Scientific Research Publishing

Engineering, 2020, 12, 194-215
https://www.scirp.org/journal/eng
ISSN Online: 1947-394X
ISSN Print: 1947-3931

Tourism Traffic Demand Prediction Using Google Trends Based on EEMD-DBN

Yi Xiao[1*], Xueting Tian[1], Ming Xiao[2]

[1]School of Information Management, Central China Normal University, Wuhan, China
[2]Network Center, Central China Normal University, Wuhan, China
Email: *yxiao@mail.ccnu.edu.cn

How to cite this paper: Xiao, Y., Tian, X.T. and Xiao, M. (2020) Tourism Traffic Demand Prediction Using Google Trends Based on EEMD-DBN. *Engineering*, **12**, 194-215.
https://doi.org/10.4236/eng.2020.123016

Received: January 29, 2020
Accepted: March 27, 2020
Published: March 30, 2020

Abstract

Predicting tourism traffic demand accurately plays an important role in making effective policies for tourist administration. It helps to distribute the resources reasonably and avoid the tourism congestions. This paper considered the noise interference and proposed a hybrid model, combining ensemble empirical mode decomposition (EEMD), deep belief network (DBN) and Google trends, for tourism traffic demand prediction. This model firstly applied dislocation weighted synthesis method to combine Google trends into a search composite index, and then it denoised the series with EEMD. EEMD extracted the high frequency noise from the original series. The low frequency series of search composite index would be used to forecast the low frequency tourism traffic series. Taking the inbound tourism in Shanghai as an example, this paper trained the model and predicted the next 12 months tourism arrivals. The conclusion demonstrated that the forecast error of EEMD-DBN model is lower remarkably than the baselines of ARIMA, GM(1,1), FTS, SVM, CES and DBN model. This revealed that nosing processing is necessary and EEMD-DBN forecast model can improve the prediction accuracy.

Keywords

Tourism Traffic Demand Forecasting, Deep Learning, Google Trends, Composite Search Index, Ensemble Empirical Mode Decomposition (EEMD), Deep Belief Network (DBN)

1. Introduction

According to data released by China Tourism Research Institute, the growth rate of inbound tourist volumes in China is relatively slow. That is to say, for a long

time in the past, the development of China's inbound tourism has been basically stagnant, which is inconsistent with the hot situation of the domestic tourism market and outbound tourism. The tourism demand forecast can provide timely basis for relevant departments to formulate effective tourism policies [1]. Modern information technology has brought great convenience to people's life, work and study. People usually turn to search engines when making travel strategies, and people's travel plans often use keyword search information as a reference [2].

In recent years, the research results of using the network search data to establish the tourism demand forecasting model are quite fruitful. The traditional econometric model and the machine learning method will be limited by historical data when forecasting the tourism demand, compared with the network search data. Search has instantaneity and subjectivity and more accurately the needs of tourists can be reflected. In fact, as early as 2009, the predictive power of web search data has been confirmed, for example, the application of Google Trends in all walks of life is considered effective [3].

Tourism is a related industry and has been greatly affected by emergencies; it has been difficult to solve the impact of emergencies on the tourism industry. The forecasting model cannot be adjusted in real time according to the changes in tourism dynamics, for example, the congestion of tourists at the entrance of Jiu Zhaigou Scenic Spot in Sichuan, China, on October 2, 2013 and Shanghai Bund stampede, on December 31, 2014 etc. These incidents have led that the number of tourists in local tourist attractions is difficult to dredge, and the quality of service of scenic spots has declined. It has become a hot spot of concern to all sectors of society. These problems indicate that the spatial allocation of tourism resources is crucial to the healthy development of tourism [1]. It is possible to balance tourism volumes of different tourist attractions due to the instant and efficient tourism demand forecasting model. In this way, the tourism industry will be more orderly and standardized, creating a pleasant atmosphere for China's inbound tourism.

In the field of tourism demand forecasting, research methods will vary depending on the conditions and objects of the forecast. Inbound tourists are more purposeful. Due to long distance of travel and the relatively long stay time compared to domestic tourism, the possibility of planning ahead is even greater. It is more prevalent to rely on online search to develop a travel schedule. However, the prediction model established by simply using the search data is not robust, and the combination of artificial intelligence and search data can greatly improve the accuracy of prediction [4] [5]. This paper uses EEMD to decompose the historical tourist volumes sequence of Shanghai inbound tourism and Google keyword search data respectively to eliminate the adverse effects of noise interference on the prediction results. Finally, the DBN with better convergence effect is used to predict the tourist volumes with the synthetic search index, which ensures the real-time validity of the prediction model.

2. Literature Review

2.1. Tourism Demand Forecast

In the forecast of tourism demand, there have been very rich research results in the past ten years [6]. Whether different methods or combinations have better predictive effects have always been the main direction of scholars' exploration [7]. The methods currently used in this field can be roughly divided into traditional time series models, artificial intelligence prediction methods [8] [9] and hybrid methods [10]. Among the studies of time series models, the most used ones are ARIMA models [11] [12], exponential smoothing models [13], and linear regression [14] and so on. Among them, ARMA has diverse prediction performance under different conditions, that is, it can be adjusted according to different research conditions to achieve better prediction results. ARMA has more possibilities [15]. The exponential smoothing model has also gradually evolved from primary exponential smoothing to quadratic exponential smoothing to cubic exponential smoothing to obtain more accurate predictions. In the study of tourism demand forecasting in several major source countries in Australia, we can see a comparison of several exponential smoothing models [12]. In fact, it can be observed in the research of many scholars that no one method has an absolute advantage. In general, we think that the combined model is more accurate than the single model [4] [5].

The application of artificial intelligence methods in tourism demand forecasting has begun to rise in the past 30 years. A back propagation neural network model can be applied to tourism demand forecasting [16]. As one of the international tourist cities, there have many methods for the prediction of Hong Kong's demand for inbound and outbound tourism, such as rough set theory [17]. Grey models are also widely used in tourism demand forecasting, including research on air passenger traffic [18] [19]. Genetic algorithms [20] have been developed from artificial neural networks [21]. The two major source countries of the Balearic Islands, the UK and Germany, have corresponding visitors every month, and some scholars have used genetic algorithms to conduct special research on this, which shows that genetic algorithms are also feasible in tourism demand forecasting [22]. Support vector machines can better solve practical problems such as small samples, nonlinear, high-dimensional numbers and local minimum points, adding network search data will greatly improve the accuracy of the prediction model, which predicts the passenger flow of Barbados Island, and there is a good embodiment [23].

2.2. Network Search Forecast

Since the network search data are used to successfully predict the epidemic [24], it has begun to use the search data to predict the phenomenon in many fields such as economics and social sciences. For example, scholars used Google search data to predict unemployment, housing prices, stocks [25] [26] [27], etc. Web search data has indeed contributed its valuable role in research in various fields,

especially in today's rapid development of information technology, when people are increasingly relying on online query tools we also hope to use Google search data to further study future consumer behavior [28].

The application of network search data is more and more extensive, providing a good enlightenment for the research of the tourism industry [28] [29]. However, due to the different language and cultural background of each region or country, the search intensity of multi-language source market sets and different leading search engine platforms will also affect the results of tourism demand forecasting [30]. A nonlinear auto-regressive method is combined with keyword search data to predict Malaysia's passenger volumes shows good predictive performance [28]. Since then, Google search data has also been used in the tourism demand forecasting study in the Caribbean region [31]. It indicates the effectiveness of web search data applications in forecasting tourism demand.

In China, there are numerous users of Baidu search engines [32]. Most of the scholars' research is dependent upon the Baidu search index. However, in the world, Google search engine dominates. This article takes Shanghai's inbound tourism demand forecast as an example, targeting the world's tourist groups, so it uses the data of the Google search engine.

