Received: 20 January 2023 | Revised: 14 April 2023 | Accepted: 27 May 2023

DOI: 10.1111/tgis.13085

RESEARCH ARTICLE

Transactions in GIS WILEY

Exploring the spatiotemporal relationships between search flows and travel flows

Yuzhou Chen[1] | Zhaoya Gong[2,3] | Qiwei Ma[4] | Ran Tao[1]

[1]School of Geosciences, University of South Florida, Tampa, Florida, USA

[2]School of Urban Planning & Design, Peking University Shenzhen Graduate School, Shenzhen, China

[3]Key Laboratory of Earth Surface System and Human-Earth Relations of Ministry of Natural Resources of China, Peking University Shenzhen Graduate School, Shenzhen, China

[4]Peking University Planning and Design Institute (Beijing) Co.Ltd, Beijing, China

Correspondence
Ran Tao, School of Geosciences, University of South Florida, Tampa, FL 33620, USA.
Email: rtao@usf.edu

Abstract

Numerous studies attempted to associate search engine data with travel behaviors. However, most existing studies focus on the destinations of search and travel, while ignoring the origins, which embed critical information of where the search requests were initiated and where the travelers came from. In this study, we explore the relationships between two types of intercity origin–destination flow data, namely travel flows and search flows, which, respectively, record the number of travelers and search requests from one city towards another. By comparing the two flows during holiday and non-holiday, we examine their complex spatiotemporal relationships from multiple perspectives, including time-lag effect, distance decay effect, spatial autocorrelation, network community, cities' rankings, and important factors of search and travel activities. The findings can deepen our understanding of search and travel behaviors, hence they can help decision makers to develop targeted strategies to enhance city's attractiveness, improve transportation infrastructure, and promote tourism.

1 | INTRODUCTION

Search engine data have proven to be an effective proxy for understanding travel behaviors and predicting travel demand. People enter real-time search requests into search engines like Google and Baidu on a daily, weekly, and monthly basis, resulting in a wealth of search query data (Choi & Varian, 2012). Google search data have been used to study human travel behaviors, such as forecasting tourist volume and hotel occupancy in the United States (Rivera, 2016), Italy (Emili et al., 2020), Germany (Bokelmann & Lessmann, 2019), Spain (Camacho & Pacce, 2018), and South Korea (Park et al., 2017). In China, Baidu search data have been widely used to

forecast visitor arrivals at key locations (Li et al., 2018), scenic sites (Huang et al., 2017; Li et al., 2019), and hotels (Zhang et al., 2019).

Related works that use search data to study travel behaviors and human mobility can be divided into three categories based on analytical method. First, the classic gravity model has been used for quantifying the attractiveness of cities (Guo et al., 2022), forecasting tourism demand (Orsi & Geneletti, 2013; Yin, 2020), and identifying impact factors of tourist demand (Akter et al., 2017; Xu et al., 2019). A common finding is that travel flow volumes are positively related to the popularity of destinations and inversely proportional to the travel costs, for example, time, distance, or transportation fees. Second, time-series analysis of search engine data has been proven effective in tourism demand forecasting. The auto-regression moving average (ARMA) and auto-regressive integrated moving average (ARIMA) have been widely used in predicting hotel occupancy (Pan et al., 2012; Yang et al., 2015), tourist visitation to the Forbidden City in China (Huang et al., 2017), and the number of tourists to Spain (Artola et al., 2015). Third, a trend in recent years is to use machine learning and artificial intelligence-based models for tourist activity forecasting. Support vector machine is used to predict foreign tourist visits with a high degree of accuracy, for example, international tourists visiting Indonesia (Purnaningrum & Athoillah, 2021), and artificial neural network is used for forecasting the tourism inflow from Hongkong to Macau (Hu & Song, 2020) and identifying tourist hot spots (Huang et al., 2022).

However, most studies so far focus on travel destinations only. The information about the origins, which embeds critical information of where the search requests were initiated and where the travelers came from, is often ignored in the analysis. By connecting origins and destinations, we can obtain search flow and travel flow data. The additional information provided by such flow data allows us to comprehend the full story of the search and travel behaviors. After all, both search flows and travel flows are spatial interaction (SI) phenomena. A search flow represents an exchange of information on the Internet via a search engine, while a travel flow indicates a change of someone's physical location via transportation. Analyzing flow data provides an opportunity to fully examine the spatial heterogeneity and the relationships of both types of SI phenomena. While travel flows have been widely used to study travel behaviors (Carter & Tao, 2023; Han et al., 2022; Tao & Thill, 2020; Zhou et al., 2023), only a few recent studies have exploited the value of search flows. Web search flow data have been used to quantify the cities' attractiveness and evaluate the sphere of influence of cities in China (Guo et al., 2022); to evaluate intercity migration tendency in China (Li & Xiao, 2022); to examine the information flow's spillover effects and influencing elements (Wu et al., 2021). However, the research value of search flows is still underexploited, especially their complex relationships with travel flows.

In this study, we aim to investigate the spatiotemporal relationships between search flows and travel flows using a series of exploratory spatiotemporal analyses. Our objectives encompass examining the time-lag effect to determine if individuals conduct searches before traveling; analyzing the distance decay effect to investigate the influence of distance on search and travel flows and how their correlation shifts as distance increases; evaluating spatial autocorrelation of search and travel flows to identify flow clusters in space; extracting network communities to evaluate the effect of city network structure on search and travel activities; ranking cities in the network of search and travel flows to identify distinct groups of cities with unique patterns; and performing multivariate regression to identify the key factors influencing search and travel flows. By synthesizing the findings from these exploratory spatiotemporal analyses, this study concludes with important relationships between search and travel flows. The conclusions can help decision-makers to develop targeted strategies to enhance city's attractiveness, improve transportation infrastructure, and promote tourism.

2 | DATA

We obtained search and travel flow data from Baidu Inc., China's most popular search engine. The Baidu Index, calculated based on online search and news data, reflects users' awareness and media attention over time pertaining

to certain keywords. Users' locations are determined through IP addresses, smartphone GPS coordinates, or proximity to cellular towers. Using the Baidu Index and users' locations, search flows are generated to represent the number of search activities in one region regarding keywords related to another region. Each distinct search query is recorded as a separate search flow. For example, if an individual searches for flight tickets and hotel bookings to city B and later looks up attractions on the same day, three search flows are generated. Travel flows are generated based on the change of users' locations each day. A travel flow is defined as an individual's movement from the origin to the destination. The travel flow data have already been processed by Baidu to identify the actual origin city and destination city of each trip, excluding intermediate stops such as transit cities. All travel and search flows are aggregated at the prefecture-level cities on a daily basis. Flows pertaining to subregions are also aggregated to their respective prefecture-level cities.

We collected data from two periods in 2018: the National Day holiday period from September 27 to October 11 and a non-holiday period from March 6 to 12. The National Day holiday is the second-busiest annual travel season in China, trailing the Spring Festival (Pan & Lai, 2019). Most people can take a paid leave for seven consecutive days, which allows them to participate in various activities such as tourist trips, business meetings, social events, and family reunions (Ge, 2022). This allows us to compare holiday and non-holiday travel patterns. The flows encompass 329 prefecture cities, excluding Beijing, Tianjin, Guangzhou, Chongqing, and Shanghai due to data privacy policies. Table 1 provides basic summaries of our data sets. Both travel and search flows exhibit greater distances and higher intensities during the holiday period than the non-holiday period.

Figures 1a,b depict search and travel flow volume during the holiday and non-holiday periods.[1] Both flows during holidays have a larger volume than non-holidays, and travel flow increases more than search flow during holidays. Figure 1c,d show the standardized search and travel flows during the two periods. Both search and travel flows increased significantly before the holiday and reached their peaks at the beginning of the holiday week. After the holiday, the two flows show a reverse pattern, which is consistent with the regular week.

