Investigating the Spatial-Temporal Variation of Pre-Trip Searching in an Urban Agglomeration

Jianxin Zhang, Yuting Yan, Jinyue Zhang, Peixue Liu, and Li Ma

Abstract: Search engines have been the primary tool for online information search before traveling. Timely detection and the control of peak tourist flows in scenic areas prevent safety hazards and the overconsumption of tourism resources due to excessive tourist clustering. This study focuses on the spatial-temporal interactions between the pre-trip stage and the after-arrival stage to investigate online information search behavior. Big data obtained from mobile roaming and search engines provide precise data on daytime and city scales, which enabled this paper to examine the relationship between daily tourist arrivals and their pre-trip searching from 40 cities within the Yangtze River Delta urban agglomeration. This study had several original results. First, tourists generally search for tourist information 2–8 days before arriving at destinations, while tourist volume and SVI from source cities show distance attenuation. Second, SVI is a precursor to changes in tourist volume. The precursory time rises with the increase of traffic time spatially. Third, we validated a VAR model and improved its accuracy by constructing it based on the spatial-temporal differentiation of search features. These findings would enhance the management and preservation of tourism resources and promote the sustainable development of tourism destinations.

Keywords: tourism information search; search engine; mobile roaming data; tourism forecasting; Webometrics

1. Introduction

With the widespread accessibility of the internet and smartphones, tourist information search (TIS) has turned into one of the most critical research categories in tourist behavior and destination marketing [1,2]. Modern tourists increasingly rely on the internet to choose a vacation destination, obtain destination information, plan their itinerary, find hotels, and other tourism services [3]. Search engines, as a primary context-aware tool, provide information to tourists who leave their usual resident place and travel to an unfamiliar destination in many forms, such as text reviews, pictures, videos, and even virtual maps [4,5]. Everyone’s online search behaviors leave traces on search engines, which makes search volume index (SVI) a rare real-time dynamic form of big data that provides the opportunity to study pre-trip behavior [6,7]. From the perspective of big data, analyzing tourists’ search data on search engines helps to predict visitor traffic and to timely detect and control the peak flow of tourists in scenic areas. This has become an important tool for destination managers for marketing, monitoring, and controlling tourist destinations, and it contributes to improving the level of sustainable development of tourism destinations [8,9].

Tourism demand forecasting is of great significance in the tourism industry, as it could help tourism practitioners and tourism managers to better formulate tourism plans, resource allocation, and marketing strategies to promote the sustainable development of the destination [10,11]. At the same time, forecasting tourism demand can also help tourism
enterprises and government departments to better respond to changes in the tourism market and improve the quality and efficiency of tourism services [12], thereby promoting the sustainable development of tourism.

Traditional tourism demand forecasting methods include time series analysis [13], regression analysis [14], and neural networks [15]. In recent years, with the development of big data technology [12] and artificial intelligence technology [16], tourism demand forecasting has also gradually developed towards a data-driven and intelligent direction. Researchers have adopted search engine data to predict COVID-19 [17], macroeconomic statistics [18], and tourist volume [19,20]. Some structural features of SVI, such as platform bias [1], language bias [19], and distance bias [6], have been noted recently. In addition, the rapid development of bulk data handling techniques has made mobile roaming data and timely sources available for tourist behavior research. Roaming data have been used to analyze tourist flows [21] and travel distances [22,23].

Studies have further used different precursor effects of SVI with various sources to optimize the prediction model of total tourist volume [6,19]. Compared to traditional data, search volume is timely monitored, easily collected, and provides instant feedback. In the last couple of decades, traditional sources, quarterly or annual data, are the mainstream sources (more than 90%) [24], but today, demand fluctuation serves as essential data for service departments for the perspective of short-term precision marketing and personalized services [25]. Using big data to obtain more accurate short-term tourist flow data makes it possible to revise the traditional forecasting model. The data resources resulting from online services are a breakthrough in the improvement of the accuracy of real-time prediction.

However, most of the current studies on SVI were based on time series data, which ignored the interaction between the spatial distribution data and time-series data of tourist sources. This study aimed to bridge this gap in the existing literature by combining daily search indexes and actual tourist arrivals from 40 cities around the destination in a large urban agglomeration. The remaining paper was organized as follows: firstly, we critically reviewed TIS literature and its predictive power in the literature review. Secondly, we presented the co-integration test and Granger causality test to analyze the correlation in the Section 3. Thirdly, we took the world cultural heritage site Zhongshan Scenic Area as an example and analyzed the time–space correlation of tourists’ mobility and related online searches in the Section 4. Furthermore, this section used the method to compare predictions and verify the application value of the spatial features of SVI precursory effects. Finally, in the Section 5, we present the theoretical contributions and managerial implications.

2. Literature Review

2.1. Tourism Information Search (TIS)

Hyde [26] determined that information search, plans, and reservations are unique aspects of pre-trip decision-making. The amount of information searched before the tour could predict the extent of a tourist’s planning and bookings. Most researchers have admitted TIS as a widely researched issue in predicting tourist behavior and filled the gap brought by econometric models [27]. Specifically, TIS by search engine data has been used in predicting the demand of hotels or accommodations [28], tourist arrivals [18,29], and tourist flows [6].

Researchers also revealed the demographical factors that affect search behavior [30]. The websites browsed and the keywords searched by tourists reflect their knowledge, background, and information needs [31]. In addition, various factors between different cultures and society in pre-trip online information search behavior are verified, such as cultural background, habits, economy, and technology [32].

