The BPNN, RBF, LSSVR-CV, and LSSVR-GSA models achieved higher accuracy than did the ARIMA models.

Finally, compared with LSSVR-CV, LSSVR-GSA can achieve improved forecasting performance, indicating the effectiveness of GSA for the optimization of hyper-parameters.

Although the proposed models have achieved excellent forecasting performance, there are of course, some limitations because only Chinese cruise tourism market was investigated. Cruise tourism demand in new markets, cruise routes and cruise lines is worth investigating in the future. Besides SQD and economic indexes, other types of data sources such as web-based text and social media data may be also helpful. Moreover, new models, for example, deep learning models, can be developed for forecasting cruise tourism demand.

Author contributions

Gang Xie organized the research, conducted empirical study and wrote the paper. Yatong Qian collected the data and provided materials and insights on cruise tourism demand forecasting. Shouyang Wang supervised the findings of this work.

Declaration of competing interest

None.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 71771207, 71601021), and the National Center for Mathematics and Interdisciplinary Sciences, CAS.

References

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.

Carić, H., & Mackelworth, P. (2014). Cruise tourism environmental impacts - the perspective from the Adriatic Sea. *Ocean & Coastal Management, 102*, 350–363.

Castillo-Manzano, J. I., & López-Valpuesta, L. (2018). What does cruise passengers' satisfaction depend on? Does size really matter? *International Journal of Hospitality Management, 75*, 116–118.

Castillo-Manzano, J. I., Lopez-Valpuesta, L., & Alanís, F. J. (2015). Tourism managers' view of the economic impact of cruise traffic: The case of southern Spain. *Current Issues in Tourism, 18*(7), 701–705.

Chang, K.-L., Chen, C.-M., & Meyer, T. J. (2013). A comparison study of travel expenditure and consumption choices between first-time and repeat visitors. *Tourism Management, 35*, 275–277.

Chang, Y.-T., Lee, S., & Park, H. (2017). Efficiency analysis of major cruise lines. *Tourism Management, 58*, 78–88.

Chatziantoniou, I., Degiannakis, S., Eeckels, B., & Filis, G. (2016). Forecasting tourist arrivals using origin country macroeconomics. *Applied Economics, 48*(27), 2571–2585.

Chen, C.-A. (2016). How can Taiwan create a niche in Asia's cruise tourism industry? *Tourism Management, 55*, 173–183.

Chen, J. M., Neuts, B., Nijkamp, P., & Liu, J. (2016). Demand determinants of cruise tourists in competitive markets: Motivation, preference and intention. *Tourism Economics, 22*(2), 227–253.

Chen, J. M., Petrick, J. F., Papathanassis, A., & Li, X. (2019). A meta-analysis of the direct economic impacts of cruise tourism on port communities. *Tourism Management Perspectives, 31*, 209–218.

Chiappa, G. D., Lorenzo-Romero, C., & Gallarza, M. (2018). Host community perceptions of cruise tourism in a homeport: A cluster analysis. *Journal of Destination Marketing & Management, 7*, 170–181.

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

Cruise Lines International Association. (2018). *Cruise industry outlook*. Clia. Retrieved on April 26, 2018 from the World Wide Web http://cruising.org/docs/default-source/research/clia-2018-state-of-the-industry.pdf?sfvrsn=2.

Cuhadar, M., Cogurcu, I., & Kukrer, C. (2014). Modelling and forecasting cruise tourism demand to Izmir by different artificial neural network architectures. *International Journal of Business and Social Research, 4*(3), 12–28.

Dai, T., Hein, C., & Zhang, T. (2019). Understanding how Amsterdam City tourism marketing addresses cruise tourists' motivations regarding culture. *Tourism Management Perspectives, 29*, 157–165.

Dawson, J., Johnston, M. E., & Stewart, E. J. (2014). Governance of Arctic expedition cruise ships in a time of rapid environmental and economic change. *Ocean & Coastal Management, 89*, 88–99.

Derrac, J., García, S., Molina, D., & Herrera, F. (2011). A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation, 1*(1), 3–18.

