Computers, Environment and Urban Systems 65 (2017) 66–78

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Extraction and analysis of city's tourism districts based on social media data

Hu Shao [a], Yi Zhang [b,*], Wenwen Li [a]

[a] School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ 86287-5302, USA
[b] Institution of Remote Sensing and Geographical Information Systems, School of Earth and Space Science, Peking University, Beijing, China

ARTICLE INFO

Article history:
Received 14 April 2015
Received in revised form 24 April 2017
Accepted 24 April 2017
Available online 25 May 2017

Keywords:
Tourism district
Social media
Spatial-temporal behavior

ABSTRACT

Through the perspective of tourism, a city as a tourist destination usually consists of multiple tourist attractions such as natural or cultural scenic spots. These attractions scatter in city spaces following some specific forms: clustered in some regions and dispersed in others. It is known that users organize their tours in a city not only according to the distance between different attractions but also according to other factors such as time constraints, expenses, interests, and the similarities between different attractions. Hence, users' travel tours can help us gain a better understanding about the relationships among different attractions at the city scale. In this paper, a methodological framework is developed to detect tourists' spatial-temporal behaviors from social media data, and then such information is used to extract and analyze city's tourism districts. We believe that this city space division will make significant contributions to the fields of urban planning, tourism facility providing, and scenery area constructing. A typical tourism city in China—Huangshan—is selected as our study area for experiments.

1. Introduction

Studying the spatial structure of city destinations has long been considered an important topic of tourism study (Pearce, 1998) and the key content of urban tourism planning (Dredge, 1999). According to the tourism-spatial system theory (Gunn and Var, 2002), *Tourism Destination* (as a city) consists of *Tourism Attractions* and *Related Infrastructures*. *Tourism Attractions* are the main inducement of tourist visits, which refers to physical or cultural features of a particular place that individual travelers or tourists perceive as capable of meeting one or more of their specific leisure-related needs. Such features may be ambient in nature (e.g., climate, culture, vegetation, or scenery) or they may be specific to a location (e.g., theater performance, a museum, or ceremony events). *Related Infrastructures* supply a variety of services to tourists, including transportation, accommodation, dining, entertainment, and so forth. Tourism attractions together with its surrounding-related infrastructures are called *Tourism District* (Dredge, 1999; Pearce, 2001), which can be viewed as tourists' main activity area.

As we know, tourists' entire process of traveling usually contains continuous several days. In most situations, each day's route starts and ends with dwelling places (Shoval, McKercher, Ng, & Birenboim, 2011), while the intermediary nodes usually consist of restaurants, tourist sites, gas stations, shops, and so forth to meet their essential and entertainment needs (David A. Fennell, 1996; Bob McKercher, Wong, & Lau, 2006; Rebollo and Baidal, 2003). What is more, restraints such as time and transportation (Lau and McKercher, 2006; Page and Connell, 2014) together with tourists' individual preferences (Bob McKercher et al., 2006) and limitations of knowledge to tourist destinations will also affect their daily tour schedule. All these factors together make tourists' daily travel route relatively comprehensive and concentrated. According to the definition, a tourism district is a concentrated area in city space consisting of multiple nodes, which together are "likely to be able to fulfill a variety of tourist needs and expectations" (Dredge, 1999; Pearce, 2001). Consequently, we believe that tourists' daily travel route should mainly overlap with a single tourism district and can help recognizing the tourism districts in city space.

The lack of appropriate tourist activity data is the main restriction of tourism study (Pearce, 1979, 2001; Lew and McKercher, 2002). However, after entering the digital age (Shoval and Isaacson, 2007), the popularity of location-aware devices and the development of social network service (SNS) have made it possible to access user-centered individual spatiotemporal behaviors and their context information, which are long time series, massive, and highly precise. This information could help geographers understand users' spatiotemporal behavior patterns inside urban space and their interactions with the urban environment, which may be able to fill the gap of data unavailability for the tourism study (Cranshaw, Schwartz, Hong, & Sadeh, 2012; Wood et al., 2013; Liu, Liu, Gao, Gong, Kang, Zhi, Chi & Shi, 2015).

* Corresponding author.
E-mail addresses: hu.shao@asu.edu (H. Shao), zy@pku.edu.cn (Y. Zhang), wenwen@asu.edu (W. Li).

http://dx.doi.org/10.1016/j.compenvurbsys.2017.04.010
0198-9715/

In this study, we introduce a novel data-driven method to extract and analyze a city's tourism district structure through tourists' aggregate travel behaviors. The Huangshan City in China is selected as our study area because of its popularity in the tourism market. The rest of this paper is organized as follows: Section 2 reviews related work about tourists' spatial-temporal behavior study and social media study. Section 3 provides the detailed explanation of the data and methods used in this paper. Section 4 provides the description of how our methodologies are applied in the study area of Huangshan City. Section 5 discusses about our work and what we need to do in the future.

2. Related works

Pearce summarized in his article that the spatial analysis on urban tourism can be conducted at three scales to develop a comprehensive picture of urban tourism (Pearce, 2001).

On the level of *City Space*, a city is treated as the "focus or unit of analysis" (Pearce, 2001), usually socioeconomic data and aggregate data have been used to study some specific tourism aspects of a single city (Lew, 1992; Shoval et al., 2011; Baležentis et al., 2012; Pons, Salamanca, & Murray, 2014) or the relationships among multiple cities like tourism flow (Oppermann, 1994, 1995; Yan, 2004; White and White, 2007; Xing-zhu and Qun, 2014; McKercher et al., 2006).

On the level of *Tourist Sites*, research of tourists' spatiotemporal activity within or among different tourist sites in city space is relatively easy to be conducted because of its small scale (Shoval and Isaacson, 2007). Data collecting methods such as questionnaire surveys, camera records, global positioning system (GPS), and land-based tracking systems have all been used to gather the information for analyzing, explaining, and simulating users' movement pattern (Hartmann, 1988; Keul and Kuhberger, 1997; Itami et al., 2003; O′Connor et al., 2005; Lau and McKercher, 2006; Lew and McKercher, 2006; Edwards and Griffin, 2013).

On the level of *Tourism Districts*, tourism does not occur uniformly; instead, it is concentrated in particular areas of urban space. Detailed analyses of urban tourism need to focus into the substructure to develop a comprehensive understanding of patterns, processes, and interrelationships of different parts of a city. This study falls in the scope of tourism district partitioning (Pearce, 2001). Several qualitative studies on this level have been conducted (Teo and Huang, 1995; Savage, Huang, & Chang, 2004; Pearce, 1998) to recognize and explain tourism districts according to their functions. Studies on the scale of tourism district, which is also the scale of our work, can help us gain profound insight on cities' subtourism-region functionality and hopefully bridge the gap between studies of city space and tourist sites. However, the difficulty lies in that tourist districts study requires tourists' spatiotemporal activity data to be both detailed and abundant. Unfortunately, limitations of conventional data acquisition methods such as telephone surveys or on-site investigations on tourist spatiotemporal behaviors can be time and energy consuming (Shoval and Isaacson, 2007). This has restricted the study of urban space at a regional scale. After we entered the digital age, location acquisition techniques have developed rapidly and generated voluminous yet cheap and easy-to-acquire data (Lu and Liu, 2012) such as cellular phone records, GPS data, consumption records, SNS data, and VGI (Volunteered Geographic Information) (Goodchild, 2007). These data depict humans' spatiotemporal behavior in detail, providing us great opportunities to accelerate tourism study.

First, the digital data utilities have been tested or verified in several studies on tourist behaviors. Cheng, Caverlee, Lee, and Sui (2011) explored a lot of potential applications of social media to depict users' spatial-temporal activities: through a huge volume of geo-tagged twitter data, they calculated individuals' patterns of moving length, radius of gyration, detected individuals' home locations, and discussed the factors which might affect users' mobility. Wood, Guerry, Silver, and Lacayo (2013) used Flickr data as their data source to approximate visitation rates of 836 recreational sites around the world, after comparing with official statistics from each site, they concluded that "the crowd-sourced information can indeed serve as a reliable proxy for empirical visitation rates." Ahas, Aasa, Roose, Mark, and Silm (2008) compared the passive mobile positioning data from Estonia with conventional accommodation statistics from the same time; they found that the correlation of these two datasets reached 0.99, proving that mobile positioning data have high precision for depicting users' aggregate spatiotemporal behaviors. Hawelka et al. (2014) performed a similar work on international travelers using twitter data; they validated "geo-located twitter as a proxy of global mobility behavior to a certain extent."

Several works studying city structure and tourist behaviors have also been conducted using these digital data in recent years. By applying cluster analysis on mobile positioning data, Asakura and Iryo (2007) found some topological characteristics of tourists' behavior. Donaire, Camprubí, and Galí (2014) detected different types of photographers' according to their travel photography from Flickr data. Zhai et al. (2015) revealed the popularity of restaurants in cities from social media data. Liu et al. improved the classification of land use in city space by introducing the spatial interaction pattern analysis of mobile phone users from different places. On the basis of point of interest (POI) and social media check-in data, Cranshaw et al. (2012) implemented the spatial partition of urban areas to study the "social dynamics" of cities. Using the same type of data, Yuan, Zheng, and Xie (2012) explored the main functions of a city's different regions. Hollenstein and Purves (2015) explored the "core area" and "border" of cities across the USA using user-generated Flickr data, while Hu et al. (2015) extracted urban areas of interest which attract people's attention in a city space. Yin, Cao, Han, Luo, and Huang (2011) extracted common tour trajectories of Flickr users in different cities. Hot spots of tourism destination inside the city space have also been detected through social media data (Crandall, Backstrom, Huttenlocher, & Kleinberg, 2009; Liu, Sui, Kang & Gao, 2014; García-Palomares, Gutiérrez, & Mínguez, 2015; Zhou, Xu, & Kimmons, 2015).

Differing from the related works, our methodology takes users' geo-weibo sequences and their spatial-temporal characters instead of single pieces of geo-weibos for tourism districts detection. More information will be revealed from our results.