Webometrics, as a new frame, can evaluate the phenomena happening in cyberspace of precursors, from online information searches to offline visits [33]. The current TIS study mainly focuses on time series data, revealing that different tourism-based factors have short, medium, or long timescale effects [34,35]. For example, the weather proved to be a short-term impact on tourist volume [36,37]; extreme events proved to be a medium-term
impact [8]; and scenic spots had a long-term impact [38]. In the post-COVID-19 era, examining the role of search data in travel has also become an important topic [39]. However, scant search engine studies have paid attention to the relationship between time-series data and spatial distribution. Thus, whether the search data also follow distance decay rules and affect tourists’ intention to travel temporally and geographically is one of the main focuses of this study. Validating the accuracy of prediction models from spatiotemporal data is more helpful for tourist market analysis and destination management.

2.2. Search Engine Used in the Forecasting

Tourist forecasting is a popular topic with great significance in practical management applications. The research on tourist flow forecasting has gone through the beginning stage of analyzing the factors that affect tourism development, the intermediate stage of qualitatively describing the relationship between various factors and tourist flow forecasting, and the current stage of constructing models for quantitative analysis. Travel flow forecasting models commonly used in recent research include traditional time-series prediction methods [40], econometric methods [41], artificial neural network techniques [42,43], and combined prediction methods [44,45].

The traditional time-series forecasting model is widely used in tourism demand research since it only needs observation information of the variables itself, which makes data collection and model construction less costly. Autoregressive integrated moving average models have achieved remarkable results in forecasting tourism demand [46,47]. Researchers develop and revise various models based on these studies to achieve better predictions. The prediction power of VAR and vector error correction (VEC) models, which reflect the causal relationship between the forecasting object and the influencing factor, is better than the traditional time-series forecasting model [48]. The prediction accuracy of the VAR model is related to multiple parameters, including lag order, model selection criteria, covariance matrix of residuals, etc. It is necessary to select and adjust these parameters according to specific situations in order to improve the prediction ability of the model. To further improve prediction accuracy, researchers began to combine different prediction methods to compensate for the limitations of existing models [46,49]. For example, Hu et al. adopted feature extraction technology and combined econometrics, integrated learning, and hybrid methods to build models and optimize the performance of tourism prediction with field effect tubes [50]. The forecasting comparison and correlation of multiple source regions have attracted the attention of scholars. The relationships between tourism demand and variables such as income, ticket prices, and travel costs within a country or across multiple regions are empirically studied [41].

As short-term data becomes more readily available, routine and even real-time predictive studies are becoming feasible and promising. For example, a short-term demand for casino tourism [51] and a holiday forecast of tourist flow for Mountain Huangshan [44] were studied by tourism scholars. Scholars have paid attention to the development of artificial intelligence (AI) and the fresh blood to tourism demand forecasting research [52]. Numerous studies have confirmed the predictive effectiveness of search-engine-query-volume data in social and economic domains [13,29]. Tourism-flow-affected factors, such as economic indicators, climate/weather index, search engines, and market sentiment, are incorporated into the forecasting models as explanatory variables to improve the accuracy [37,53]. Referring to research in hospitality and tourism, the pre-trip behavior of tourists reflected by SVI is a major concern [54]. The study of tourist search behavior found that tourists typically make travel decisions at least two to three weeks in advance, which is more accurate than the seasonal pattern uncovered in the previous study [1,55].

The advantage of real-time and regional search data is gradually taken into account, which improves daily tourist forecasting [56]. Google or Baidu Trends, as forecasts with a large number of users, were adopted to predict tourist volumes or hotel occupancy for various scales of destination area in an approaching time (seasonally, weekly, and daily) [28,57,58]. In particular, the space-time features and platform bias of SVI are worth
the further detailed study, which could be used to correct and aggregate SVI for preferable forecasting on a short-time scale [19].

3. Methodology

To eliminate spurious regression in regression analysis, we used the augmented Dickey-Fuller (ADF) test method in this study to examine the stability and co-integration of time-series data. Furthermore, as a statistical method for testing causality between time series, this study used Granger causality tests to investigate the relationship between tourist volume and SVI in each city and demonstrate the causal relationship in the temporal dimension. Finally, the impulse response function was utilized to examine the dynamic influence of SVI on the tourism flow in each city by creating a vector autoregression model (VAR) with each city’s tourist arrivals and SVI data. VAR is a multivariate time series analysis model that is used to describe the linkage relationship between multiple variables. The effect of random perturbations from a standard deviation of scenic spot SVI on a dependent variable of visitor arrivals was observed with greater precision. More precisely, the effect of random perturbations to the standard deviation of a scenic spot’s SVI on the dependent variable of visitor arrivals was identified. To investigate the influence of SVI on the spatial distribution characteristics of tourist flow in the precursory days, we examined the lag of impulse response peak (LRP) in the function graphs of various cities, which refers to the time delay between the point of application of an impulse input signal and the peak of the output response signal, and it is an important parameter used to characterize the behavior of a system and to design control algorithms. This study used LRP to indicate the most significant days for the number of tourist lags behind the SVI.

Following that, we built SVI data to forecast the world heritage site’s tourist flow. To evaluate and compare the prediction model’s performance, we first examined the variable of the tourist flow time series. Then, we used historical data from the Yangtze River Delta’s tourist flow time series to anticipate tourist flow using an autoregressive moving average model (ARMA), which represents time series data as a linear combination of autoregressive terms and moving average terms, taking into account both the historical data of the time series itself and the influence of random error terms. The SVI data were introduced as an explanatory variable to the tourist flow time series prediction, and a VAR prediction model was developed. VAR, which has been widely employed in the economics and tourism literature, treats time series as endogenous in a dynamic process [53,59,60]. The above methods mainly used Arcgis and Stata for data analysis.