3 | ANALYTICAL FRAMEWORK

Figure 2 presents our analytical framework. First, we calculate Spearman's rank correlation coefficient between search and travel flows to serve as the benchmark for further analysis. Then, we investigate the spatiotemporal relationships between search and travel flows through a series of exploratory spatiotemporal analyses, including evaluating the time-lag effect, distance decay effect, spatial flow autocorrelation, community network detection, rank cities' connection, and identifying influential factors. Specifically, we assess whether people search before travel by examining the time-lag effect, that is, whether search flows are followed by a corresponding volume of

TABLE 1 Basic summary of data.

	Holiday		Non-holiday	
Variables	**Travel flow**	**Search flow**	**Travel flow**	**Search flow**
Cities	329	329	329	329
OD pairs	108,288	81,383	97,220	71,621
Distance (mean)	860.4 mi	828.0 mi	823.6 mi	801.4 mi
Distance (medium)	778.9 mi	741.8 mi	745.2 mi	717.3 mi
Flow intensity (mean)	2165.3	830.7	884.7	423.1
Flow intensity (std)	16,882.2	986.8	7101.1	560.1
Flow intensity (max)	1,108,362	32,923	532,948	9911
Flow intensity (min)	1	57	1	57

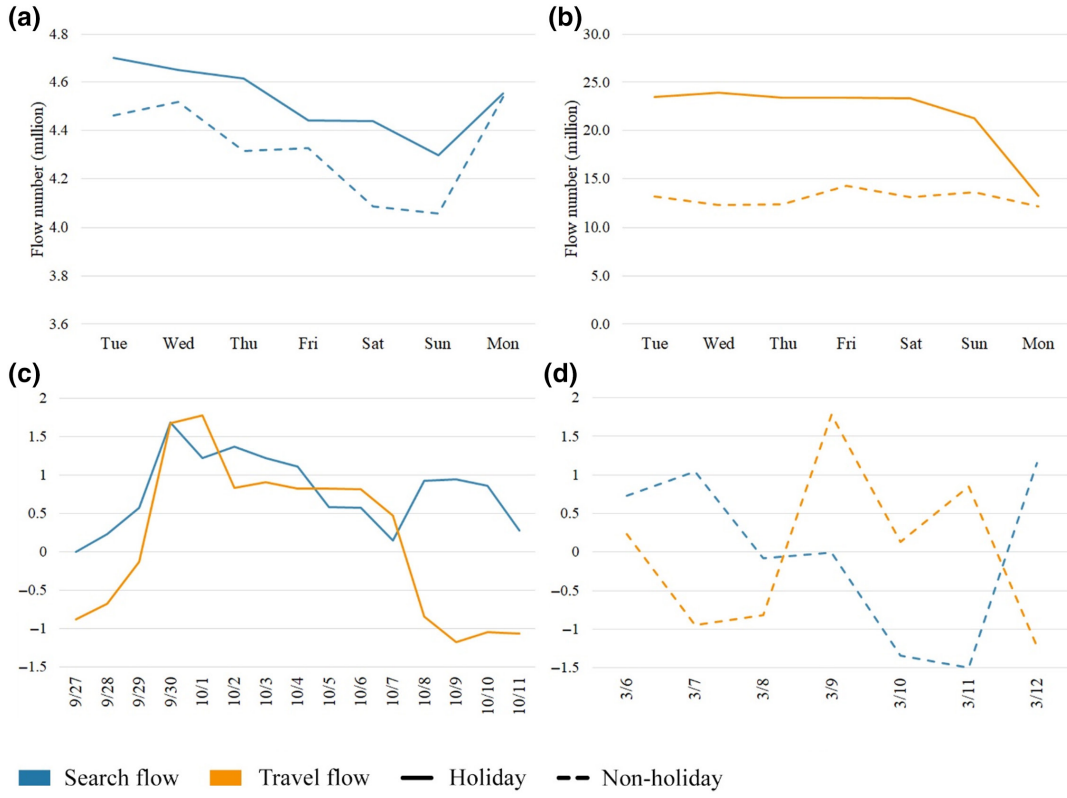

FIGURE 1 (a) Raw search flow in the holiday and non-holiday week; (b) raw travel flow in the holiday and non-holiday week; (c) standardized travel and search flow in the holiday week; and (d) standardized travel and search flow in the non-holiday week.

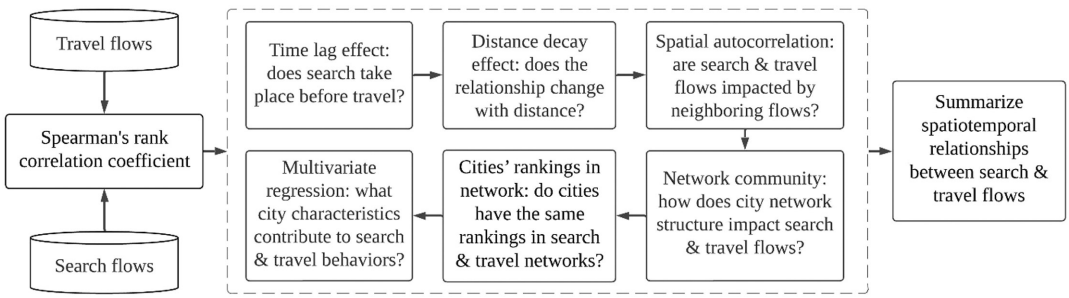

FIGURE 2 Research framework.

travel flows shortly thereafter. We then analyze distance decay to understand how distance impacts travel flows, search flows, and their correlation. After that, we evaluate the spatial autocorrelation of both flows using the FlowLISA method to determine if neighboring flows influence search and travel patterns. Next, we detect the network structure and rank cities' connections to search and travel flows to reveal the group of cities with the dense connections and each city's position in the search and travel networks. Lastly, we employ multivariate regression to identify the factors influencing search and travel flows between cities and interpret the underlying reasons for the distinct patterns we observe.

To understand how city size and type influence search and travel behaviors, we categorize all 329 cities into four groups: Big Tourist City (BTC), Big Non-Tourist City (BNTC), Small Tourist City (STC), and Small Non-Tourist

City (SNTC). The criteria for classification are based on city size and the weight of tourism industry in the local economy. Big cities include provincial capitals and sub-provincial cities, while tourist cities are those where tourism income constitutes over 30% of the gross regional product (GRP). The number of cities in each group is 6, 25, 25, and 273, respectively.

4 | RESULTS

4.1 | Correlations between search and travel flows

We calculate Spearman's rank correlation coefficients between the search and travel flows for each OD pair of cities. The results listed in Table 2 indicate that the correlations between search and travel flows are generally strong and positive. We use the correlation between search and travel flow for all cities as a benchmark for comparing the results of specific city groups. First, there are notable differences between big cities and small cities. Big cities, including both BTCs and BNTCs, exhibit stronger correlations during both holiday and non-holiday compared with small cities. In contrast, the correlations for STCs and SNTCs are both lower than the benchmark, indicating a weaker relationship between search and travel flows. Second, tourist cities have stronger correlations than non-tourist cities. BTCs exhibit the strongest correlation during both holiday and non-holiday, and STCs also exhibit stronger correlations than SNTCs. Third, when comparing holiday and non-holiday correlations for the same city group, the correlations during holidays are always stronger. There are significant differences between holiday and non-holiday for big cities, whereas the difference is moderate for small cities.

4.2 | Time-lag effect

The concept of time-lag effect is illustrated in Figure 3, where X_n is the travel flow on day n, and Y_n is the search flow on day n. If X_4, X_5, and X_6 and Y_3, Y_4, and Y_5 are highly similar, there might exist a 1-day time-lag between search and travel. The concepts of the two- and three-day time-lag effects are shown in green and orange colors, respectively. We adopt the dynamic time warping (DTW) method (Müller, 2007) to measure the similarity between time series of travel and search flow volume for each city pair. DTW is an algorithm for measuring similarity between two sequences, which may vary in time or speed. The basic idea is to align two sequences by warping the time axis to match them as well as possible. We measure the similarity multiple times for time-lag situations. If a city pair shows the highest similarity at m-day lag, we assign it to this category. We conduct a sensitivity analysis by exploring the correlation changes with time-lag between search and travel flows. The result shows that the

TABLE 2 Correlation coefficients between search and travel flows.