3. Data and methods

3.1. Study area

In this paper, we selected Huangshan City as our study area, which is a typical tourism city belonging to Anhui Province in the Southeast of China (Fig. 1, a). The main income resource of Huangshan City is its tourism product. Huangshan City's agreeable weather, four distinct seasons, and innumerable nature and culture landscapes help it attract 30 million tourists every year, which is almost 20 times the local population (1.47 million, 2013).[1]

Huangshan City is named by its famous scenic spot Mountain Huangshan (Mt. Huangshan), or literally translated as Yellow Mountain, and ranks among the Great Wall and the famous Terracotta Army as one of China's most luring tourist attractions (Fig. 1, b). In the northwest of Xiuning County, there lies another mountain named Qiyun (literally "Cloud-High Mountain"). Not as famous as Mt. Huangshan, Mt. Qiyuan mainly attracts visitors from the local area. The largest man-made lake in Anhui province, Taiping Lake, is also located north of Huangshan City.

[1] http://ah.anhuinews.com/system/2012/01/13/004705855.shtml
http://www.newshs.com/a/20140329/00104.htm
http://whc.unesco.org/en/list/547
http://www.huangshantour.com/english/
http://www.china-mike.com/china-tourist-attractions/huangshan/
http://www.chinahighlights.com/huangshan/tours.htm

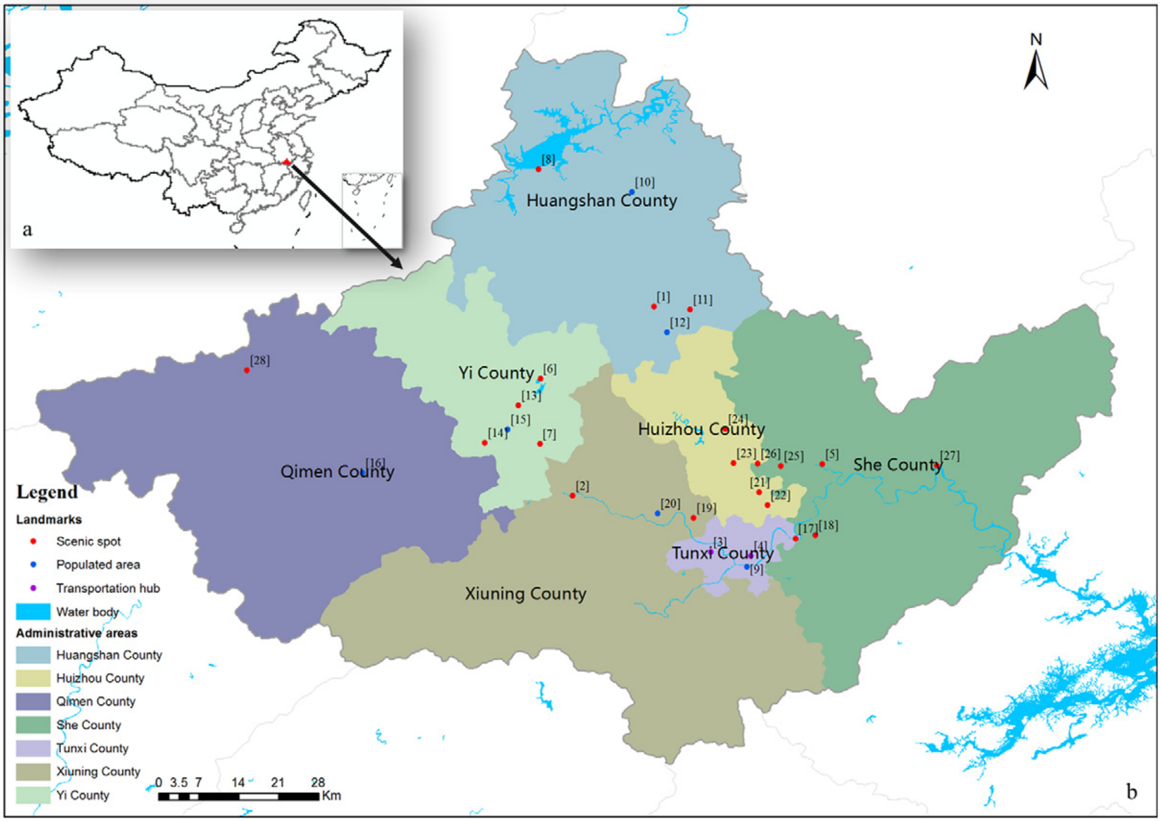


Fig. 1. General situation of Huangshan City. Subgraph **a.** Huangshan City's location in China. Subgraph **b.** Seven counties of Huangshan City and landmarks of Huangshan City. Information about these landmarks can be found in Table 1.

Besides the nature landscapes, many of Huangshan City's cultural landscapes also attract millions of tourists from all around the world every year. In particular, Xidi and Hongcun Ancient Villages have hundreds of years of history. Huizhou ancient town, which lies about 20 km northeast of Huangshan City's downtown, also has a rich history and attracts many visitors for a one-day trip.

From Fig. 1, b we can find that Huangshan City's scenic spots are scattered around its area, while its railway station and international airport are both located in Huangshan City's downtown—Tunxi County (In China, the administrative level of "city" is higher than "county," usually a city consists of several counties). Hence, Tunxi County plays the role of center service region for tourists here.

In the following sections, we introduce our methodology for extracting and analyzing the city's tourism districts in detail. In general, our methodology consists of four steps: (1) collecting social media data of our study area, including POI data, users' profile data, and geo-weibo (micro-blog with lon-lat coordinate) data; (2) extracting users' tourism related geo-weibos and organizing them day by day; (3) dividing the city space into grids and applying a community detection algorithm to find strongly associated grids; and (4) using pattern analysis methods to extract and explain the tourism districts of the city. The workflow of our study is represented in Fig. 2.

3.2. Sina Weibo data

In this research, we use social media data collected from Sina-Weibo to study people's tourist behaviors in Huangshan City. Sina-Weibo is one of the most popular SNSs in China after its establishment in 2009 (Guo, Li, & Tu, 2011).

We implemented a crawler to retrieve three types of data through Sina Weibo's application programming interface (APIs), which are POI data, user-profile data, and users' geo-weibo data.

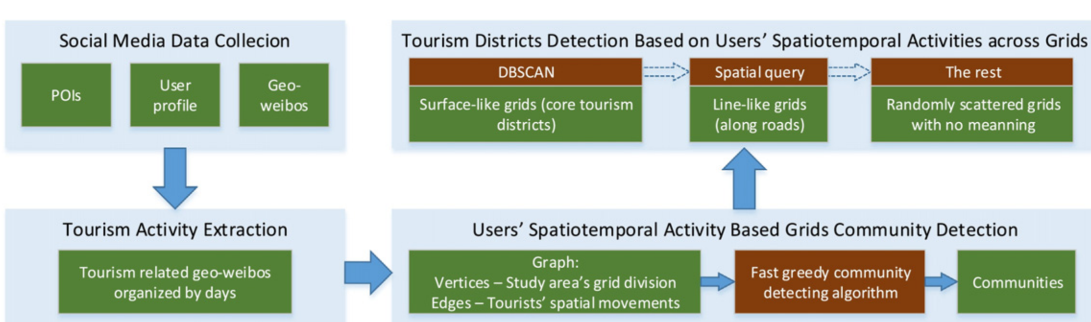


Fig. 2. Proposed workflow for tourism pattern analysis.

Table 1
Landmarks of Huangshan city.

Id	Name	Type
[1]	Mt. Huangshan	Scenic spot (AAAAA level)
[2]	Mt. Qiyun	Scenic spot (AAAA level)
[3]	Huangshan tunxi international airport	Transportation hub
[4]	Huangshan railway station	Transportation hub
[5]	Huizhou ancient town	Scenic spot (AAAA level)
[6]	Hongcun ancient village	Scenic spot (AAAAA level)
[7]	Xidi ancient village	Scenic spot (AAAAA level)
[8]	Taiping Lake	Scenic spot (AAAA level)
[9]	Huangshan City's downtown	Populated area
[10]	Huangshan County's downtown	Populated area
[11]	Jade Valley	Scenic spot (AAAA level)
[12]	Tangkou Town	Populated area
[13]	Gui Park	Scenic spot (AAAA level)
[14]	Nanping scenic area	Scenic spot (AAAA level)
[15]	Yi County's downtown	Populated area
[16]	Qimen County's downtown	Populated area
[17]	Huashan cave	Scenic spot (AAAA level)
[18]	Mt. Bawang scenic area	Scenic spot (AAA level)
[19]	Guchengyan	Scenic spot (AAA level)
[20]	Xiuning County's downtown	Populated area
[21]	Former site of the New Fourth Army's Headquarters	Scenic spot (AAA level)
[22]	Huizhou culture park	Scenic spot (AAA level)
[23]	Qiankou village	Scenic spot (AAAA level)
[24]	Chengkan village	Scenic spot (AAAA level)
[25]	Tangyue memorial archway park	Scenic spot (AAAA level)
[26]	Tangmo village	Scenic spot (AAAA level)
[27]	Shanshui scenic area of Xin'an River	Scenic spot (AAAA level)
[28]	Mt. Gu'niujiang	Scenic spot (AAAA level)

3.2.1. POI data

All the POIs falling inside the boundary of Huangshan City are retrieved. Each record of POI contains the attributes of *poi_id, title (name), address (text form), lon-lat coordinates, category code, category name, number of check-in users,* and *number of check-in micro-blogs*. Table 2 presents a sample of the POI data.