We selected cities that have the same response delay time as one type based on the results of the spatial-temporal divergence between online search and visitor flow and developed a bivariate VAR prediction model using tourist volume and SVI for each city [1,61]. The tourist flow variable is denoted by KL-n (n denotes the response lag horizon). For example, the total city tourist flow in ‘response lag period two’ is denoted by KL-2, and the total city tourist flow in ‘response lag period three’ is denoted by KL-3. Similarly, each city’s total tourist volume was computed by adding the forecasted outcomes for each city’s tourist flow (variable ZY). The SVI variable represents the SVI of the city’s scenic area in which the response lags. It is denoted by BD-n (n denotes the response lag period). For example, BD-2 denotes the city’s SVI in ‘response lag period two’ and ‘response lag period three’ by BD-3. As a result, the following estimation equation is derived:

\[ Y_t = A_1Y_{t-1} + \ldots + A_nY_{t-n} + \epsilon_t; \ Y_t = [KL-n \ BD-n] \]

where \(n\) denotes the response lag period for the association between various categories of tourist volume and SVI; \(KL-n\) and \(BD-n\) denote the response lag period for tourist flow between source city and SVI, respectively. The total tourist flow \(VF_{var}\) is equal to the sum of the expected values for the \(n\) types of source cities listed above, \(Y_i\). That is,

\[ VF_{var} = \sum_{i=1}^{n} Y_{ti}, \]

where \(n\) denotes the response lag period for the association between various categories of tourist volume and SVI; \(KL-n\) and \(BD-n\) denote the response lag period for tourist flow between source city and SVI, respectively. The total tourist flow \(VF_{var}\) is equal to the sum of the expected values for the \(n\) types of source cities listed above, \(Y_i\). That is,
where $ZY$ denotes the total tourist volume from the scenic spot’s source cities, $Y_t$ denotes the expected value of each city’s tourist volume, and $n$ denotes the response lag period type.

4. Empirical Results

4.1. Data Resource

At the end of June 2017, China’s Ministry of Industry and Information Technology reported that the country had 1.36 billion smartphone subscribers. That is, nearly every Chinese individual owns a cell phone. The collection of large amounts of data from search engines and mobile devices in every city expands the possibilities for additional research and more practical applications.

Zhongshan Scenic Area is located in the Chinese city of Nanjing, Jiangsu Province (Figure 1). It is home to two world-famous attractions: the Ming Xiaoling Mausoleum, a UNESCO World Heritage Site, and Sun Yat-sen’s Mausoleum. Its primary long-term reliable supply of tourists is the Yangtze River Delta’s 40 cities (local users were excluded). Zhongshan attracted 10.8 million tourists from the Yangtze River Delta during the study period (1 April 2016 to 31 April 2017), accounting for 84.6 percent of total tourists.

![Figure 1. The location of the Yangtze River Delta and Zhongshan Scenic Area, Nanjing, China.](image-url)

4.1.1. Online Search Keywords

Baidu (www.baidu.com) is China’s most popular search engine. Baidu Index, similar to Google Trends, is a critical record of Internet users’ search volume on Baidu. It utilizes frequency weighting to determine the popularity of each keyword on the Baidu platform. It reflects the actual volume of specific searches, with no upper limit and consistent processing, unlike Google’s 0–100 trend processing [62]. In prior comparative research, the Baidu index was deemed to be more appropriate for China due to its larger market share in the Chinese language [63]. The Baidu Index can rapidly and accurately display online TIS fluctuation...
for various cities. Thus, the SVI can readily be adjusted to match the daily tourist flow in various cities.

We used the destination-related keywords from tourists in the cities of the Yangtze River Delta in the Baidu Index “https://index.baidu.com (accessed on 3 June 2017)” as tourist SVI. When picking keywords for the search, statistic objects connected to the Zhongshan Scenic Area’s primary attractions were employed. The benchmark keywords were “Nanjing Sun Yat-sen Mausoleum”, “Zhongshan Mausoleum Ticket”, “Ming Xiaoling Mausoleum”, “Sun Yat-sen Mausoleum”, and “Spirit Valley Temple” (all in Chinese). These were the five most frequently used keywords connected to the study site. Then, we extracted each term associated with the Zhongshan Scenic Area from the Baidu Index between 1 April 2016 and 31 April 2017 and combined them to create the overall index for each city. They were denoted by the abbreviations ‘province-city initials + BD’ (for example, Hangzhou SVI recorded as ZJ-HZBD, Huzhou SVI recorded as ZJ-hzBD, and Shanghai SVI as SHBD). The study assumed that travelers search for a large number of keywords in a concentrated manner before a trip and thus aggregated five keywords as a complete indicator.

4.1.2. Tourist Volume Data

Existing data on tourist volume research are primarily derived from two sources. Firstly, international organizations or local tourism bureaus’ statistical yearbooks are consulted. Unfortunately, these statistics are deficient in relevance, specificity, and timeliness. Secondly, tourism researchers develop survey programs to meet specific study objectives. However, there are shortcomings, such as a small sample size that cannot accurately reflect the unique tourist conditions when gathered by sampling surveys and other ways.

To solve both of these data gaps, we collected data using a new system called the Nanjing Smart Tourism System. The Nanjing Smart Tourism System, which was developed in collaboration with telecom operators, is capable of collecting daily tourist data for the Zhongshan Scenic Area. It utilizes geolocation data from tourists’ mobile phones to track and record their trips to the scenic area. As a result, these data are large-scale, with a high degree of relevance and timeliness. From 1 April 2016 to 31 April 2017, we obtained the real-time daily number of tourists visiting the Zhongshan Scenic Area from cities in the Yangtze River Delta (including Anhui province, Jiangsu province, Zhejiang province, and Shanghai City). They are denoted by ‘province-city initials + Y’ (e.g., the Hangzhou tourist volume is recorded as ZJ-HZY, Huzhou tourist volume as ZJ-hzY, and Shanghai tourist volume as SHY).