Category	Variable one	Variable two	Correlation (holiday)	Correlation (non-holiday)	*p*-value
All cities (AC)	All travel flow	All search flow	0.690	0.647	<2.2e−16
Big tourist cities (BTC)	To BTC travel flow	To BTC search flow	0.814	0.766	<2.2e−16
Big non-tourist cities (BNTC)	To BNTC travel flow	To BNTC search flow	0.802	0.721	<2.2e−16
Small tourist cities (STC)	To STC travel flow	To STC search flow	0.654	0.624	<2.2e−16
Small non-tourist cities (SNTC)	To SNTC travel flow	To SNTC search flow	0.625	0.608	<2.2e−16

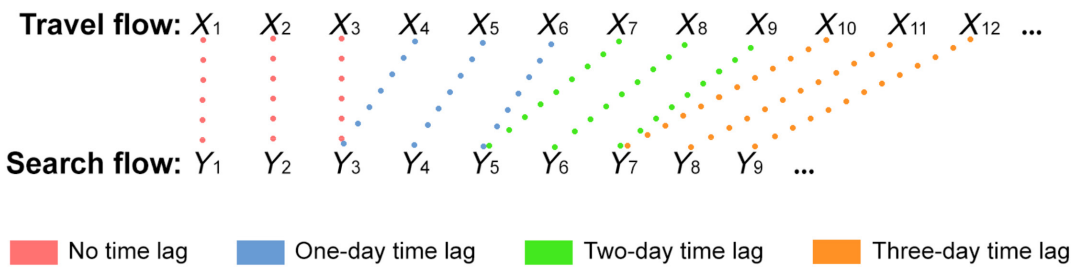

FIGURE 3 Illustration of different time lags.

correlation declines sharply after a three-day lag. Consequently, we examine and discuss the time-lag effect up to 3 days, that is, $m \in \{0, 1, 2, 3\}$.

Table 3 summarizes the results. For all city pairs during holiday, 4982 (or 30% of) city pairs are categorized as no time lag, while 14%, 15%, and 41% of city pairs belong to 1-day, 2-day, and 3-day lag. During non-holiday period, the distribution (30, 21, 19, and 30%) is significantly different from the holiday suggested by the chi-squared test. It indicates that more people tend to search the destination city 3 days ahead of their holiday travel.

For BTC during the holidays, the percentage of each respective category is 24, 14, 16, and 46%, indicating that more people tend to search their destinations 3 days ahead of their holiday travel to BTC. For BNTC during holiday, the percentage distribution (24, 11, 14, and 51%) is similar to BTC. The trend for holiday travel to big cities can be attributed to their popularity as travel destinations, as travelers need to book transportation and lodging in advance. Data show that over 37% of search and travel flow pairs are related to big cities. This high travel demand may necessitate travelers to make advanced trip planning. For both STC and SNTC, the difference between holiday and non-holiday is relatively small, indicating that advanced trip planning is not as prevalent for traveling to small cities during holiday.

4.3 | Distance decay effect

The distance decay effect refers to the SI between two locations decreasing as the distance between them increases. We calculate the distance decay effect as the coefficient of distance variable in the classic gravity model. The gravity model is one of the most commonly used SI models in population migration, travel behavior, and urban space research due to its simplicity and explicit adoption of the distance decay effect. The simplified gravity model can be expressed as Equation (1):

$$T_{ij} = kP_iP_jd_{ij}^{\beta} \tag{1}$$

where T_{ij} is the intensity between city i and city j, P_i and P_j represent the population of the two cities, and d_{ij} is the distance between them. k is the constant coefficient, and β is the distance decay factor which we will use to describe the distance decay effect of both search and travel behaviors.

Table 4 lists the results. First, it is unsurprising to find that search flows are less affected by distance than travel flows, as search is an online activity. Another study found that flows on Weibo, a popular Chinese social media platform, exhibit a similar distance decay effect of −0.331 (Wang et al., 2018). Second, both search and travel flows have a stronger distance decay effect during holidays than non-holidays. This pattern is observed for every city group. It is a counterintuitive finding as we typically would assume that people travel to and search for further destinations during holidays when they have more time to spend on road. Third, among the city groups, we found that BNTC has significantly weaker distance decay effects than the other cities. The better transportation

TABLE 3 Results of time-lag effect based on DTW.

	All cities	BTC	BNTC	STC	SNTC	All cities	BTC	BNTC	STC	SNTC
	Holiday					Non-holiday				
No time-lag	4982 (30%)	244 (24%)	1242 (24%)	290 (33%)	3206 (32%)	5812 (30%)	318 (33%)	1525 (29%)	355 (31%)	3614 (31%)
One-day lag	2405 (14%)	139 (14%)	574 (11%)	160 (19%)	1532 (16%)	4077 (21%)	198 (21%)	1187 (23%)	228 (20%)	2464 (20%)
Two-day lag	2582 (15%)	157 (16%)	705 (14%)	145 (17%)	1575 (16%)	3595 (19%)	172 (18%)	937 (18%)	213 (19%)	2273 (19%)
Three-day lag	6901 (41%)	471 (46%)	2643 (51%)	270 (31%)	3517 (36%)	5707 (30%)	266 (28%)	1597 (30%)	344 (30%)	3500 (30%)
Chi-square		0.0006	<2.2e−16	<2.2e−16	<2.2e−16	<2.2e−16	<2.2e−16	<2.2e−16	0.452	<2.2e−16

Note: For chi-square test results, BTC, BNTC, STC, and SNTC are both compared with all cities during the holiday, while cities during non-holiday are compared with their corresponding holiday.

TABLE 4 Results of distance decay effect.

Category	Variable (T_{ij})	Distance decay effect	
		Holiday	Non-holiday
All cities	Travel flow	−2.136	−1.844
	Search flow	−0.404	−0.335
Big tourist city (BTC)	To BTC travel	−2.138	−1.893
	To BTC search	−0.443	−0.384
Big non-tourist city (BNTC)	To BNTC travel	−1.926	−1.678
	To BNTC search	−0.352	−0.310
Small tourist city (STC)	To STC travel	−2.057	−1.724
	To STC search	−0.413	−0.331
Small non-tourist city (SNTC)	To SNTC travel	−2.206	−1.917
	To SNTC search	−0.439	−0.319

accessibility of BNTCs certainly weakens distance's impedance on travel. However, this finding is not held for BTCs, which might be because of its small sample size.

We further explore how the correlation between search and travel flows changes with distance. Figures 4a,b show the trend of correlation changes at the distance interval of every 200 km. In general, we observe that the correlation weakens in a non-linear fashion as distance increases. For all cities, 1400 km is a key threshold. Beyond 1400 km, the correlation quickly drops lower than the benchmarks obtained in Section 4.1. For BTC during holidays, a few long-distance (over 3600 km) flows disturb the overall correlation, which mainly connects to Lhasa. Meanwhile, we observe an increase in correlation for BNTC at distances longer than 3600 km, with the highest correlation observed for distances exceeding 4000 km. These long-distance flows are mainly connected to big cities in Fujian, Zhejiang, and Heilongjiang. We also found that for BTCs and STCs, the correlation fluctuates more radically due to small group size.

