ISSN 2616-5902 Vol. 7, Issue 2: 72-81, DOI: 10.25236/AJBM.2025.070210

Research on the relationship and prediction between Baidu Index and city tourist number—Take Guangzhou City as an example

Xiaojie Chen*

School of Economics, Guangxi University, Nanning, China
cxiaojie1008@163.com
**Corresponding author*

Abstract: *This paper analyses the relationship between the Baidu index and tourist numbers in Guangzhou, establishing an ARMA model for monthly tourist numbers and making predictions. Seven Baidu keywords, including "Guangzhou food," are added as explanatory variables to build a VAR model for comparison. The results show a long-run equilibrium and Granger causality between tourist numbers and Baidu keywords. The PC-based VAR model's prediction accuracy is 47% higher than the ARMA model. In comparison, the mobile-end VAR model outperforms the PC model by 3% in prediction accuracy but explains 3% more of the variance in tourist number changes. These findings can support the decision-making of relevant departments.*

Keywords: *Baidu index; ARMA model; VAR model; Number of tourists*

1. Introduction

With the rapid growth of China's tourism industry, the number of tourists in cities and scenic spots continues to rise. Forecasting actual data based on the Baidu index has been widely used in economic, social and other fields. In the social field, there are early warnings of virus infection outbreaks[1], spatio-temporal differences and influencing factors of public attention to the epidemic[2], etc. In the economic field, there are stock market correlations[3], household inflation expectations[4], etc. At present, most of the research based on Baidu index prediction focuses on the field of tourism, exploring the influencing factors of tourism, the prediction of the number of tourists, the spatial-temporal characteristics of tourism information flow and other topics. Ding Xin et al. took Xiamen City as an example to analyze the influencing factors of tourism network attention[5], while Kang Junfeng et al. took Shanghai as an example to predict its future tourism trend[6]. In terms of the current results, most of the current results are based on the traditional econometric model, and the selection of keywords for explanatory variables in the model is relatively simple, and few studies distinguish the data of mobile terminal and PC terminal.

This paper takes Guangzhou as an example, combines the cointegration test and Granger causality test in econometrics, analyzes the relationship between the Baidu index and the number of tourists in Guangzhou, and uses the number of tourists in Guangzhou (domestic) to establish the autoregressive moving average of Guangzhou's monthly tourist scale. ARMA) model and make a prediction, and then add the Baidu index of "Guangzhou food", "Guangzhou transportation", "Guangzhou Scenic spots", "Guangzhou hotels", "Guangzhou specialities", "Guangzhou weather" and "Guangzhou snacks" extracted from the six elements of tourism "food, accommodation, transport, travel and entertainment" and correlation test as explanatory variables. The vector auto regression (VAR) model and autoregressive moving average model are established to compare the prediction accuracy.

2. Research Methodology

2.1 ARMA model

The Autoregressive Moving Average (ARMA) model is one of the essential methods in time series analysis. The general form of an ARMA(p,q) model can be expressed as:

$$y_t = \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad (1)$$

ISSN 2616-5902 Vol. 7, Issue 2: 72-81, DOI: 10.25236/AJBM.2025.070210

where p denotes the order of the autoregressive model, q denotes the order of the moving average model, $\emptyset_i(i = 1, 2, \ldots, p)$, $\theta_j(j = 1, 2, \ldots, q)$ are the coefficients to be determined in the model, ε_t is the error term, and y_t is the observed value. The AR and MA models are special cases of the ARMA (p,q) model. When $p = 0$, $\mathrm{ARMA}(0, \mathrm{q}) = \mathrm{MA}(\mathrm{q})$. When $q = 0$, $\mathrm{ARMA}(\mathrm{p}, 0) = AR(\mathrm{q})$.

2.2 VAR model

The Vector Autoregressive (VAR) model regresses each time series on the lagged values of all variables in the system, modelling each endogenous variable as a function of these lags. It is useful for analysing variable interactions, the impact of stochastic disturbances, and forecasting interconnected economic time series. The VAR model in this study is expressed as follows:

$$\mathrm{Z_t} = \mathrm{A_1 Z_{t-1}} + \mathrm{A_2 Z_{t-2}} + \cdots\cdots + \mathrm{A_p Z_{t-p}} + \varepsilon_\mathrm{t} \quad (2)$$

where Z_t is a vector consisting of endogenous variables, i.e., $Z_t = (Y_t, BD_{1t}, BD_{2t} \cdots BD_{(k-1)\,t})$. $Y_t, BD_{1t}, BD_{2t} \cdots BD_{(k-1)\,t}$ represent the number of tourists in Guangzhou and the Baidu search indices for different keywords, respectively. ε_t is a k-dimensional random disturbance vector, p is the number of lags, and $A_1, A_2, \cdots\cdots, A_p$ are $k \times k$ matrices of estimated parameters.