4.1.3. Distribution Characteristics

The population of a city in the Yangtze River Delta can have an effect on tourism statistics. The relative tourist arrivals from the source cities were calculated by normalizing the tourist volume and population of each city. This information highlighted the travelers’ geographical characteristics (Figure 2). Tourists are primarily drawn from surrounding cities, and the quantity of tourists in a given space is often diminished by geographical distance.

The SVI of tourists in 40 cities has various characteristics, and the number of online searches is first determined by the regional internet user population [61]. Assuming that other factors remained constant, the standardized processing of SVI was performed on the result of the daily SVI value of the Zhongshan Scenic Area divided by the number of city Internet users in each city of the Yangtze River Delta (the number of city Internet users = city population* network penetration rate) (Figure 3). Along with the geographic attenuation of tourism network search volume, it demonstrated the distribution characteristics, with Shanghai, Hefei, and Hangzhou as the core cities.
Figure 2. Spatial distribution of tourist arrivals to the Zhongshan scenic area in the Yangtze River Delta.
4.2. Spatial Disparity of SVI Precursory Effects

Firstly, the ADF test was performed to verify daily tourist data and index data from each city. Each city’s SVI is zero-order and monostable, allowing for the application of the Granger causality test. Due to the Granger causality test’s increased sensitivity to the lag period chosen, the optimal lag order is typically calculated using methods such as...
the Akaike Information Criterion (AIC) and Schwarz Criterion. Table A1 contains the test results.

When analyzing whether the daily SVI was a Granger cause of tourist volume, all test findings rejected the null hypothesis that the daily SVI may be regarded as a forerunner to the shift in the daily tourist amount. When analyzing whether the daily tourist volume was the Granger cause of the SVI of the day, the test findings of eight cities contradicted the initial hypothesis, showing that the daily tourist quantity was the Granger cause of the SVI. The daily online search in this type of city could affect the real number of tourists, and the number of daily tourists can also cause changes in the web search. The two showed mutual correlation and causation.

Using the tourist flow from each city as a dependent variable, a standard deviation of random disturbances was applied by corresponding SVI [60]. The peak period in the impulse response figure indicates the precursory duration of SVI (Figure 4). The peak period of visitor volume response to the SVI was 2–8 days in the 40 cities of the Yangtze River Delta region, showing that tourists generally search for tourist information 2–8 days before arriving at the picturesque place. For example, tourists from 12 cities in Jiangsu Province search for information 2–4 days before arriving at the picturesque area. The antecedent time is shorter than that of tourists from other provinces. Tourists from Zhejiang Province often search for tourist information 3–8 days before arriving at the scenic area.

Figure 4. The relationship between driving distance and the lag of impulse peak tourist arrival response to search volume index from 40 cities.

The spatial distribution map of online tourist search lead days (Figure 5) revealed a clear dispersion of urban visitors within 150 km of Nanjing, including Maanshan, Changzhou, and Wuhu in Anhui Province, and Zhenjiang and Yangzhou in Jiangsu Province. Cities 200–250 km apart have provincial-level utility impacts, and the distance affects the sensitivity of the online search time from the province’s source cities. The sensitivity is lower than that of the province’s source cities. For urban travelers, the pre-trip search period of 250–600 km increases with distance. The distance between cities in Zhejiang Province is greater than the distance between cities in Nanjing, and the search time for tourists is between four and eight days.
To further demonstrate how tourist pre-trip seeking features vary with distance, we considered simply the time of arrival of each city in Nanjing as an independent variable (designated as \( X \)) and the number of response periods (designated as \( Y \)) in each city. A linear regression equation was developed for the correlation (where \( R = 0.992 \), the probability of the \( F \) statistic is less than 0.05, and the model is significant):

\[
\ln(Y) = -2.2685 + 0.0656\ln(X) \quad (3)
\]

According to Equation (3), traffic time has a considerable positive effect on the duration of the pre-trip search, with an elasticity coefficient of 0.0656. As a unit’s traffic time increases, so does the search time, which indicates that potential tourists must collect information for an extended period before arriving at the picturesque place.

**Figure 5.** Spatial distribution of the lag of impulse peak tourist arrival response to search volume index from 40 cities.

To further demonstrate how tourist pre-trip seeking features vary with distance, we considered simply the time of arrival of each city in Nanjing as an independent variable (designated as \( X \)) and the number of response periods (designated as \( Y \)) in each city. A linear
A regression equation was developed for the correlation (where $R = 0.992$, the probability of the $F$ statistic is less than 0.05, and the model is significant):

$$\ln(Y) = -2.2685 + 0.0656\ln(X) \quad (3)$$

According to Equation (3), traffic time has a considerable positive effect on the duration of the pre-trip search, with an elasticity coefficient of 0.0656. As a unit’s traffic time increases, so does the search time, which indicates that potential tourists must collect information for an extended period before arriving at the picturesque place.

We observed that SVI, as a precursor, varies with distance but not entirely with driving distance. Zhongshan, Xuzhou, and Suzhou in the Yangtze River Delta’s source markets have only high-speed rail connections to Nanjing in the northern cities (the speed is about twice that of self-driving cars). Soochow, Wuxi, Shanghai, and Hangzhou are hub cities in the Yangtze River Delta tourism traffic network. Transit between these cities and Nanjing is more convenient than transportation between equidistant cities. The lead time of tourist search activity does not alter precisely as physical distance increases but rather represents the integrated distance. We noticed that SVI as a precursor is different for distance but is not entirely for driving distance.

4.3. Application Verification of Precursory Effects

We optimized the prediction of daily tourist volume using the precursory effects and their spatial features mentioned earlier. In this process, we compared the spatial feature optimized VAR model with the traditional ARIMA model.

