IOP Conf. Series: Materials Science and Engineering **231** (2017) 012046 doi:10.1088/1757-899X/231/1/012046

Research of the distribution of tourists' attributes based on internet data: A case study of Kunming

Bingyang Chen[1,2], Kun Yang[1,2] , Jiasheng Wang[1,2,*]

[1]School of Information Science and Technology, Yunnan Normal University, Kunming 650500, China,
[2]The Engineering Research Center of GIS Technology in Western China Ministry of Education of China, Kunming 650500, China

[*]Corresponding author

Abstract. With the development of the era of big data, the ever-growing user trajectory provides the basis for studying multi-scale tourist activity law. This paper selected 17 famous tourist attractions in Kunming. Sina Microblog, Ctrip Travel, Lvmama Travel Network and other platforms were used to extract 139727 records between Oct. 2015 and Sep. 2016. The methods of data mining and clustering analysis were used to explore the activity characteristics of tourists with different attributes in scenic spot and the activity differences of different age tourists in different scenic spots affected by season, not only considered gender, geographical, check-in time and other factors, but also the introduced age attributes. At the same time, the scenic area is divided into "Adolescent active pattern ","Young and middle-aged women active pattern"," Middle-aged and old men active pattern" and " General active pattern" according to different tourists' activities law of different gender and age in spatial perspective. Research shows that female tourists are mainly distributed in the Green Lake Park, Nanping Street, Dounan Flower Market and other attractions, elderly male tourists are mainly distributed in Expo Park, Jindian area. Foreign tourists accounted for 86.32% of the total tourists, reflecting the rapid development of tourism in Kunming. The spatial distribution of tourist attractions has an impact on the distribution of tourists' attributes. The number of tourists of Shilin, Jiuxiang, Guandu Ancient Town are accounted for 36.38% of the total tourists, which shows that the spatial distribution of tourist attributes is consistent with the development of key tourist areas in Kunming.

1 Introduction

With the development of the information age, modern people can express their feelings and communicate with the outside world through the Internet. Especially in the process of tourism, tourists express their feelings and pictures and so on in the scenic spots. The study of a large number of tourists' behaviour data can promote the rapid development of tourism. The traditional travel log is difficult to obtain large-scale information while Internet data with sample data sufficiently large, accurate, real-time data, data source diversity, can support user attribute distribution features research effectively.

With the increasing variety of data sources and the maturity of data processing methods, spatial-temporal data analysis has attracted more and more attention in the field of urban geography [1,2,3]. For example, based on the taxi GPS positioning data [4], bus card data[5] , mobile phone signalling data [6,7]and so on to study the traffic characteristics of internal city [8],the traffic hot spot distribution [9], the

temporal-spatial characteristics of traffic flow [10], the resident trip characteristics [11,12] and the population mobility [13]. Many scholars have done in-depth research on the behaviour characteristics of users. Such as Wang Bo [14] with sina micro-blog location service data, take Nanjing city as an example, the regional division of urban activities by cluster analysis from the time, space, activity three aspects; Li Xiang [15] find out the differences of different gender and geographical preference for scenic spot, the low and peak seasons in scenic spots and the distribution similarity of the low and peak seasons from the user's gender, region and check-in time.

Generally speaking, most of the existing researches on user behaviour ignore the user's age, and study the relationship between the distribution of user attributes and the nature of regional space is less. In view of this, this study selected Kunming city as the research area, the introduction of age attributes, from gender, geographical, age and release time and other aspects of information mining. Then, the difference of activity characteristics of tourists was analysed with different attributes in the scenic area and the difference of tourists' activities in different scenic spots affected by age and season. At the same time, the types of scenic spots were discussed based on the differences of tourists' gender and age attributes, it analysed the corresponding relationship between tourist attribute distribution and spatial characteristics of different types of scenic spots. The study provides a reference for the future study on the distribution of tourists' attributes, it provides the basis for the management and planning of scenic spots, and it provides a reference for tourists to select the scenic spots, thus promoting the development of tourism economy in Kunming.

