ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20191231071057
Original scientific paper

The Role of Internet Search Index for Tourist Volume Prediction Based on GDFM Model

Yiran LI, Mengyao XU, Xuan WEN, Daomeng GUO

Abstract: Tourist volume is increasing with the expansion of the scale of tourism, and improving the prediction of tourist volume is helpful for tourism managers to make decisions. Internet search index can be applied to predict the behavior of users, which is widely used in the study of tourist volume prediction and infectious disease prediction. However, the high dimension and correlation of Internet search index tends to reduce the accuracy of the models, which increases the average prediction error of common time-series models. The dynamic factor model (DFM) proposed in our study can be used to solve the problem. This study selects 23 variables and introduces the generalized dynamic factor model (GDFM) to predict tourist volume. The model cannot only reduce the dimensionality of high-dimensional Internet search index data, but also reflects the dynamic correlation between Internet search index data. The results show that the prediction accuracy is improved in our method, and the prediction accuracy of tourist volume is improved by over 10%, with an average error of only 4.3% when compared with the neural network (NN) model. Our study not only provides implications for decision-makers to predict tourist volume timely and accurately, but also helps companies understand tourist' behavior and make the best strategic decisions.

Keywords: big data analysis; generalized dynamic factor model; internet search index; tourist volume prediction

1 INTRODUCTION

Accurate prediction of tourist volume can provide a basis for tourism managers' decision-making, which is accurate and effective for making scientific decisions. Moreover, it can enable local tourism-related catering, hotel and other service industries to make scientific planning in advance, such as infrastructure, reception capacity and other aspects. These services could enhance the tourist experience. Hence, it is necessary to predict tourist volume timely and accurately.

The Internet search index can dynamically monitor the search scale of keywords and the change of public opinions. It can generate user portrait [1] and predict uses' demand [2] by mining the Internet search index. For example, tourists tend to search for local information about weather and traffic when making tourism planning. In addition, they will consult relevant information such as hotels, scenic spots and travel companies to help make decisions about the travel destination and itinerary planning. Therefore, it is possible to accurately predict the number of tourists who visit the destination by capturing and analyzing the search trends about concerns of the tourist destination.

Although some existing studies have applied the Internet search index to the prediction of the number of tourists, the results showed that the average error was relatively high [3-6]. In this study, a feasible variable selection method named generalized dynamic factor model (GDFM) was proposed based on existing studies. GDFM is widely used in economic and financial cycle analysis for it can extract a small amount of useful information from a high-dimensional data set, and these extracted common factors can be used for variable prediction, economic index construction and structural analysis [7]. This study combined multiple keywords related to tourism, such as catering, accommodation, travel, shopping and entertainment, and introduced the GDFM to generate the predictive model. This method cannot only process a lot of high-dimensional Internet search data, but also reflect the dynamic correlation among the Internet search trend data.

In order to verify the prediction power of the proposed method, we collected Internet search index data to empirically test the number of tourists in Beijing from January 2014 to December 2016. The empirical results indicated that our method was significantly better than the prediction results based on neural network and HoltWinters models.

The remainder of this paper is organized as follows: Section 2 reviews the literature about the prediction method for tourist volume and factors influencing the number of tourists. Section 3 introduces the methodology used in this study. Then, data acquisition and empirical analysis are provided in Section 4. Finally, Section 5 summarizes the results and highlights the future research direction.

2 LITERATURE REVIEW

2.1 Prediction Methods for Tourist Volume

Most studies used multiple regression analysis, autoregressive integrated moving average (ARIMA) model and neural networks to predict the number of tourists. UysalandEIRoubi [8] compared the usefulness of artificial neural networks and multiple regression in the prediction of visitor numbers, and they found the powerful prediction of artificial neural networks in the number of tourists, with the average prediction error rate of 3.23%. Unhapipat [9] used ARIMA $(0,0,0) \times (1,1,0)_{12}$ model to forecast the number of international tourists in Bumthang, Bhutan from 2012 to 2016, with 91% prediction accuracy. However, these studies rely too much on historical data. The reliable prediction results were dependent on the quantity of historical data. A seasonal autoregressive integrated moving average (SARIMA) model was established by Chang and Liao [10], and the seasonal model SARIMA $(1,1,1) \times (1,0,0)_{12}$ by considering the rising trend and seasonality of the sequence, and the mean absolute percentage error was 8.9%. In addition, there are gray system theory [11] and the synthetic index approach [12] to predict the number of tourists. Park et al. [13] used the index of Google search engine to make a short-term prediction of the number of Japanese visitors to South Korea, and believed that the prediction effect of Google augmented model was better than the ordinary time-series models. Hence, Internet search index data has been used to

predict the number of tourists, but the average error rate was high and the prediction accuracy was poor.

The existing studies always rely on historical data to predict the tourist volume, but historical data has a strong delay, and its prediction granularity is large. They ignore the important question that dynamic data can better reflect the characteristics of tourist industries. In addition, compared with the common time-series models, the artificial neural network has higher prediction accuracy, but it has high algorithm complexity and is strongly dependent on trends of the raw data. Hence, this research tried to use Internet search index to reflect the dynamic process of tourist volume. Additionally, gray system theory and neural network model have the best prediction results in the existing researches, therefore, the neural network model was also constructed in our study to compare the prediction results with GDFM model.