4.4 | Spatial autocorrelation

Like most other geographic phenomena, spatial flows also obey the first law of geography (Tobler, 1970): "everything is related to everything else, but near things are more related than distant things." Evaluating spatial autocorrelation of flow data can help us understand to what extent flows share similar values as other geographically adjacent flows. Flows' association can be attributed to their origins, destinations, or both. In this study, we evaluate spatial autocorrelation of search flows and travel flows separately using FlowLISA (Tao & Thill, 2020), calculated as Equation (2):

$$FI_{(i,j)} = \left(n\left(f_{(i,j)} - \bar{f}\right) \sum_{(u,v)\neq(i,j)} w_{ij,uv}\left(f_{(u,v)} - \bar{f}\right) \right) \Big/ \sum_{i,j}^{n} \left(f_{(i,j)} - \bar{f}\right)^2 \quad (2)$$

where $FI_{(i,j)}$ is the local Moran's I statistic or the spatial autocorrelation measure of flow between origin i and destination j. $f_{(i,j)}$ represents the volume of flow between regions i and j. n is the total number of flows in the study area. $\bar{f}$ is the average volume of all flows. $w_{ij,uv}$ is the spatial flow weight between $f_{x(i,j)}$ and $f_{y(u,v)}$. In a simple binary configuration, $w_{ij,uv}$ equals one if $f_{(i,j)}$ and $f_{(u,v)}$ have adjacent origins and adjacent destinations, or if they have their origins (or destinations) be the same region while destinations (origins) adjacent. FlowLISA was extended from the local Moran's I (Anselin, 1995) to measure spatial autocorrelation for flow data. It allows researchers to quickly identify significant local flow patterns, such as flow clusters and flow outliers.

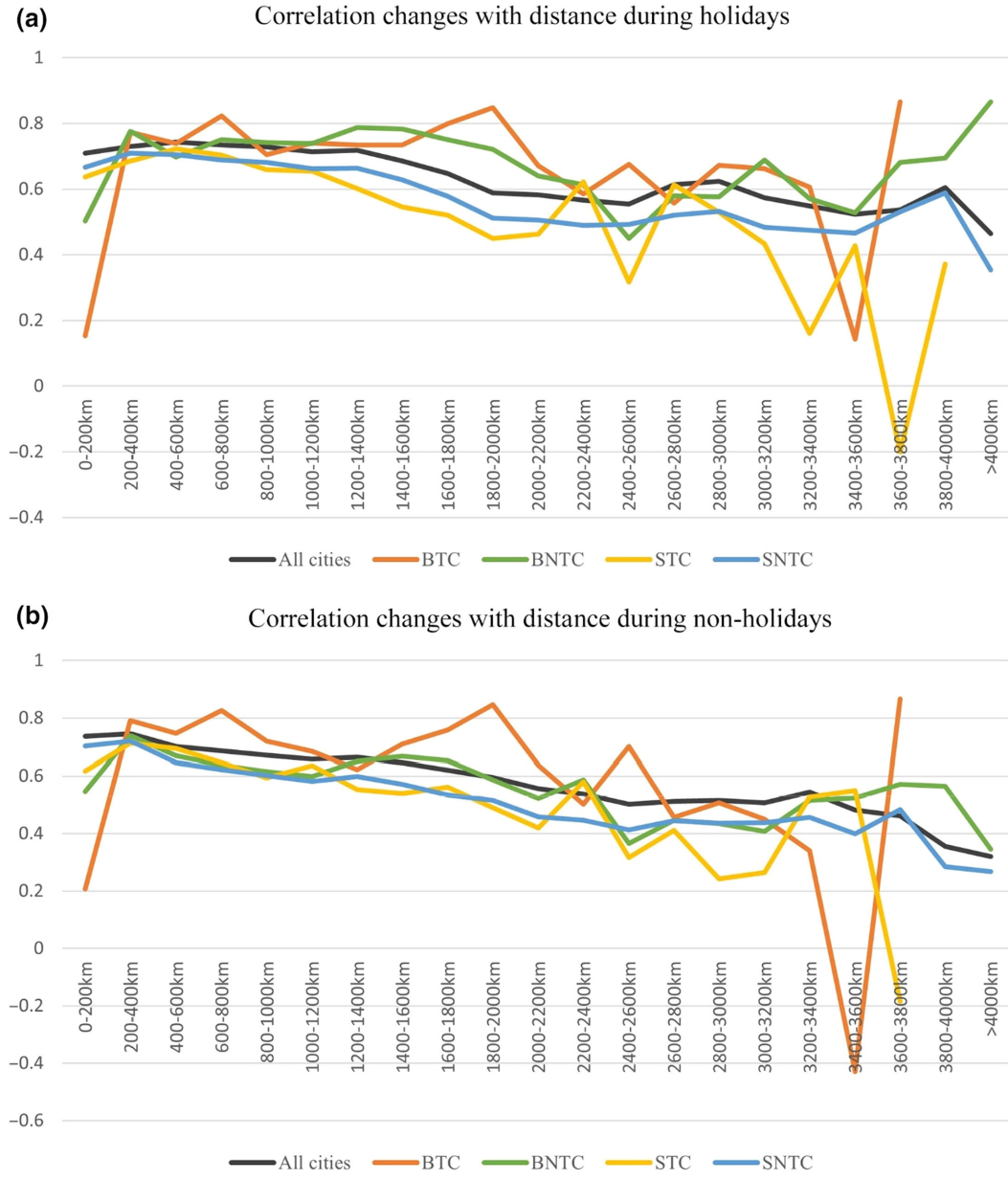

FIGURE 4 Correlation changes with distance: (a) holidays and (b) non-holidays.

Figures 5a,b display FlowLISA results of travel flows during holidays[2] and non-holidays at 0.001 significance level. High-high (HH) flows are the sole type of significant local patterns, which represent high-value flows surrounded by their high-value neighboring flows. HH flows are also known as flow clusters that exhibit strong spatial autocorrelation. Most HH flows are connected with big cities. Some big cities, such as Guiyang, Xi'an, Wuhan, and Chengdu, serve as both the origins and destinations of HH flows. Cities such as Shenzhen and Changsha, serve only as the origins of HH flows. In contrast, cities like Zhengzhou serve solely as destinations. During holidays, travel flows exhibit stronger spatial autocorrelation as we observe more HH flows. Cities with abundant attractions such as Chengdu and Xi'an have many more HH flows during holidays. Zhengzhou, as a transportation hub city and not known for abundant tourist resources, exhibits strong spatial autocorrelation