2 Data and method

2.1 Data acquisition and pre-processing

This paper is based on the Internet data, and in the spring city Kunming is studied as an example. First of all, it chose 17 major tourist attractions in Kunming (Stone Forest, Jiuxiang, Guandu Old Town, Cuihu Park, Daguan Park, Nationalities Village, Nanping Street, Golden Horse and Jade Rooster, Expo Park, Dian Lake, Xishan, Dounan Flower Market, Military Academy, Anning Hot Spring, Andy Scenic Spot, Panlong Temple, Safari Park) ,recorded their latitude and longitude; Then, it called Sina micro-blog access to nearby locations API interface ,to ensure coverage of all attractions searched for radius with 5km.It selected from November 1, 2015 to October 31, 2016 a year time span, and it collected the user's ID, gender, location (attribution), birthday, check-in spot, check-in time and other 6 data. The statistics of relevant data are shown in table 1.

Table 1: The data of tourists' check-in

id	*gender*	*location*	*birthday*	*check-in spot*	*check-in time*
51xxxx7272	female	Sichuan Chengdu	1991-2-12	Golden Horse and Jade Rooster	2016-3-12 19:00:12
21xxxx6572	male	Guizhou Bijie	1997-6-14	Stone Forest	2016-9-18 15:25:18
...	...	...	...	...	...

The scenic spot records were extracted manually from Ctrip Travel, Lvmama and other tourism platforms, and then received a total of 139727 data.

Table 2: Scenic spots and check-in number

scenic spots	*check-in number*	*scenic spots*	*check-in number*
Stone Forest	20262	Dian Lake	6016
Jiuxiang	11970	Xishan	5043
Guandu Old Town	10201	Dounan Flower Market	4672
Cuihu Park	8348	Military Academy	3946

Daguan Park	8292	Anning Hot Spring	3291
Nationalities Village	7812	Andy Scenic Spot	2502
Nanping Street	7695	Panlong Temple	2446
Golden Horse and Jade Rooster	7537	Safari Park	1567
Expo Park	6845		

Due to the scope of the scenic spots available check-in point too much, this mixed with the data of ordinary users, so the first is data pre-processing. In order to distinguish between tourists and ordinary users check-in data, remove the enterprise companies, schools, banks and other places where tourists check-in less likely; The user data for the location is not clear and the check-in point on the boundary of the study area is eliminated. Finally, 115280 records were extracted, as shown in table 2.

2.2 Research methods

In this study, the study area is defined as 17 large tourist attractions in Kunming. The methods of data mining and clustering analysis were used to explore the activity laws of tourists with different attributes in scenic spots. Specific methods are as follows:

The Z-test is used for information mining, reflect the different regions (local and foreign), gender (male, female) tourists of different types of attractions of the preference differences, and its quantitative representation. Calculate the mean value (μ_0), standard deviation (σ) of the whole release number, sample mean value ($\bar{X}$) and sample number (n).The specific calculation formula is shown in formula (1) and formula (2).

$$\sigma = \sqrt{\frac{1}{\text{n-1}}\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad (1)$$

$$\text{Z} = \frac{\bar{X}\text{-}\mu_0}{\frac{\sigma}{\sqrt{n}}} \quad (2)$$

Data normalization processing to obtain X^*. The chi-square test is used to reflect the different age, season tourists of different types of attractions of the preference differences, and calculate the chi-square test(χ^2) , formula as follows:

$$X^* = \frac{X_i\text{-min}}{\text{max - min}} \quad (3)$$

$$\chi^2 = \sum_{i}^{n}\frac{\left(f_i^0 - f_i^e\right)^2}{f_i^0} \quad (4)$$

With f_i^0 represents theory release number, f_i^e represents actual release number, X^* is the data obtained after the actual release number normalization.

After the sample standard deviation (S) is got, the difference in activity of tourists affected by age and season in several scenic spots samples is analysed.

$$S = \sqrt{\frac{1}{\text{n}}\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad (5)$$

$$\text{Z} = \frac{\bar{X}_1\text{-}\bar{X}_2}{\sqrt{\frac{\text{S}_1^2}{n_1} + \frac{\text{S}_2^2}{n_2}}} \quad (6)$$

With $\bar{X}_1, \bar{X}_2$ is the mean value of sample 1, sample 2, S_1, S_2 is the standard deviation of sample 1, sample 2, n_1, n_2 is the capacity of sample 1, sample 2.

In addition, according to the gender and age of the two attributes of the impact on the different degree of tourist behaviour of the hierarchical clustering analysis, and then the distribution of tourist attractions is classification discussed.