4.3.1. ARIMA Model

According to the ADF tests, the tourist volume sequence lacked a unit root and was a smooth time series. Thus, the ARIMA($p,d,q$) model’s parameter ‘$d$’ was set to 0. Due to trailing phenomena in the autocorrelation function (ACF) and partial correlation function (PACF) diagrams of the tourist volume sequence, the ARMA ($p,q$) model was constructed. We increased the lag duration to obtain a more accurate prediction model and compared the AIC, SC, and residual sequence correlation of $p$ and $q$ for numerous parameters. When $p = 6$ and $q = 5$, the AIC and SC values for the model were the smallest, and the residual sequence was unrelated. The best estimate model for daily tourist volume ($VF_{arma}$) using ARMA (6, 5) was built, yielding the following estimated results:

$$VF_{arma} = \sum_{i=1}^{6} AR(i)YZ(-i) + \sum_{i=1}^{5} MA(i)\epsilon(-i) + C \quad (4)$$

The goodness of fit for Equation (4) is 0.7124; the AIC is 21.561; the SC is 20.483; and the reciprocal of the AR model roots are 0.87, 0.25 $- 0.21i$, 2.25 $+ 0.21i$, and $-0.33 + 0.88i$, respectively, with $-0.33 - 0.88i$, $-0.08 - 0.54i$, and $-0.08 + 0.54i$ within the unit circle. The daily tourist volume forecast’s root mean square error (RMSE) was 5077.382 for the sample period.

4.3.2. Vector Autoregressive Model

According to the Equation, the response lag period of visitor volume to network search consists of six situations (1). We created a VAR estimating model for KL-2 and BD-2 and received the following expected results by setting the lag period to 2:

$$Y_t = [25.98 \ 0.23 \ 0.11 \ 1.53 \ Y_{t-1} + [ -7.52 \ 0.18 \ 0.37 \ 20.73 \ Y_{t-2} + \epsilon_t ] \quad (5)$$

$$Y_t = [KL-2 \ BD-2 ]$$

$$Y_t = [KL-2 \ BD-2 ] \quad (6)$$
Then, using a VAR estimating model for KL-3 and BD-3 and a lag period of 3, we obtained the estimated values as follows:

\[
Y_t = \begin{bmatrix} 45.76 & 1.44 & -0.22 & 15.23 \\ -11.12 & 30.58 & -4.02 & 25.26 \end{bmatrix} Y_{t-1} + \begin{bmatrix} 23.47 & -0.33 & 4.92 & 12.62 \\ -11.12 & 30.58 & -4.02 & 25.26 \end{bmatrix} Y_{t-2} + \epsilon_t ;
\]

\[
Y_t = [KL-3 \ BD-3]
\]

We constructed daily tourist volume estimate formulae for cities that lag behind by 4, 5, 6, and 8 days and then added the tourist volume from each source city in Zhongshan to obtain the total estimated tourist volume (VF\text{var}). After forecasting tourist volume using Equations (1) and (2), the root mean square error (RMSE) of the prediction result was determined to be 4158.382, which is less than the ARMA model. The tourist volume outside the sample in December 2017 was forecast to further assess the VAR model’s forecasting power. Figure 6 illustrates the prediction results and the actual tourist volume. V is the actual number of tourists, while VF denotes the ARMA and VAR model forecast results, respectively. The VAR model’s predicted value was closer to the actual value, and the RMSE of the forecasted data for the inspection period was 3414.48, which is less than the ARMA model without SVI (4271.85). Therefore, the VAR model optimized based on spatiotemporal patterns is generally more predictable.

\[
\begin{align*}
V_{\text{actual}} & \quadVF_{\text{arma}} \quadVF_{\text{var}} \\
\end{align*}
\]

Figure 6. Comparison of the actual number of tourists (V) with the forecast number in the vector autoregression (VF\text{var}) and autoregressive moving average models (VF\text{arma}) in December 2017.

5. Conclusions

Cyberspace interacts with physical worlds on a spatial and material level [64]. On the internet, time and distance do not vanish but are recreated and reflected in geocyberspace [65]. This study examined online information search behavior by examining the spatial-temporal interactions between the pre-trip and post-arrival stages. The spatial-temporal link between multivariate models was investigated. In comparison to a previous study on the characteristics of China’s domestic pre-tour searching behavior [48,54], this paper quantified the length of tourist flow lags as distance increased for different cities ranging from 1 to 8. By investigating the spatial-temporal interactions between the pre-trip stage and the after-arrival stage, this paper uncovered the temporal and spatial deviations of pre-trip searching and validated their potential application in tourism forecasting, making an original contribution to the study of online information search behavior.

The contributions of this study are manifold. Predictability and interpretability were the focus of this study. Firstly, we attempted to integrate time and space into a tourism...
demand forecasting model. Few attempts have been made to deduce the information search flow’s time-space features. Previous research has utilized this indicator to forecast destinations. However, this method ignores the tourism market’s internal spatial-temporal variation [57]. Using cellular signaling monitoring and search engine data from big data platforms, we investigated the temporal and spatial differentiation of visitors’ TIS. A time and accuracy-optimized approach was proposed for forecasting tourist volume from many sources. Compared to the existing research results, the model constructed in this paper has wider usability.