3. Methodology Formulation

3.1. Principle of EEMD (Ensemble Empirical Mode Decomposition)

EEMD is an improved algorithm of EMD (Experimental Mode Decomposition) [33], which effectively solves the problem that EMD relies on local number of extreme data information. EMD is generally used for the decomposition of the original sequence from the data itself, because EMD is decomposed according to its own characteristics, and no other prior conditions are needed, so it is more used in noise processing and prediction. However, when the signal is not stable or contains anomalous events, EMD cannot show its superiority [34]. When the signal is disturbed by an abnormal event (such as pulse interference); mode mixing phenomenon occurs. In order to compensate for the shortcomings of EMD in modal decomposition, EEMD can be effectively solved to solve the model aliasing phenomenon. EEMD can make the decomposition scale more uniform, suppress the influence of abnormal events on the signal, and make the prediction more accurate.

The basic methods of EEMD are as follows:

Step 1: Calculate the sequence (set to $P_{(t)}$) local number of extreme data point using EMD, the maximal value constitutes the upper envelope $m_{(t)}$, and the minimum value constitutes the envelope $n_{(t)}$, and the mean $z_{(t)}$ of the upper and lower envelopes at any point is zero.

Step 2: Subtract the mean of the upper and lower envelopes with the sequence to get $R_{(t)}$.

$$R_{(t)} = \frac{P_{(t)} - \left(m_{(t)} + n_{(t)}\right)}{2} \tag{1}$$

Verify that $R_{(t)}$ satisfies the IMF. If not, repeat steps 1 and 2 until $R_{(t)}$ satisfies the IMF condition, and treat $R_{(t)}$ as an IMF separated from $P_{(t)}$ one by one. In the above process, the finite number of IMF_i components and the sum of the remainders $u_{(t)}$ and $Y_{(t)}$ are decomposed one by one from high frequency to low frequency by multiple screenings.

$$Y_{(t)} = \sum_{i=1}^{N} IMF_i + u_{(t)} \tag{2}$$

Step 3: Add random white noise to the sequence $P_{(t)}$, and equalize the abnormal events, so that the abnormal event mode is mixed into the random white noise mode during the EMD decomposition process, and then normalized, and the random white noise is applied by applying the EMD pair. The subsequent signal is decomposed to obtain an IMF_i component.

Step 4: Get IMF_i integration after decomposition (adding a new random normal distribution white noise sequence)

$$P_{(t)} = \sum_{i,j=1}^{n} IMF_{i_{(t)},j_{(t)}}, i = 1,\cdots,N; j = 1,\cdots,n \tag{3}$$

Step 5: Preset a threshold k. If the integrated value in the fourth step is less than k, it is removed as noise. If the integrated value in the fourth step is greater than k, the IMF_i is reset, and Q is an entropy function.

$$K_i\left[IMF_{i_{(t)}}, P_{(t)}\right] = Q\left[IMF_i\right] - Q\left[IMF_{i_{(t)}}, P_{(t)}\right] \tag{4}$$

3.2. Compositions of Google Search Keyword Variables

The era of big data brings new opportunities for the establishment of tourism demand forecasting models. The dependence of users on search engines can provide important data for tourism demand forecasting. In this paper generalized dynamic factor model (GDMF) [35], which can process high-dimensional data, is used to combine keyword variables. The unique advantage of GDMF is that variable data can be updated in real time, and known variables can be interpreted by common parts of unknown variables, aggregated into travel-related indices [36].

Forni (2004) proposed the idea of using VAR to represent the model of GDFM [37]. The traditional factor model is composed of the sum of s common factor k_t and special factor j_t. On this basis, GDFM gives common factors. Partially multiplied by the load matrix of m^*n, denoted as α, then the observed variable can be expressed as:

$$X - \alpha k_t + j_t \tag{5}$$

The matrix transformation of k_t can be expressed as:

$$k_t = pk_{t-1} + q\delta_t \tag{6}$$

where $\delta_t = (\delta_{1t}, \delta_{2t}, \delta_{3t}, \cdots, \delta_{nt})$ is a s-dimensional common component.

Tourism demand has many uncertainties and is able to be influenced by policies and media indices. The network search can reflect the tourists' decision-making behavior motives, but due to the influence of some emergencies, the search volume of one or several keywords in a certain period will be extremely high or very low, and these data are abnormal. So we create a standard scale for the search data when synthesizing the keywords. When the data show a maximum value beyond the standard scale, the method of taking the mean value is used to process the abnormal data. Using EEMD to decompose tourist volumes sequence n IMF components, the same approach is applicable to the Google keyword search index sequence.

3.3. DBN (Deep Belief Network) Prediction Model

Hinton proposed the Deep belief network [38], and the initial parameters of the model are obtained through unsupervised training methods. Compared with the traditional neural network model, DBN does not need a large number of supervised signals, and is not easy to fall into local minimum, which can greatly improve the convergence efficiency. The DBN network model is stacked by multiple restricted Boltzmann machine (RBM) layers. The DBN model with the best effect is obtained through repeated training on multiple layers RBM. The most basic RBM consists of a hidden layer and a visible layer.

Hinton proposed the idea of training each layer of RBM separately [39], which is, extracting input data features in the hidden layer of the first layer RBM, using it to train the second layer RBM, repeating this process until DBN all RBMs stacked in the model are trained. It is assumed that the established DBN model has a total of c hidden layers and d visible layers, and the values of the i^{th} visible layer and the j^{th} hidden layer are: $<v_i, h_j>$, and the bias of the two is: a_i, b_j, then the parameter θ of the RBM is:

$$\theta = (\omega_{ij}, a_i, b_j) \tag{7}$$

Then, the energy of the RBM is expressed as:

$$E(v, h : \theta) = -\sum_{i=1}^{c} a_i v_i - \sum_{j=1}^{d} b_j h_j - \sum_{i=1}^{c}\sum_{j=1}^{d} a_i \omega_{ij} h_j \tag{8}$$

ω_{ij} is the symmetric connection weight between the visible layer v_i and the hidden layer h_j. The probability of the binary state of the hidden layer v_i being set to 1 and the probability of the binary state of the visible layer h_j being set to 1 are calculated [39]. In general, the algorithm of contrast divergence is used to represent the log likelihood gradient of RBM, and the weight and offset parameters are updated by calculation, for detailed algorithm, see Hinton's work [40].

In this paper, we propose an EEMD-DBN prediction model whose structure is shown in **Figure 1**.

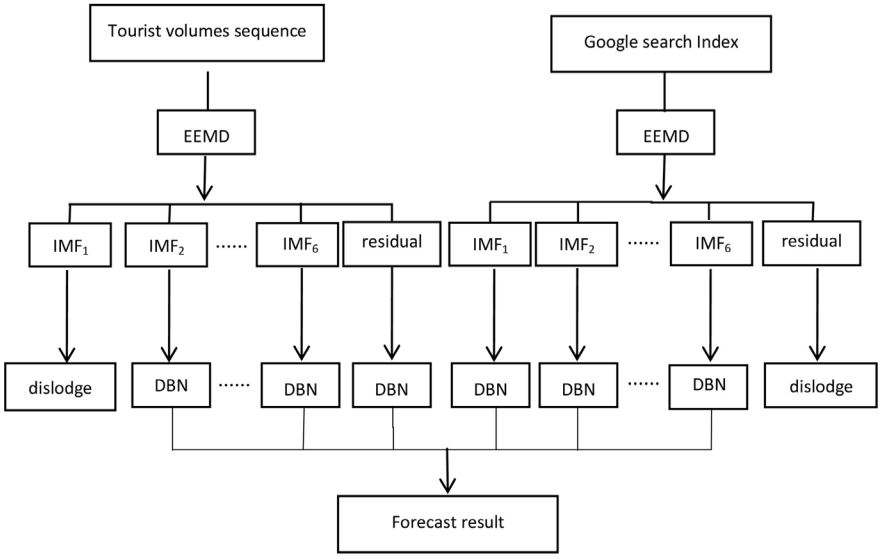


Figure 1. The structure of EEMD-DBN prediction model.