3. Empirical analysis

3.1 Selection of Baidu keywords and data

Baidu, the largest Chinese search engine, provides the Baidu index, which tracks search volumes for keywords, offering insights into public interest. This index is available for both PC and mobile search data. This paper uses Baidu index data for tourism-related keywords such as "food," "accommodation," and "transport" to analyze Guangzhou's tourist numbers. Seven keywords were selected: "Guangzhou food," "Guangzhou map," "Guangzhou transportation," "Guangzhou attractions," "Guangzhou hotels," "Guangzhou specialities," and "Guangzhou snacks." Monthly data from February 2017 to May 2022 on Guangzhou's tourist numbers and Baidu search volumes were used for econometric modeling and prediction. The descriptive statistics of the variables are shown in Table 1.

Table 1: Description of variables

Variables	Description	Variables	Description
Y	Tourist number in Guangzhou (Domestic)	JDyd	Keywords "Guangzhou hotel" mobile terminal Baidu index
mspc	Keywords "Guangzhou food" pc end Baidu index	tcpc	Keywords "Guangzhou specialty" Baidu index on the pc side
msyd	Keywords "Guangzhou food" mobile terminal Baidu index	tcyd	Keywords "Guangzhou specialty" mobile terminal Baidu index
jtpc	Keywords "Guangzhou Transportation" Baidu index on the pc side	tqpc	Keywords "Guangzhou weather" pc Baidu index
jtyd	Keywords "Guangzhou Transportation" mobile terminal Baidu index	tqyd	Keywords "Guangzhou weather" mobile terminal Baidu index

ISSN 2616-5902 Vol. 7, Issue 2: 72-81, DOI: 10.25236/AJBM.2025.070210

jdpc	Keywords "Guangzhou scenic spots" Baidu index on pc	xcpc	Keywords "Guangzhou snacks" Baidu index on pc
jdyd	Keywords "Guangzhou scenic spots" mobile terminal Baidu index	xcyd	Keywords "Guangzhou snacks" mobile terminal Baidu index
JDpc	Keywords "Guangzhou hotel" Baidu index on the pc side		

3.2 Model Test

3.2.1 Unit root and cointegration tests

Given that the sample data in this study are time series, it is essential to ensure stationarity and avoid spurious regression. Therefore, unit root and cointegration tests were conducted before establishing the econometric model. The unit root test was performed using the ADF (Augmented Dickey-Fuller) test, and the results are presented in Table 2.

Table 2: Unit root test

Variables	Order of difference	ADF	P value	Conclusion
Y	0	-3.4201	0.06134	Stationary
mspc	1	-5.1995	0.01	Stationary
jtpc	0	-4.2473	0.01	Stationary
jdpc	0	-3.4298	0.05981	Stationary
JDpc	0	-3.4396	0.05827	Stationary
tcpc	0	-3.2995	0.08039	Stationary
tqpc	0	-3.4765	0.05244	Stationary
xcpc	1	-5.0958	0.01	Stationary
msyd	1	-4.8349	0.01	Stationary
jtyd	1	-4.706	0.01	Stationary
jdyd	1	-5.1458	0.01	Stationary
JDyd	0	-4.1443	0.01	Stationary
tcyd	1	-4.3186	0.01	Stationary
tqyd	0	-3.9097	0.01972	Stationary
xcyd	1	-4.6592	0.01	Stationary

ISSN 2616-5902 Vol. 7, Issue 2: 72-81, DOI: 10.25236/AJBM.2025.070210

As shown in Table 2, the original series for Guangzhou's domestic tourist numbers, Guangzhou attractions (PC), Guangzhou hotels (PC), Guangzhou specialities (PC), Guangzhou weather (PC), Guangzhou hotels (mobile), and Guangzhou weather (mobile) are stationary (I(0)). Other variables become stationary after first-order differencing (I(1)), meeting the prerequisites for cointegration analysis. The Johansen cointegration test was used to examine long-term equilibrium relationships between Guangzhou's domestic tourist numbers and the Baidu Index of seven keywords on PC and mobile platforms. The results are in Table 3.