Sina Weibo officially provides a two-layer category architecture for its POI data (which is recorded in the field of categories): the first layer includes 14 categories which cover the most common types of places such as *shopping service, exclusive shop, gas station, outdoor tourism, other location*, and so forth. The second layer divides the first layer into 268 subcategories, giving each POI a more detailed description. Under such a hierarchy, tourism-related POIs are mostly included in the group of "Outdoor Tourism" (at first level), including subcategories such as *national level scenic spot, temple, world heritage, ski resort, zoo*, and so forth. In fact, there are also some POIs in the dataset been classified as "Unknown" by the website. To increase the accuracy of data, we manually went through these records and picked out all the tourism-related POIs by their names and locations. After manually filtering and correcting, we finally obtained 1264 direct tourism-related POIs. These POIs comprise 5.7% of all the POIs of Huangshan City (18,458 records). Fig. 3 shows the distribution of tourism-related POIs and other POIs in Huangshan city.

Table 2
Example of a POI record (the content in italic were translated from Chinese).

Field	Value	Description
poiid	B2094653D464A6FD409B	Unique id of the POI
title	*Wolonggu Tourism Region of Dazhang Mountain, Wuyuan.*	Title (name) of the POI
address	*Dazhang-mountain town, Wuyaun county*	Address of this POI
lon	117.75653	Longitude coordinate
lat	29.51371	Latitude coordinate
categories	194,240	Category code of this POI, 194 is the second-level category and 240 is the first-level category
category_name	*national level scenic spot*	Name of this POI's second-level category
checkin_user_num	771	Number of users who have ever checked-in here
checkin_num	871	Number of check-in weibos about this POI. One user may have multiple check-in weibos of one place, hence checkin_num ≥ checkin_user_num

3.2.2. Users' profile data

The profile of users who have checked-in at the POIs recently (within a 1-year timeframe) could be retrieved through the POI. User profile data include details of a user's information such as *registration place, gender, number of friends, number of followers, number of published micro-blogs*, and *account creation time*.

3.2.3. Geo-weibos in Huangshan City

Once the users' information is acquired in step 2, their geo-weibos could be retrieved through the API. There are two types of geo-weibos: check-in and noncheck-in. If the GPS module is turned on while user is posting a weibo, the lon-lat coordinates of user's current location will be recorded, and this weibo will become the geo-weibo. During the process, the Sina Weibo user interface (UI) will push a list of POIs to user for selection. The POIs are ranked according to their distances to user's current location. The select of POI is optional. Once a POI is selected, the geo-weibo become check-in geo-weibo. Hence, check-in geo-weibos carry more information than the normal geo-weibos.

Every record of geo-weibo includes attributes such as *id, associated user_id, text content, lon-lat coordinates, POI information (in case of check-in geo-weibo), publish time, number of reposts*, and *number of comments.*

We obtained 39,150 user profiles who have published at least one geo-weibo in Huangshan city during our study period: from July 1, 2012 to July 1, 2013. A total of 216,547 geo-weibos were posted in Huangshan City during this year. Fig. 4 shows all the geo-weibos in Huangshan City with their point density as background.

For each user u, we calculated the number of geo-weibos he/she posts in each day d, which is called the weibo post of each user-day - $n_{u,d}$.

Fig. 5 shows the distribution of $n_{u,d}$ of our dataset. According to the statistics, 91,170 (42.1%) geo-weibos were posted solely in the user days, while the remaining 57.9% geo-weibos belong to the user days when multiple geo-weibos were posted.

According to the official user group statistics report[2] from the Sina Weibo company, the amount of male and female users of Sina Weibo are about the same (50.1% vs 49.9%). However, younger generation individuals (born after 1980s) are the main users of Sina Weibo. Weibo users are mainly concentrated on highly educated groups—70.8% of the users have a degree of bachelor or higher (vs. this index of the whole China is 34.5%). The bias of demographic is very common among SNS data source such as Twitter or Facebook. For the reasons of protecting users' privacy, Sina Weibo does not provide the information such as users' income, education level through the public API. As we have discussed in the last section, social media data are a reliable data source for spatial-temporal behavior research (Wood et al., 2013).

We acquired a statistics report of the tourism market distribution (the number of tourists from each province of China) of Huangshan City from Huangshan Travel and Tourism Committee. Since our study

[2] http://data.weibo.com/report/reportDetail?id=76

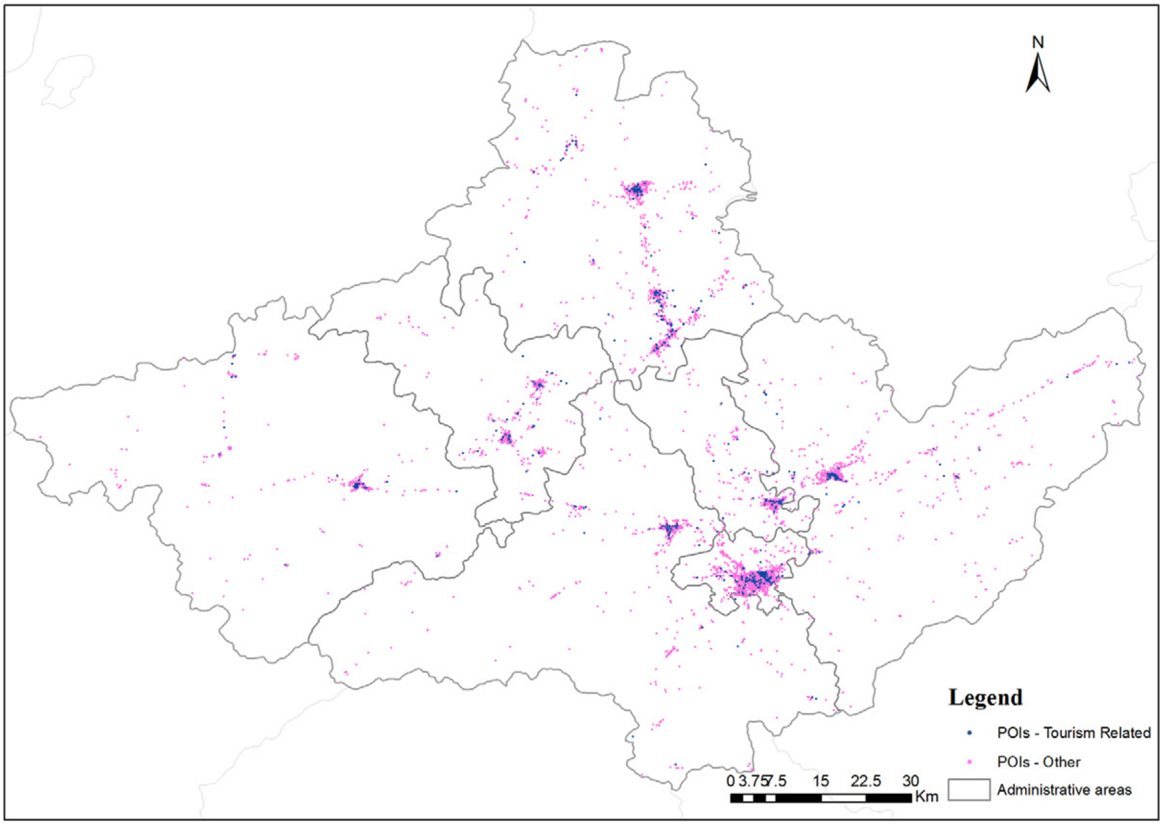


Fig. 3. Distribution of POIs in Huangshan city.

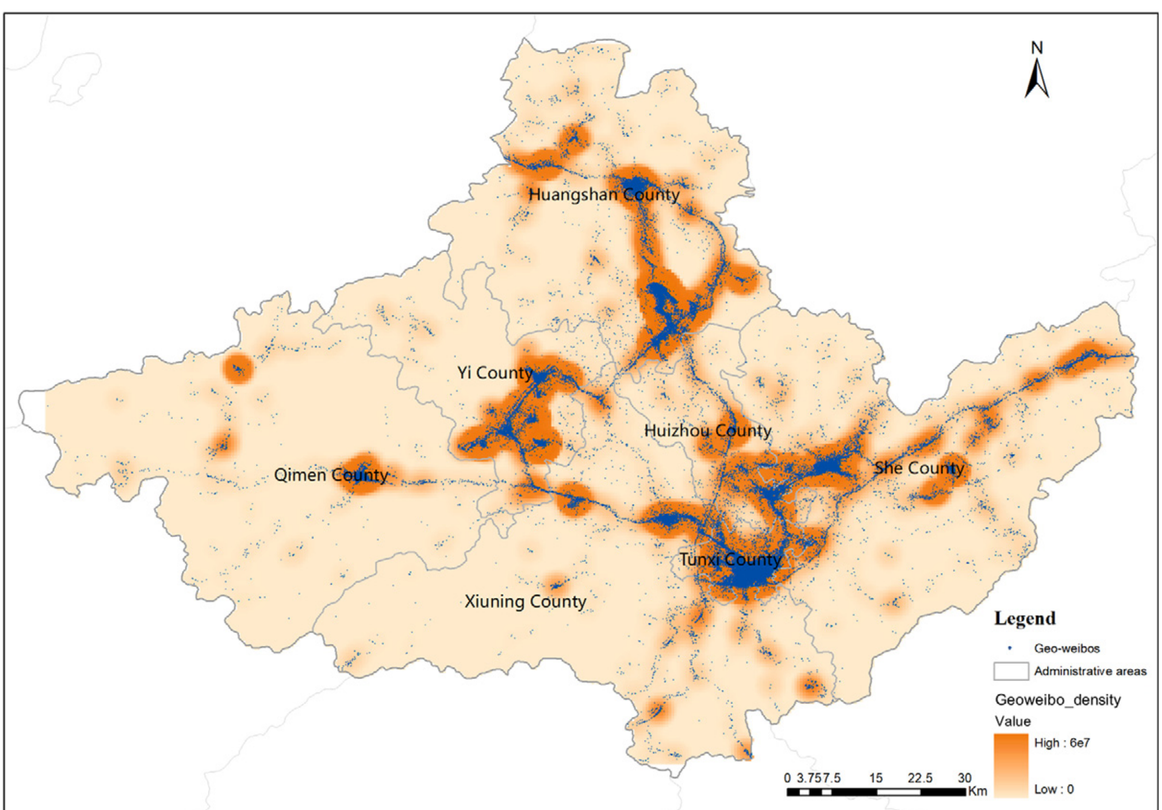


Fig. 4. Distribution of geo-weibos with their point density distribution in Huangshan City from July 1, 2012 to July 1, 2013.