4. Experimental Study

4.1. Data Set

4.1.1. Tourist Volumes Data

Shanghai is an important pillar of China's national economic development; and it attracts many tourists from home and abroad to visit here with rich tourism resources. The regional differences in China's inbound tourism development have gradually narrowed, but it is undeniable that Shanghai still has a huge impetus to the progress of China's inbound tourism. According to the 2017 National Statistical Report on National Economic and Social Development, as of the end of 2017, there were 99 A-level scenic spots in Shanghai, including 3 scenic spots in 5A, 50 scenic spots in 4A, and 46 scenic spots in 3A. It has become one of the cities with the most inbound tourist volumes in China.

The data selected in this paper are the number of monthly inbound tourists from 2004 to 2018 in Shanghai, and divide the data into two parts: the training set and the predictive test set. In order to ensure the validity of the prediction model, 2004-2017 was selected as the sample data for the training and establishment of the prediction model, and the 2018 tourist data were used as the prediction set. Baidu search engine is more widely used in China, so it is more suitable for China's domestic tourism demand forecast. Globally, Google's users are more extensive, accounting for about 66.7% of the world's total [29]. So this article uses the search data of Google search engine. **Figure 2** is a time-series map of monthly inbound tourist traffic in Shanghai from 2004 to 2017. As can be seen from the figure, the time series of tourist traffic shows an overall slow upward trend and cyclical fluctuations in a certain period of time. And in the third quarter and fourth quarter of 2010, there is a peak feature, which is inseparable from the occurrence of the Shanghai World Expo in the second quarter of 2010. The monthly data used in this paper has a total of 168 data points, which better

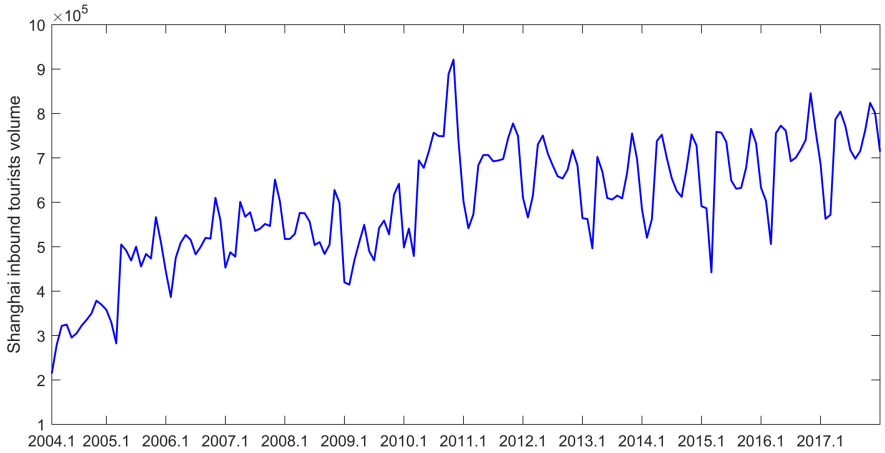


Figure 2. Tourist volumes sequence in Shanghai.

reflects the long-term trend of Shanghai tourist volumes. Compared with the annual data, the time series of monthly data changes more significantly, which can provide a more detailed basis for tourism destination tourism decisions.

4.1.2. Keyword Search Data

Compared with short-distance travel, inbound tourists will stay at the destination for a relatively long time, so people have a tendency to use search engines to develop relevant travel plans in advance. This includes travel route planning, travel hotel ordering, and travel destination information inquiry. This series of behaviors is basically done by means of search engines, so this paper uses keyword search index to continue to improve the forecasting accuracy of tourism demand forecasting model. A key step in the synthesis of online search index is the selection of search keywords. In terms of keyword selection, there is currently no mature program and theoretical system. This article uses a more common method of directly selecting keywords.

Firstly, the 50 common keywords related to tourism are selected for search volume search. According to the Google search volume ranking, the first 16 keywords are retained, and the Pearson correlation test is performed on these 16 keywords, and the correlation with the tourism is the largest 5 keywords as research samples as **Figure 3**. There is some abnormal value in the network search volume of the 5 keywords we finally determined. Directly eliminating the abnormal data will lead to the lack of research samples. Therefore, this paper performs the mean processing on the abnormal data values to ensure the integrity of the experimental data.

4.2. Data Analysis

It can be seen from the time series of inbound tourist traffic in Shanghai that the fluctuation of passenger tourism volumes data is more obvious, and there is a large amount of data to be processed. The peak of the tourism volumes sequence is from September to October of 2010. As we all know, from May to June 2010,

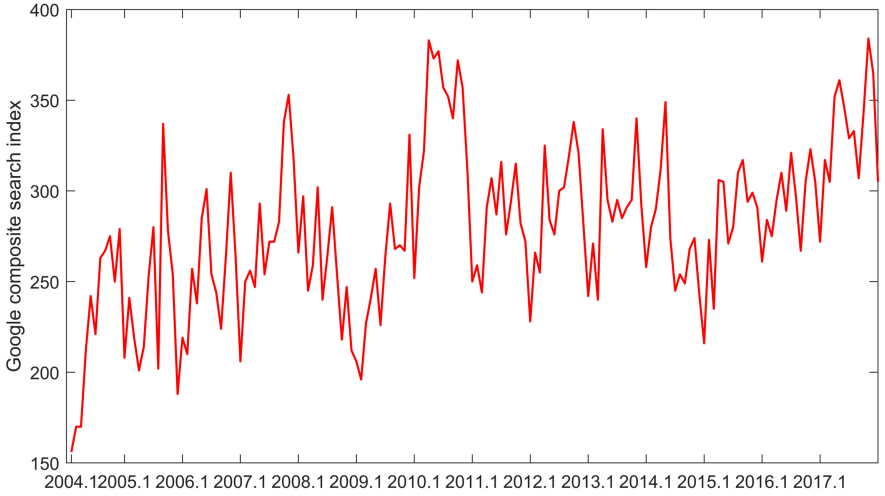


Figure 3. Keyword search data sequence.

Shanghai hosted the World Expo, which played a positive role in Shanghai's tourism development to a certain extent, which led to the rapid growth of Shanghai's inbound tourism volumes in a short period of time. In order to reduce the impact of abnormal events on the forecasting model, we performed a simple noise reduction process on the tourism volumes time series to eliminate the interference of the data peaks and valleys. As shown in **Figure 4**, the time series after smoothing the linear trend factor of the volume data is relatively flat.

Determine keywords according to the eight characteristics of tourism activities, eat, live, travel, travel, purchase, entertainment, determine the basic keywords, such as: travel to Shanghai, Shanghai attractions, Shanghai hotels, Shanghai flights, etc., then enter the basic keywords for Google search volume Inquire. When entering the basic keywords on Google, Google will intelligently recommend keywords related to it, and then record relevant keywords recommended by Google, and sort out 50 keywords related to inbound tourism in Shanghai, the 16 keywords with the largest search volume are retained, and the Pearson correlation coefficient between the tourism volumes and the search keyword is calculated. From the analysis results, we can see that there is a negative correlation between some keywords and the tourism volumes, and these keywords are eliminated. Based on the tourist psychology motivation angle and the Pearson correlation coefficient calculation results, as shown in **Table 1**, finally, choose the keyword with Pearson correlation coefficient above 0.3 as experimental data.