2.2 Factors Influencing the Tourist Volume

At present, there are few studies that predict the tourist volume by Internet search index. However, the prediction based on Internet search index in economic and social behavior has become a hot topic. Kholodilinet al. [14] pointed out that Internet search index can be used to predict consumption and unemployment rates. They found the prediction models based on Google search index were far more accurate than others. Ripberger [15] used Internet search index to measure public attention and produced great results. Ginsberg et al. [16] found that the search volume of several keywords related to influenza in Google had a strong correlation with the visit number of relevant patients. They built a surveillance model based on Google search index, which could predict the outbreak trend of influenza two weeks earlier than the traditional detection method. Hence, Internet search index records the search concerns and demands of the public, reflects the behavioral trends, and provides powerful dynamic data for tourist volume prediction.

Although there are few studies on the factors influencing the number of tourists, most researchers agreed that per capita disposable income [17-19] and per capita gross domestic product (GDP) [20-22] had a significant impact on the tourist volume. Eeckels et al. [23] used spectral analysis to examine the relationship between cyclical component of GDP and tourist volume. Their findings pointed out the importance of tourism industries and supported the tourism-led economic growth hypothesis. Yang et al. [24] applied multilevel models to investigate the factors that affect the domestic tourism demand of urban and rural visitors in China, and the results indicated that there was a co-integration relationship among the number of tourists, individual income and average income over the city. Ding [25] used R software to build a multiple linear regression model, and found that there is a significant positive correlation between GDP, per capita consumption of tourists and tourist volume.

In addition, other scholars confirmed that the tourism conditions, destination characteristics, transport characteristics, macro-economic conditions and unforeseeable circumstances would affect the number of tourists [26-28]. Kim et. al [29], using tourist spending as a regulating factor, believed that tourist destination image, tourist motivation and perceived quality were associated with tourist satisfaction and revisit intention. Combined with the existing studies, there is a correlation between per capita disposable income and per capita GDP. Hence, this study put Internet search index and per capita disposable income into the model, and then selected effective variables for data analysis from catering and accommodation, objective conditions and entertainment-related.

3 METHODOLOGY

3.1 Generalized Dynamic Factor Model

The factor model was proposed by British psychologist Charles E. Spearman [30] to define and measure intelligence. The purpose of factor analysis is to describe the correlation between variables using a small number of potential and unobservable factors. Suppose $X_t = (X_{1t}, X_{2t}, \ldots, X_{Nt})'$ is a set of data with relevance, where X_{it} represents the observation value of the variable i in group t, $i = 1, \ldots, N$, $t = 1, \ldots, T$. The factor model assumes that the correlations between variables are for the presence of some unobservable common factor F_t. Specifically, the factor model has the following form:

$$X_{it} = \lambda_i^{'} F_t + \varepsilon_{it} \tag{1}$$

Where F_t. is the common factor vector of $r \times 1$ dimensional, the elements will influence at least two variables. λ_i is the load coefficient of factors, and ε_{it} is a heterogeneous part of X_{it}.

The classic factor analysis is categorized as a static factor model because these models examine contemporaneous co-movements among the observations. However, the static factor model is mainly used to process cross-section data, not appropriate for time series data because changes in certain factors might lead or lag changes in the examined variables.

Forni et al. [31] proposed the generalized dynamic factor model, and they argued that "dynamic" and "approximate" are two important characteristics for a factor model to solve the time series data. Firstly, analyzing the time series data is a typical dynamic problem. For others, the model must allow the heterogeneity part to be a cross-sectional correlation. The orthogonality assumption of the heterogeneity part is unrealistic for most typical dynamic problems. Therefore, the generalized dynamic factor model is better suited for our study. It consists of two parts: a common component and a special component. A generalized dynamic factor model is represented as follows:

$$X_{it} = x_{it} + \varepsilon_{it}, i \in N, t \in \mathrm{Z} \tag{2}$$

$$x_{it} = b_{iq}(L) u_t, i \in N, t \in \mathrm{Z}, q \in N \tag{3}$$

Where $u_t = (u_{1t}, u_{2t}, \ldots, u_{qt})'$ is a q-dimensional white noise sequence, and it contains all possible variables that may affect x_t in the GDFM. L is the lag operator, $u_t = \Psi(L)\, u_{t-1} + \eta_t$.

The model satisfies the following two basic assumptions:

(1) u_{it} is orthogonal to each other and orthogonal to ξ_{it}.
(2) ξ_{it} is weakly correlated and some covariances are allowed. Where u_{it} is called the common factor, and ξ_{it} is called the special factor. Some researchers believe that for the given time t, the dimension of the variable is finite. Under this assumption, the model can be represented as follows:

$$x_{it} = \lambda_{i1}F_{1t} + \lambda_{i2}F_{2t} +, ..., \lambda_{ir}F_{rt} + \xi_{it} \tag{4}$$

$$F_t = (F_{1t}, ..., F_{rt})' = N(L)u_t \tag{5}$$

where F_t is the main component of u_t, and it contains all information of u_t on the time series and is orthogonal to F_t with each other, and λ_{it} is the eigenvalue vector of x_{it}.

