International Journal of Geo-Information

MDPI

Article

Regionalization Analysis and Mapping for the Source and Sink of Tourist Flows

Qiushi Gu [1], Haiping Zhang [2,3,4,*], Min Chen [2,3,4] and Chongcheng Chen [5]

1 School of Humanities, Southeast University, Nanjing 210096, China
2 Key Laboratory of Virtual Geographic Environment, Nanjing Normal University, Ministry of Education, Nanjing 210023, China
3 Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China
4 State Key Laboratory Cultivation Base of Geographical Environment Evolution (Jiangsu Province), Nanjing 210023, China
5 Key Lab of Spatial Data Mining & Information Sharing, Ministry of Education, Fuzhou University, Fuzhou 350108, China
* Correspondence: 161301028@stu.njnu.edu.cn; Tel.: +86-18765829126

Received: 10 June 2019; Accepted: 20 July 2019; Published: 23 July 2019

Abstract: At present, population mobility for the purpose of tourism has become a popular phenomenon. As it becomes easier to capture big data on the tourist digital footprint, it is possible to analyze the respective regional features and driving forces for both tourism sources and destination regions at a macro level. Based on the data of tourist flows to Nanjing on five short-period national holidays in China, this study first calculated the travel rate of tourist source regions (315 cities) and the geographical concentration index of the visited attractions (51 scenic spots). Then, the spatial autocorrelation metrics index was used to analyze the global autocorrelation of the travel rates of tourist source regions and the geographical concentration index of the tourist destinations on five short-term national holidays. Finally, a heuristic unsupervised machine-learning method was used to analyze and map tourist sources and visited attractions by adopting the travel rate and the geographical concentration index accordingly as regionalized variables. The results indicate that both source and sink regions expressed distinctive regional differentiation patterns in the corresponding regional variables. This study method provides a practical tool for analyzing regionalization of big data in tourist flows, and it can also be applied to other origin-destination (OD) studies.

Keywords: geographical regionalization; tourist flow data; cartographic generalization; geospatial interaction; regional analysis; geographic concentration index of scenic spots

1. Introduction

Tourism, a major cross-regional activity, can reflect interregional spatial interaction patterns and regional characteristics to some extent [1,2]. The travel origin-destination (OD) chain formed by tourists moving from one area to another can be regarded as a tourist flow unit, and the travel OD chains of all tourists constitute tourism flow together [3]. Generally, the tourist origin region can be regarded as the source region, whereas the destination can be perceived as the sink region. When many tourists travel from the same source region to the same destination, a source and sink are formed [4,5]. Tourist flow can reflect regional tourist movement patterns, and through these movement patterns, the driving factors of tourist behavior patterns and the regional features of tourism resources can be further explored [6]. In recent years, with the continuing development of the Internet and mobile communication technologies, a large sample of big data on tourist flow has emerged [7–10]. With the emergence of big data on tourist flow and the increasing attention of scholars on the study of travel

behavior patterns, increasingly more scholars focus on tourism big data, especially among geographers who focus on space and region. Geographic information systems (GIS) and machine learning are being increasingly introduced into the study of tourism behavior.

Tourist travel motivation is usually affected by traffic conditions, economic conditions, cultural differences, and regional climate [10–13]. The length of vacation also affects travel distance and stay duration [14]. For example, tourists in cold regions often choose to travel to resorts with warm climates [15]; tourists tend to choose travel routes with convenient transportation [16]; tourists with good economic conditions may be more inclined to choose more expensive destinations [17,18]; and cultural differences may lead tourists to choose either destinations that are different from their native cultures or mysterious but famous scenic spots [19–21]. In a word, there are many factors that affect tourism flow. The remarkable spatial self-similarity and regional differentiation of these factors, as well as the regional imbalance of tourism resources, result in a regional differentiation of source and sink regions. In addition, the ratio of tourists to the local source market population varies at different times.