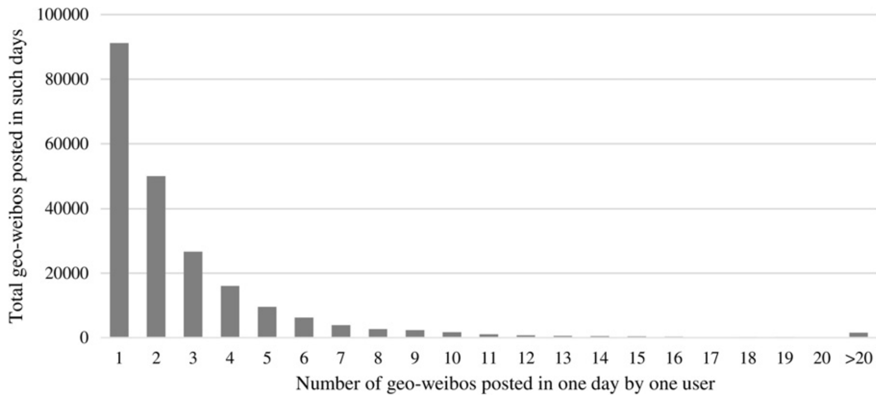


Fig. 5. Statistics of geo-weibo post in one day by users.

period is from July 1, 2012 to July 1, 2013, a correlation analysis between the actual number of tourists and number of Weibo posts in 2012 and 2013 for different provinces was conducted. It was found that these two datasets are highly correlated (with a coefficient value higher than 0.95). This result indicates that Sina weibo data are a good representative of the actual tourists in Huangshan City.

All the data we used in the experiments come from public APIs of Sina Weibo. No sensitive information of users such as real name, occupation, age, income, or home address is provided. Because we are trying to find the tourism districts instead of isolated spots, during the following experiments, we will only focus on tourists' aggregated behaviors.

3.3. Tourism activities extraction

As the most important characteristic of social media data, Sina Weibo truly records many aspects of users' daily life: what they like, what they are thinking, what they are doing, and so forth. As for geo-weibos, they give us more hints of people's activity type. If a user posts a check-in geo-weibo at a cinema, he/she might be going to watch a movie. However, if a user posts a check-in geo-weibo at a scenic area, he/she might be enjoying a trip. Indeed, POI information has been used in many researches as the reference of users' activity types (Huang, Li, & Yue, 2010, Spinsanti et al., 2010, Ye et al. 2011, Furletti, Cintia, Renso, & Spinsanti, 2013).

In this research, we post two basic assumptions of users' tourism activities:

Assumption1. If a user is posting a check-in geo-weibo at a tourism-related POI, such as a national park or mountain, this user is currently involved in some tourism activities.

Assumption2. If we detect a user who is involved in some tourism activities using assumption1, then his/her whole day's activities may have a close relationship with a trip, no matter whether he/she is in restaurants, hotels, vehicles, or ticket halls.

According to the above assumptions, we can extract users' tourism activities using the strategy as follows:

First, we arranged all of an individual's geo-weibos by “day,” which is

$$W_u = D_u^1 \cup D_u^2 \cup \ldots \cup D_u^n$$

Here, W_u is the set of all of user u's geo-weibos, and D_u^i is the set of all of user u's geo-weibos in day i. Day 1 begins with July 1, 2012. We split every day at 4 am. According to our data, this is the weakest time of people's activities within a day.

We could check every user's weibo set on each day to find whether there exists a tourism-related geo-weibo. Because according to Assumption1, such a geo-weibo reveals user's tourism activities. If it exists, according to Assumption2, all of this day's geo-weibos are supposed to be tourism-related. The rule is written formally as follows:

$$\{\exists \text{weibo}_p | \text{weibo}_p \in D_u^p \textit{ and } \text{weibo}_p \textit{ is tourism}-\textit{related}\} \\ => D_u^p \textit{ is tourism}-\textit{related}$$

Here, D_u^p denotes user u's weibo set at day p.

By using these POIs, we could extract 25,880 users who have been involved in tourism activities in Huangshan City. During their traveling days, 77,048 geo-weibos were posted.

3.4. Community detection on geo-weibos

Finding the tourism districts of a city is the main target of this study. As we discussed before, when a traveler is planning his/her visiting route of a day, all the essential and entertainment needs should be met with the restraints of time, transportation, and budget being considered. These factors make all the stops of a traveler's daily route meaningful. Currently, this important information could be uncovered from their geo-weibos. We could build a connection of different spots according to users' daily check-in geo-weibos. If we could aggregate this information from thousands of users' data, tourism communities can be detected. Such communities will reflect the compactness of tourists' spatial movement.

The community detection algorithm is employed to execute this task. Community detection emerges from graph analysis methods. In the graph theory, a graph (or network) consists of vertices and edges, which connect the vertices to each other (Trudeau, 2013). Every edge has a weight, reflecting the strength of the relationship between the two vertices it connects. The greater the weight, the stronger the two vertices are connected. If there is no edge between two nodes, the weight can be viewed as 0. A partition is a division of a graph in clusters such that each vertex belongs to one and only one cluster/subgraph, and all the edges remain the same. Community detection is a way of partitioning a graph to achieve the goal of maximizing connections within the same cluster and minimizing connections across different clusters. By generating this partition, vertices in the same cluster usually share similar features or functions, forming clusters as communities. (Girvan and Newman, 2002, Fortunato, 2010). Community detection is widely applied in fields including social networks (Newman, 2001, Girvan and Newman, 2002, Java, Song, Finin, & Tseng, 2007), computer networks (Krishnamurthy and Wang, 2000, Flake, Lawrence, Giles, & Coetzee, 2002, Hopcroft, Khan, Kulis, & Selman, 2003), biology, and ecology (Holme and Huss, 2003, Croft et al., 2006, Sundaresan, Fischhoff, Dushoff, & Rubenstein, 2007, Shimono and Beggs, 2014). Fig. 6 shows an example of community detection using graph partition.

The performance of community detection can be evaluated by modularity Q (Clauset, Newman, & Moore, 2004). Let us assume a graph that

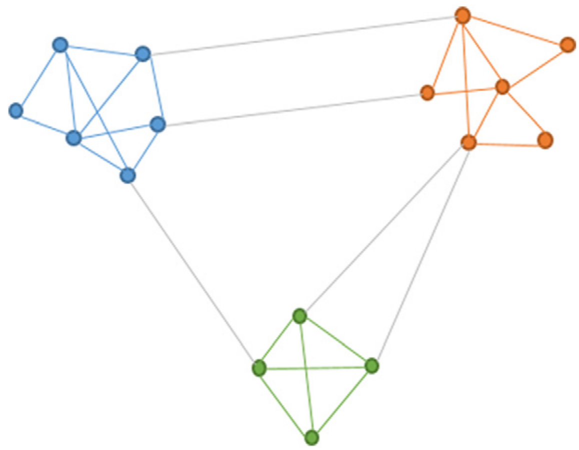


Fig. 6. Graph partition example. Three communities are identified in this case.

contains n vertices and m edges; we used A_{vw} to represent whether two vertices v and w are connected:

$$A_{vw} = \begin{cases} 1 \text{ } if \text{ } vertices \text{ } \boldsymbol{v} \text{ } and \text{ } \boldsymbol{w} \text{ } are \text{ } connected \\ 0 \text{ } otherwise \end{cases}$$

We also introduce another function $\delta(a,b)$ to calculate whether two values are equal:

$$\delta(a,b) = \begin{cases} 1 \text{ } if \text{ } a = b \\ 0 \text{ } otherwise \end{cases}$$

Then, modularity can be represented as:

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v \times k_w}{2m} \right] \delta(C_v, C_w)$$

In this function, k_v means the number of edges connected with vertex v (or v's degree), C_v means the index of community that vertex v belongs to. Here δ function acts as a filter to guarantee that only the vertices belonging to the same communities are calculated. A_{vw} represents whether vertex v and w are connected. According to the concept of configuration models (Van Der Hofstad, 2009), if we generate a random graph by maintaining the vertices number as n, edges number as m, and each vertex's degree as before, while reconnecting the vertices randomly, then the possibility of vertex v and w are connected in this random graph is $\frac{k_v \times k_w}{2m}$. Consequently, the modularity refers to the clusterness of a graph's partition in comparison with the randomly divided graph: the greater the value of modularity, the tighter the subgraphs are intra-connected, so the better a graph is divided. In practice, a modularity value greater than 0.3 means a significant detected community structure in the graph (Clauset et al., 2004).

Indeed, community detection itself is a hot research topic. Many algorithms have been designed to implement graph partition fast and efficiently (Radicchi, Castellano, Cecconi, Loreto, & Parisi, 2004, Newman, 2004, Wu and Huberman, 2004, Clauset et al., 2004). In our research, we adopted the fast greedy community detection algorithm (Clauset et al., 2004) to detect city's spatial structure of tourism.

The fast greedy community detection algorithm can be described in the following three steps:

Step 1. Treat every vertex of grid as a community, calculate the increase of modularity: ΔQ_{ij} for every pair of communities.
Step 2. Select the largest ΔQ_{ij}, join the corresponding communities, and recalculate ΔQ_{ij} between this new community with all the other communities.
Step 3. Repeat step 2 until there is only one community left.

It can be provided that there is only one peak of modularity during the process of community joining. By using a sophisticated data structure and heuristic technique, fast greedy community detection algorithm could approximately run in linear time, which is very suitable for our dataset.