3 Tourist attributes information mining

After the statistical analysis of the tourist activity laws, it can be found that female tourists were more than male in full year, foreign tourists were more than the local; The main groups were young and middle-aged tourists, its behaviour was affected by the difference in low and peak seasons. Which is similar with Zhang Ziang [16], further explore the differences in the behaviour of tourists with different attributes of a single attraction.

3.1 Information mining based on "gender" and "location"

In the 122386 records extracted, of which the male was 52308, female was 70078, the ratio of male to female was 42.74:57.26. Expo Park as a sample for a total of 6845 data, the ratio of male to female was 48.5:51.5. Assuming that the sample is the same as the whole trend, the female tourists are more than the male. After calculating the standard deviation and mean value of the whole male tourists, the Z-test parameters is got Z_{gender}=8.07>1.96, so the original assumption is not established. There is a difference between the sample and the total data of female tourists are more than male tourists. It can be seen that the attractions of female tourists are still more, but men prefer to go to Expo Park.

Region factors also use the Z-test. The ratio of local tourists and foreign tourists in sample is 17.56:82.44. The Z-test parameters is got Z_{region}=7.84>1.96, in the region is also contrary to the whole trend. It can be seen that local tourists prefer Expo Park. The degree of the difference of tourists' activities affected by gender and location is shown in Table 3.

Table 3: Gender and location impact on tourists' activities

scenic spots	Z_{gender}	Z_{region}	*scenic spots*	Z_{gender}	Z_{region}
Stone Forest	1.71	1.26	Xishan	1.72	2.09
Jiuxiang	1.68	1.52	Dounan Flower Market	1.79	1.98
Guandu Old Town	1.63	0.96	Military Academy	0.82	1.24
Cuihu Park	0.83	0.38	Anning Hot Spring	1.98	0.97
Daguan Park	1.88	2.04	Andy Scenic Spot	1.21	8.16
Nationalities Village	1.85	1.94	Panlong Temple	7.12	6.52
Nanping Street	0.95	0.41	Safari Park	6.83	7.23
Golden Horse and Jade Rooster	0.94	0.37	Dian Lake	1.64	0.34
Expo Park	8.07	7.84			

As can be seen from table 3, the Z_{gender} of Expo Park, Andy Scenic Spot, Panlong Temple and other places is greater than 1.96, description of this kind of attractions more preference for male tourists; At the same time the Z_{region} of this kind of scenic spots as well as the Anning hot spring is greater than 1.96, local tourists are more likely to visit these scenic spots can we get. It shows that there are some differences in the distribution of tourists' attributes and the differences in the overall situation.

3.2 Information mining based on "age" and "season"

Also take Expo Park as an example, put 10-70 years of age take 10 years as a unit is divided into 6 age groups. The age of tourists mainly concentrated in the 40-50 age range, release times up to 1412 times, the minimum number of published 10-20 age range, a total of 563 times. First of all, the data were normalized and then calculated Chi-square statistic, the χ^2_{age} of Expo Park is 12.52, this shows that the Expo Park sample compared with the overall, tourist behaviour are greatly influenced by the age property, mainly reflected in the middle and old people in this region a little more. This method can also be used to investigate whether the sample is affected by season, divide the year into 12 months, χ^2_{season} of Expo Park is 3.28, compared to the χ^2_{season}(11.35) of the Anning hot springs, Expo Park is limited by seasonal factors.

The Z-test was used to analyse the degree of difference between tourists' behaviour in the sample by age and season. The Expo Park and the Dounan flower market tourists behaviour affected by age, seasonal difference degree can be got after calculation. Z_{age}=2.55, Z_{season}=3.34, the two are greater than

1.96, therefore, the differences of two scenic spots are affected by the age and seasonal attributes are large. While the Expo Park and the Panlong Temple tourists' behaviour affected by age, seasonal difference degree was 0.58 and 0.73 respectively, it can be seen that the similarity between Expo Park and Panlong temple is high. So it can be used to measure the similarity of tourist behaviour in different scenic spots .The difference of tourists' activities in the scenic area affected by age and season is shown in Table 4.