From the PC-based test results: The trace statistic for "no cointegration" is 115.96, exceeding the critical value of 52.00, indicating at least one cointegrating relationship. For "at most one cointegration," the trace statistic is 69.03, exceeding 46.45, confirming at least two cointegrating relationships. For "at most two cointegrations," the trace statistic is 53.80, exceeding 14.90, confirming at least three. For "at most three cointegrations," the trace statistic is 39.90, exceeding 34.40, confirming at least four. For "at most four cointegrations," the trace statistic is 17.38, below 28.14, indicating four cointegrating relationships.

These results confirm four cointegrating relationships among PC-based Baidu Index variables and tourist numbers, indicating a long-term equilibrium. Similarly, the mobile-based test confirms a long-term relationship between mobile Baidu Index variables and tourist numbers.

Table 3: Cointegration tests for the variables

Null hypothesis	PC side Johansen cointegration test results		Mobile terminal Johansen cointegration test results	
	Trace statistic quantity	0.05 critical value	Trace statistic quantity	0.05 critical value
no cointegration relationship	115.96	52.00	126.23	52.00
at most one cointegrating relationship	69.03	46.45	81.57	46.45
at most two cointegrating relationships	53.80	40.30	56.06	40.30
at most three cointegrating relationships	39.90	34.40	33.94	34.40
at most four cointegrating relationships	17.38	28.14	——	——

3.2.2 Gram causality test

According to the results in Table 4, the cointegration relationship between independent and dependent variables confirms Granger causality. Using AIC, SC, HQ, and FPE criteria, the optimal lag length was set to 5. Granger causality tests revealed that on the PC platform, unidirectional causality exists for keywords like "Guangzhou Map," "Guangzhou Traffic," "Guangzhou Attractions," "Guangzhou Specialties," and "Guangzhou Snacks," while bidirectional causality exists for "Guangzhou Hotels" and "Guangzhou Weather." On the mobile platform, bidirectional causality exists for "Guangzhou Cuisine," "Guangzhou Hotels," and "Guangzhou Snacks," with unidirectional causality for the remaining keywords.

ISSN 2616-5902 Vol. 7, Issue 2: 72-81, DOI: 10.25236/AJBM.2025.070210

These findings demonstrate predictive and interactive relationships between tourist numbers and keyword search trends across platforms.

Table 4: Granger causality test results

Granger causality	F value	P value	Conclusion
mspc does not Granger cause Y	1.9161	0.1238	accept
Y does not Granger cause the mspc	0.3644	0.8327	accept
dtpc does not Granger cause Y	2.9981	0.02793	reject
Y does not Granger cause dtpc	1.7243	0.1609	accept
jtpc does not Granger cause Y	2.3288	0.07012	reject
Y does not Granger cause jtpc	1.4114	0.2452	accept
jdpc does not Granger cause Y	2.4341	0.06064	reject
Y does not Granger cause jdpc	0.5464	0.7025	accept
JDpc does not Granger cause Y	5.7967	0.0007292	reject
Y does not Granger cause JDpc	4.22	0.005413	reject
tcpc does not Granger cause Y	4.9663	0.002061	reject
Y does not Granger cause tcpc	1.1639	0.3391	accept
tqpc does not Granger cause Y	5.5667	0.0009686	reject
Y does not Granger cause tqpc	2.6027	0.04806	reject
xcpc does not Granger cause Y	1.7991	0.1453	accept
Y does not Granger cause xcpc	2.0925	0.09712	reject
msyd does not Granger cause Y	4.6828	0.001681	reject
Y does not Granger cause msyd	2.2638	0.06498	reject
jtyd does not Granger cause Y	3.1592	0.01623	reject
Y does not Granger cause jtyd	0.6514	0.662	accept
jdyd does not Granger cause Y	8.0384	2.024e-05	reject
Y does not Granger cause jdyd	1.3474	0.263	accept
JDyd does not Granger cause Y	3.8806	0.005442	reject
Y does not Granger cause JDyd	2.5829	0.03955	reject
tcyd does not Granger cause Y	2.8424	0.02644	reject
Y does not Granger cause tcyd	1.2944	0.2841	accept