To apply this algorithm, we need to transform our study area into the graph space by following steps:

1. Divide the area of Huangshan City into 500×500 m^2 grids. Each grid is represented by a vertex. Fig. 7 presents how the area of Huangshan City is divided into 201,000 grids. The grids containing geo-weibos are filled with blue, and the others are left hollow.
2. Build the connections/edges among the grids. We use tourists' daily movements to substitute edges among these grids: we have arranged user's geo-weibos by day and filtered the tourism-related days. Therefore, user's geo-weibos of days of interest can be represented as:

$$W_i^p = \{w_1, w_2, \ldots, w_n\}$$

Here i refers to $user_i$ and p refers to day_p. Each geo-weibo in W_i^p is associated with a grid it falls in. Therefore, we can get the set of grids $G_i^p = \{g_1, g_2, \ldots, g_n\}$ from W_i^p. According to the assumption2, we can construct undirected edge among each pair of grids. As many as C_n^2 edges will be constructed except those pairs who contain the same grid.
3. Execute the community detection algorithm on this dataset.

3.5. Tourism districts detection based on the community detection results

After the last step of community detection process, we can go a step further to combine the detection result with the background data of Huangshan City to get a more profound view of the attributes of these activities and how they interact with each other. By doing this, we will be able to extract the city's tourism districts.

We could classify the detected communities into three different categories according to their geometric shapes, which are surface-like, line-like, and point-like, respectively. Every shape has its own tourism meaning: surface-like regions represent tourists' spatial behavior in scenic areas or tourism service centers. Line-like regions indicate tourists' movement along transportation lines such as roads or railways. Point-like regions refer to users' shift between scenic areas and/or tourism service centers. The detailed description of these categories is listed in Table 3. In our study, we use different kinds of methods to extract these three kinds of spatial distribution patterns of grids in a community.

For the surface-like distributed grids, we can use the algorithm of density based spatial clustering of applications with noise (DBSCAN) to extract them (Ester, Kriegel, Sander, & Xu, 1996). DBSCAN is a spatial clustering algorithm, which is based on the clustering objects' spatial distribution density—densely and closely distributed objects will be clustered into the same cluster. Two parameters are needed for DBSCAN: (1) the distance threshold - *Eps*, if point p is within *Eps* of point q, then p is considered as q's neighborhood. (2) The minimum number of neighborhoods a point needs to have to be treated as a "core point" of the cluster—*Minpts*. In DBSCAN, if point p meets the standard of *Minpts*, then p will be treated as a core point, and all points within the distance of *Eps* of point p will be classified into the same cluster of point p.

Once we set the *Eps* and *Minpts*, DBSCAN will repeatedly check all points in the dataset and cluster points according to the rules we mentioned. In our study, to apply DBSCAN to the grids, we use the central points of the grids to represent them. We set Eps as two times of diagonal neighborhood distance (Fig. 8) and set Minpts as 5.

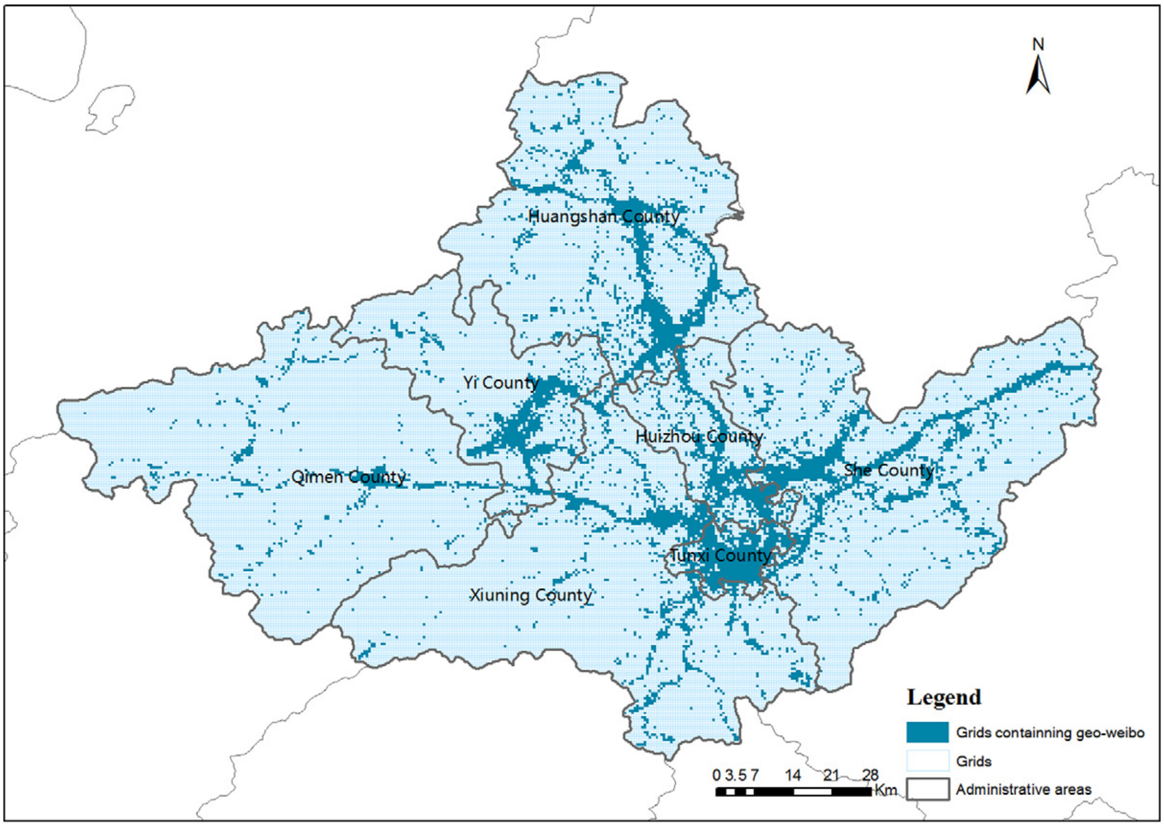


Fig. 7. Grids of Huangshan City, our study area is divided into 201,000 grids with the diameter of 500 × 500 m; blue solid grids means that geo-weibos fall in these rectangles.

For the line-like distributed grids, we assume that these points are mainly distributed along roads during tourists' commuting to or away from destinations, so we can simply use the roads' buffer region to spatially select the grids distributing along the roads. In our study, we access the road data from Open Street Map and only use roads with the level of *motorway, primary, secondary or tertiary*. The buffer distance of roads is set as 500 m, the same as the grid's side length.

For the point-like distributed grids, after the extraction of surface-like and line-like distributed grids, the rested points which scattered sparsely in the study area can be treated as point-like distributed grids.

On the basis of the recognition and analysis of the communities, we can go a step further to extract the city's tourism districts by following below steps:

Step 1. Determine the integral surfaces clustered by grids from the same communities in space, at the same time merge the embedded grids from other activity regions. These grids refer to the interaction between different tourism regions but are supposed to belong to the local tourism district.

Step 2. Recognize the grids distributing along transportation lines. These grids represent the tour when tourists were heading to or leaving tourism regions, so we can safely delete these grids.

Step 3. Recognize the remaining grids scattering solely and far away from any tourism districts. These grids may refer to users' casual behaviors or spatial movement to some low-level scenic spots or communities, which are not recorded in the background material. Because these behaviors are quite rare and not typical, we can also delete these grids. Before removing the isolating grids, double check if any of them containing plenty of geo-weibos. If yes, find the reason and preserve them if necessary.

4. Results

By applying the community-detecting algorithm, the grids in Huangshan city are clustered into 14 distinct communities. Table 4 demonstrates the details of these communities. The biggest six communities contain the immense majority of the total grids. From the field of "mean geo-weibo per grid," we can find that geo-weibos are also concentrated on these big communities, leaving several isolated weibos to the rest of the communities. We assume that communities from 7 to 14 may be generated from several users' isolated and random activities, which have no representativeness. Hence, for the following analysis, we will focus on the communities of 1–6. The overview of these communities is shown in Fig. 9.

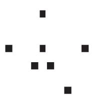
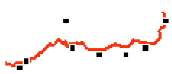
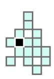
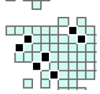
From Fig. 9, we can find that every community has its own dominant region in space, which is formed by concentrated grid members, appearing as relatively "whole surface." Each community also includes a portion of members locating far away from their own dominant region, some even within other regions, revealing shapes of disordered points or lines in the city. The modularity of the community detection result is 0.398. According to Clauset et al.'s practice, modularity greater than 0.3 means a relatively good value as a graph partition result, indicating the community structure of this user-spatial-behaviors-based city space division is significant (Clauset et al., 2004).

Details of the community detection results are presented in Fig. 10 by different communities separately.

Combining the community detection results in Fig. 10 and the background information about Huangshan City, we can find some interesting features among these communities:

For all the six communities, they share a common pattern that the domain regions are consist of important scenic areas together with populated areas nearby: C1.: Mt. Huangshan scenic area and a small town

Table 3
Summary of grid distribution patterns.

Element	Type	Example	Spatial distribution features	Tourism related features	Tourism activity	Extraction method
Surface	Clustering		Containing relatively large number of grids. The densely distributed grids in space could form an integrated surface. Region may be embedded by grids from other elements.	(1) Tourism attraction: consists of the area of scenic spot. Tourists' activity level is usually high, which can be reflected by the geo-weibo number. (2) Tourism service center: cities, towns, or densely populated counties who play a role of tourism service center, supplying accommodations, food and transportation for visitors. (3) Group of tourism attractions: Along with the development of the tourism industry, spatial adjacent tourism attractions and/or tourism service centers may merge into big and integrated tourism areas.	Tourism activities.	DBSCAN
	Dispersing		Grids in this type if region are sparsely scattered in space. Grids from other regions are rarely embedded there.	Tourists' activity level is not very high. Tourism attraction here has a low level of attracting ability, which may need to be developed further.		
Line			Grids distribute along a line object (such as road or river) in space.	This type of grid mainly spreads along the roads connecting tourism attractions with other tourism attractions, tourism service center, or densely populated region (tourism market). These kinds of grids may cluster in highways' exits and service areas. What's more, the number of line type grids could reflect scenic area's attracting ability.	During the route of heading to or leaving the scenic area.	Road buffer based spatial query
Point	Embedded	(1)  (2) 	Grids are embedded in other region's surface. The number of grids varies from little (1) to many (2).	Tourism distributing center or transit station, such as the assembly points of package tours, airport, train station, and so forth. This type of grids can be used to refer to the relation strength between two areas.	Tourists' transit process form one place to another.	The left grids after former steps
	Isolated		Grids scatter in isolation	(1) Scenic spots of low level (2) Small residence community (3) Places tourists reach by chance	Tourists' fortuitous activity.	