Tourism flow is one of the main research areas in tourism geography. At present, the research on tourism flow includes the following aspects. (1) In terms of research perspective, a large number of scholars focus on the theoretical framework and concepts of tourism flow, and they contribute greatly to the theoretical framework of tourist flow [22,23]. Some scholars attach importance to the spatial effects of tourism flow and investigate the moving track of tourists from aspects of spatial structure and spatial patterns [24,25]. Some scholars focus on the temporal and spatial evolution characteristics of tourism flow and analyze the changing process from the perspective of space and time [26]. Other scholars are concerned about the socioeconomic effects of tourism flows, such as high-speed rail effects [27,28]. In addition, the industrial development and economic growth brought by tourism have become a research hotspot [29]. (2) In terms of research data, questionnaires, interviews, and online data have been the main data of tourism flow research for some time [30,31]. In recent years, GPS-based tracking data and base station signaling data have greatly improved the accuracy and quantity of data [32–34]. Today, with the big data era approaching, the acquisition of data on tourism flow has become easier. (3) In terms of research methods, traditional statistical analysis [35], spatial analysis [36,37], and gravity models [38] are included. Recently, social network-related data analysis methods, such as the complex network method [39,40] and the GIS spatial interaction method [41,42], have received increasing academic attention. (4) Visualization is the core tool for processing tourism big data [43]. Intelligent and scientific analysis together with vivid visual expression are regarded as an important way to enhance respondents' perception of tourism flow information and knowledge. At present, the research on visualizing tourism big data is mainly based on existing GIS analysis methods [44–46], and few new methods that fully consider the characteristics of tourism flow have been proposed.

Scholars have made significant research progress in terms of data acquisition methodology innovation. In terms of studies on tourism source and sink, scholars mainly focused on the characteristics of spatial structure, main influencing factors, and the formation mechanism of source or sink [47,48]. However, few researchers consider regional aggregation and spatial differentiation for both source and sink regions. This is mainly limited by the fact that it was difficult to obtain detailed nationwide tourist flow statistics in the past, and many advanced regional division methods have not been introduced into the analysis of tourist flows [49]. Some scholars analyze tourism flow from the perspective of spatial structure and influencing mechanisms by using GIS technology and complex network analysis models. However, they fail to fundamentally focus on the localization characteristics of source and sink regions, which make it difficult to have a deeper understanding of the regional characteristics of tourism flows. This study collected tourism flow data for 51 scenic spots in Nanjing during five short-period national holidays in 2018 and calculated the travel rate for each origin city and the geographical concentration index for each scenic spot in Nanjing. Then, by using a machine-learning method, geographical regionalization and mapping were undertaken to geographically divide regions and map the source and sink regions.

2. Study Area and Data Description

For an individual tourist, the place of departure constitutes the source region, whereas the destination is the sink region. Each source region and its corresponding sink region together constitute a complete tourism flow. Therefore, in this study, the source region is mainland China, and the sink region is Nanjing, as shown in Figure 1. Nanjing is the capital city of Jiangsu Province in Eastern China, with a total area of 6587 square kilometers. In 2017, the built-up area was 1398.69 square kilometers, with a resident population of 8.335 million, an urban population of 6.859 million, and an urbanization rate of 82.3%. Nanjing is a famous historical city in China, representing the political, economic, and cultural center of many past dynasties. Nanjing is one of the major tourist cities in China, and it attracts a large number of foreign tourists every year.