Secondly, our research proposed a new method for predicting daily visitor numbers at tourist attractions based on accurate data from mobile roaming data and search engines. At the same time, we integrated a variety of research methods, making the research results more general and scalable and allowing for more accurate tourism forecasting models. We used the Zhongshan Scenic Area in Nanjing as a case study to examine daily tourist flow and SVI data in 40 Yangtze River Delta source cities over 640 days. The Granger causality test, the VAR model, and the impulse response function were used to determine the geographical distribution disparities between nearby source cities, web search data, and the spatiotemporal connection between tourist flow and SVI. Finally, using these correlations, we used ARMA and VAR models to forecast tourist arrivals in a scenic location. We finish by saying the following:

First, tourist volume and SVI from source cities demonstrate attenuation due to distance. There is a direct correlation between tourist volume and SVI daily statistics. That is to say, SVI is predictive of changes in visitor numbers. According to Buhalis, search engines are the first choice for families to collect information when making travel plans, and SVIs can be used as an important indicator for the destination prediction of tourist behavior [66]. Among the source area, tourists from major source cities are the main predictive of SVI.

Second, when traffic time increases spatially, the precursory time for tourists’ online searches increases. The SVI results on the lead time for changes in tourist flow can be interpreted as the time required for tourists to gather knowledge prior to traveling in order to acquire various search behavior characteristics based on spatial cognition. For example, tourist search query data about “hotels and flights” can significantly improve the forecast accuracy of local tourist numbers [29]. Additionally, it demonstrates that travelers traveling across short distances spend less time preparing for information than long distances.

Finally, the prediction model developed using the spatial-temporal differentiation of search features exhibits high predictability. The forecasting model for scenic tourist flow that incorporates SVI is more accurate than the classic time series forecasting model that does not incorporate SVI. It is capable of estimating the inter-daily tourist volume, which is critical for monitoring and managing the tourist flow in the scenic site.

Thus, when monitoring the tourist volume, the tourist destination can choose the best way to control the tourist volume and disseminate information according to the tourist search behavior pattern in the past [67]. At the same time, travel agencies and destination management organizations (DMOs) should narrow the gap between travel intentions and actual arrivals to improve the destination’s capacity to manage tourist flows and promote the healthy and sustainable development of destinations. These insights can be used to improve the accuracy of tourism advertisements and their efficacy. In terms of future research implications, we can select additional prediction models based on the particular spatial-temporal correlations between tourist volume and SVI at the city and daily scales, hence optimizing forecast accuracy rather than at the country and monthly scales. Additionally, a prospective tourist from a variety of cities may be interested in a variety of factors during their pre-trip research, such as scenery or service facilities. Spatial and temporal differentiation is a vital task that requires further investigation.
6. Limitations

The regional disparities in online search prior to tourist visits, on the other hand, show the statistical association between overall search results and visitor volume in the research. Although we analyzed the tourist flow and SVI data among 40 cities in the Yangtze River Delta, it is possible to continue investigating the longer distance source or even international tourist flow. In addition, only the effect of the traffic time variable on the relationship between SVI and tourist flow was studied to investigate the influencing elements of the temporal coupling relationship between SVI and tourist flow.

Moreover, in the post-epidemic era, the importance of web search has been revealed. Thus, the spatial-temporal pattern alterations in the post-COVID-19 era must be monitored in the future.

Author Contributions: Conceptualization, P.L. and J.Z. (Jianxin Zhang); methodology, L.M.; data curation, P.L.; writing—original draft preparation, L.M.; writing—review and editing, P.L. and J.Z. (Jinyue Zhang); visualization, Y.Y.; supervision, P.L.; project administration, P.L.; funding acquisition, P.L. and J.Z. (Jianxin Zhang). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Award No. 42001145 to Peixue Liu) and MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Award No. 22YJA760106 to Jianxin Zhang).

Institutional Review Board Statement: This study does not involve human subjects or animals. Therefore, no Institutional Review Board approval was required.

Informed Consent Statement: All participants provided informed consent prior to their participation in this study. They were provided with detailed information about the study objectives, procedures, potential risks, and benefits, and they voluntarily agreed to participate.

Data Availability Statement: Not available to the public. The data presented in this study are available only on reasonable request from the corresponding author.

Acknowledgments: The authors would like to appreciate the anonymous reviewers for their insightful comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Granger causality test results.