Table 4: Age and season impact on tourists' activities

scenic spots	χ^2_{age}	χ^2_{season}	*scenic spots*	χ^2_{age}	χ^2_{season}
Stone Forest	2.56	0.82	Xishan	8.64	3.64
Jiuxiang	2.47	0.97	Dounan Flower Market	8.23	9.56
Guandu Old Town	2.71	1.09	Military Academy	4.88	1.36
Cuihu Park	3.92	10.03	Anning Hot Spring	6.23	3.28
Daguan Park	7.96	3.82	Andy Scenic Spot	1.54	11.35
Nationalities Village	8.17	0.97	Panlong Temple	11.87	3.63
Nanping Street	4.63	2.67	Safari Park	13.26	3.58
Golden Horse and Jade Rooster	4.67	2.82	Dian Lake	8.21	1.98
Expo Park	12.52	5.73			

As can be seen from table 4, the Expo Park, Andy Scenic Spot, Panlong Temple and other age statistics (χ^2_{age}) are greater than 10, it shows that the age distribution trend of these attractions are different from the overall; By calculating the Zage value of the three is less than 1.96, indicating that tourists in this kind of attractions affected by age are more similar. Seasonal statistics (χ^2_{season}) of Cuihu Park, Dian Lake, Anning Hot Spring is also relatively close, by calculating Z value can be found these scenic spots' difference degree are affected by the season are similar, whose difference with the whole is bigger.

4 Analysis on the distribution characteristics of tourists' attributes

The law of tourists' activities from the perspective of gender and age are studied (Table 5).

Table 5: Distribution of tourists' gender and age

scenic spot	Z_{gender}	χ^2_{age}	*scenic type*
Dian Lake	1.72	8.23	Adolescent active pattern
Xishan	1.79	8.64	
Daguan Park	1.88	7.96	
Nationalities Village	1.85	8.17	
Safari Park	1.64	8.21	
Cuihu Park	0.83	3.92	Young and middle-aged women active pattern
Military Academy	1.98	6.23	
Nanping Street	0.95	4.63	
Golden Horse and Jade Rooster	0.94	4.67	
Dounan Flower Market	0.82	4.88	
Expo Park	8.07	12.52	Middle-aged and old men
Andy Scenic Spot	7.12	11.87	

Panlong Temple	6.83	13.26	active pattern
Stone Forest	1.71	2.56	
Jiuxiang	1.68	2.47	General active
Guandu Old Town	1.63	2.71	pattern