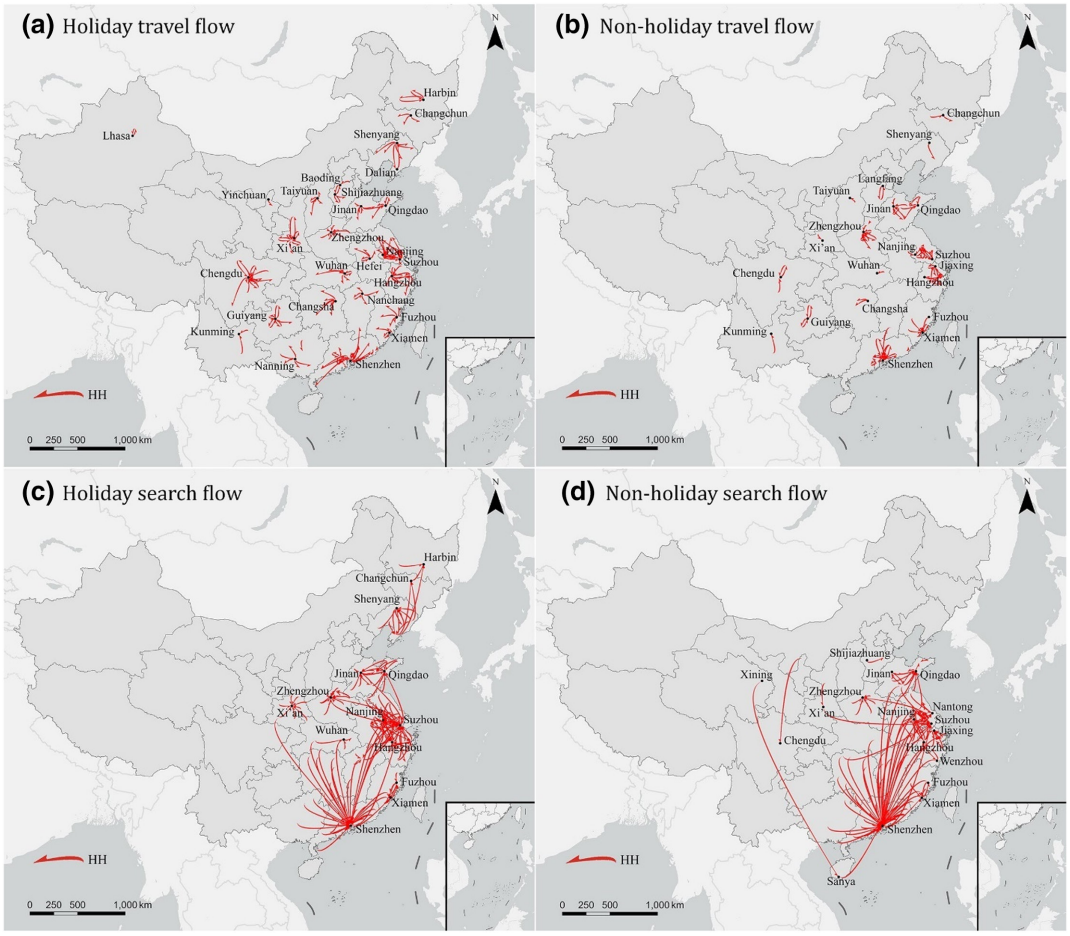

FIGURE 5 FlowLISA results at 0.001 significance level: (a) holiday travel flow; (b) non-holiday travel flow; (c) holiday search flow; and (d) non-holiday search flow.

during both holiday and non-holiday. It is worth noting that many HH flows tend to be short due to the strong distance decay effect on travel.

Figures 5c,d display FlowLISA results of search flows. The HH flows are more concentrated in a few cities, and they are longer than their travel flow counterparts. Shenzhen, as the largest immigrant city, is the primary destination of search flows during both holiday and non-holiday periods. Northeast China is observed with strong spatial autocorrelation of search flows during holidays, but the HH flows are constrained within the region.

4.5 | Network community

The connections between cities suggest that to understand cities we must see them not only as places in space but also as spaces of flow (Batty, 2013). Communities in complex networks refer to dense groups of nodes (cities) that have strong connections to each other. We can identify the regionalization delineated by flow distribution among cities by detecting the community structures of networks of two types of flows. We can further examine the alignment of regionalization of travel and search flows and compare the network structure during holidays and non-holidays. We use the well-known Louvain algorithm (Blondel et al., 2008) to extract network communities.

The Louvain algorithm works by optimizing a modularity function, which measures the quality of the division of the network into communities. The Louvain algorithm is known for its efficiency and ability to handle large networks. The algorithm is based on a hierarchical approach that recursively merges smaller communities into larger ones, until a global maximum of the modularity function is reached. The equation for the modularity function used in the Louvain method is:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{3}$$

where Q is the modularity of the partition, A_{ij} is the weight of the edge between nodes i and j, k_i and k_j are the sum of the weights of the edges attached to nodes i and j, m is the sum of the weights of all edges in the network, c_i and c_j are the community assignments of nodes i and j, and the δ-function $\delta(c_i, c_j)$ is defined as 1 if $c_j = c_j$, and 0 otherwise.

Figure 6 displays the Louvain results. For search flows during holidays, we observe communities that comprise non-adjacent geographic regions, for example, Community 4 comprises Jiangsu, Zhejiang, Anhui, Henan, and the 1000-mile-away Three-River-Source National Park (marked with a red star) in Figure 6a. Such search activities that overcome a long distance during holidays are probably due to tourist trips between the non-adjacent regions.

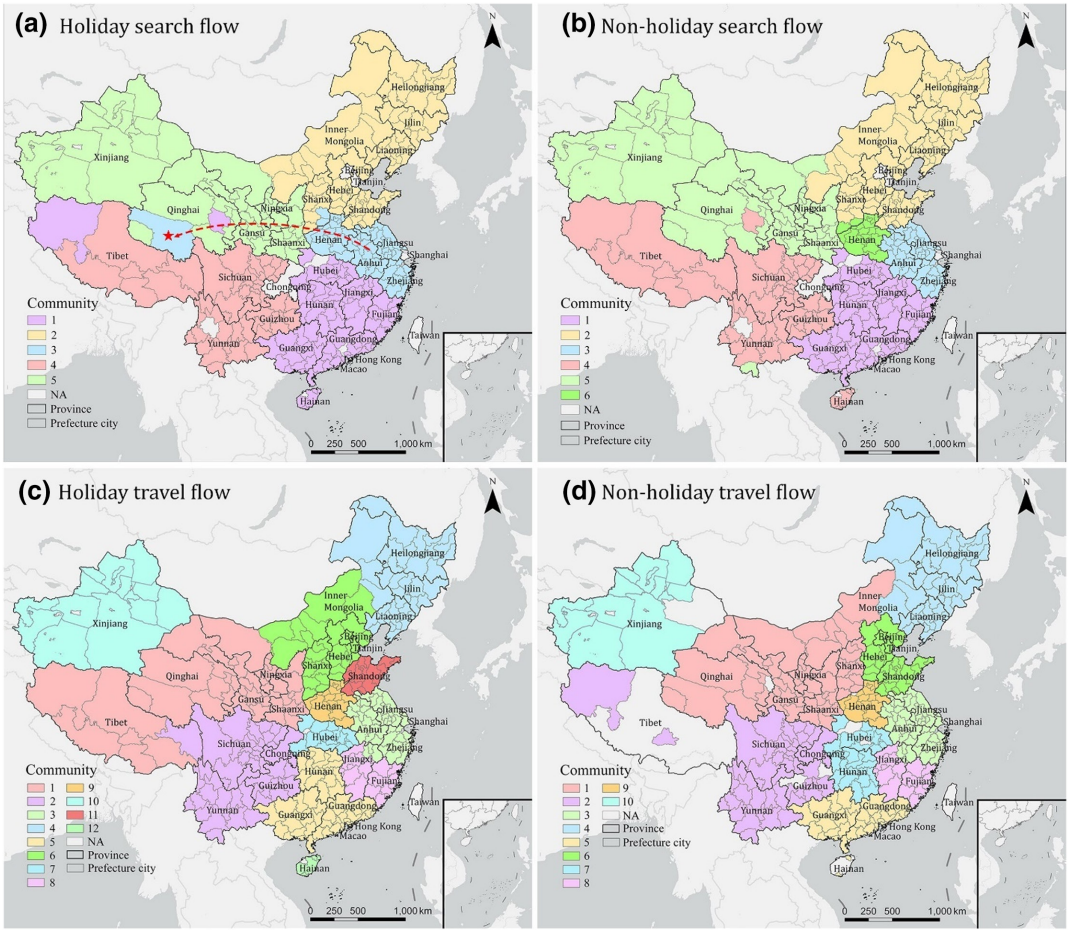

FIGURE 6 Louvain results: (a) holiday search flow; (b) non-holiday search flow; (c) holiday travel flow; and (d) non-holiday travel flow.

During non-holidays, there are significant intra-provincial search activities within Henan, which makes itself a standalone community.

We detected more communities using travel flows, and none of which comprises non-adjacent geographic regions. The provinces of Shandong, Hubei, and Henan form communities that align with their respective provincial boundaries during holidays, indicating their residents' preference for intra-provincial travel. Hunan joins the community of Guangxi and Guangdong during holidays, which could be attributed to the homecoming travel of migrant labors from Guangdong. The differences between the structures of search and travel networks verify the distance decay effect: travel flows are more constrained within the region, whereas search flows can well connect more distant and even non-adjacent regions.

4.6 | Cities' rankings in networks

We examine whether a city's ranking in the search network is as high as in the travel network. Inspired by the method presented in Wu and Yao (2021), we calculate the proportion of network connections related to each city to all network connections. This is done by summing the flow intensity of each city connected with other cities and dividing it by the overall total flow intensity, as shown in Equation (4):

$$PNC_i = \sum_{j=1}^{n} S_{ij} / S_{\text{total}} \tag{4}$$

where PNC_i is the proportion of network connections related to city i to all network connections S_{total}, n is the total city numbers, and S_{ij} is the network connection between city i and j. Then, we rank the results in ascending order, so that a city with a lower rank has a lower proportion. We obtained two separate rankings using search and travel flows, respectively.