<table>
<thead>
<tr>
<th>Granger Causality</th>
<th>F</th>
<th>p Value</th>
<th>Granger Causality</th>
<th>F</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>JS-SZBD/ → JS-SZY</td>
<td>3.0242</td>
<td>0.0264</td>
<td>JS-TZY/ → JS-TZBD</td>
<td>2.3728</td>
<td>0.1817</td>
</tr>
<tr>
<td>JS-SZBD/ → JS-SZY</td>
<td>3.8721</td>
<td>0.0305</td>
<td>JS-TZBD/ → JS-TZY</td>
<td>3.3643</td>
<td>0.0394</td>
</tr>
<tr>
<td>JS-WXY/ → JS-WXBD</td>
<td>2.3936</td>
<td>0.0277</td>
<td>JS-SQY/ → JS-SQBD</td>
<td>2.4824</td>
<td>0.1134</td>
</tr>
<tr>
<td>JS-WXBD/ → JS-WXY</td>
<td>2.4024</td>
<td>0.0219</td>
<td>JS-SQBD/ → JS-SQY</td>
<td>1.7216</td>
<td>0.0341</td>
</tr>
<tr>
<td>JS-ZJY/ → JS-ZJBD</td>
<td>1.9947</td>
<td>0.1313</td>
<td>JS-YCY/ → JS-YCBD</td>
<td>1.9985</td>
<td>0.0972</td>
</tr>
<tr>
<td>JS-ZJBBD/ → JS-ZJY</td>
<td>3.6285</td>
<td>0.0211</td>
<td>JS-YCBD/ → JS-YCY</td>
<td>3.3845</td>
<td>0.0133</td>
</tr>
<tr>
<td>JS-CZY/ → JS-CZBD</td>
<td>1.3323</td>
<td>0.1633</td>
<td>JS-HAY/ → JS-HABD</td>
<td>1.4735</td>
<td>0.07937</td>
</tr>
<tr>
<td>JS-CZBD/ → JS-CZY</td>
<td>3.3824</td>
<td>0.0216</td>
<td>JS-HABD/ → JS-HAY</td>
<td>3.2678</td>
<td>0.0482</td>
</tr>
<tr>
<td>JS-NTY/ → JS-NTBD</td>
<td>1.7834</td>
<td>0.0749</td>
<td>JS-LYGY/ → JS-LYGBD</td>
<td>2.5571</td>
<td>0.1113</td>
</tr>
<tr>
<td>JS-NTBD/ → JS-NTY</td>
<td>3.6296</td>
<td>0.0137</td>
<td>JS-LYGBD/ → JS-LYGY</td>
<td>3.4737</td>
<td>0.0081</td>
</tr>
<tr>
<td>JS-YZY/ → JS-YZBD</td>
<td>3.3721</td>
<td>0.0831</td>
<td>JS-XZBDB/ → JS-XZBD</td>
<td>1.0890</td>
<td>0.0213</td>
</tr>
<tr>
<td>JS-YZBD/ → JS-YZY</td>
<td>4.0013</td>
<td>0.0125</td>
<td>JS-XZBDB/ → JS-XZY</td>
<td>2.9386</td>
<td>0.0050</td>
</tr>
<tr>
<td>AH-CZY/ → AH-CZBD</td>
<td>4.1127</td>
<td>0.0721</td>
<td>AH-HNBD/ → AH-HNB</td>
<td>4.2078</td>
<td>0.0538</td>
</tr>
<tr>
<td>AH-CZBD/ → AH-CZY</td>
<td>4.3382</td>
<td>0.0031</td>
<td>AH-HNBBD/ → AH-HNY</td>
<td>4.3681</td>
<td>0.0120</td>
</tr>
<tr>
<td>AH-MASY/ → AH-MASBD</td>
<td>4.3297</td>
<td>0.1772</td>
<td>AH-HBBBD/ → AH-HBBD</td>
<td>4.2976</td>
<td>0.0900</td>
</tr>
<tr>
<td>AH-MASBD/ → AH-MASY</td>
<td>4.4019</td>
<td>0.0043</td>
<td>AH-HBBBD/ → AH-HBY</td>
<td>4.3174</td>
<td>0.0087</td>
</tr>
<tr>
<td>AH-WHY/ → AH-WHBD</td>
<td>3.9782</td>
<td>0.1522</td>
<td>AH-BZBBD/ → AH-BZBD</td>
<td>4.3933</td>
<td>0.1103</td>
</tr>
</tbody>
</table>
Table A1. Cont.

<table>
<thead>
<tr>
<th>Granger Causality</th>
<th>F</th>
<th>p Value</th>
<th>Granger Causality</th>
<th>F</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AH-WHBD → AH-WHY</td>
<td>4.1973</td>
<td>0.0381</td>
<td>AH-BZBD → AH-BZY</td>
<td>3.8962</td>
<td>0.0064</td>
</tr>
<tr>
<td>AH-HFY/D → AH-HFBD</td>
<td>4.3215</td>
<td>0.0226</td>
<td>AH-FY/D → AH-FY/D</td>
<td>3.8586</td>
<td>0.1074</td>
</tr>
<tr>
<td>AH-HFBD/D → AH-HFY</td>
<td>3.7829</td>
<td>0.0097</td>
<td>AH-FY/D → AH-FY/D</td>
<td>4.6274</td>
<td>0.0048</td>
</tr>
<tr>
<td>AH-XCY/D → AH-XCBY</td>
<td>3.7751</td>
<td>0.1030</td>
<td>AH-AQY/D → AH-AQY/D</td>
<td>3.9856</td>
<td>0.0937</td>
</tr>
<tr>
<td>AH-XCBY/D → AH-XCY</td>
<td>4.3970</td>
<td>0.0042</td>
<td>AH-AQY/D → AH-AQY/D</td>
<td>3.8775</td>
<td>0.0219</td>
</tr>
<tr>
<td>AH-BBY/D → AH-BBBD</td>
<td>3.5823</td>
<td>0.0830</td>
<td>AH-LAY/D → AH-LABD</td>
<td>4.0294</td>
<td>0.1143</td>
</tr>
<tr>
<td>AH-BBBD/D → AH-BBY</td>
<td>3.8952</td>
<td>0.0052</td>
<td>AH-LABD/D → AH-LAY</td>
<td>4.1395</td>
<td>0.0005</td>
</tr>
<tr>
<td>AH-SZY/D → AH-SZBD</td>
<td>4.0173</td>
<td>0.0077</td>
<td>AH-CZY/D → AH-CZBD</td>
<td>3.9402</td>
<td>0.1047</td>
</tr>
<tr>
<td>AH-SZBD/D → AH-SZY</td>
<td>4.1196</td>
<td>0.0036</td>
<td>AH-CZBD/D → AH-CZY</td>
<td>3.6844</td>
<td>0.0031</td>
</tr>
<tr>
<td>AH-TLY/D → AH-TLBD</td>
<td>3.8925</td>
<td>0.0823</td>
<td>AH-HS/D → AH-HSBD</td>
<td>4.1186</td>
<td>0.2901</td>
</tr>
<tr>
<td>AH-TLBD/D → AH-TLY</td>
<td>4.3168</td>
<td>0.0271</td>
<td>AH-HSBD/D → AH-HS</td>
<td>4.3921</td>
<td>0.0048</td>
</tr>
<tr>
<td>ZJ-JHZY/D → ZJ-HZBD</td>
<td>3.0121</td>
<td>0.0038</td>
<td>ZJ-JHY/D → ZJ-HJB/</td>
<td>3.0642</td>
<td>0.1031</td>
</tr>
<tr>
<td>ZJ-HZBD/D → ZJ-HZY</td>
<td>3.0793</td>
<td>0.0029</td>
<td>ZJ-HJB/D → ZJ-JHY</td>
<td>3.9941</td>
<td>0.0002</td>
</tr>
<tr>
<td>ZJ-hzY/D → ZJ-hzBD</td>
<td>2.9947</td>
<td>0.1125</td>
<td>ZJ-NBY/D → ZJ-NBBD</td>
<td>2.8749</td>
<td>0.0036</td>
</tr>
<tr>
<td>ZJ-hzBD/D → ZJ-hzY</td>
<td>3.1185</td>
<td>0.0038</td>
<td>ZJ-NBBD/D → ZJ-NBY</td>
<td>2.9951</td>
<td>0.0047</td>
</tr>
<tr>
<td>ZJ-JXY/D → ZJ-JXBD</td>
<td>3.0275</td>
<td>0.0921</td>
<td>ZJ-ZSY/D → ZJ-ZSBD</td>
<td>3.0625</td>
<td>0.0722</td>
</tr>
<tr>
<td>ZJ-JXBD/D → ZJ-JXY</td>
<td>2.9952</td>
<td>0.0089</td>
<td>ZJ-ZSBD/D → ZJ-ZSY</td>
<td>3.0431</td>
<td>0.0311</td>
</tr>
<tr>
<td>ZJ-SXY/D → ZJ-SXBD</td>
<td>2.9875</td>
<td>0.0802</td>
<td>ZJ-TZY/D → ZJ-TZBD</td>
<td>2.9742</td>
<td>0.0814</td>
</tr>
<tr>
<td>ZJ-SXBD/D → ZJ-SXY</td>
<td>3.0174</td>
<td>0.0044</td>
<td>ZJ-TZBD/D → ZJ-TZY</td>
<td>2.9873</td>
<td>0.0207</td>
</tr>
<tr>
<td>ZJ-QZY/D → ZJ-QZBD</td>
<td>3.0293</td>
<td>0.0723</td>
<td>ZJ-WZY/D → ZJ-WZBD</td>
<td>3.0616</td>
<td>0.0563</td>
</tr>
<tr>
<td>ZJ-QZBD/D → ZJ-QZY</td>
<td>2.9981</td>
<td>0.0117</td>
<td>ZJ-WZBD/D → ZJ-WZY</td>
<td>2.8979</td>
<td>0.0081</td>
</tr>
<tr>
<td>ZJ-LSY/D → ZJ-LSBD</td>
<td>3.0783</td>
<td>0.0824</td>
<td>SHY/D → SHBD</td>
<td>3.7826</td>
<td>0.0041</td>
</tr>
<tr>
<td>ZJ-LSBD/D → ZJ-LSY</td>
<td>2.9433</td>
<td>0.0307</td>
<td>SHBD/D → SHY</td>
<td>3.5741</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