Diedrich, A. (2010). Cruise ship tourism in Belize: The implications of developing cruise ship tourism in an ecotourism destination. *Ocean & Coastal Management, 53*, 234–244.

Dwyer, L., & Forsyth, P. (1998). Economic significance of cruise tourism. *Annals of Tourism Research, 25*(2), 393–415.

Gabe, T. M., Lynch, C. P., & McConnon, J. C., Jr. (2006). Likelihood of cruise ship passenger return to a visited port: The case of Bar Harbor, Maine. *Journal of Travel Research, 44*(3), 281–287.

Gibson, P., & Parkman, R. (2019). *Cruise operations management: Hospitality perspectives* (3rd ed.). New York: Routledge.

Gounopoulos, D., Petmezas, D., & Santamaria, D. (2012). Forecasting tourist arrivals in Greece and the impact of macroeconomic shocks from the countries of tourists' origin. *Annals of Tourism Research, 39*(2), 641–666.

Henthorne, T. L. (2000). An analysis of expenditures by cruise ship passengers in Jamaica. *Journal of Travel Research, 38*, 246–250.

Huang, X., Zhang, L., & Ding, Y. (2017). The Baidu Index: Uses in predicting tourism flows - a case study of the Forbidden City. *Tourism Management, 58*, 301–306.

Hung, K. (2018). Understanding the cruising experience of Chinese travelers through photo-interviewing technique and hierarchical experience model. *Tourism Management, 69*, 88–96.

Hung, K., Wang, S., Guillet, B. D., & Liu, Z. (2019). An overview of cruise tourism research through comparison of cruise studies published in English and Chinese. *International Journal of Hospitality Management, 77*, 207–216.

Klein, R. A. (2011). Responsible cruise tourism: Issues of cruise tourism and sustainability. *Journal of Hospitality and Tourism Management, 18*, 107–116.

Kollwitz, H., & Papathanassis, A. (2011). Evaluating cruise demand forecasting practices: A Delphi approach. In P. Gibson, A. Papathanassis, & P. Milde (Eds.), *Cruise sector challenges* (pp. 39–55). Gabler Verlag.

Larsen, S., & Wolff, K. (2016). Exploring assumptions about cruise tourists' visits to ports. *Tourism Management Perspectives, 17*, 44–49.

Li, S., Chen, T., Wang, L., & Ming, C. (2018). Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu index. *Tourism Management, 68*, 116–126.

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

Lv, S.-X., Peng, L., & Wang, L. (2018). Stacked autoencoder with echo-state regression for tourism demand forecasting using search query data. *Applied Soft Computing Journal, 73*, 119–133.

MacNeill, T., & Wozniak, D. (2018). The economic, social, and environmental impacts of cruise tourism. *Tourism Management, 66*, 387–404.

Mak, J. (2008). Taxing cruise tourism: Alaska's head tax on cruise ship passengers. *Tourism Economics, 14*(3), 599–614.

Müller, U. (2013). Discussion of "nowcasting US GDP: The role of ISM business surveys". *International Journal of Forecasting, 29*(4), 659–663.

Niavis, S., & Tsiotas, D. (2018). Decomposing the price of the cruise product into tourism and transport attributes: Evidence from the Mediterranean market. *Tourism Management, 67*, 98–110.

Pan, B., & Yang, Y. (2017). Forecasting destination weekly hotel occupancy with big data. *Journal of Travel Research, 56*(7), 957–970.

Paoli, C., Vassallo, P., Dapueto, G., Fanciulli, G., Massa, F., Venturini, S., & Povero, P. (2017). The economic revenues and the emergy costs of cruise tourism. *Journal of Cleaner Production, 166*, 1462–1478.

Park, S., Lee, J., & Song, W. (2017). Short-term forecasting of Japanese tourist inflow to South Korea using Google trends data. *Journal of Travel & Tourism Marketing, 34*(3), 357–368.

Pavlić, I. (2013). Cruise tourism demand forecasting - the case of Dubrovnik. *Tourism and Hospitality Management, 19*(1), 125–142.