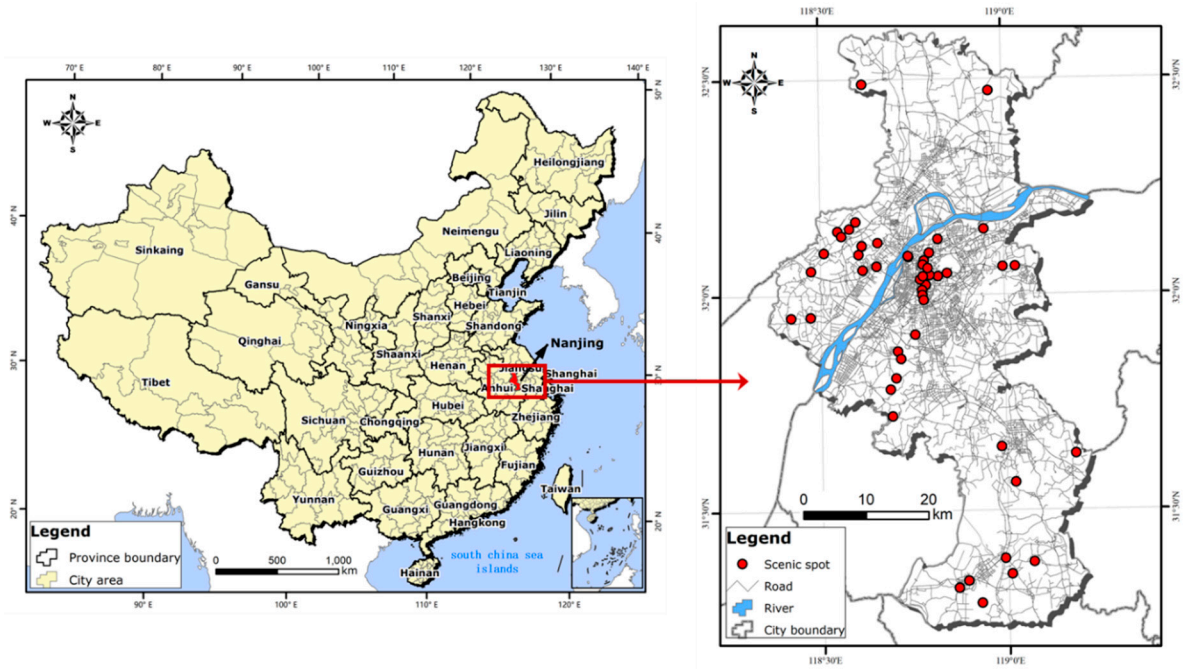


Figure 1. Study area (all cities of mainland China and all scenic spots in Nanjing).

This study is based on tourist tracing data during five national holidays in 2018 provided by the Nanjing tourism big data monitoring platform. The source regions are prefecture-level cities in mainland China, while the destination is Nanjing. The original data record the tourist flows from domestic cities to 51 scenic spots of Nanjing. In this study, data were collected based on Table 1, which describes the details of five short-period holidays, including their official names, duration time, and the total number of tourists during each holiday over five short-period holidays.

Table 1. Data description of tourist flow data.

Short-Period Holidays	Periods of Time	Total Number of Tourists
New Year's Day	2017/12/30–2018/01/01	224.40 million
Tomb-Sweeping Day	2018/04/05–2018/04/07	426.68 million
International Labor Day	2018/04/29–2018/05/01	493.04 million
Dragon Boat Festival	2018/06/16–2018/06/18	268.06 million
Mid-Autumn Festival	2018/09/22–2018/09/24	630.53 million

Note: These data are still sample data, instead of being from the whole population. They were obtained from real-name mobile base stations monitored by China Mobile, which has the largest number of users in China.

Table 2 shows the structure of tourist flow data. This table includes tourists' departure place, destination, visiting time, and the scenic spots they visited. Through geocoding, we can obtain

geographical coordinates of origin cities and destination cities. In addition, through the visited date, we can determine which holidays the record corresponds to. Because this study focuses on the tourism flow from all domestic cities to Nanjing, the detailed attributes of scenic spots were not involved in the analysis; instead, their attributes were aggregated and abstracted to the city level of Nanjing.

Table 2. Data structure of tourist flow data.