ISSN 2616-5902 Vol. 7, Issue 2: 72-81, DOI: 10.25236/AJBM.2025.070210

tqyd does not Granger cause Y	1.4144	0.2383	accept
Y does not Granger cause tqyd	2.0511	0.09042	reject
xcyd does not Granger cause Y	3.6998	0.007135	reject
Y does not Granger cause xcyd	2.514	0.04403	reject

3.3 Establishment and analysis of forecasting model

3.3.1 ARMA model

Based on the autocorrelation (ACF) and partial autocorrelation (PACF) plots of Guangzhou's tourist numbers, several ARMA(p, q) models were compared. Using AIC, SC, and RMSE criteria, the ARMA(5, 0, 4) model was identified as optimal, with lower AIC and SC values, higher goodness-of-fit, and lower RMSE, indicating strong predictive performance. Detailed results are shown in Table 5.

$$Y = \sum_{i=1}^{5} AR(i)Y(-i) + \sum_{i=1}^{4} MA(i)\varepsilon(-i) + C \quad (3)$$

Table 5: ARMA model for tourist numbers in Guangzhou

Variables	parameter estimates	Variables	parameter estimates
AR(1)	0.5556 (0.3789)	MA(1)	0.0786 (0.3899)
AR(2)	-0.1944 (0.3332)	MA(2)	0.2947 (0.2649)
AR(3)	0.5729 (0.3020)	MA(3)	-0.6199 (0.3258)
AR(4)	0.4168 (0.3654)	MA(4)	-0.5337 (0.3070)
AR(5)	-0.3902 (0.2134)	C	-0.5339 (0.1466)

The AIC value for Equation (1) is 618.38, with an in-sample root mean square error (RMSE) of 82.93. An ACF test and Ljung-Box test on the residuals indicate no significant autocorrelation, as all p-values are greater than 0.5, confirming that the residuals are white noise. Thus, this model can be used to forecast Guangzhou's domestic tourist numbers for January to October 2022. The forecast results, shown in Figure 1, suggest a fluctuating upward trend in domestic tourist numbers for Guangzhou in the near future.

ISSN 2616-5902 Vol. 7, Issue 2: 72-81, DOI: 10.25236/AJBM.2025.070210

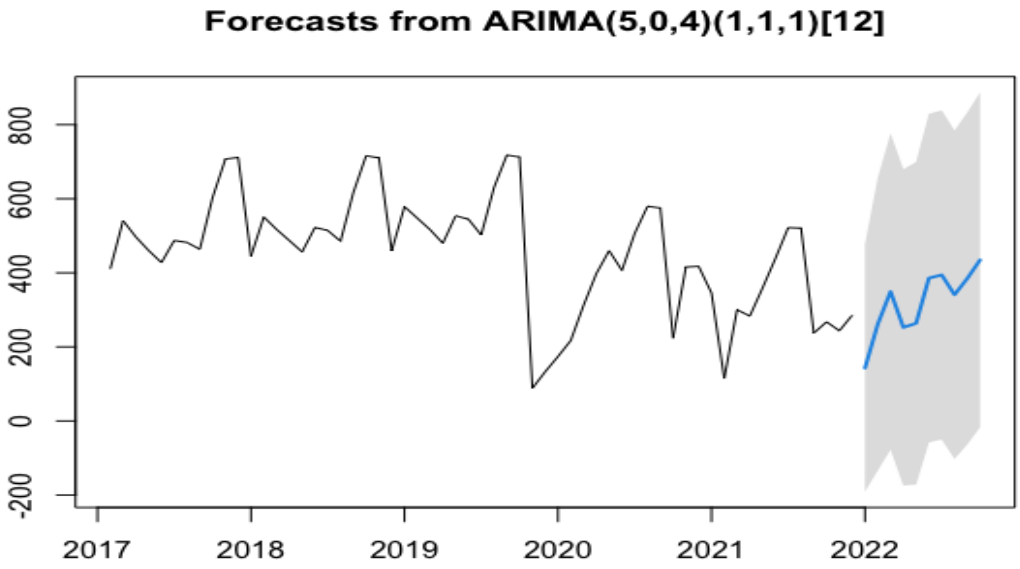

Figure 1: ARMA model predicts

3.3.2 VAR Model

$$Y = \sum_{i=1}^{5} C(1,i)Y(-i) + C\sum_{i=1}^{5} C(2,i)mspc(-i) + \sum_{i=1}^{5} C(3,i)jtpc(-i) + \sum_{i=1}^{5} C(4,i)jdpc(-i) + \sum_{i=1}^{5} C(5,i)JDpc(-i) + \sum_{i=1}^{5} C(6,i)tcpc(-i) + \sum_{i=1}^{5} C(7,i)tqpc(-i) + \sum_{i=1}^{5} C(8,i)xcpc(-i)\,C(5) \quad (4)$$