Tangkou; C2.: Hongcun ancient village, Xidi ancient village, together with Yi County's downtown; C3.: Tunxi County's downtown, Xiuning County's downtown and some scenic spot around the downtown areas, such as Mt. Qiyun, and so forth; C4.: Taiping Lake scenic area, Jade Valley and Huangshan County's downtown; C5.: Huizhou Ancient Town, the Shanshui scenic area of Xin'an River and She County's downtown; and C6.: Mt. Gu'niujiang and Qimen County's downtown. Such results make good sense because the communities are extracted from tourists' daily traveling routes during which their essential and entertainment needs are met. Therefore, from the results, we now know that each community acts as a relatively independent unit for tourists' daily trips. Inside each community, the scenic areas are attractions for tourists, and the populated area plays a role of service center for the scenic areas.

Hollow grids in Fig. 10 are those distributed along the roads. It is intuitive to infer that the related geo-weibos are published during tourists' tour heading to or leaving from the scenic areas. Therefore, these lines can also be used to indicate the relation between the tourism destinations with other tourism service centers.

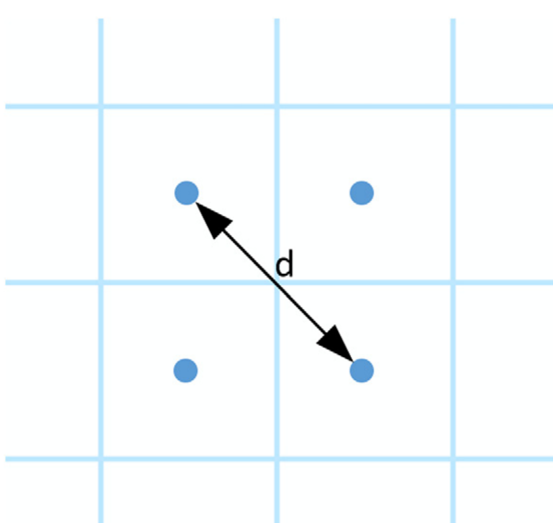


Fig. 8. Distance threshold of DBSCAN, we set Eps as two times of diagonal neighborhood distance (2d). In our study, it is set as 1414 m.

Table 4
Statistics of communities.

Community Id	Grids number	Geo-weibo number	Mean Geo-weibos per grid
1	452	33702	74.40
2	597	20798	34.84
3	631	13640	21.62
4	306	3906	12.76
5	350	4300	12.29
6	50	681	13.62
7	4	6	1.50
8	3	2	0.67
9	3	2	0.67
10	3	3	1.00
11	3	2	0.67
12	3	2	0.67
13	3	2	0.67
14	3	2	0.67
summary	2411	77048	31.94

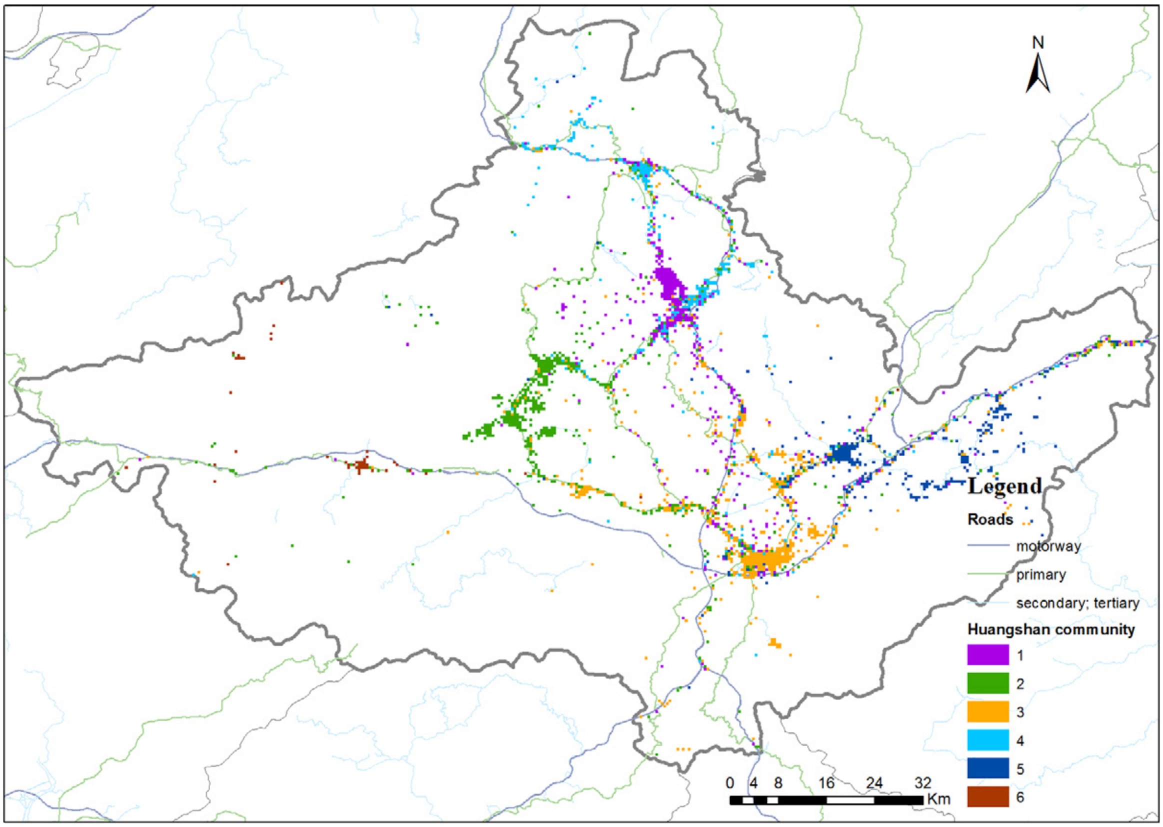


Fig. 9. Communities detected in Huangshan City after applying the community detection algorithm.

There are some grids belonging to one community but being embedded into the domain region of other communities (Fig. 9). For example, some grids of C1 locate in Tunxi County's downtown(C3) and Huangshan County's downtown(C4). These grids indicate that how tourists transfer among different regions: the main transportation hubs mainly locate in Tunxi. Tourists from other cities usually arrive at Tunxi downtown first before heading to Mt. Huangshan. These grids can also represent the location of hotels where the tourists finish their visiting of one community and begin visiting another community the next day. In general, such grids represent the strong connections among different communities.

The grids that spread randomly in the communities are colored in gray. Related geo-weibos in these grids are generated by a small number of individuals and do not show any significant meaning for our study.

By following the steps mentioned in Section 3.5 to delete the grids distributed along the roads and merge the grids embedded in the domain areas of each community, we can obtain the final tourism districts of Huangshan City, which are shown in Fig. 11. These six districts come from the biggest six communities and are named after the major scenic spots or administrative regions they belong to[3]:

1. **Huangshan Scenic Area**: this district consists of: (1) the core scenic area of Mt. Huangshan; (2) less important regions spreading around the border of the scenic core; and (3) the local tourism service center—Tangkou Town. Although only 18.98% of grids in total belong to this region, the geo-weibo published in this region occupied 43.75% of the total, which indicates Mt. Huangshan scenic area is the most important tourism product of Huangshan City.
2. **Ancient Villages in Southern Anhui**: the scenic areas of Xidi ancient village, Hongcun ancient village, Nanping ancient village, Guiy Garden together with Yi County's downtown link up into a single stretch, representing a big and integral cultural tourism region. The core tourism products of this region are its cultural sights—Xidi and Honcun Ancient Villages.
3. **The Tourism Vacation Belt around Tunxi County**: this tourism district center at Huangshan City's downtown—Tunxi County. Tunxi County serves as the tourism service center of Huangshan City: the train station, long-distance bus station, and airport of Huangshan are all located here. Besides supplying the basic services for tourists from other tourism regions, this region also plays a role of tourism region itself, which supplies various and integral tourism products for tourists. For the downtown, it can supply amusement and leisure services. In surrounding areas, also there are many AAAA and AAA state-level cultural and natural spots such as Mt. Qiyun, Chengkan village, Qiankou village, Tangyue memorial archway park, Huashan cave, Guchengyan, former site of the New Fourth Army's Headquarters, Huizhou culture park, Mt. Bawang scenic area, and so forth. According to tourists' activity preferences, Mt. Qiyun and Xiuning County (at northwest of Tunxi downtown) are also classified into this tourism region even though they are located a little far away from Tunxi County.
4. **The Tourism Vacation Belt around Gantang Town**: this tourism district lies in the north of Huangshan City, the core of the region is Gantang County, which indeed is the secondary tourism service center of Huangshan City. This region also involves two parts of scenic spots lying in the north and south of Gantang Town. The north part is mainly occupied by Taiping Lake, and the south part consists of scenic spots such as Jade Valley, Shimen Canyon, Nine-Dragon Waterfall, and so forth. The fact that part of this region comes from Mt. Huangshan scenic area infers that Gantang town indeed plays a role of service center for Mt. Huangshan.
5. **Huizhou Ancient Town**: the core of this tourism district is occupied by Huizhou ancient town scenic area and the downtown of Huizhou

[3] Information about the administration areas, landmarks and tourism districts of Huangshan city can be find in Figure 1, Figure 10, Figure 11 and Table 1