3.2 Estimation Method

The traditional principal component analysis (PCA) achieved data dimensionality reduction by transforming the original related variables into several uncorrelated variables by a linear transformation. Stock and Watson [32] completed the proof process of PCA under a weaker assumption, which is further generalized to obtain generalized principal component analysis (GPCA). Then, the estimation method proposed by Forni et al. [33] effectively implements the generalized principal component estimation of the dynamic factor model. Some studies compared the estimation method of principal component, dynamic principal component and generalized principal component using Monte Carlo Simulation (MCS) and actual data prediction, and the results were different when the sample size was small. However, many researchers believe that the prediction results will be stable when using the GPCA method in a dynamic factor model.

The GPCA algorithm assumes that the sample set $X = \left\{x_j \in R^{\mathrm{k}}\right\}_{j=1}^{N}$ in k-dimensional space R^{k}, locating in nunknown linear subspace $\mathrm{S} = \left\{S_i \in R^{\mathrm{k}}\right\}_{i=1}^{n}$ respectively. S_i and X_i are what we need to figure out, where $X_i = \left\{x_j \in X : x_j \in S_i\right\}$. The algorithm can be represented in the following equations:
(1) Assuming the dimension of subspace S_i is $\mathrm{K} - k_i$, and S_i can be represented by k_i linear equations:

$$S_i = \left\{x \in R^{\mathrm{k}} : B_i^{\mathrm{T}} x = 0\right\} = \left\{x \in R^{\mathrm{k}} : \bigcap_{j=1}^{k_i}\left(b_{ij}^{\mathrm{T}} x = 0\right)\right\} \tag{6}$$

where $B_i = \left(b_{i1}, ..., b_{ik_i}\right) \in R^{K \times k_i}$ is a basis for the orthogonal complement of S_i. Each point x has to be a member of a subspace, and it can be shown in the following way:

$$\begin{aligned} &\bigcup_{i=1}^{n}\bigcap_{j=1}^{k_i}\left(b_{ij}^{\mathrm{T}} x = 0\right) \Leftrightarrow \bigcap_{\sigma}\bigcup_{i=1}^{n}\left(b_{\sigma(i)}^{\mathrm{T}} x = 0\right) \Leftrightarrow \\ &\Leftrightarrow \bigcap_{\sigma}\prod_{i=1}^{n}\left(b_{\sigma(i)}^{\mathrm{T}} x = 0\right) \Leftrightarrow \bigcap_{\sigma}^{\sigma} p_n\sigma(x) = 0 \end{aligned} \tag{7}$$

Where σ represents a way of choosing a combination of $b_{\sigma(i)}$.
(2) Assuming $x_j \in S_i$ and it has the following form:

$$Dp_n\sigma\left(x_j\right) = \sum_{i=1}^{n}\left(b_{\sigma(i)}\right)\prod_{l \neq i}\left(b_{\sigma(l)}^{\mathrm{T}} x_j\right) \tag{8}$$

When taking a sample of $x_{j(i)}$ for each of subspaces S_i, and taking the derivative of all the polynomials $\bigcap_{\sigma} p_n\sigma\left(X\right)$ at point $x_{j(i)}$, we can get all the perpendicular vectors b_{ij} for S_i. Then, $\widehat{X_l} = \left\{x \in X : d\left(x, S_i\right) = \min_j d\left(x, S_i\right)\right\}$ represents the classification result of sample points, where $d(x, St)$ represents the distance from point x to the subspace St.

4 DATA ACQUISITION AND RESULTS ANALYSIS

4.1 Experimental Object Selection

We chose the tourist volume in Beijing in this study, for it can reflect the national conditions of China. In this paper, the data of the tourist volume came from the official website of the Beijing Tourism Development Commission (lyw.beijing.gov.cn). The total number of tourists was subtracted to the number of inbound tourists in order to exclude the influence of the inbound tourist volume. In addition, we can only obtain the data of the tourist volume in each month, so in order to build the model, the monthly data was changed into daily data through sliding average processing. Additionally, the overall research framework is illustrated in Fig. 1.

4.2 Data Acquisition and Pre-Processing

According to the results of the *CNZZ* data center (www.cnzz.com), the usage rate of various search engines in the market was different in August 2014, such as Baidu was 56.33%, 360 search was 29.01% and new Sogou was 12.75%. This study collected the number of tourists in Beijing from 2014 to 2016, assuming that the utilization rate of each search engine remains unchanged during this period.

Table 1 Categories of keywords related to tourism

ID	Keywords category
1	Catering and accommodation (C)
2	Objective conditions (O)
3	Entertainment-related (E)

Firstly, we selected three categories of tourism in Beijing, namely, catering and accommodation, objective conditions, entertainment-related, as shown in Tab. 1. These three categories not only reflect the demand of travelers but also represent the relevant industries on the supply side. Secondly, a large number of relevant keywords were selected in each category. The coincidence degree and similarity of some keywords were relatively high, and each search engine provided the function of merge processing, hence we combined some keywords with a high correlation and calculated their search index. The keywords in each variable can be seen in Tab. 2. Thirdly, if the number of certain keywords was very low

(for example, hotel group buying), the search engine would not provide search trend data. Hence, the unusable keywords were eliminated in this process. Then, we obtained the Internet search index data from January 1, 2014 to December 31, 2016, including computer terminal and mobile terminal. Finally, the final search index data was weighted averaged according to the utilization ratio in Baidu index, 360 index and Sogou index respectively. The statistics of the data are provided in Tab. 3.