where *Y* represents the number of domestic tourists in Guangzhou, *mspc*, *jtpc*, *jdpc*, *JDpc*, *tcpc*, *tqpc and xcpc* denote the Baidu search indices for the keywords "Guangzhou food," "Guangzhou transportation," "Guangzhou attractions," "Guangzhou hotels," "Guangzhou specialities," "Guangzhou weather," and "Guangzhou snacks," respectively. *C (j, i)* represents the estimated coefficient of parameter *j* (where j=1,2,…7 corresponding to *Y*, *mspc*, *jtpc*, *jdpc*, *JDpc*, *tcpc*, *tqpc*, *xcpc*), and *i*=1,2,3,4,5 is the lag order.

The VAR model's adjusted R^2 for the PC side is 0.58, with a root mean square error of 43.8239. For the mobile side, the adjusted R^2 is 0.64, and the root mean square error is 42.53552. The model is stable, as all inverse roots of the characteristic equation lie within the unit circle. The detailed estimation results are shown in Tables 6 and 7.

Table 6: Estimation results of the PC-side VAR model

Variables	Parameter estimates	Variables	Parameter estimates	Variables	Parameter estimates
C(1,1)	0.066 (0.361)	C(3,5)	0.083 (0.099)	C(6,4)	-0.073 (0.113)
C(1,2)	0.474 (0.391)	C(4,1)	0.018 (0.061)	C(6,5)	-0.063 (0.074)
C(1,3)	-1.108 (0.324)	C(4,2)	0.014 (0.058)	C(7,1)	-0.0002 (0.001)
C(1,4)	0.528 (0.542)	C(4,3)	0.017 (0.057)	C(7,2)	0.0003 (0.001)

ISSN 2616-5902 Vol. 7, Issue 2: 72-81, DOI: 10.25236/AJBM.2025.070210

C(1,5)	-0.216 (0.532)	C(4,4)	0.018 (0.060)	C(7,3)	-0.001 (0.001)
C(2,1)	-0.041 (0.043)	C(4,5)	0.077 (0.048)	C(7,4)	-0.0002 (0.001)
C(2,2)	-0.088 (0.050)	C(5,1)	-0.067 (0.174)	C(7,5)	-0.001 (0.001)
C(2,3)	0.016 (0.064)	C(5,2)	-0.212 (0.203)	C(8,1)	-0.209 (0.145)
C(2,4)	-0.029 (0.059)	C(5,3)	-0.036 (0.179)	C(8,2)	-0.149 (0.130)
C(2,5)	-0.060 (0.064)	C(5,4)	-0.090 (0.197)	C(8,3)	0.053 (0.138)
C(3,1)	0.055 (0.153)	C(5,5)	0.326 (0.179)	C(8,4)	-0.019 (0.134)
C(3,2)	0.068 (0.094)	C(6,1)	0.124 (0.085)	C(8,5)	-0.433 (0.156)
C(3,3)	-0.130 (0.080)	C(6,2)	0.198 (0.100)	C(9)	-21.328 (18.333)
C(3,4)	0.070 (0.107)	C(6,3)	0.122 (0.074)		

Table 7: Estimation results of the mobile-side VAR model

Variables	Parameter estimates	Variables	Parameter estimates	Variables	Parameter estimates
C(1,1)	-0.354 (0.467)	C(3,5)	0.111 (0.071)	C(6,4)	-0.003 (0.012)
C(1,2)	-1.479 (0.447)	C(4,1)	0.012 (0.006)	C(6,5)	-0.0004((0.010)
C(1,3)	-0.886 (0.538)	C(4,2)	0.021 (0.009)	C(7,1)	-0.0001 (0.0001)
C(1,4)	-0.117 (0.478)	C(4,3)	0.001 (0.011)	C(7,2)	0.0001(0.0001)