Using the standard scale we set as the limit, the abnormal data values in the search volume of the 5 keywords with the most relevant correlation are processed. When the search volume of keywords is lower or higher than the standard scale, many researchers will adopt the method of directly eliminating such abnormal data, but this will also cause excessive cleaning damage to the data. According to the seasonal characteristics of the tourism industry itself, this paper averages the other annual data of the month in which the data of the

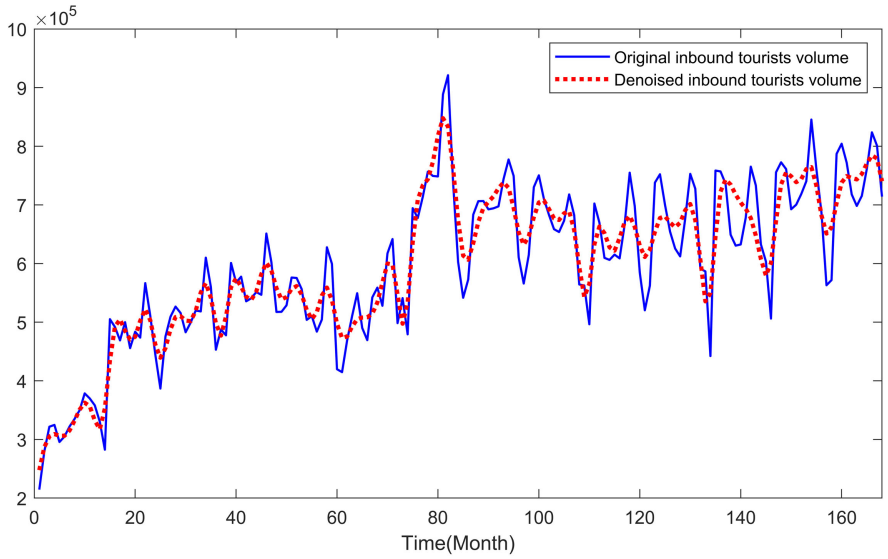


Figure 4. Tourism volumes sequence noise reduction.

Table 1. Search query correlation coefficient of Google keyword.

Keyword	Correlation Coefficient
Flight to Shanghai	0.597
Shanghai weather	0.128
Shanghai traffic	−0.420
Weather in Shanghai China	0.348
Shanghai tourist attractions	−0.183
China time Shanghai	0.656
Shanghai flight	0.479
Shanghai scene	−0.292
Shanghai hotel booking	0.024
Bund Shanghai	0.393
Shanghai cuisine	−0.312
Shanghai food	0.178
Shanghai airport	−0.272
Shanghai airport arrivals	0.045
Shanghai restaurant	0.124
Shanghai visa	0.053

abnormal data is taken to ensure the integrity of the data information.

The basic method of GDFM index composite is to use the weighted idea to sum up the common components of the variables. First, use the variance contribution rate to determine the number of factors, and then calculate the common components of the multidimensional stationary search data s_{it}, and finally add the search index.

$$s_{it} = \sum_{i=1}^{n} b_{it}\left(L^{k}\right) g_{nt} \tag{9}$$

n is the determined number of factors, L is the lag operator, and k is the number of lag operators. According to the Forni that when $n = 4$ and $k = 5$, the model works best [35].

After processing the abnormal data of the keyword search volume, the effect of the abnormal value on the authenticity of the search index is reduced. However the tourism is a relatively relevant industry, emergencies or other urban activities not related to tourism have a certain impact on tourism search, which has led to a significant increase or decrease in the overall network search of tourist destinations over time. For example, the SARS incidents in 2003 which affected the web search of tourist destinations. Abnormal data processing is for individual data, as shown in **Figure 5**. We still need to optimize the predictive power of the search index by simple noise reduction processing.

It can be seen that the composite search index does reflect the trend of tourist volumes and shows a certain lead time, as shown in **Figure 6**, which is also consistent with people's behavioral motives. At the peak of the tour, people seem more willing to use the search data in advance to help them make travel decisions.

4.3. Evaluating Indicator

In order to investigate the prediction ability of the established model from different angles, this paper selects the mean absolute percentage error (MAPE), the mean square error (MSE), the mean absolute error (MAE) and the fitting coefficient R^2 as the evaluation indicators of the model from different angles. Measure the predictive affect of the model. Among them, $\overline{x}_i$ represents the model simulation output value, that is, the predicted tourist volume; x_i represents the actual number of visitors, and n is the number of test data.

The fitting coefficient R^2 represents the degree of fitting of the predicted value curve to the actual value curve. The value of R^2 can measure the fitting ability of

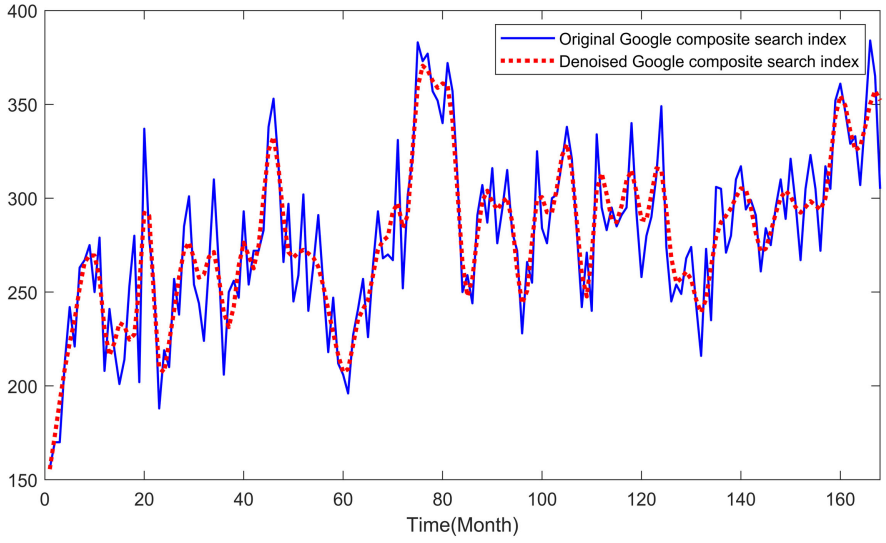


Figure 5. Search of key word index noise reduction.

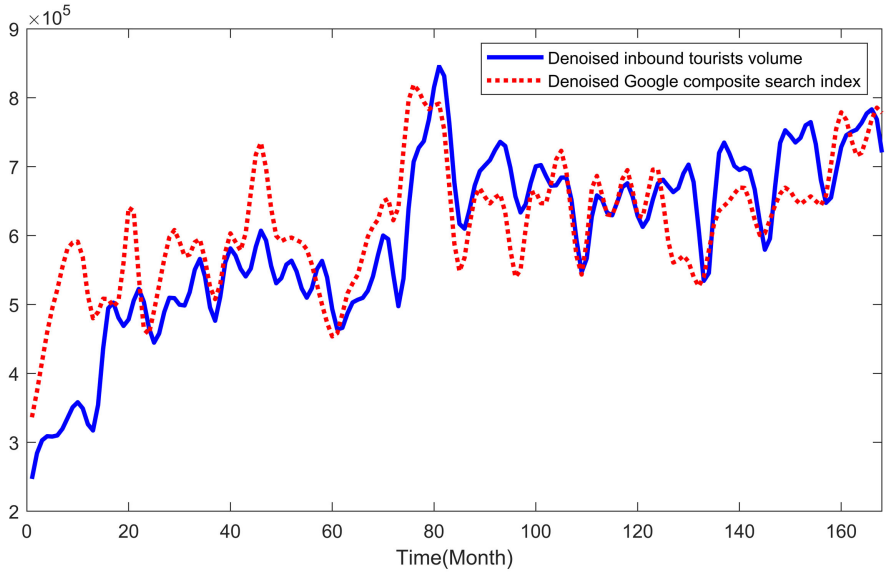


Figure 6. Fitting of tourism volumes sequence and composite search index.

the model, $R^2 \in [0,1]$, and the closer the value of R^2 is to 1, the model is explained the stronger the fitting ability.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2} \tag{10}$$

The MAE measures the accuracy of the prediction by calculating the difference between the predicted value and the true value data. The smaller value of MAE we have the higher prediction accuracy.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}_i| \tag{11}$$

MAPE is an evaluation criterion used to explain the relative error of the prediction model, which can well evaluate the prediction ability of the model.

$$\text{MAPE} = \frac{\sum_{i=1}^{n}|x_i - \bar{x}_i|/x_i}{n}100\% \tag{12}$$

RMSE represents the square root of the ratio of the sum of the predicted value to the true value and the ratio of the experimental number, used to estimate the degree of deviation between the predicted value and the true value.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_i)^2} \tag{13}$$

4.4. EEMD Noise Reduction

In the forecast research of tourism demand in recent years, there are many combined forecasting models, especially the combination of network search index and time series model. However, these studies lack the intensity of noise

processing. The modal decomposition of the tourist volumes time series and the keyword search index allows the Shanghai inbound tourist volumes and the keyword search index to obtain a uniformly distributed decomposition scale, so that smooth the interference of abnormal events.