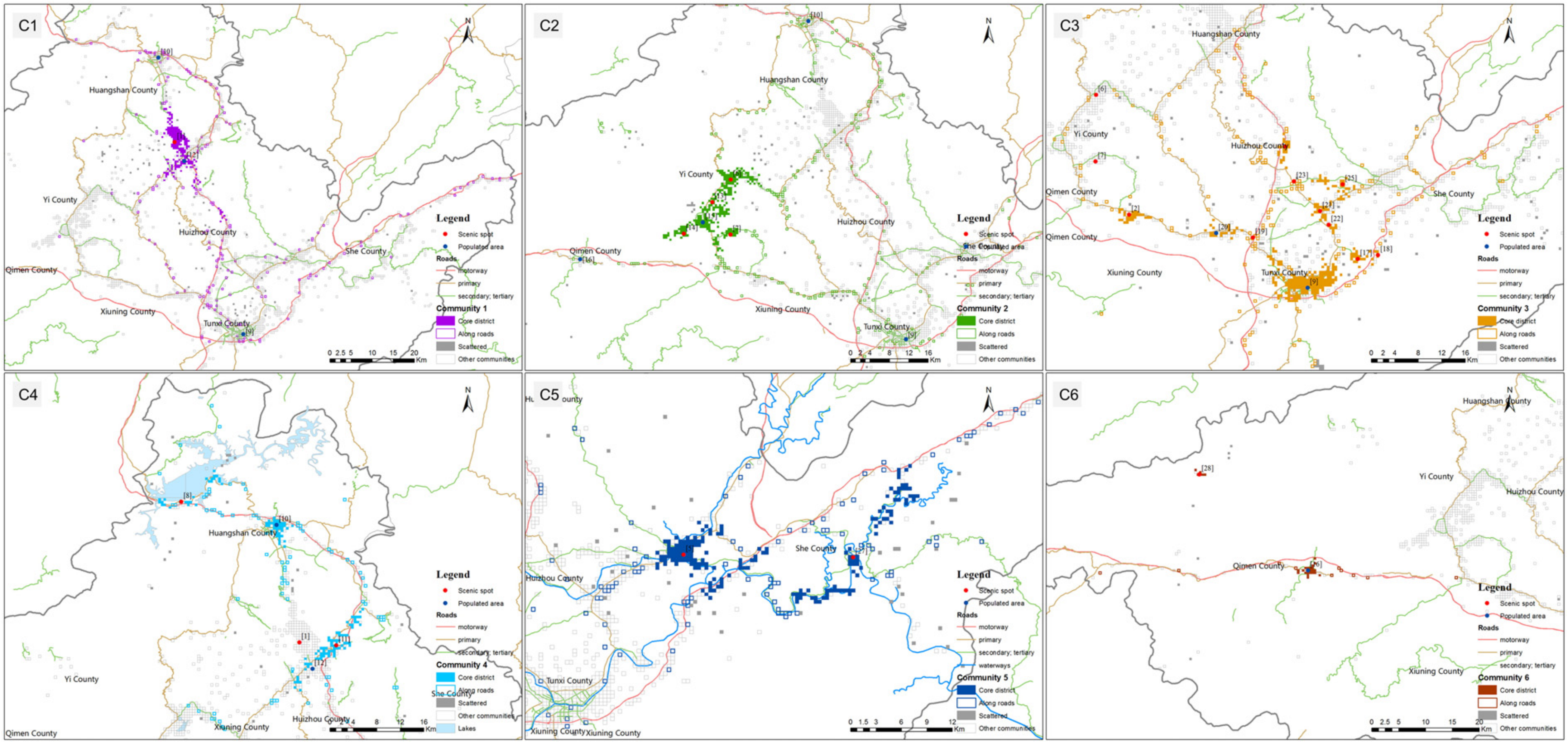


Fig. 10. Spatial patterns of the biggest six communities. The grids in the core districts of these communities are presented with solid colors. Hollow grids are those distributing along roads. Scattered grids and grids from other communities are colored gray. Red dots represent scenic spots and blue dots represent populated areas.

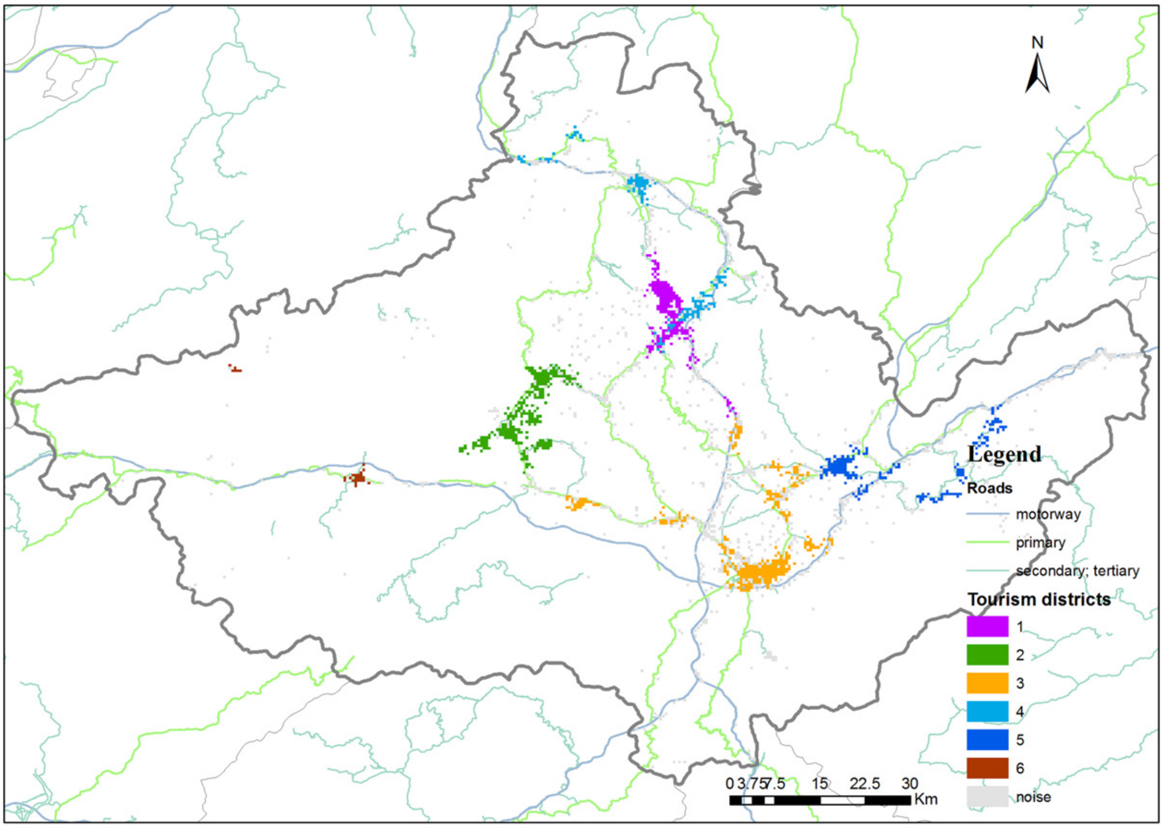


Fig. 11. Tourism districts of Huangshan City. After merging the embedded grids of each core district and removing the noise grids, we can get the final core tourism districts of Huangshan City.

County. This region also involves the scenic spots belt along the upper reach of Xin'an River.

6. **Gu'niujiang Scenic Area**: this is the smallest tourism district we detected in Huangshan City, lying solely in the west of Huangshan City and seems relatively independent of other districts. This region consists of Guniujiang scenic area and its service center—the downtown of Qimen County.

5. Discussion and conclusion

From the results of the extracted tourism districts in the previous section, we can find that each tourism district's scenic spots are concentrated in space and consistent in type (natural, cultural, or recreational). Each tourism district has its core tourism attraction and service center. The scope of each tourism district appropriately overlaps with tourists' daily accessible area. There is not a scenic area been divided into two tourism districts. We validated our results by comparing them with the tourism planning products provided by Huangshan Travel and Tourism Committee and third parties[4,5]. In addition, because our results come from tourists' activities and their interaction with the environment, they could reveal much more information about the tourism districts such as common sequences among districts by tourists, the temporal pattern of their movements, and the difference of movements between tourists and local residents. We could even monitor the development of tourism districts in a long term as a city is continuously expanding or being developed. In this paper, we mainly introduce the methodological pipeline for extracting the tourism districts. The work of analyzing how tourists moving among these districts will definitely be conducted in the next step. Consequently, we believe that the knowledge discovered from this new data source could help urban planners get more comprehensive and accurate understanding of the city's tourism status.

To sum up, our study makes a major contribution to the data-driven science and the tourism studies in the following aspects. First, we introduced social media data to tourism study and developed a comprehensive methodology framework for tourism districts extraction. Although there are some concerns about the representativeness of social media data in the field of people's activity study, several previous researches and our preliminary research have shown that social media data are a reliable data source for tourism research (Ahas et al., 2008; Wood et al., 2013). Second, our data-driven method bridges tourists' spatiotemporal activities with city's tourism spatial structure, which makes the detection result meaningful. Third, our quantitative methodology is able to delineate explicit boundaries of each tourism district and can be used to evaluate the importance of each district. Fourth, our methodology can be used for long-term monitoring of the development and reshaping process of tourism districts along with the development of the city. Hence, it has practical application in guiding the urban plan and city tourism development or evaluating the planning result.

Acknowledgements

This work was supported by funds from the National Science Foundation of China grants 41625003.

References

Ahas, R., Aasa, A., Roose, A., Mark, U., & Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, *29*(3), 469–486.

Asakura, Y., & Iryo, T. (2007). Analysis of tourist behaviour based on the tracking data collected using a mobile communication instrument. *Transportation Research Part A: Policy and Practice*, *41*(7), 684–690.

[4] http://www.hsta.gov.cn/xinxi/html/150602115628.html
[5] http://wenku.baidu.com/view/4e0063593b3567ec102d8a55.html?pn=51

Baležentis, T., Kriščiukaitienė, I., Baležentis, A., & Garland, R. (2012). Rural tourism development in Lithuania (2003 − 2010)—A quantitative analysis. *Tourism Management Perspectives*, *2*, 1–6.