Attribute Name	Attribute Type	Attribute Description
Origin_city	string	Source city of tourist
Destination_city	string	Nanjing city
Visited_date	datetime	Date of visit
Scenic_spot	string	Visited scenic spots in Nanjing

3. Methodology

3.1. Travel Rates from Origin to Nanjing

Regionalization refers to the classification of spatial objects according to one or a few spatial location-related features [50]. The principle of this classification follows the idea that attributes of geographical objects shall be similar in the same category, whereas the attributes of geographical objects shall be quite different in different categories. There are five attributes in this study, that is, the five ratio attributes for the respective five holidays. The number of tourists from each origin is related to the population of each origin city, and in this study, we try to eliminate this side effect caused by population. Therefore, the attribute is the ratio of tourists from origin cities to the origin population. The five ratio attributes were used as five variables for modeling, analysis, and regionalization mapping. In this study, *ASet* represents the set of attributes that participated in the spatial data analysis, and for each feature, $ASet = \{A_1, A_2, ..., A_n\}$ contains the set $x = \{a_1, a_2, \ldots, a_n\}$, where n equals 5. In Equation (1), P_i presents the total population of the city i, and P_{i_t} presents the number of tourists from the city i to Nanjing during the holiday t. The ratio of tourist flow from the city i during the period t is

$$a_t = \frac{P_{i\_t}}{P_i} \ (t = 1, 2, 3, 4, 5). \tag{1}$$

For multivariate geographical zoning, it is necessary to define the principle of proximity between regional units, that is, what kind of relationship between two regions is called proximity. In GIS modeling, many kinds of proximity relations are defined, such as the queen rule with common edges, rook rule with common points and edges, inverse distance weight rule, and k-nearest [51]. Considering the continuity of regions, this paper did not use rules such as k-nearest, which do not have topological adjacency, and instead used the queen rule to conceptually model the spatial relationship.

3.2. Geographic Concentration Index of Scenic Spots

The large number of attractions available in Nanjing allows for various personal selection preferences. This regional uneven development pattern could be measured by geographical concentration indexes. This paper calculates the geographical concentration index for each scenic spot. In this study, the geographical concentration index reflects the tourist concentration degree when the tourists come from different places, that is, the degree of imbalance between different regions. Thus, *GSet* represents the set of attributes in the Nanjing scenic spot s, and for each scenic spot, $GSet = \{G_1, G_2, ..., G_n\}$ contains the set $y = \{g_1, g_2, ..., g_n\}$, where $n = 5$. Then, the geographical concentration index of a scenic spot g_t during the holiday t is [52]

$$g_t = 100 \times \sqrt{\sum_{i=1}^{n} (X_i/T)^2}, \tag{2}$$

where n is the total of origin cities, T is the total tourists of the destination, and X_i is the total tourist volume of the i-th origin city. From the formula, the result of the calculation is influenced not only by whether the distribution of tourists is balanced or not but also by the regional distribution of the number of origin cities. In this study, the number of origin cities is fixed at 315, so the G value reflects whether the distribution of tourists is balanced. According to Equation (2), the larger the value of G is, the more unbalanced the tourist distribution is; in contrast, the smaller the value of G is, the more balanced the tourist distribution is. Therefore, this study divides the regions into zones by travel rates in source places, whereas in sink places, the geographical concentration indices of each of the five short-term national holidays were used to undertake regional division.

3.3. Moran's I Index of Global Spatial Autocorrelation

Moran's I index is a spatial statistical model for detecting and measuring spatial autocorrelation, including global spatial autocorrelation and local spatial autocorrelation [53]. In this paper, the global Moran's index is used to measure whether there is spatial autocorrelation or local self-similarity in the tourist travel rate of tourists from different prefecture-level cities to Nanjing on different short-period holidays. The Moran's I index formula used to measure global spatial autocorrelation is as follows [54]:

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}, \tag{3}$$

where n represents the number of prefecture-level cities or the number of scenic spots, X_i is the travel rate of the prefecture-level city i or the geographical concentration index of the scenic spot i, and the spatial weight matrix w_{ij} represents the geographical proximity relationship between any two elements (prefecture-level city) i and the prefecture-level city j, or between scenic spot i and scenic spot j.