Pavlic, I., Svilokos, T., & Tolic, M. S. (2015). Tourism, real effective exchange rate and economic growth: Empirical evidence for Croatia. *International Journal of Tourism Research, 17*(3), 282–291.

Perea-Medina, B., Rosa-Jiménez, C., & Andrade, M. J. (2019). Potential of public transport in regionalisation of main cruise destinations in Mediterranean. *Tourism Management, 74*, 382–391.

Rashedi, E., Nezamabadi-pour, H., & Saryazdi, S. (2009). Gsa: A gravitational search algorithm. *Information Sciences, 179*(13), 2232–2248.

Raun, J., Ahas, R., & Tiru, M. (2016). Measuring tourism destinations using mobile tracking data. *Tourism Management, 57*, 202–212.

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

Sanz-Blas, S., Buzova, D., & Carvajal-Trujillo, E. (2019). Familiarity and visit characteristics as determinants of tourists' experience at a cruise destination. *Tourism Management Perspectives, 30*, 1–10.

Seidl, A., Guiliano, F., & Pratt, L. (2007). Cruising for colones: Cruise tourism economics in Costa Rica. *Tourism Economics, 13*(1), 67–85.

Sun, X., Feng, X., & Gauri, D. K. (2014). The cruise industry in China: Efforts, progress and challenges. *International Journal of Hospitality Management, 42*, 71–84.

Sun, X., Gauri, D. K., & Webster, S. (2011). Forecasting for cruise line revenue management. *Journal of Revenue and Pricing Management, 10*(4), 306–324.

Sun, X., Kwortnik, R., & Gauri, D. K. (2018). Exploring behavioral differences between new and repeat cruisers to a cruise brand. *International Journal of Hospitality Management, 71*, 132–140.

Sun, S., Wei, Y., Tsui, K.-L., & Wang, S. (2019). Forecasting tourist arrivals with machine learning and internet search index. *Tourism Management, 70*, 1–10.

Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters, 9*(3), 293–300.

Tsamboulas, D., Moraiti, P., & Koulopoulou, G. (2013). How to forecast cruise ship arrivals for a new port-of-call destination. *Transportation Research Record: Journal of the Transportation Research Board, 2330*(1), 24–30.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Vogelvang, B. (2005). *Econometrics: Theory and applications with eviews*. Pearson Education.

Wang, Y., Jung, K.-A., Yeo, G.-T., & Chou, C.-C. (2014). Selecting a cruise port of call location using the fuzzy-AHP method: A case study in east Asia. *Tourism Management, 42*, 262–270.

Wondirad, A. (2019). Retracing the past, comprehending the present and contemplating the future of cruise tourism through a meta-analysis of journal publications. *Marine Policy, 108*, 103618.

Wu, H.-C., Cheng, C.-C., & Ai, C.-H. (2018). A study of experiential quality, experiential value, trust, corporate reputation, experiential satisfaction and behavioral intentions for cruise tourists: The case of Hong Kong. *Tourism Management, 66*, 200–220.

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

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

Gang Xie is an Associate Professor at Academy of Mathematics and Systems Science, Chinese Academy of Sciences. Dr. Xie received his PhD in Management Science and Engineering from Huazhong University of Science and Technology, Wuhan, Hubei Province, China, in 2006. His research interests are focused in economic forecasting, business intelligence and tourism management.

Yatong Qian, is a master student at the Institute of System Science, Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences, China. Her main research interests include big data techniques, economic forecasting and tourism management.

Shouyang Wang is a professor at Academy of Mathematics and Systems Science, CAS. He received a PhD degree in Operations Research from the Institute of Systems Science, Chinese Academy of Sciences in 1986. He is currently the academician at the International Academy of Systems and Cybernetics and the Third World Academy of Sciences. He has won such awards as Jr. Walter Scott Award (2014), the Green Group Awards (2008–2012), the International Society of Multiple Criteria Decision Making Chairmanship Award, and the Fudan Management Excellence Award. His current research interests include financial optimization, financial engineering and forecasting sciences.