The core idea of EEMD is to add Gaussian white noise to the signal and perform ensemble averaging. The two important parameters of EEMD are the ratio k of white noise to the standard deviation of the original signal amplitude, and the average number of times M, however there is no specific calculation method for the values of k and M. Combined with the experience of the researchers and the experiments in this paper, and for the data characteristics, we take k = 0.2 and M = 100 as the benchmark experimental values, and then adjust them continuously in the experiment to get the best EEMD model with the best decomposition effect.

When performing the EEMD test of the first iteration, the Gaussian white noise sequence $f_{(t)}$ was added to the Shanghai inbound tourist volumes time series $P_{(t)}$ and the keyword search index $I_{(t)}$, and n trials were performed. After that, the nth pending tourism volumes time series and the network search index sequence are obtained.

$$P_{n(t)} = p_{(t)} + lf_{n(t)} \tag{14}$$

$$I_{n(t)} = I_{(t)} + lf_{n(t)} \tag{15}$$

In further experiments, the EEMD parameter values continuously adjusted. The decomposed tourist volumes time series and the search index sequence contain 6 IMF components and one residual. The amplitude and fluctuation of the IMF component are different. It can also be seen from the figure that the amplitude and fluctuating frequency of the first IMF are always the largest and the wavelength is relatively short. The first IMF component obtained after decomposition is removed as noise, and the remaining IMF_2, IMF_3, IMF_4, IMF_5, IMF_6 and smooth trend residuals are summed as an experimental sequence of tourist volumes and keyword search respectively.

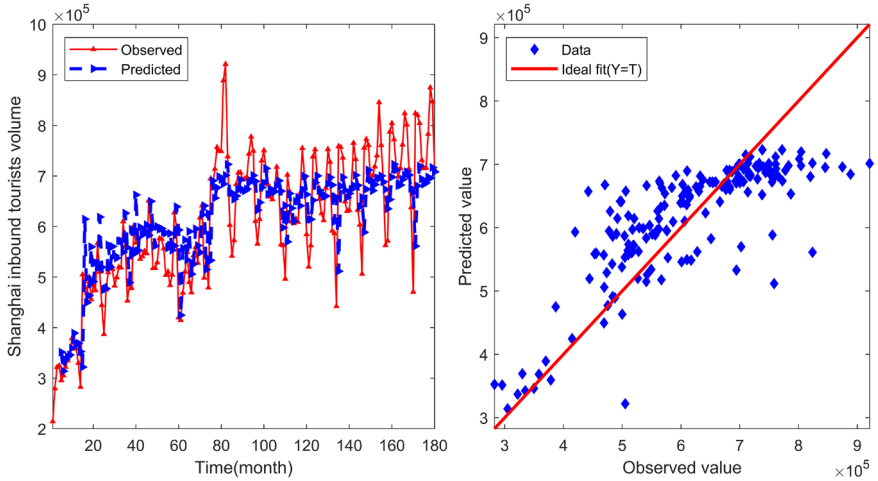
4.5. DBN Forecast

We are required to determine the number of hidden layer nodes and the number of hidden layers in the DBN. There is not any exact rule about the number of nodes in the DBN input layer and hidden layer. Basing on the experience of researchers, this paper sets the number of layers of DBN to N = 3 and the number of neurons is set at intervals of 5 for each hidden layer. The number of nodes in the layer is taken as an integer in [100, 1000], and then each additional layer of hidden layer is added to determine the optimal value of the number of neurons in the second hidden layer, and the experiment is repeated to achieve the highest accuracy. It component RBM of the DBN needs to optimize the feature extraction ability through training. Therefore, the DBN also needs a weight; the weight can determine the influence factor of the maximum probability of the training

sample. We define the learning rate of the DBN model to 0.1, the number of iterations. Set to200, by repeating the training, it is determined that the final DBN structure is a 3-layer RBM composition, and the number of hidden neurons in the first layer and the second layer is 20, 15, respectively.

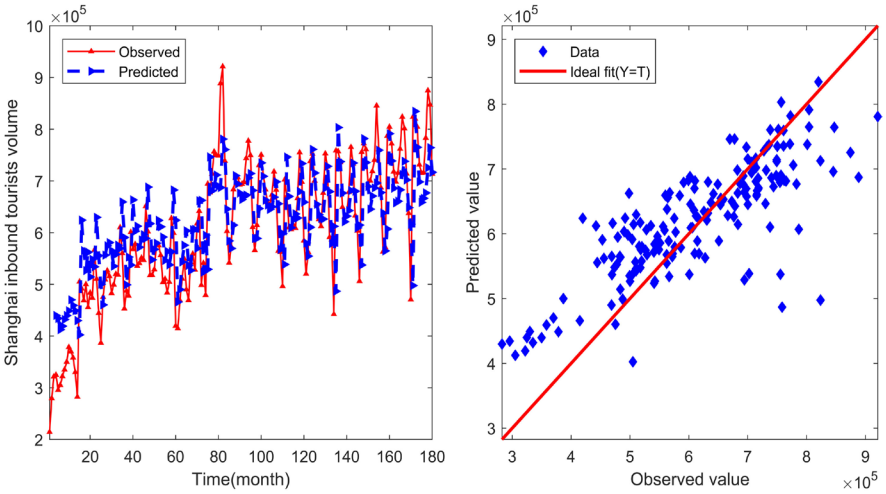
In this paper, the established DBN model is used for prediction and compared with support vector machine, ARIMA, GM(1,1), fuzzy time series and cubic exponential smoothing model. The group (a) of Figures 6-11 shows the fit between the predicted value and the actual value, and the group (b) of Figures 7-12 shows the degree of dispersion between the predicted value and the observed value.

It can be seen from the comparison of the (a) group images of the six different



(a) (b)

Figure 7. Forecasts of the monthly Shanghai inbound tourist volume using GDFM-ARIMA: (a) the predicted series; and (b) the scatter of the predicted series.



(a) (b)

Figure 8. Forecasts of the monthly Shanghai inbound tourist volume using GDFM-GM(1,1): (a) the predicted series; and (b) the scatter of the predicted series.

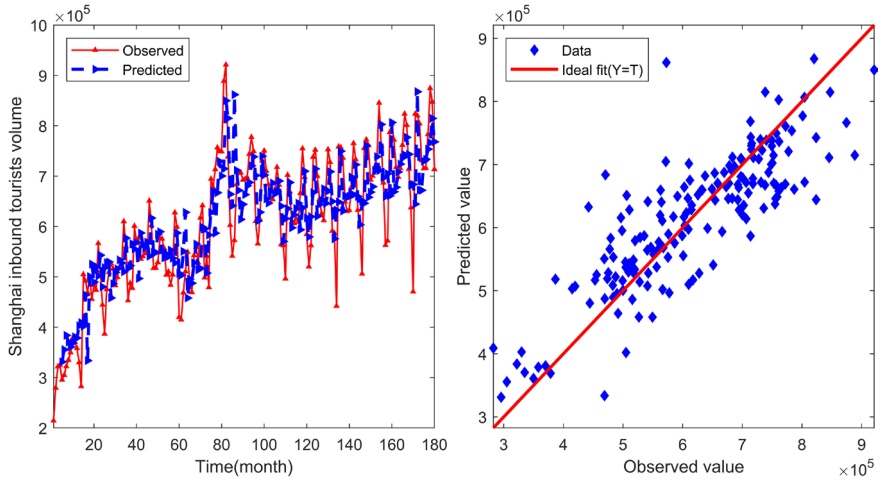


Figure 9. Forecasts of the monthly Shanghai inbounds tourist volume using GDFM-FTS: (a) the predicted series; and (b) the scatter of the predicted series.