Under the assumption of normality and randomization, Moran's I calculates the mean or expected value of observations based on basic regional units (prefecture-level cities or scenic spots) and finally as a statistical significance test, Z-score was performed by detecting the spatial autocorrelation of each regionalized variable for each short period holiday. The formula for calculating the Z-score is as follows [55]:

$$Z = \frac{I - \frac{-1}{n-1}}{\sqrt{Var(I)}}. \tag{4}$$

The variance value of Moran's I here is given under the assumption of standardization, and the specific calculation formula is as follows:

$$Var(I) = \frac{n[n^2 + 3 - 3n)B + 3A^2 - nC] - K[\left(n^2 - n\right)B + 6A^2 - 2nC]}{(n-1)(n-2)(n-3)A^2}, \tag{5}$$

whereas

$$A = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij},\ B = \sum_{i=1}^{n}\sum_{j=1}^{n} {w_{ij}}^2,\ C = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} w_{ij}\right)^2,\ \mathrm{K} = \frac{n\sum_{i=1}^{n}(x_i - \overline{x})^4}{n\sigma^4}. \tag{6}$$

If the expected value of Moran's I is less than zero, and the observed value of Moran's I is larger than the expected value, it indicates that the regionalized variable presents an aggregated distribution pattern and it is consistent with positive spatial autocorrelation. If the observation of Moran's I is less than the expected value, it indicates that it presents a discrete distribution pattern and it is consistent with negative spatial autocorrelation. In this paper, global spatial autocorrelation is used to measure whether the regional variables have spatial aggregation pattern or not either at source level or at the attraction level. If there is a significant spatial clustering pattern, it provides a prerequisite for subsequent regionalization and regional mapping.

3.4. Geographical Regionalization Modeling

After calculating the travel rates for each origin city and the geographical concentration index for each scenic spot in Nanjing, as well as identifying the adjacency relation among the cities and among the scenic spots, a minimum spanning tree was constructed that could reflect both the spatial structure of geographical units and the attribute values involved in the calculation. This study used the attribute vectors x_i and x_j of objects i and j to measure the difference between i and j, linking the cost d (i, j) and the edge (v_i, v_j) to correlate the different measurement contexts. Once there is a comparable scale for attributes, such as standard deviation, usually the square of the Euclidean distance between x_i and x_j will be selected [56]:

$$d(i,\ j) = d(x_i,\ x_i) = \sum_{l=1}^{n} \left(x_{il} - \ x_{jl}\right)^2. \tag{7}$$

The path from node v_1 to node v_k is a sequence of nodes $(v_1, v_2, ..., v_k)$ connected by edges $(v_1, v_2, ..., v_k)$, ..., (v_{k-1}, v_k). For any node v_i and v_j, there is at least one path to connect them. A spatial cluster is a subset of the connected nodes, and our goal is to divide graph G into disjoint cluster C $(G1, \ldots, G_c)$. It is difficult to deal with this optimization issue because this function is based on the intracluster uniformity measure. In this paper, the attribute is expected to be a random variable with continuous variation, which could produce only a minimum spanning tree among the different measures. Finally, the minimum span tree is segmented. The best segmentation policy is to try to keep the highest similarity of feature attributes within the group and the maximum difference of feature attributes between the groups.

As an unsupervised machine-learning method was used for this study, the number of clusters needed to be predetermined. It depends on prior knowledge, and it could be tested by the Calinski–Harabasz pseudo-F test [57]. The Calinski–Harabasz pseudo-F statistic is a ratio reflecting the intercluster variance and intracluster variance. In other words, it can properly reflect similarities within the group and the differences between the groups. The formula is as follows:

$$F = \frac{\left(\frac{R^2}{n_c - 1}\right)}{\left(\frac{1 - R^2}{n - n_c}\right)}, \tag{8}$$

whereas

$$R^2 = \frac{SST - SSE}{SST}, \tag{9}$$

where SST reflects the differences between groups, and SSE reflects the similarities within a group. The two formulas are as follows:

$$SST = \sum_{i=1}^{n_c} \sum_{j=1}^{n_i} \sum_{k=1}^{n_v} \left(v_{ij}^k - \overline{v^k}\right)^2, \tag{10}$$

$$SSE = \sum_{i=1}^{n_c} \sum_{j=1}^{n_i} \sum_{k=1}^{n_v} \left(v_{ij}^k - \overline{v_t^k}\right)^2, \tag{11}$$

where n is the number of regions; n_i is the number of features in group I; n_c is the total number of regions; n_v is the number of variables participating in the regionalization; v_{ij}^k is the value of the k variable of j factor in group i; and $\overline{v^k}$ is the average of the variable k; $\overline{v_t^k}$ is the average of the variable k in group i. This paper uses the Calinski–Harabasz pseudo-F test to determine and evaluate the optimal number of groups.