Table 2 Variables and keywords collected in this study

ID	Variable name	Keywords
Catering and accommodation (C)		
1	Snack	Beijing snacks + Beijing specialties
2	Snack_introduction	Beijing snack street + Beijing snack street guide
3	Restaurant	Hotels in Beijing + restaurants in Beijing
4	Hotel_book	Beijing guesthouse reservation
5	Hotel_book2	Beijing hotel reservation
6	Hotel_group	Beijing hotel group purchase
Objective conditions (O)		
7	Subway	Beijing subway query + Beijing subway fare
8	Bus	Beijing bus route + Beijing bus fare
9	Plane	Beijing air ticket + Beijing flight inquiry
10	Train	Beijing train ticket + Beijing train timetable
11	Weather	Beijing weather
12	Temperature	Beijing air temperature
13	Traffic	Beijing transportation
Entertainment-related (E)		
14	Scenery	Scenic spots in Beijing
15	Map	Beijing map
16	Strategy	Beijing travel strategy
17	Tour	Beijing tourist groups
18	Ticket	Beijing tickets
19	Shop	Shopping in Beijing
20	Bar	Bars in Beijing
21	Night_life	Night life in Beijing
22	Concert	Vocal concerts in Beijing
23	Jewelry	Jewelry fairs in Beijing

Table 3 Descriptive statistics

ID	Variable name	Mean	Std. Dev.	Min	Max
1	Snack	2703.503	318.71792	1965.17	3841.57
2	Snack_introduction	753.0819	217.06426	312.08	1862.42
3	Restaurant	925.8516	149.54593	582.43	1385.59
4	Hotel_group	215.3794	98.00173	57.75	458.2
5	Hotel_book	526.9016	152.10363	23	1057.54
6	Hotel_book2	214.2682	57.38744	68.43	450.79
7	Weather	221876.4186	102904.9746	41692.77	564287.84
8	Temperature	804.7586	249.49893	446.09	2311.86
9	Traffic	810.1934	158.83926	538.9	1475.88
10	Subway	855.3658	261.22782	539.37	2104.25
11	Bus	315.73	84.3878	166.37	619.82
12	Plane	260.9423	43.55852	152.97	359.58
13	Train	372.5969	57.7691	258.77	604.85
14	Scenery	1547.6979	368.0017	830.86	2652.93
15	Ticket	120.2619	29.54825	26.13	204.28
16	Tour	2850.612	1206.35458	1487.16	6356.07
17	Strategy	224.5244	89.89609	118.51	571.78
18	Map	13852.1437	2678.96662	6335.53	21698.36
19	Shop	264.9843	42.22601	189.01	478.13
20	Bar	320.0695	35.27428	232.37	388.52
21	Night_life	183.8677	15.20014	143.83	219.77
22	Concert	430.1981	96.65769	233.25	793.13
23	Jewelry	214.1334	59.19715	117.28	418.06