Figures 7a,b show the results during holidays and non-holidays. There was no obvious seasonal change of city rankings by comparing holiday and non-holiday rankings, though the actual travel and search volume fluctuated. Most cities have comparable rankings in the two networks, except for three distinct groups of cities. Group 1, circled by blue dashed lines, includes cities that rank much higher in the search network than in the travel network. Most cities in Group 1 are tourist cities, such as Lhasa, Huangshan, and Zhangjiajie. Their reputation helps generate tremendous search volume. However, their relatively moderate rankings in the travel network may be because of high travel cost and low search-to-travel conversion rate, that is, travelers conduct intensive search activities before making one trip.

Group 2, circled by green dash lines, includes cities that rank much higher in the travel network than in the search network. Group 2 includes SNTCs such as Ordos, which are not famous to be search targets but are popular travel destinations due to their importance in the regional economy and the transportation network. Surprisingly, Group 2 also includes STCs such as Xinzhou, Yiyan, and Jinzhong. As tourist cities, they are not as famous as the STCs in Group 1 to attract many searches. But they rank high in the travel network by attracting many regional travelers.

Group 3, circled by yellow dashed lines, includes cities that rank the highest in both networks. Group 3 mainly consists of big cities such as Wuhan, Chengdu, and Xi'an. They attract many travelers for various purposes, including leisure, business, and family reunions. They also attract a comparable volume of search inquiries. Interestingly, we found two small cities, namely Baoding and Jiaxing, also belong to group 3. This could be attributed to two reasons. First, both cities have abundant tourist resources to attract many travelers and searchers. Second, Baoding's adjacency to Beijing and Jiaxing's adjacency to Shanghai help them stand out from other small cities by attracting a large volume of travel and search flows.

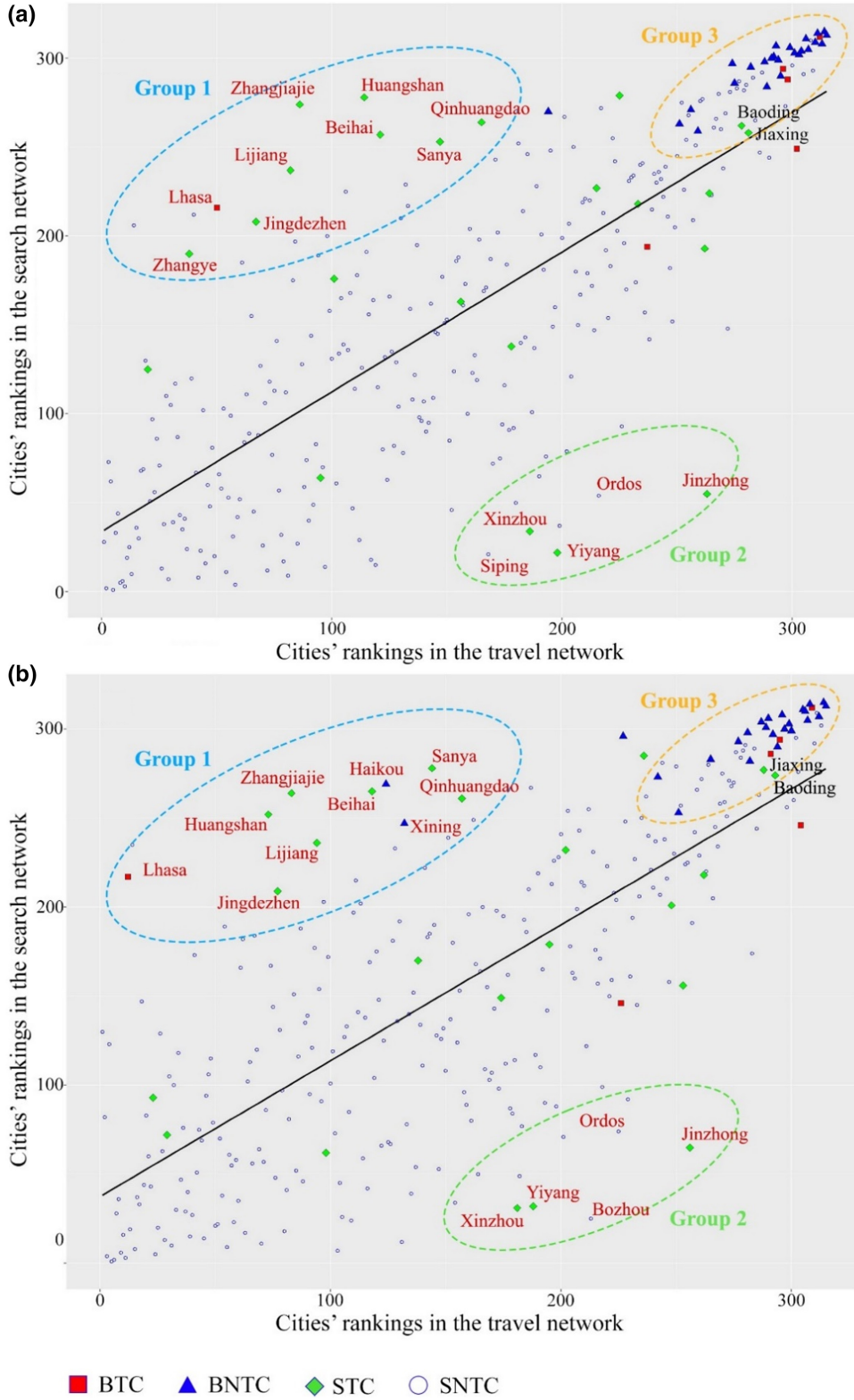

FIGURE 7 Ranks of cities in networks: (a) holidays and (b) non-holidays.

4.7 | Multivariate regression models

Search and travel activities are related with a city's attractiveness, which can be broken down into factors such as city type, economic development, traffic convenience, and recreation and shopping services (Ezmale, 2012; Guo et al., 2022). To find out which factors are crucial contributors to generating more search and travel flows, we construct a multivariate regression model comprising multiple categories of city features, including economics, traffic convenience, recreation, tourism, and administrative function. These features contain binary dummy variables for BTC, BNTC, STC, high-speed rail (Hsr), and airports (Air). In our city classification system, there are three primary categories: BTC, BNTC, and STC. Cities are assigned a value of 1 if they fall into one of these categories and a value of 0 if they do not. SNTC is excluded due to its relatively poor performance. We also incorporate numerical variables for GRP, the proportion of the tertiary sector (P-Third), and average distance (AvgDist) of a city to all other cities to represent the location factor.

On the one hand, BTC, BNTC, and STC are the three most significant objectives in our prior analysis. On the other hand, big cities tend to have larger populations, more diverse economic activities, and superior transportation infrastructure, potentially making them more appealing to travelers. Meanwhile, tourist cities may possess unique cultural, natural, or entertainment attractions that draw visitors from other places. Hsr and Air capture the effect of transportation infrastructure on search and travel behaviors. Hsr and Air can provide convenient and fast transportation options for travelers, stimulating economic activity in the cities they serve. GRP is a vital economic indicator used to describe the level of economic development in a region. It helps examine the impact of economic development on search and travel behavior. P-Third explores the influence of service industries, including retail, hospitality, and entertainment, on search and travel behavior, as these factors may significantly attract tourists to a city. AvgDist captures the impact of location on search and travel behavior. Cities closer to other cities or transport hubs may be more accessible and attractive to travelers than cities that are farther away.