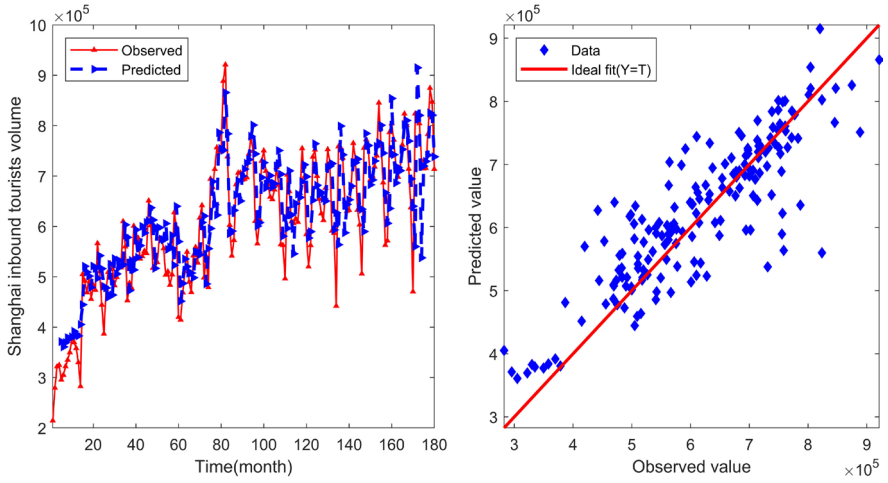


Figure 10. Forecasts of the monthly Shanghai inbound tourist volume using GDFM-SVM: (a) the predicted series; and (b) the scatter of the predicted series.

prediction models that the predicted values of the ARIMA model show obvious convergence characteristics, and ARIMA has greater limitations in dealing with non-stationary time series. The prediction ability of the GM(1,1) model usually shows a large volatility, which is affected by the smoothness of the tourist volumes data series, the prediction effect of the GM(1,1) model seems to be less than ideal. FTS usually optimizes the uncertainty of data and solves fuzzy problems. When forecasting separately it often does not really work out well. SVM is a widely used forecasting model in tourism demand forecasting in recent years. However, the kernel function selection of SVM model is a very difficult problem, and it is computationally complex and often sensitive to data loss. The cubic exponential smoothing is based on an exponential smoothing model and a quadratic exponential smoothing model. It is commonly used in China's domestic

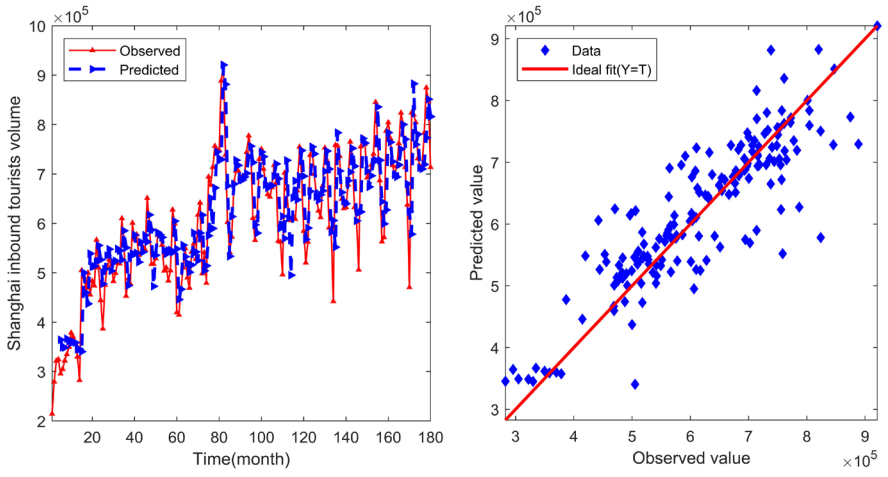


Figure 11. Forecasts of the monthly Shanghai inbound tourist volume using GDFM-CES: (a) the predicted series; and (b) the scatter of the predicted series.

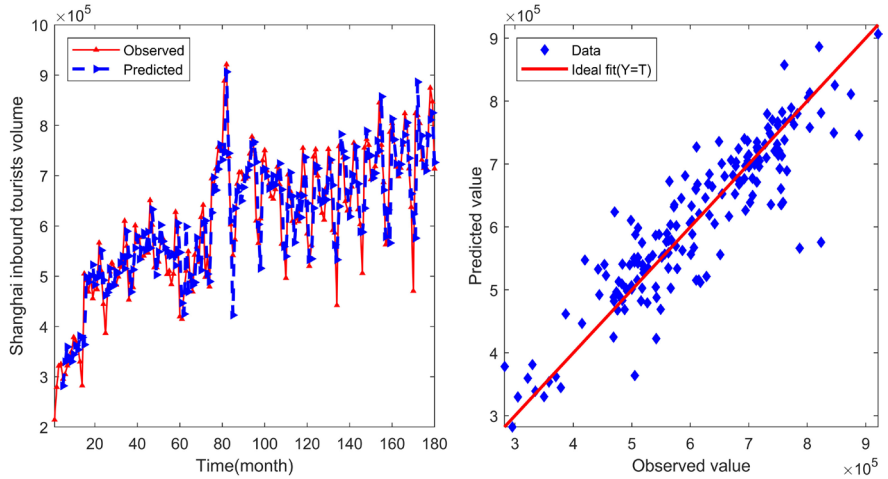


Figure 12. Forecasts of the monthly Shanghai inbound tourist volume using GDFM-DBN: (a) the predicted series; and (b) the scatter of the predicted series.

tourism demand forecast because it has good predictive ability for seasonal time series. From the figure, the fitting effect of the cubic exponential smoothing and DBN is relatively trustworthy.

The group image of (b) is a good representation of the degree of dispersion between the ideal prediction and the actual predictions of the six different models. The discrete trend of the DBN model is the most stable and the least discrete, that is, the predicted value of the DBN model is closely surrounding the actual value. Combining the final results of the (a) group image with the (b) group image, we can conclude that the DBN has better predictive power than other models.

4.6. Comparison of Different Forecasting Methods

The comparison experiment was set up in two groups. The first group predicted

the tourist volumes without adding the composite search index, and the second group joined the Google search index to predict the tourist volumes. The results of both groups were evaluated by MAE, MAPE, RMSE, and R^2 as shown in Table 2. The MAE, MAPE, RMSE, and R^2 values predicted by the first group of ARIMA models were 10.16828, 0.14032, 1.27639, and 0.62679, respectively; the MAE, MAPE, RMSE, and R^2 values predicted by GM(1,1) were 8.47406, 0.1159, 1.21326, and 0.63701, respectively; The MAE, MAPE, RMSE, and R^2 values predicted by fuzzy time series are 7.77446, 0.11376, 0.99984, and 0.70957, respectively; the MAE, MAPE, RMSE, and R^2 values predicted by SVM are 7.95296, 0.11417, 1.13246, and 0.74384, respectively; cubic exponential smoothing model prediction The MAE, MAPE, RMSE, and R^2 values were 7.31596, 0.10397, 0.99589, and 0.76065, respectively; the MAE, MAPE, RMSE, and R^2 values predicted by DBN were 5.8507, 0.08417, 0.90995, and 0.80651, respectively. Obviously, the evaluation indicators of the DBN model are better than other models, especially the values of MAE and MAPE are nearly half smaller than ARIMA, and the R^2 value is also the closest to 1.