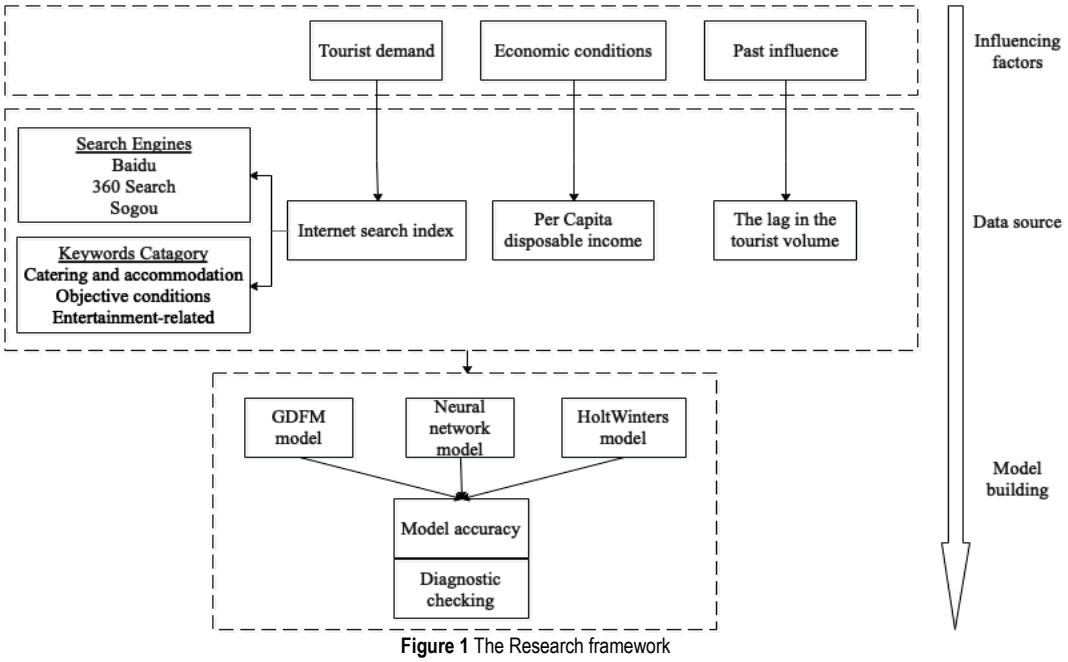

Figure 1 The Research framework

4.3 Experimental Results

4.3.1 Variable Selection

This study adopted the dynamic principal component analysis (GPCA) method proposed by Forni et al. [33] to extract variables. We extracted the required common factors from the multiple variables of catering and accommodation, objective conditions and entertainment-related.

The cumulative contribution rates of these three factors can be seen in Fig. 2 - Fig. 4.

It can be seen from Fig. 2 - Fig. 4 that the first four factors (C1-C4) of the catering and accommodation have explained 90% of the variables, the first five factors (O1 - O5) of the objective conditions represented nearly 90% of

the variables and the first five factors (E1-E5) of entertainment-related have explained over 80% of the variables. Therefore, we retained the first four factors of the catering and accommodation, the first five factors of the objective conditions and the first five factors of entertainment-related.

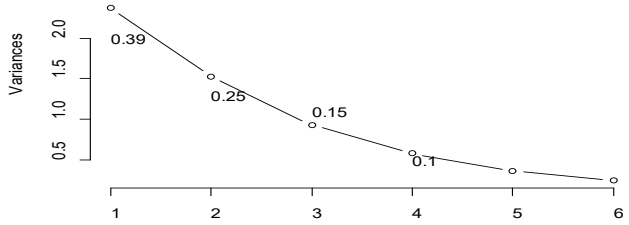

Figure 2 The cumulative contribution rates of catering and accommodation

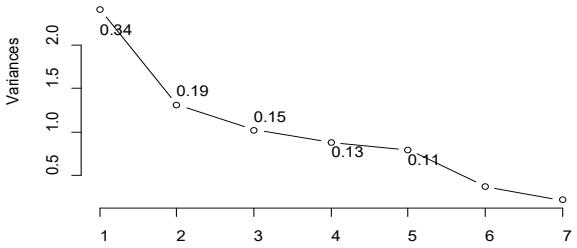

Figure 3 The cumulative contribution rates of objective conditions

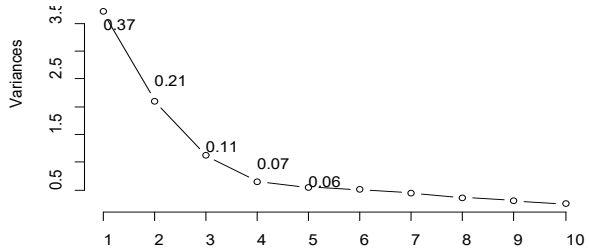

Figure 4 The cumulative contribution rates entertainment-related

4.3.2 Model Estimation

The per capita disposable income and the lag in the tourist volume were put into the model as special factors. In order to ensure the rationality of the prediction model, we assume that we only know the number of tourists a week ago, and the lag term was set as a 7-order lag. We made a stepwise regression to the model, and the estimated parameters of the model and *P*-values are given in Tab. 4. The results show that the adjusted *R*-squared is 0.95, indicating the fitting effect of the model is good.

Table 4 The model for the number of tourists to Beijing

Variables	Estimate	*P*-value
C1	0.57	0.002
C2	0.75	0.001
C3	−0.44	< 0.001
C4	−0.64	0.03
O1	−0.50	0.001
O2	−0.52	< 0.001
O3	−1.64	< 0.001
O4	−0.44	0.009
O5	0.65	< 0.001
E1	0.87	0.008
E2	−0.31	0.137
E3	−0.41	0.079
E4	1.03	< 0.001
E5	1.74	< 0.001
Per capita disposable income	0.98	0
P-value of *F*-statistic < 0.001		
Adjust *R*-squared = 0.95		

4.3.3 Diagnostic Checking

In order to ensure the authenticity and accuracy of the model, the model residuals analysis is performed. The result is shown in Fig. 5, indicating that the residual of the model basically conforms to the normal distribution. The abscissa represents the residual and the ordinate represents the dependent variable in Fig. 6. The results show that there is no significant correlation between the residual and the dependent variable, indicating that the independent variable has been extracted well and meets the independence test.