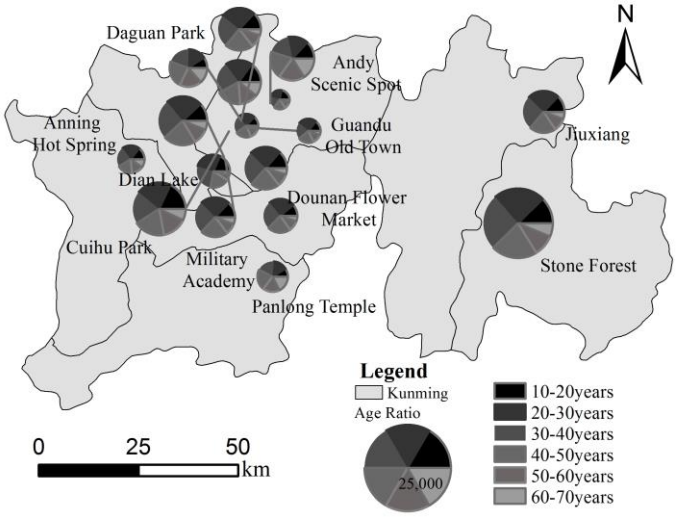

Figure 1: Tourists' age attribute distribution

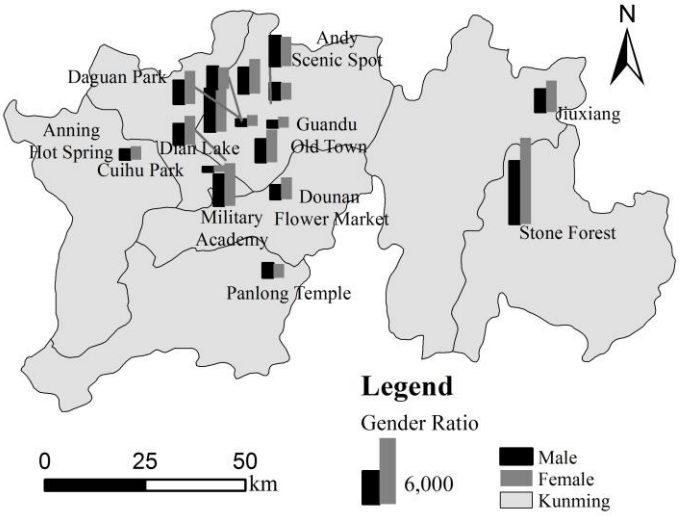

Figure 2: Tourists' gender attribute distribution

Hierarchical cluster analysis on the distribution character of tourists' attributes according to the gender and age distribution table, and it will be divided into four categories. Dian Lake, Xishan, Daguan Park, Nationalities Village, Safari Park are Adolescent active pattern; Cuihu Park, Military Academy, Nanping Street, Golden Horse and Jade Rooster, Dounan Flower Market are belong to Young and middle-aged women active pattern; Expo Park, Andy Scenic Spot, Panlong Temple are part of Middle-aged and old men active pattern; Stone Forest, Jiuxiang and Guandu Old Town are General active pattern.

(A)Adolescent active pattern

In this kind of attractions in the 20-30 age range of the largest proportion of tourists, tourist activity is higher in winter. Tourists generally in the early morning come to the Shore of Dian Lake to watch gulls, and then climb up the Xishan to watch the panorama of Dian Lake, after a short break at the foot of Xishan, come to the northern of the Daguan Park, Nationalities Village and Safari Park. Because

climbing requires a certain amount of physical strength, and strong stimulation amusement items in Daguan Park, at the same time, there are strange costumes of ethnic minorities, the existence of a variety of rare animals, these have a great appeal to teenagers.

(B) Young and middle-aged women active pattern

The main tourists are women in such attractions, and in the 20-50 ages range performance is more obvious. This is because the Nanping Street not only has the characteristics cate of Yunnan but also a variety of antiques. After the tourists buy souvenirs here mostly come to Lake Park, Golden Horse and Jade Rooster to admire the scenery. Visible the release of women's core areas are the place with food, beauty, shopping. It can be seen from Figure 2 male tourists raise significantly in Military Academy, the Z-test value in Table 5 also shows the male tourists prefer it, but because it is located in the Cui Lake, Nanping Street, Golden Horse and Jade Rooster and other extreme to attract young female tourist attractions nearby, female tourists are also covered in this scenic spots. It shows that the spatial distribution of tourist attractions affects the distribution of tourists' attributes. The huge gaps between male and female tourists check-in number in Dounan Flower Market response to women like flowers.

(C) Middle-aged and old men active pattern

This kind of scenic spots are indoor attractions, the Expo Park and Andy Scenic Spot is near, so tourists of the two attractions are similar. Panlong temple fair of Panlong Temple held regularly to attract many elderly tourists. This kind of scenic spot is mainly attracted by the old and middle-aged tourists, a large proportion of tourists in the 40-60 age range, and most tourists are male. Mainly because of these attractions are of historical significance. Therefore, it can be in this kind of scenic spot sales of senile male in favour of tourism products to meet the consumption demand of tourists.

(D) General active pattern

Such scenic spots for the general public favourite attractions, most tourists are female. It mainly composed of 20-50 age range. The basic distribution trend is similar to that of the whole tourist in Kunming, affected by age, gender and other attributes are small. The Stone Forest, Guandu Town, Jiuxiang are outdoor attractions, their distribution are similar. These scenic spots because of the unique scenery and low physical requirements of tourists, at the same time, it is also the key tourism development area of Kunming, it suitable for a variety of attributes tourists come to visit.

5 Conclusion

This study takes the combination of data mining and clustering analysis, taking Kunming as an example, the distribution characteristics of tourists' attributes are analysed with the Internet data, the main conclusions are as follows.

There is a certain difference between the law of tourist activities of the single scenic spots and the law of the whole tourist activities in Kunming. More women than men in the whole tourist, and the most dense 20-50 age range, the most frequent activities in summer and winter, but there are still some scenic spots and the overall trend of the opposite. Such as Expo Park, Andy Scenic Spot, Panlong Temple attractions are mainly elderly tourists, most tourists for local tourists and not affected by the low season. At the same time, the Military Academy for male tourists favourite attractions, but because it near the attractions where female tourists can have cate , enjoy beauty and go shopping, and it is visited by female in passing. It shows that the spatial distribution of scenic spots has some influence on the distribution of tourists' attributes.