Looking at the results of the second set of experiments, after joining the Google keyword search index, we used six different models to predict the tourist volumes as shown in Table 3. The MAE, MAPE, RMSE, and R^2 values predicted by ARIMA after adding the search index are 10.40777, 0.15558, 1.30863, and 0.64662, respectively; the MAE, MAPE, RMSE, and R^2 values predicted by GM(1,1) are 7.03853, 0.09941, 0.85096, and 0.64336 respectively; MAE, MAPE, RMSE, and R^2 values predicted by fuzzy time series are 8.02626, 0.10963, 1.05467, and 0.72485, respectively; the MAE, MAPE, RMSE, and R^2 values predicted by SVM are 6.28613, 0.09048, 1.00487, and 0.779, respectively; The MAE, MAPE, RMSE, and R^2 values predicted by the smoothing model were 5.99157, 0.084653, 0.91868, and 0.78252, respectively; the MAE, MAPE, RMSE, and R^2 values predicted by DBN were 5.04885, 0.071167, 0.65961, and 0.82518, respectively. Compared with the results of the first set of experiments, it can be found that after adding the Google keyword search index, the MAE, MAPE, and RMSE values of GM(1,1), SVM, CES, and DBN are smaller than those when the search index is not added, also improved in R^2. Obviously, the DBN model still maintains the most predictive performance, and the indicators are ahead of other forecasting models. This shows that the DBN model does show a good forecasting

Table 2. Forecasting performance evaluation of six models.

	MAE (10^4)	MAPR	RMSE (10^5)	R^2
ARIMA	10.16828	0.14032	1.27639	0.62679
GM(1,1)	8.47406	0.1159	1.21326	0.63701
FTS	7.77496	0.11376	0.99984	0.70957
SVM	7.95296	0.11417	1.13246	0.74384
CES	7.31596	0.10397	0.99589	0.76065
DBN	5.8507	0.08417	0.90995	0.80651

Table 3. Forecasting performance evaluation of six models with search index.

	MAE (10^4)	MAPR	RMSE (10^5)	R^2
Index-ARIMA	10.40777	0.15558	1.30863	0.64662
Index-GM(1,1)	7.03853	0.09941	0.85096	0.64336
Index-FTS	8.02626	0.10963	1.05467	0.72485
Index-SVM	6.28613	0.09048	1.00487	0.779
Index-CES	5.99157	0.084653	0.91868	0.78252
Index-DBN	5.04885	0.071167	0.65961	0.82518

effect in terms of tourist volumes forecasting, and the Google search index can indeed optimize the forecast of tourist volumes.

4.7. Stability Test and Granger Causality Test

The granger causality test is to prove the effectiveness of Google keyword search data on tourism demand forecasting model. Before conducting the Granger causality test, we must first ensure that the tourist volumes sequence and the Google search index sequence are stationary. According to the unit root test results, the Google search index is stationary, at the 1% significance level (**Table 4**). The tourist volumes sequence shows the non-stationary, state during the same period, so the differential processing has to be done. Under the second-order differential level, the tourist volumes sequence is stationary (**Table 5**). Further, Granger causality test between variables is shown in **Table 6**. According to the results of the Granger causality test, Google search data is the cause of tourists' travel behavior. Then it proves that the Google search index has predictive ability for inbound tourism in Shanghai. Therefore, it is feasible to use the search data to predict the tourism volumes. The tourism volumes sequence is recorded as Y_t, and the keyword search index is recorded as I_t.

Table 4. Augmented Dickey-Fuller unit root test on Google search index.

		t-statistic	Prob
Augmented Dickey-Fuller test statistic		−4.447879	0.0002
Test critical values:	1% level	−3.467205	
	5% level	−2.877636	
	10% level	−2.575430	

Table 5. Augmented Dickey-Fuller unit root test on tourism volumes (2nd difference).

		t-statistic	Prob
Augmented Dickey-Fuller test statistic		−12.42683	<0.0001
Test critical values:	1% level	−3.469933	
	5% level	−2.878829	
	10% level	−2.576067	

Table 6. Standard granger causality tests.

null hypothesis	F-statistic	Prob
Y_t does not Granger cause I_t	7.39027	0.0001
I_t does not Granger cause Y_t	3.06659	0.0295

5. Conclusions

Network technology is constantly upgrading, and has achieved good popularity, becoming an indispensable part of people's daily lives. With the advent of the 5G era, web search may penetrate deeper into our daily lives, especially in terms of travel. The EEMD decomposition method adopted in this paper overcomes the large noise defects in the traditional index composite, making the keyword search index play the most important role in tourism demand forecasting. However, we have to admit that the selection of keywords is something we need to explore further, although it is now possible to use high-speed computers to extract keywords, and such methods will greatly improve the accuracy of keywords, but this technology extremely high hardware requirements and therefore no universality. Accurately finding keywords that best reflect the motivation of tourist guests will further optimize the tourism demand forecasting model. Five comparative forecasting models selected in this paper are widely used in tourism demand forecasting. The maturity of artificial intelligence technology will also bring new opportunities and challenges to tourism demand forecasting.

Tourism is a comprehensive industry that will not only be affected by force major such as weather, natural disasters, but also subjective factors such as politics, economy, culture and even religion. In the traditional forecast of tourism demand, the research using quantitative methods accounts for the majority, which also neglects the tourism behavior caused by people's subjective consciousness to some extent, and the keyword search data, reflects the subjective behavior of people. Although the keyword information has been applied to some extent, it is not deep enough. Therefore, accurate extraction of keywords and qualitative analysis is another challenging problem that we need to work hard to solve.

Acknowledgements

This research is supported by the Fundamental Research Funds for the Central Universities under Grant No. CCNU19ZN024 and the Humanities and Social Sciences Layout Foundation of the Ministry of Education of China under Grant No. 20YJA740047.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Clerides, S. and Adamou, A. (2010) Prospects and Limits of Tourism-Led Growth:

The International Evidence. *Review of Economic Analysis*, **2**, 287-303. https://doi.org/10.2139/ssrn.1495747

[2] Fesenmaier, D., Xiang, Z., Pan, B. and Law, R. (2011) A Framework of Search Engine Use for Travel Planning. *Journal of Travel Research*, **50**, 587-601. https://doi.org/10.1177/0047287510385466

[3] Choi, H. and Varian, H. (2009) Predicting the Present with Google Trends. Technical Report, Google Inc. https://doi.org/10.2139/ssrn.1659302

[4] Chan, C., Witt, S., Lee, Y. and Song, H. (2010) Tourism Forecast Combination Using the CUSUM Technique. *Tourism Management*, **31**, 891-897. https://doi.org/10.1016/j.tourman.2009.10.004

[5] Song, H. and Li, G. (2008) Tourism Demand Modeling and Forecasting: A Review of Recent Research. *Tourism Management*, **29**, 203-220. https://doi.org/10.1016/j.tourman.2007.07.016

[6] Shen, S., Li, G. and Song, H. (2011) Combination Forecast of International Tourism Demand. *Annals of Tourism Research*, **38**, 72-89. https://doi.org/10.1016/j.annals.2010.05.003

[7] Peng, B., Song, H. and Crouch, G. (2014) A Meta-Analysis of International Tourism Demand Forecasting and Implications for Practice. *Tourism Management*, **45**, 181-193. https://doi.org/10.1016/j.tourman.2014.04.005