Table 5 presents the regression results for search and travel flows during the departure phase of holidays and non-holidays. For travel flows, we find that GRP, BTC, BNTC, and AvgDist significantly influence travel flows during holidays. For search flows, the results during holidays and non-holidays are similar. BTC has the greatest impact on search and travel flows during holidays and non-holidays, followed by GRP and BNTC for non-holidays. Generally, the attractiveness of a city is proportional to its size, as GRP, BTC, and BNTC are all significant factors. It is worth noting that BNTC's coefficient is much higher during holidays than non-holidays, but it is not the case for search. This implies that travelers to BNTCs during holidays do not necessarily search for their destinations.

For search flows, Hsr and P-Third are additional significant factors. Surprisingly, cities with high-speed rail attract significantly more search flows but not travel flows. This may be due to the data collection mechanism: a traveler typically creates many search flows before creating one travel flow (Pan & Fesenmaier, 2006). The proportion of the tertiary sector (P-Third) contributes to more search flows as well, indicating that strong service sector can attract more search inquiries towards the city.

5 | DISCUSSIONS AND CONCLUSIONS

Through a series of spatiotemporal analyses of search and travel flows, we have obtained the following important findings. The correlation between search and travel flows is overwhelmingly positive. The correlation is stronger for big cities or tourist cities during holidays, which makes search flow a good indicator for holiday travels. However, for small cities or non-tourist cities, using search flows to infer travel flows is less effective.

The time-lag analysis reveals that there exists an up-to-three-day delay from search to travel flows. Time-lag effect is stronger for big cities during holidays. Their popularity as travel destinations makes travelers book their transportation and lodging in advance. However, advanced trip planning is not as prevalent for traveling to small cities during holidays. Distance decay effect is present in both search and travel flows, and it is more pronounced for travel

TABLE 5 Multivariate regression results during holiday and non-holiday.

	Travel flow (holidays)		Travel flow (non-holidays)		Search flow (holidays)		Search flow (non-holidays)	
Variable	Coefficient	Standard error	Coefficient	Standard error	Coefficient	Standard error	Coefficient	Standard error
Grp	0.748***	0.030	0.771***	0.038	0.705***	0.027	0.715***	0.026
Hsr	0.010	0.028	0.002	0.031	0.063**	0.023	0.056*	0.022
Air	−0.044	0.027	−0.048	0.031	0.031	0.022	0.019	0.021
P-Third	0.045	0.032	0.044	0.036	0.080**	0.026	0.070**	0.025
BTC	1.059***	0.187	1.031***	0.210	0.818***	0.151	0.732***	0.147
BNTC	0.623***	0.118	0.429**	0.132	0.770***	0.095	0.821***	0.092
STC	0.142	0.096	0.127	0.107	0.024	0.077	0.065	0.075
AvgDist	−0.096***	0.027	−0.064*	0.031	−0.085***	0.022	−0.058**	0.021
Adjust R^2	0.823		0.777		0.885		0.891	
F value	153		115.3		251.9		269.5	

*Represent 5% significance level; **Represent 1% significance level; ***Represent 0.1% significance level.

flows than search flows, during holidays than during non-holidays. Among the city groups, BNTC has significantly weaker distance decay effects than the other cities, thanks to their decent transportation accessibility. We also found the correlation between search and travel flows weakens as distance increases, particularly beyond 1400 km.

The FlowLISA results reveal strong spatial autocorrelation, which is easier to be observed among search and travel flows linking to big cities. Some travel flow clusters are bidirectional, while some are unidirectional, for example, HH flows originated from Shenzhen and Changsha, and HH flows bound to Zhengzhou. The demographic and economic structure is a potential factor, for example, cities with large number of immigrant laborers such as Shenzhen are more likely to be the origins rather than destinations of travel flows. In contrast, search flow clusters are more concentrated in a fewer number of big cities, possibly due to the Matthew effect on online activities.

From the network's perspective, we detected more communities using travel flows than using search flows. Most communities are formed by several adjacent administrative regions. Only a few search-flow communities comprise non-adjacent geographic regions. Cities' rankings in search network and travel network are generally comparable. We primarily discussed three special groups of cities: Group 1 is mainly formed by tourist cities, which rank much higher in the search network than in the travel network; Group 2 comprises both SNTCs and STCs, rank much higher in the travel network than in the search network; Group 3 cities rank the highest in both networks, which consists of big cities and two small cities, namely Baoding and Jiaxing.

The multivariate regression model shows that BTC has the greatest impact on search and travel flows during both holiday and non-holiday periods, followed by GRP and BNTC for the non-holiday period. These results indicate that the attractiveness of a city is directly proportional to its size. There are many more travelers to BNTCs during holidays, who do not necessarily search for their destinations. Moreover, having high-speed rail or a strong service sector can help a city attract significantly more search flows but not travel flows. High-speed rail connectivity can enhance a city's accessibility and convenience for travelers, potentially sparking interest and prompting individuals to research information about the city. Similarly, a strong service sector may signal a high standard of living, making the city more appealing to prospective visitors who might search for details on the city's services, job opportunities, or general living conditions. However, the presence of high-speed rail or a robust service sector does not guarantee an increase in actual travel to the city. Factors such as travel expenses, accommodation availability, and tourist attractions also contribute to a city's overall attractiveness and influence people's decision to visit.

Some of the findings are nonintuitive and even surprising. For instance, both search and travel flows have a stronger distance decay effect during holidays than non-holidays, implying that people do not search for or travel to further destinations despite having more time to spend on trips. We also observed network communities with non-adjacent geographic regions using search flows, indicating that some people have strong search interest in distanct regions. With cities' rankings, we found tourist cities like Xinzhou, Yiyan, and Jinzhong with lower-than-expected search rankings, and small cities like Baoding and Jiaxing have higher-than-expected search and travel rankings. In both cases, their locations and proximity to big cities help them attract many nearby travelers. In conclusion, this study has explored the complex relationships between search and travel flows, provided valuable insights into the factors that shape city attractiveness and interconnectedness, and filled the gap in the literature as most related studies use search indexes of trip destinations while omitting the origins. By examining correlation, time-lag effect, distance decay effect, spatial autocorrelation, network community, and cities' rankings, we have identified key patterns and driving factors of search and travel activities. The findings of this study have important implications for urban planning and tourism management. Understanding the factors that influence search and travel flows can help policymakers and urban planners develop targeted strategies to enhance city attractiveness, improve transportation infrastructure, and promote tourism. Furthermore, recognizing the roles of different city types and their positions in search and travel networks can inform the development of regional cooperation and integration initiatives that leverage the unique strengths of each city. For instance, Xi'an, a BTC located in Shaanxi province, experiences a high volume of search flows targeting the city itself during holidays. As a good example, Xi'an has implemented a series of incentive policies to attract these potential visitors, such as offering free entry to some attractions, and granting a 50% discount to those who travel by high-speed railway

from neighboring cities (National Day Golden Week, 2018). These policies have significantly increased the attractiveness of Xi'an during holiday periods, promoting the development of the tourism industry. Similar policies can be adopted by local officials, such as Three-River-Source National Park, which attracted a large volume of search flows but not travel flows through our analysis.

The study's limitation is that our data do not include all prefecture-level cities, and the lack of data of five big cities including Beijing and Shanghai will affect the accuracy of the experimental results to some extent. Another data limitation is that some variables in the regression analysis are categorical rather than numerical, such as Hsr and Air, which might affect the model interpretation. The future direction could be to explore the relationships between search flows and other types of flow data to find more meaningful applications of search engine data, such as investment flow, labor migration flow, and cargo flow. Additionally, by incorporating additional variables, such as social media data or tourist preferences, to better understand the factors that drive search and travel behavior.

ACKNOWLEDGMENTS

The authors would like to thank the journal editors and anonymous reviewers for their valuable comments and suggestions.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID


Yuzhou Chen https://orcid.org/0000-0002-1556-084X

ENDNOTES

[1] We selected the week from October 2nd to October 8th in the holiday period to match with the non-holiday week.

[2] We selected data in the departure phase from September 27th to October 2nd to focus on where people travel to.

REFERENCES

Akter, H., Akhtar, S., & Ali, S. (2017). Tourism demand in Bangladesh: Gravity model analysis. *Tourism: An International Interdisciplinary Journal*, *65*(3), 346–360.