Cranshaw, J., Schwartz, R., Hong, J. I., & Sadeh, N. (2012). *The livehoods project: Utilizing social media to understand the dynamics of a city.*

Cheng, Z., Caverlee, J., Lee, K., & Sui, D. Z. (2011). Exploring millions of footprints in location sharing services. *ICWSM*, *2011*, 81–88.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, *70*(6), 066111.

Crandall, D. J., Backstrom, L., Huttenlocher, D., & Kleinberg, J. (2009, April). *Mapping the world's photos* (pp. 761–770)Proceedings of the 18th international conference on world wide web. ACM.

Croft, D. P., James, R., Thomas, P. O. R., Hathaway, C., Mawdsley, D., Laland, K. N., & Krause, J. (2006). Social structure and co-operative interactions in a wild population of guppies (*Poecilia reticulata*). *Behavioral Ecology and Sociobiology*, *59*(5), 644–650.

Donaire, J. A., Camprubí, R., & Galí, N. (2014). Tourist clusters from Flickr travel photography. *Tourism Management Perspectives*, *11*, 26–33.

Dredge, D. (1999). Destination place planning and design. *Annals of Tourism Research*, *26*(4), 772–791.

Edwards, D., & Griffin, T. (2013). Understanding tourists' spatial behaviour: GPS tracking as an aid to sustainable destination management. *Journal of Sustainable Tourism*, *21*(4), 580–595.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd. Vol. 96.* (pp. 226–231) (No. 34).

Fennell, D. A. (1996). A tourist space-time budget in the Shetland Islands. *Annals of Tourism Research*, *23*(4), 811–829.

Flake, G. W., Lawrence, S., Giles, C. L., & Coetzee, F. M. (2002). Self-organization and identification of web communities. *Computer*, *35*(3), 66–70.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3), 75–174.

Furletti, B., Cintia, P., Renso, C., & Spinsanti, L. (2013, August). *Inferring human activities from gps tracks* (pp. 5)Proceedings of the 2nd ACM SIGKDD international workshop on urban computing. ACM.

García-Palomares, J. C., Gutiérrez, J., & Mínguez, C. (2015). Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS. *Applied Geography*, *63*, 408–417.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), 7821–7826.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*(4), 211–221.

Gunn, C. A., & Var, T. (2002). *Tourism planning: Basics, concepts, cases.* Psychology Press.

Guo, Z., Li, Z., & Tu, H. (2011, October). *Sina microblog: An information-driven online social network* (pp. 160–167)Cyberworlds (CW), 2011 international conference on. IEEE.

Hartmann, R. (1988). Combining field methods in tourism research. *Annals of Tourism Research*, *15*(1), 88–105.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, *41*(3), 260–271.

Hollenstein, L., & Purves, R. (2015). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, *1*, 21–48.

Holme, P., & Huss, M. (2003). *Discovery and analysis of biochemical subnetwork hierarchies.* (arXiv preprint q-bio/0309011).

Hopcroft, J., Khan, O., Kulis, B., & Selman, B. (2003, August). *Natural communities in large linked networks* (pp. 541–546)Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining. ACM.

Hu, Y., Gao, S., Janowicz, K., Yu, B., Li, W., & Prasad, S. (2015). Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems*, *54*, 240–254.

Huang, L., Li, Q., & Yue, Y. (2010, November). *Activity identification from GPS trajectories using spatial temporal POIs' attractiveness* (pp. 27–30)Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks ACM.

Itami, R., Raulings, R., MacLaren, G., Hirst, K., Gimblett, R., Zanon, D., & Chladek, P. (2003). RBSim 2: Simulating the complex interactions between human movement and the outdoor recreation environment. *Journal for Nature Conservation*, *11*(4), 278–286.

Java, A., Song, X., Finin, T., & Tseng, B. (2007, August). *Why we twitter: Understanding microblogging usage and communities* (pp. 56–65)Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on web mining and social network analysis. ACM.

Keul, A., & Kuhberger, A. (1997). Tracking the Salzburg tourist. *Annals of Tourism Research*, *24*(4), 1008–1012.

Krishnamurthy, B., & Wang, J. (2000). On network-aware clustering of web clients. *ACM SIGCOMM Computer Communication Review*, *30*(4), 97–110.

Lau, G., & McKercher, B. (2006). Understanding tourist movement patterns in a destination: A GIS approach. *Tourism and Hospitality Research*, *7*(1), 39–49.

Lew, A., & McKercher, B. (2006). Modeling tourist movements: A local destination analysis. *Annals of Tourism Research*, *33*(2), 403–423.

Lew, A. A. (1992). Place representation in tourist guidebooks: An example from Singapore. *Singapore Journal of Tropical Geography*, *12*(2), 124–137.

Lew, A. A., & McKercher, B. (2002). Trip destinations, gateways and itineraries: The example of Hong Kong. *Tourism Management*, *23*(6), 609–621.

Liu, Y., Sui, Z., Kang, C., & Gao, Y. (2014). Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. *PloS one*, *9*(1), e86026.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., Chi, G., & Shi, L. (2015). Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers*, *105*(3), 512–530.

Lu, Y., & Liu, Y. (2012). Pervasive location acquisition technologies: Opportunities and challenges for geospatial studies. *Computers, Environment and Urban Systems*, *36*(2), 105–108.

McKercher, B., Wong, C., & Lau, G. (2006). How tourists consume a destination. *Journal of Business Research*, *59*(5), 647–652.

Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, *98*(2), 404–409.

Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E*, *69*(6), 066133.

Oppermann, M. (1994). Regional aspects of tourism in New Zealand. *Regional Studies*, *28*(2), 155–167.

Oppermann, M. (1995). A model of travel itineraries. *Journal of Travel Research*, *33*(4), 57–61.

Page, S., & Connell, J. (2014). Transport and tourism. *The Wiley Blackwell companion to tourism* (pp. 155–167).

Pearce, D. G. (1979). Towards a geography of tourism. *Annals of Tourism Research*, *6*(3), 245–272.

Pearce, D. G. (1998). Tourist districts in Paris: Structure and functions. *Tourism Management*, *19*(1), 49–65.

Pearce, D. G. (2001). An integrative framework for urban tourism research. *Annals of Tourism Research*, *28*(4), 926–946.

Pons, A., Salamanca, O. R., & Murray, I. (2014). Tourism capitalism and island urbanization: Tourist accommodation diffusion in the Balearics, 1936–2010. *Island Studies Journal*, *9*(2), 239–258.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(9), 2658–2663.

Rebollo, J. F. V., & Baidal, J. A. I. (2003). Measuring sustainability in a mass tourist destination: Pressures, perceptions and policy responses in Torrevieja, Spain. *Journal of Sustainable Tourism*, *11*(2–3), 181–203.

Savage, V. R., Huang, S., & Chang, T. C. (2004). The Singapore river thematic zone: Sustainable tourism in an urban context. *The Geographical Journal*, *170*(3), 212–225.

Shimono, M., & Beggs, J. M. (2014). Functional clusters, hubs, and communities in the cortical microconnectome. *Cerebral Cortex*, bhu252.

Shoval, N., & Isaacson, M. (2007). Tracking tourists in the digital age. *Annals of Tourism Research*, *34*(1), 141–159.

Shoval, N., McKercher, B., Ng, E., & Birenboim, A. (2011). Hotel location and tourist activity in cities. *Annals of Tourism Research*, *38*(4), 1594–1612.

Spinsanti, L., Celli, F., & Renso (2010). *Where you go is who you are: Understanding people's activities by places visited* (pp. 38–52)In: BMI 2010 - 5th workshop on behaviour monitoring and interpretation, CEUR workshop proceedings. BMI.

Sundaresan, S. R., Fischhoff, I. R., Dushoff, J., & Rubenstein, D. I. (2007). Network metrics reveal differences in social organization between two fission–fusion species, Grevy's zebra and onager. *Oecologia*, *151*(1), 140–149.

Teo, P., & Huang, S. (1995). Tourism and heritage conservation in Singapore. *Annals of Tourism Research*, *22*(3), 589–615.

Trudeau, R. J. (2013). *Introduction to graph theory.* Courier Corporation.

Van Der Hofstad, R. (2009). Random graphs and complex networks. Retrieved from http://www.win.tue.nl/rhofstad/NotesRGCN.pdf.

White, N. R., & White, P. B. (2007). Home and away: Tourists in a connected world. *Annals of Tourism Research*, *34*(1), 88–104.

Wood, S. A., Guerry, A. D., Silver, J. M., & Lacayo, M. (2013). Using social media to quantify nature-based tourism and recreation. *Scientific Reports*, *3*.

Wu, F., & Huberman, B. A. (2004). Finding communities in linear time: A physics approach. *The European Physical Journal B-Condensed Matter and Complex Systems*, *38*(2), 331–338.

Xing-zhu, Y., & Qun, W. (2014). Exploratory space–time analysis of inbound tourism flows to China cities. *International Journal of Tourism Research*, *16*(3), 303–312.

Yan, L. J. (2004). A cross-cultural study of international tourist flows in China's tourism regions. *Tourism Geographies.*

Ye, M., Janowicz, K., Mulligann, C., & Lee, W. C. (2011, November). *What you are is when you are: The temporal dimension of feature types in location-based social networks* (pp. 102–111)Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems. ACM.

Yin, Z., Cao, L., Han, J., Luo, J., & Huang, T. S. (2011, April). Diversified trajectory pattern ranking in geo-tagged social media. *SDM* (pp. 980–991).

Yuan, J., Zheng, Y., & Xie, X. (2012, August). *Discovering regions of different functions in a city using human mobility and POIs* (pp. 186–194)Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. ACM.

Zhai, S., Xu, X., Yang, L., Zhou, M., Zhang, L., & Qiu, B. (2015). Mapping the popularity of urban restaurants using social media data. *Applied Geography*, *63*, 113–120.

Zhou, X., Xu, C., & Kimmons, B. (2015). Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform. *Computers, Environment and Urban Systems*, *54*, 144–153.