ISSN 2616-5902 Vol. 7, Issue 2: 72-81, DOI: 10.25236/AJBM.2025.070210

C(1,5)	0.076 (0.431)	C(4,4)	0.002 (0.009)	C(7,3)	0.00004(0.0001)
C(2,1)	0.002 (0.011)	C(4,5)	-0.009 (0.006)	C(7,4)	-0.00001(0.0001)
C(2,2)	0.008 (0.011)	C(5,1)	0.057 (0.039)	C(7,5)	-0.0002(0.0001)
C(2,3)	0.008 (0.010)	C(5,2)	-0.113 (0.043)	C(8,1)	-0.045 (0.041)
C(2,4)	0.018 (0.011)	C(5,3)	-0.016 (0.058)	C(8,2)	0.021 (0.040)
C(2,5)	0.028 (0.009)	C(5,4)	0.024 (0.069)	C(8,3)	-0.036 (0.037)
C(3,1)	0.027 (0.050)	C(5,5)	0.038 (0.054)	C(8,4)	-0.003 (0.042)
C(3,2)	0.046 (0.063)	C(6,1)	-0.0002 (0.011)	C(8,5)	-0.048 (0.036)
C(3,3)	0.061 (0.059)	C(6,2)	0.023 (0.012)	C(9)	7.816 (15.592)
C(3,4)	0.170 (0.061)	C(6,3)	0.026 (0.014)		

4. Conclusion and Discussion

First, a long-term equilibrium exists between Guangzhou's domestic tourist numbers and the Baidu indices for seven keywords ("Guangzhou cuisine," "Guangzhou traffic," "Guangzhou attractions," "Guangzhou hotels," "Guangzhou specialities," "Guangzhou weather," "Guangzhou snacks") on both PC and mobile platforms, satisfying the cointegration condition. Second, there is a unidirectional or bidirectional Granger causality relationship between the Baidu indices and Guangzhou's domestic tourist numbers, with significant differences between mobile and PC indices for certain keywords. Third, the PC-end VAR model improves forecasting accuracy of domestic tourist numbers by 47% compared to the ARMA model, while the mobile-end VAR model improves accuracy by 49%, indicating that network search data outperforms traditional models. Fourth, the mobile-end VAR model outperforms the PC-end VAR model with a 3% improvement in forecasting accuracy and a 4% increase in goodness of fit, suggesting superior predictive ability. The mobile-end model also explains 3% more variance in tourist numbers and Baidu indices compared to the PC-end model, though the difference is minimal.

In summary, the VAR model surpasses the traditional ARMA model in both explanatory and predictive power, with the mobile-end model slightly outperforming the PC-end model. Mobile platforms offer portability and quicker access, aligning with travel decision-making, while PC platforms provide detailed information for in-depth searches. Future research could combine their strengths to enhance tourism forecasting accuracy.

ISSN 2616-5902 Vol. 7, Issue 2: 72-81, DOI: 10.25236/AJBM.2025.070210

References

[1] Wu H C, Wu C, Lu Q B, et al. Early warning research on norovirus outbreak based on Baidu Index[J]. Chinese Journal of Preventive Medicine, 2021, 22(2): 120–124.

[2] Sun Y T, Xiao F, Zhou Y, Spatiotemporal differences and influencing factors of public attention during the COVID-19 pandemic: An analysis based on Baidu Search Index[J]. Tropical Geography, 2020, 40(3): 375–385.

[3] Lin J J, Tang Y, Zhou X L, et al. Research on the time-frequency domain relationship among northbound funds, Baidu Index, and stock markets: From the perspective of co-skewness[J]. Chinese Journal of Management Science, 2022, 30(1): 20–31.

[4] Dai Y Y, Cai D H, Zhang Y. Research on the relationship between internet search and residents' inflation expectations: Empirical analysis based on depositors' surveys and Baidu Index[J]. Shanghai Finance, 2020(11): 42–51.

[5] Ding X, Wang J Q, Li Y Q, Spatiotemporal characteristics and influencing factors of online attention to tourist destinations based on Baidu Index: A case study of Xiamen[J]. Resource Development & Market, 2018, 34(5): 709–714.

[6] Kang J F, Guo X Y, Fang L. Forecasting tourism trends based on the spatiotemporal distribution of Baidu Index: A case study of Shanghai[J]. Journal of Southwest China Normal University (Natural Science Edition), 2020, 45(10): 72–81.