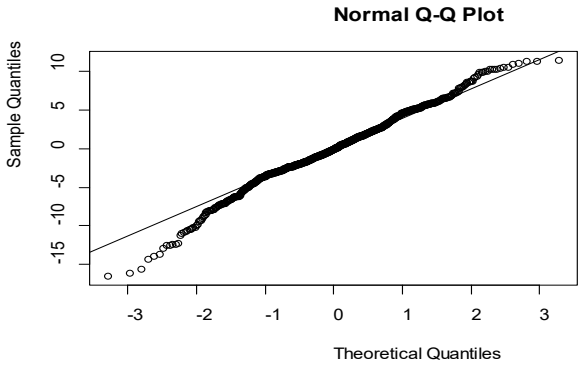

Figure 5 Residuals test

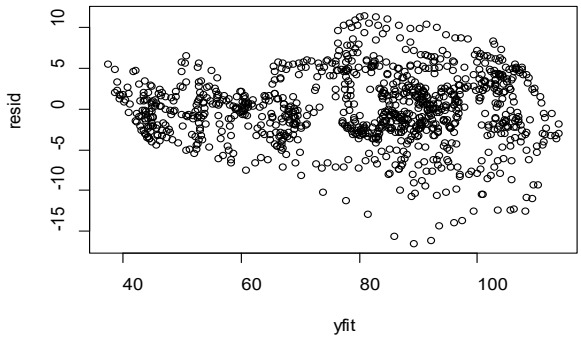

Figure 6 Independence test

4.3.4 Forecasting Using GDFM

We used the first 1000 samples as the training data to fit the model, and predicted the tourist volume in the next week with the fitted model. The solid line in Fig. 7 was the true value, the point was the predicted value, and the dashed line was the confidence interval of 85%.

Then, we compared the predicted results with neural network model (Fig. 8) and the smooth predicted value of HoltWinters model (Fig. 9), and the results can be seen in Tab. 5.

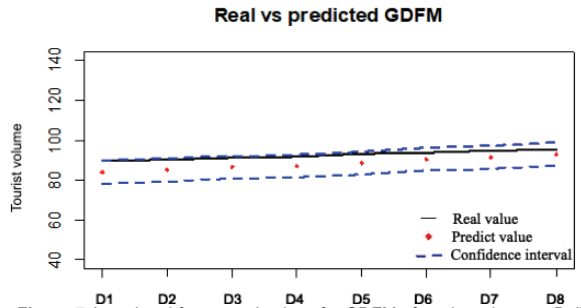

Figure 7 Actual and forecasted values for GDFM of tourist volume to Beijing from 27, Sep 2016-4, Oct 2016.

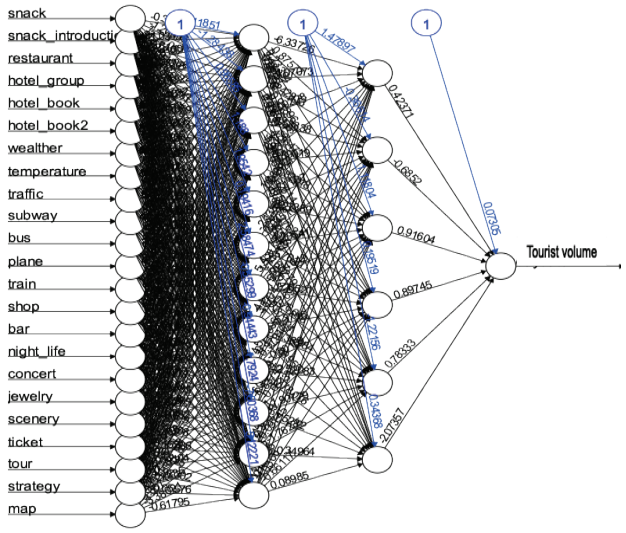

Figure 8 Forecasting model for neural network of tourist volume

Table 5 Prediction performance comparison of each model in tourism volume

Date	Real value	GDFM	Neural network	HoltWinters
28 Sep, 2016	90.55	85.40	81.40	88.84
29 Sep, 2016	91.60	86.90	79.64	88.84
30 Sep, 2016	92.07	87.02	78.48	88.84
1 Oct, 2016	93.37	88.86	77.91	88.84
2 Oct, 2016	93.72	90.47	76.14	88.84
3 Oct, 2016	94.89	91.58	76.65	88.84
4 Oct, 2016	95.60	92.97	75.65	88.84
RMSE		4.19	15.54	4.59

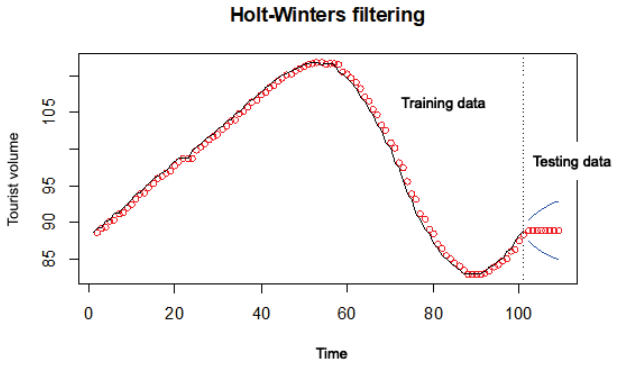

Figure 9 Forecasting model for HoltWinters of tourist volume.

It can be seen from Tab. 5 that the performance of GDFM model is significantly better than the neural network model and HoltWinters model. In terms of prediction accuracy, the accuracy of the GDFM model was 95.6%, the Root Mean Square Error (RMSE) was 4.19, and the error rate was 4.39%. However, the RMSE of the neural network model and HoltWinters model were 15.54 and 4.59. In terms of predictive stability, the confidence interval of the GDFM model in the seventh day of prediction period was still narrow, while the confidence interval of the neural network model was obviously wide and HoltWinters model performed worse.