From the perspective of gender and age, the scenic spots in Kunming can be divided into four categories: “Adolescent active pattern”, “Young and middle-aged women active pattern”, “Middle-aged and old men active pattern” and “General active pattern”. The preference differences of different attributes of tourists to different types of scenic spots are studied. So as to analyses the spatial characteristics of tourist attractions. The study find that female tourists like attractions where have food, beauty, shopping, older men prefer to have historic, antique attractions, teenagers like to challenge exciting attractions; The attractions of "General active pattern" are the key region of tourism development, the number of tourists accounted for 36.38% of the total number of tourists, the spatial distribution of tourist attributes is consistent with the development of key tourist areas in Kunming.

Acknowledgements
This research was financially supported by the National Natural Science Foundation of China (41501436).

References

[1] Li Deren, Yao yuan, Shao Zhenfeng. “Big data in smart cities” , Geomatics and Information Science of Wuhan University, 39,pp.631-640,(2014).

[2] Qin Xiao, Xiong Lifang, Zeng Feng, Zhu Shoujia. “Methods in urban temporal and spatial behaviour research in the Big Data Era”, Progress in geography, 32, pp.1352-1361,(2013).

[3] Chai Yanwei, Ta Na.“Progress in space-time behavior research in China”, Progress in Geography,32,pp.1362-1373,(2013).

[4] Huang Xiaoting, Li Wenxuan, Zhang Haiping, Qing Qianlong. “Evaluation of Tourist Temporal-spatial Behavior based on GPS Data ”, Tourism Tribune,31,pp. 41-49,(2016).

[5] LI Fangzheng,LI Wanyi,LI Xiong. “Research on Urban Greenway Planning based on Big Data of Bus Smart Card ”,Urban Development Studies,pp.27-33,(2015).

[6] Li Qingquan, Zhou Baoding. “Smartphone-based individual indoor spatiotemporal behavior analysis”, Progress in Geography, 34,pp. 457-465,(2015).

[7] Ding Liang, Niu Xinyi, Song Xiaodong. “Measuring the employment center system in Shanghai central city: A study using mobile phone signaling data” , Acta Geographica Sinica,71,pp. 484-499,(2016).

[8] SCHOLZ R W,LU Y. “Detection of dynamic activity patterns at a collective level from large-volume trajectory data”, International Journal of Geographical Information Science,28,pp. 946-963 ,(2014).

[9] FERREIRA N,POCO J,VO H T,et al. “Visual exploration of big spatio-temporal urban data: A study of New York city taxi trips” IEEE Transactions on Visualization and Computer Graphics,19,pp. 2149-2158,(2013).

[10] Wu Jiansheng, Huang Li, Liu Yu, et al. “Traffic Flow Simulation Based on Call Detail Records”, Acta Geographica Sinica, 67,pp. 1657-1665,(2012).

[11] CALABRESE F,COLONNA M,LOVISOLO P,et al. “Real-time urban monitoring using cell phones: A case study in Rome”, IEEE Transactions on Intelligent Transportation Systems,12,pp. 141-151,(2011).

[12] EAGLE N,PENTLAND A,LAZER D. “Inferring friendship network structure by using mobile phone data”, Proceedings of the National Academy of Sciences,106,pp. 15274-15278,(2009).

[13] GONZALEZ M C,HIDALGO C A,BARABASI A-L. “Understanding individual human mobility patterns” , Nature,453,pp. 779-782,(2008).

[14] Wang Bo, Zhen Feng, Zhang Hao. “The Dynamic Changes of Urban Space-time Activity and Activity Zoning Based on Check-in Data in Sina Web” , Scientia Geographica Sinica,35,pp. 151-160,(2015).

[15] Li Xiang, Zhang Jing, Jiang Nan,et al. “An Evaluation Method of Scenic Spots Based on Location Check-in Data and Classified Information of Scenic Spots ”, Journal of Geomatics Science and Technology,pp. 405-411,(2015).

[16] Zhang Ziang, Huang Zhenfang, Jin Cheng, et al. “Research on Spatial-temporal characteristics of scenic tourist activity based on sina microblog: a case study of Nanjing Zhongshan mountain national park”, Geography and Geographic Information Science,31,pp. 121-126,(2015).