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical Analysis*, *27*(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Artola, C., Pinto, F., & de Pedraza García, P. (2015). Can internet searches forecast tourism inflows? *International Journal of Manpower*, *36*, 103–116. https://doi.org/10.1108/IJM-12-2014-0259

Batty, M. (2013). *The new science of cities*. MIT Press. https://doi.org/10.7551/mitpress/9399.001.0001

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Bokelmann, B., & Lessmann, S. (2019). Spurious patterns in Google trends data—An analysis of the effects on tourism demand forecasting in Germany. *Tourism Management*, *75*, 1–12. https://doi.org/10.1016/j.tourman.2019.04.015

Camacho, M., & Pacce, M. J. (2018). Forecasting travellers in Spain with Google's search volume indices. *Tourism Economics*, *24*(4), 434–448. https://doi.org/10.1177/1354816617737227

Carter, L. C., & Tao, R. (2023). Evaluating COVID-19's impacts on Puerto Rican's travel behaviors. *Geo-Spatial Information Science*, 1–11. https://doi.org/10.1080/10095020.2022.2161426

Choi, H., & Varian, H. (2012). Predicting the present with Google trends. *Economic Record*, *88*, 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x

Emili, S., Figini, P., & Guizzardi, A. (2020). Modelling international monthly tourism demand at the micro destination level with climate indicators and web-traffic data. *Tourism Economics*, *26*(7), 1129–1151. https://doi.org/10.1177/1354816619867804

Ezmale, S. (2012). Strategies for enhancing attractiveness of the cities in Latgale region. *European Integration Studies*, *6*, 121–127. https://doi.org/10.5755/j01.eis.0.6.1601

Ge, C. (2022). *National day holiday travel tips*. https://www.chinanews.com.cn/cj/2022/09-28/9862732.shtml

Guo, H., Zhang, W., Du, H., Kang, C., & Liu, Y. (2022). Understanding China's urban system evolution from web search index data. *EPJ Data Science*, *11*(1), 20. https://doi.org/10.1140/epjds/s13688-022-00332-y

Han, B., Zhu, D., Cheng, C., Pan, J., & Zhai, W. (2022). Patterns of nighttime crowd flows in tourism cities based on taxi data—Take Haikou prefecture as an example. *Remote Sensing*, *14*(6), 1413. https://doi.org/10.3390/rs14061413

Hu, M., & Song, H. (2020). Data source combination for tourism demand forecasting. *Tourism Economics*, *26*(7), 1248–1265. https://doi.org/10.1177/1354816619872592

Huang, X., Jagota, V., Espinoza-Muñoz, E., & Flores-Albornoz, J. (2022). Tourist hot spots prediction model based on optimized neural network algorithm. *International Journal of System Assurance Engineering and Management*, *13*(1), 63–71. https://doi.org/10.1007/s13198-021-01226-4

Huang, X., Zhang, L., & Ding, Y. (2017). The Baidu Index: Uses in predicting tourism flows—A case study of the Forbidden City. *Tourism Management*, *58*, 301–306. https://doi.org/10.1016/j.tourman.2016.03.015

Li, H., & Xiao, Z. (2022). Comparisons and predictions of intercity population migration propensity in major urban clusters in China: Based on use of the Baidu index. *China Population and Development Studies*, *6*(1), 55–77. https://doi.org/10.1007/s42379-022-00103-2

Li, K., Lu, W., Liang, C., & Wang, B. (2019). Intelligence in tourism management: A hybrid FOA-BP method on daily tourism demand forecasting with web search data. *Mathematics*, *7*(6), 531. https://doi.org/10.3390/math7060531

Li, S., Chen, T., Wang, L., & Ming, C. (2018). Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu index. *Tourism Management*, *68*, 116–126. https://doi.org/10.1016/j.tourman.2018.03.006

Müller, M. (2007). Dynamic time warping. In M. Muller (Ed.), *Information retrieval for music and motion* (pp. 69–84). Springer. https://doi.org/10.1007/978-3-540-74048-3_4

National Day Golden Week. (2018). Favorable policies for the public at various attractions in Xi'an. https://www.sohu.com/a/256644774_100159420

Orsi, F., & Geneletti, D. (2013). Using geotagged photographs and GIS analysis to estimate visitor flows in natural areas. *Journal for Nature Conservation*, *21*(5), 359–368. https://doi.org/10.1016/j.jnc.2013.03.001

Pan, B., & Fesenmaier, D. R. (2006). Online information search: Vacation planning process. *Annals of Tourism Research*, *33*(3), 809–832. https://doi.org/10.1016/j.annals.2006.03.006

Pan, B., Wu, D. C., & Song, H. (2012). Forecasting hotel room demand using search engine data. *Journal of Hospitality and Tourism Technology*, *3*, 196–210. https://doi.org/10.1108/17579881211264486

Pan, J., & Lai, J. (2019). Spatial pattern of population mobility among cities in China: Case study of the National Day plus Mid-Autumn Festival based on Tencent migration data. *Cities*, *94*, 55–69. https://doi.org/10.1016/j.cities.2019.05.022

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

Purnaningrum, E., & Athoillah, M. (2021, March). SVM approach for forecasting international tourism arrival in East Java. *Journal of Physics: Conference Series*, *1863*, 012060. https://doi.org/10.1088/1742-6596/1863/1/012060

Rivera, R. (2016). A dynamic linear model to forecast hotel registrations in Puerto Rico using Google trends data. *Tourism Management*, *57*, 12–20. https://doi.org/10.1016/j.tourman.2016.04.008

Tao, R., & Thill, J. C. (2020). BiFlowLISA: Measuring spatial association for bivariate flow data. *Computers, Environment and Urban Systems*, *83*, 101519. https://doi.org/10.1016/j.compenvurbsys.2020.101519

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*(Suppl. 1), 234–240. https://doi.org/10.2307/143141

Wang, Z., Ye, X., Lee, J., Chang, X., Liu, H., & Li, Q. (2018). A spatial econometric modeling of online social interactions using microblogs. *Computers, Environment and Urban Systems*, *70*, 53–58. https://doi.org/10.1016/j.compenvurbsys.2018.02.001

Wu, C., Zhuo, L., Chen, Z., & Tao, H. (2021). Spatial spillover effect and influencing factors of information flow in urban agglomerations—Case study of China based on Baidu search index. *Sustainability*, *13*(14), 8032. https://doi.org/10.3390/su13148032

Wu, K., & Yao, C. (2021). Exploring the association between shrinking cities and the loss of external investment: An intercity network analysis. *Cities*, *119*, 103351. https://doi.org/10.1016/j.cities.2021.103351

Xu, L., Wang, S., Li, J., Tang, L., & Shao, Y. (2019). Modelling international tourism flows to China: A panel data analysis with the gravity model. *Tourism Economics*, *25*(7), 1047–1069. https://doi.org/10.1177/1354816618816167

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

Yin, L. (2020). Forecast without historical data: Objective tourist volume forecast model for newly developed rural tourism areas of China. *Asia Pacific Journal of Tourism Research*, *25*(5), 555–571. https://doi.org/10.1080/10941665.2020.1752755

Zhang, B., Pu, Y., Wang, Y., & Li, J. (2019). Forecasting hotel accommodation demand based on LSTM model incorporating internet search index. *Sustainability*, *11*(17), 4708. https://doi.org/10.3390/su11174708

Zhou, M., Yang, M., & Chen, Z. (2023). Flow colocation quotient: Measuring bivariate spatial association for flow data. *Computers, Environment and Urban Systems*, *99*, 101916. https://doi.org/10.1016/j.compenvurbsys.2022.101916

How to cite this article: Chen, Y., Gong, Z., Ma, Q., & Tao, R. (2023). Exploring the spatiotemporal relationships between search flows and travel flows. *Transactions in GIS*, *27*, 1338–1356. https://doi.org/10.1111/tgis.13085