5 CONCLUSION

Our analysis proposed a powerful prediction framework to apply the Internet search index to the prediction of tourism volume. This study indicated that Internet search index could reflect the general public's concerns and future plans, hence it was feasible and significant in forecasting study. Meanwhile, it is crucial for decision makers and business managers to predict tourism volume timely and accurately. Using Internet search index will help companies make the best strategic decisions during peak or off-peak travel.

Secondly, we introduced a prediction model which builds the model using the main components of Internet search index with GDFM. The prediction model was compared with neural network model and HoltWinters model. Meanwhile, the empirical results indicated that GDFM model had the best performance among the three prediction models mentioned in this paper. There are three advantages for using GDFM model as the prediction model in this study: (1) Large data sets usually have a complex correlation. GDFM model achieved the reduction of dimensions based on the information from various aspects. In addition, it not only improved the prediction accuracy, but also avoided the omission of important variables; (2) It has contingency and uncertainty when fitting the model using neural network. However, GDFM model has a more stable effect; (3) GDFM model has a good performance when predicting data in a long period of time. Several models perform well in short-term forecasts, but the confidence interval increases with the increase of the forecast period. It can be seen from the prediction results of GDFM that the confidence interval of the generalized dynamic factor model in the seventh day of prediction period is still narrow and maintains a higher prediction accuracy.

Finally, the study also has limitations. For one thing, slightly different keywords for Internet search index may lead to different prediction results, and the selection of keywords should be standardized. For another, a linear GDFM model was built around the components of the Internet search index in this paper, and the applicability of the model for the non-linear model should be discussed in the future research.

6 REFERENCES

[1] Luo, L., Liao, C., Zhang, F., Zhang, W., Li, C., Qiu, Z., & Huang, D. (2018). Applicability of internet search index for asthma admission forecast using machine learning. *The International journal of health planning and management*, *33*(3), 723-732. https://doi.org/10.1002/hpm.2525

[2] Wang, X., Ye, Q., Zhao, F., & Kou, Y. (2018). Investor sentiment and the Chinese index futures market: Evidence from the internet search. *Journal of Futures Markets*, *38*(4), 468-477. https://doi.org/10.1002/fut.21893

[3] Yang, X., Pan, B., Evans, J. A., & 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

[4] Beldona, S. (2005). Cohort analysis of online travel information search behavior: 1995-2000. *Journal of Travel Research*, *44*(2), 135-142. https://doi.org/10.1177/0047287505278995

[5] Buhalis, D. & Law, R. (2008). Progress in information technology and tourism management: 20 years on and 10 years after the Internet-The state of eTourism research. *Tourism management*, *29*(4), 609-623. https://doi.org/10.1016/j.tourman.2008.01.005

[6] 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. https://doi.org/10.1016/j.tourman.2014.07.014