Notes: The null hypothesis test, H0: A B, implies ‘it does not cause yt’.

References
1. Yang, X.; Pan, B.; Evans, J.A.; Lv, B. Forecasting Chinese tourist volume with search engine data. Tour. Manag. 2015, 46, 386–397. [CrossRef]
14. Liu, A.; Lin, V.S.; Li, G.; Song, H. Ex Ante Tourism Forecasting Assessment. J. Travel Res. 2020, 61, 64–75. [CrossRef]
16. Hu, M.; Song, H. Data source combination for tourism demand forecasting. Tour. Econ. 2019, 26, 1248–1265. [CrossRef]
24. Gunter, U.; Oender, I. Forecasting international city tourism demand for Paris: Accuracy of uni- and multivariate models employing monthly data. Tour. Manag. 2015, 46, 123–135. [CrossRef]
32. Huertas, A.; Miguel, O.-M. Do tourists seek the same information at destinations? Analysis of digital tourist information searches according to different types of tourists. Eur. J. Tour. Res. 2022, 32, 3211. [CrossRef]
33. Xie, G.; Li, X.; Qian, Y.; Wang, S. Forecasting tourism demand with KPCA-based web search indexes. Tour. Econ. 2021, 27, 721–743. [CrossRef]
38. Bi, J.-W.; Han, T.-Y.; Yao, Y. Collaborative forecasting of tourism demand for multiple tourist attractions with spatial dependence: A combined deep learning model. Tour. Econ. 2023. [CrossRef]
42. Claveria, O.; Torra, S. Forecasting tourism demand to Catalonia: Neural networks vs. time series models. Econ. Model. 2014, 36, 220–228. [CrossRef]
47. Jiao, E.X.; Chen, J.L. Tourism forecasting: A review of methodological developments over the last decade. Tour. Econ. 2019, 25, 469–492. [CrossRef]
51. Gu, X.; Sheng, L.; Lei, C. Specialization or diversification: A theoretical analysis for tourist cities. Cities 2022, 122, 1381–1384. [CrossRef]


56. Camacho, M.; Jose Pacce, M. Forecasting travellers in Spain with Google’s search volume indices. Tour. Econ. 2018, 24, 434–448. [CrossRef]


58. Li, X.; Law, R. Forecasting Tourism Demand with Decomposed Search Cycles. J. Travel Res. 2020, 59, 52–68. [CrossRef]


61. Yao, L.; Ma, R.; Wang, H. Baidu index-based forecast of daily tourist arrivals through rescaled range analysis, support vector regression, and autoregressive integrated moving average. Alex. Eng. J. 2021, 60, 365–372. [CrossRef]


63. Xie, G.; Qian, Y.; Wang, S. Forecasting Chinese cruise tourism demand with big data: An optimized machine learning approach. Tour. Manag. 2021, 82, 104208. [CrossRef]


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.