[8] Xiao, Y., Liu, J., Xiao, J., Hu, Y., Bu, H. and Wang, S.Y. (2015) Application of Multiscale Analysis-Based Intelligent Ensemble Modeling on Airport Traffic Forecast. *Transportation Letters*: *The International Journal of Transportation Research*, **7**, 73-79. https://doi.org/10.1179/1942787514Y.0000000035

[9] Xiao, Y., Liu, Y., Liu, J.J., Xiao, J. and Hu, Y. (2016) Oscillations Extracting for the Management of Passenger Flows in the Airport of Hong Kong. *Transportmetrica A*: *Transport Science*, **12**, 65-79. https://doi.org/10.1080/23249935.2015.1099576

[10] Xiao, Y., Liu, J.J., Hu, Y., Wang, Y.F., Lai, K.K. and Wang, S.Y. (2014) A Neuro-Fuzzy Combination Model Based on Singular Spectrum Analysis for Air Transport Demand Forecasting. *Journal of Air Transport Management*, **39**, 1-11. https://doi.org/10.1016/j.jairtraman.2014.03.004

[11] Lim, C. and McAleer, M. (2001) Monthly Seasonal Variations: Asian Tourism to Australia. *Annals of Tourism Research*, **28**, 68-82. https://doi.org/10.1016/S0160-7383(00)00002-5

[12] Lim, C. and McAleer, M. (2002) Time Series Forecasts of International Travel Demand for Australia. *Tourism Management*, **23**, 389-396. https://doi.org/10.1016/S0261-5177(01)00098-X

[13] Martin, A. and Witt, F. (1989) Forecasting Tourism Demand: A Comparison of the Accuracy of Several Quantitative Methods. *International Journal of Forecasting*, **5**, 1-13. https://doi.org/10.1016/0169-2070(89)90059-9

[14] Gounopoulos, D., Petmezas, D. and 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**, 641-666. https://doi.org/10.1016/j.annals.2011.09.001

[15] Chu, F.L. (2009) Forecasting Tourism Demand with ARMA-Based Methods. *Tourism Management*, **30**, 740-751. https://doi.org/10.1016/j.tourman.2008.10.016

[16] Chen, J., Chen, Z.X., Xing, L. and Fu, X.D. (2005) Forecasting of Yunnan's International Tourism Demand Based on BP Neural Network. *Journal of Kunming Teachers College*, **27**, 89-91.

[17] Goh, C. and Law, R. (2003) Incorporating the Rough Sets Theory into Travel Demand Analysis. *Tourism Management*, **24**, 511-517. https://doi.org/10.1016/S0261-5177(03)00009-8

[18] Samagaio, A. and Wolters, M. (2010) Comparative Analysis of Government Forecasts for the Lisbon Airport. *Journal of Air Transport Management*, **16**, 213-217. https://doi.org/10.1016/j.jairtraman.2009.09.002

[19] Nguyen, L.T., Shu, H.M., Huang, F.Y. and Hsu, M.B. (2013) Accurate Forecasting Models in Predicting the Inbound Tourism Demand in Vietnam. *Journal of Statistics and Management Systems*, **16**, 25-43. https://doi.org/10.1080/09720510.2013.777570

[20] Chen, K.Y. and Wang, C.H. (2007) Support Vector Regression with Genetic Algorithms in Forecasting Tourism Demand. *Tourism Management*, **28**, 215-226. https://doi.org/10.1016/j.tourman.2005.12.018

[21] Burger, C., Dohnal, M., Kathrada, M. and Law, R. (2001) A Practitioners Guide to Time Series Method for Tourism Demand Forecasting: A Case Study of Durban, South Africa. *Tourism Management*, **22**, 403-409. https://doi.org/10.1016/S0261-5177(00)00068-6

[22] Alvarez-Diaz, M., Mateu-Sbert, J. and Rosselló-Nadal, J. (2009) Forecasting Tourist Arrivals to Balearic Island Using Genetic Programming. *International Journal of Computational Economics and Econometrics*, **1**, 65-75. https://doi.org/10.1504/IJCEE.2009.029153

[23] Jackman, M. and Naitram, S. (2015) Nowcasting Tourist Arrivals in Barbados: Just Google It! *Tourism Economics*, **21**, 1309-1313. https://doi.org/10.5367/te.2014.0402

[24] Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. and Brilliant, L. (2009) Detecting Influenza Epidemics Using Search Engine Query Data. *Nature*, **457**, 1012-1014. https://doi.org/10.1038/nature07634

[25] Askitas, N. and Zimmermann, K.F. (2009) Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly*, **55**, 107-120. https://doi.org/10.3790/aeq.55.2.107

[26] Beracha, E. and Wintoki, M.B. (2013) Forecasting Residential Real Estate Price Changes from Online Search Activity. *Journal of Real Estate Research*, **35**, 283-312.

[27] Da, Z., Engelberg, J. and Gao, P. (2011) In Search of Attention. *Journal of Finance*, **66**, 1461-1499. https://doi.org/10.1111/j.1540-6261.2011.01679.x

[28] Bangwayo-Skeete, P.F. and Skeete, R.W. (2015) Can Google Data Improve the Forecasting Performance of Tourist Arrivals? Mixed-Data Sampling Approach. *Tourism Management*, **46**, 454-464. https://doi.org/10.1016/j.tourman.2014.07.014

[29] Yang, X., Pan, B., Evans, J.A. and Lv, B. (2015) Forecasting Chinese Tourist Volume with Search Engine Data. *Tourism Management*, **46**, 386-397. https://doi.org/10.1016/j.tourman.2014.07.019

[30] Theologos, D., Eleni, M. and Bing, P. (2018) Google Trends and Tourists' Arrivals: Emerging Biases and Proposed Corrections. *Tourism Management*, **66**, 108-120. https://doi.org/10.1016/j.tourman.2017.10.014

[31] Yassin, I.M., Zabidi, A., Salleh, M.K.M. and Khalid, N.E.A. (2013) Malaysian Tourism Interest Forecasting Using Nonlinear Auto-Regressive (NAR) Model. *Proceedings of IEEE* 3*rd International Conference on System Engineering and Technology*, Shah Alam, 19-20 August 2013, 32-36. https://doi.org/10.1109/ICSEngT.2013.6650138

[32] Kennedy, A.F. and Hauksson, K.M. (2012) Global Search Engine Marketing:

Fine-Tuning Your International Search Engine Results. Que Publishing, Indianapolis.

[33] Huang, N.E., Shen, Z., Long, S.R., *et al*. (1998) The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis. *Proceedings of the Royal Society A*: *Mathematical*, *Physical and Engineering Sciences*, **454**, 903-995. https://doi.org/10.1098/rspa.1998.0193

[34] Chen, C.F., Lai, M.C. and Yeh, C.C. (2012) Forecasting Tourism Demand Based on Empirical Mode Decomposition and Neural Network. *Knowledge-Based Systems*, **26**, 281-287. https://doi.org/10.1016/j.knosys.2011.09.002

[35] Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000) The Generalized Dynamic-Factor Model: Identification and Estimation. *Review of Economics and Statistics*, **82**, 540-554. https://doi.org/10.1162/003465300559037

[36] Amstad, M. and Potter, S. (2009) Real Time Underlying Inflation Gauges for Monetary Policymakers. FRB of New York Staff Report, 420. https://doi.org/10.2139/ssrn.1532280

[37] Forni, M., Giannone, D., Lippi, F. and Reichlin, L. (2004) Opening the Black Box Structural Factor Models versus Structural VARS. Discussion Papers of CEPR.

[38] Hinton, G.E., Osindero, S. and The, Y.W. (2006) A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, **18**, 1527-1554. https://doi.org/10.1162/neco.2006.18.7.1527

[39] Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. and Kingsbury, B. (2012) Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, **29**, 82-97. https://doi.org/10.1109/MSP.2012.2205597

[40] Hinton, G. (2002) Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, **14**, 1771-1800. https://doi.org/10.1162/089976602760128018