[7] Armeanu, D., Andrei, J. V., Lache, L., & Panait, M. (2017). A multifactor approach to forecasting Romanian gross domestic product (GDP) in the short run. *PloS one*, *12*(7), e0181379. https://doi.org/10.1371/journal.pone.0181379
[8] Uysal, M. & El Roubi, M. S. (1999). Artificial neural networks versus multiple regression in tourism demand analysis. *Journal of Travel Research*, *38*(2), 111-118. https://doi.org/10.1177/004728759903800203
[9] Unhapipat, S. (2018). ARIMA model to forecast international tourist visit in Bumthang, Bhutan. *In Journal of Physics: Conference Series*, *1039*(1), 012023. https://doi.org/10.1088/1742-6596/1039/1/012023
[10] Chang, Y. W. & Liao, M. Y. (2010). A seasonal ARIMA model of tourism forecasting: The case of Taiwan. *Asia Pacific journal of Tourism research*, *15*(2), 215-221. https://doi.org/10.1080/10941661003630001
[11] Liu, X., Peng, H., Bai, Y., Zhu, Y., & Liao, L. (2014). Tourism flows prediction based on an improved grey GM (1, 1) model. *Procedia-Social and Behavioral Sciences*, 138, 767-775. https://doi.org/10.1016/j.sbspro.2014.07.256
[12] Duro, J. A. (2018). Seasonality of tourism: A new decomposition. *Tourism Economics*, *24*(5), 615-621. https://doi.org/10.1177/1354816618768319
[13] 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. https://doi.org/10.1080/10548408.2016.1170651
[14] Kholodilin, K. A., Podstawski, M., Siliverstovs, B., & Bürgi, Constantin Rudolf Salomo. (2009). Google searches as a means of improving the nowcasts of key macroeconomic variables. *Social Science Electronic Publishing*. https://doi.org/10.2139/ssrn.1507084
[15] Ripberger, J. T. (2011). Capturing curiosity: Using internet search trends to measure public attentiveness. *Policy studies journal*, *39*(2), 239-259. https://doi.org/10.1111/j.1541-0072.2011.00406.x
[16] Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012-1014. https://doi.org/10.1038/nature07634
[17] Dai, B., Jiang, Y., Yang, L., & Ma, Y. (2017). China's outbound tourism-Stages, policies and choices. *Tourism Management*, 58, 253-258. https://doi.org/10.1016/j.tourman.2016.03.009
[18] Liu, Q., Zhang, X., Huang, S., Zhang, L. & Zhao, Y. (2020). Exploring Consumers' Buying Behavior in a Large Online Promotion Activity: The Role of Psychological Distance and Involvement. *Journal of Theoretical and Applied Electronic Commerce Research*, *15*(1), 66-80. https://doi.org/10.4067/S0718-18762020000100106
[19] Liu, Q., Zhang, X., Zhang, L. & Zhao, Y. (2019). The interaction effects of information cascades, word of mouth and recommendation systems on online reading behavior: An empirical investigation. *Electronic Commerce Research*, *19*(3), 521-547. https://doi.org/10.1007/s10660-018-9312-0
[20] Qing, F. & Gennian, S. (2016). Effects of Per Capita GDP and Urbanization on Domestic Tourism Development in China's Eight Regions. *Areal Research and Development*, (4), 18.
[21] Xu, W., Liu, L., Zhang, Q. & Liu, P. (2018). Location decision-making of equipment manufacturing enterprise under dual channel purchase and sale mode. *Complexity*, 1-16. https://doi.org/10.1155/2018/3797131
[22] Xu, W. & Yin, Y. (2018). Functional objectives decision-making of discrete manufacturing system based on integrated ant colony optimization and particle swarm optimization approach, *Advances in Production Engineering & Management*, *13*(4), 389-404. https://doi.org/10.14743/apem2018.4.298
[23] Eeckels, B., Filis, G., & Leon, C. (2012). Tourism income and economic growth in Greece: empirical evidence from their cyclical components. *Tourism Economics*, *18*(4), 817-834. https://doi.org/10.5367/te.2012.0148
[24] Yang, Y., Liu, Z. H., & Qi, Q. (2014). Domestic tourism demand of urban and rural residents in China: Does relative income matter? *Tourism Management*, 40, 193-202. https://doi.org/10.1016/j.tourman.2013.05.005
[25] Ding, H. (2018). An Analysis of Domestic Tourism Consumption Based on R Software. *In MATEC Web of Conferences*, *228*, 05006). https://doi.org/10.1051/matecconf/201822805006
[26] Zeng, B. & He, Y. (2019). Factors influencing Chinese tourist flow in Japan-a grounded theory approach. *Asia Pacific Journal of Tourism Research*, *24*(1), 56-69. https://doi.org/10.1080/10941665.2018.1541185
[27] Dilogini, K.,Shivany, S., & Kumara A. (2019). Analysis of Relation Between Customer Behavior and Information Technology Market. *Journal of System and Management Sciences*, *9*(1), 87-104.
[28] Safayet, MD A., Islam, MD H. & Ahmed, S. (2018). A Case Study on Risk Management in Existing Construction Project in Bangladesh. *Journal of Logistics, Informatics and Service Science*, *5*(1), 1-16.
[29] Kim, M. J., Jung, T., Kim, W. G., & Fountoulaki, P. (2015). Factors affecting British revisit intention to Crete, Greece: high vs. low spending tourists. *Tourism Geographies*, *17*(5), 815-841. https://doi.org/10.1080/14616688.2015.1062908
[30] Lovie, A. D. & Lovie, P. (1993). Charles Spearman, Cyril Burt, and the origins of factor analysis. *Journal of the History of the Behavioral Sciences*, *29*(4), 308-321. https://doi.org/10.1002/1520-6696(199310)29:4<308::AID-JHBS2300290402>3.0.CO;2-P
[31] Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and statistics*, *82*(4), 540-554. https://doi.org/10.1162/003465300559037
[32] Stock, J. H. & Watson, M. (2008). The evolution of national and regional factors in US housing construction. *Volatility and time series econometrics: essays in honor of Robert F. Engle.*
[33] Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2004). The generalized dynamic factor model consistency and rates. *Journal of Econometrics*, *119*(2), 231-255. https://doi.org/10.1016/S0304-4076(03)00196-9

Contact information:

Yiran LI, PhD
(Corresponding author)
School of Information Management, Wuhan University,
No. 299 Bayi Road, Wuchang District, Wuhan City, Hubei Province, China
E-mail: yiran_li94@sina.com

Mengyao XU, MD
School of Information Management, Wuhan University,
No. 299 Bayi Road, Wuchang District, Wuhan City, Hubei Province, China
E-mail: 2018201040026@whu.edu.cn

Xuan WEN, PhD
School of Information Management, Wuhan University,
No. 299 Bayi Road, Wuchang District, Wuhan City, Hubei Province, China
E-mail: Xuan_W113@163.com

Daomeng GUO, PhD
(Corresponding author)
1) School of Economics and Management, Hubei Engineering University,
No. 272 Jiaotong Road, Xiaonan District, Xiaogan City, Hubei Province, China
2) School of Information Management, Wuhan University,
No. 299 Bayi Road, Wuchang District, Wuhan City, Hubei Province, China,
E-mail: guodaomeng@whu.edu.cn