4. Results

4.1. Spatial Autocorrelation Measurement of Regional Variables

As shown in Table 3, the travel rates were calculated with the global Moran's *I* index for the five short-period national holidays. In Table 3, the Z-score of the five short-period national holidays are larger than 2.58. It means that, at the 99% confidence level, the travel rates in all five short-period national holidays presented a significant spatial agglomeration pattern, especially during International Labor Day. Through the above spatial autocorrelation analysis, it can be concluded that the travel rates in all five short-period national holidays presented a regional similarity at the level of the prefecture-level city. This result provides a prerequisite for further regionalization analysis among origin cities.

Table 3. Global Moran's *I* index and its statistical significance for travel rates in prefecture-level cities.

Short-Period Holidays	Global Moran's *I* Index	Z-Score
New Year's Day	0.327 ***	19.56
Tomb-Sweeping Day	0.333 ***	18.71
International Labor Day	0.371 ***	21.37
Dragon Boat Festival	0.323 ***	18.14
Mid-Autumn Festival	0.347 ***	18.98

(*) 90% confidence interval; (**) 95% confidence interval; (***) 99% confidence interval.

The spatial autocorrelation of the geographic concentration index of all scenic spots during a particular holiday is used to uncover whether it presents spatial autocorrelation in all scenic spots in Nanjing, and whether it presents an agglomeration pattern at the broader level. Table 4 shows the overall distribution pattern of the geographic concentration index calculated by the global Moran's *I* for each short-period national holiday. Obviously, the spatial autocorrelation test value Z-score of the scenic spot concentration index during all holidays is larger than 2.58, which indicates that the geographical concentration index of the scenic spot during different short-period national holidays has spatial similarity and an agglomeration effect, and all the significance levels exceed 99%.

Table 4. Global Moran's *I* index and its statistical significance for the geographical concentration index of scenic spots.

Short-Period Holidays	Global Moran's *I* Index	Z-Score
New Year's Day	0.202 ***	3.23
Tomb-Sweeping Day	0.287 ***	4.47
International Labor Day	0.214 ***	3.43
Dragon Boat Festival	0.303 ***	4.76
Mid-Autumn Festival	0.294 ***	4.55

(*) 90% confidence interval; (**) 95% confidence interval; (***) 99% confidence interval.

It is clear that a significant spatial agglomeration effect exists on both the source and sink, and this aggregation also has its complexity due to mutual intersection features. For example, for the A area, the travel rate may be high on New Year's Day and International Labor Day, but the travel rate may be very low on other short-period national holidays. As both tourist sources and sinks used five regionalization variables, it is difficult to express the regionalization characteristics and the regulations of the tourist flow for the five holidays using simple analytical methods. Therefore, the alternative geographical mapping method could scientifically and vividly present the similarities within the regions and differences between the regions.

4.2. Regionalization of Tourist Flow Intensity

The geographically mapping method was followed by using the travel rate of five holidays in each origin city as a regionalized variable. Through the pseudo-*F* test, we got the peak value when the pseudo-*F* test value was 13. This means that the entire region can be divided into 13 regions. Those 13 divisions could maintain the highest number of similarities within the broader region and the largest differences between various regions. Meanwhile, since Nanjing and other cities in Jiangsu Province have comparatively higher travel rates than the cities outside Jiangsu, Nanjing and other cities in Jiangsu Province are regarded as two separate and independent geographical regions, and they will not be partitioned in regionalization analysis. As indicated in Figure 2, mainland China is divided into 15 geographical regions, coded with Roman numerals from I to XV.

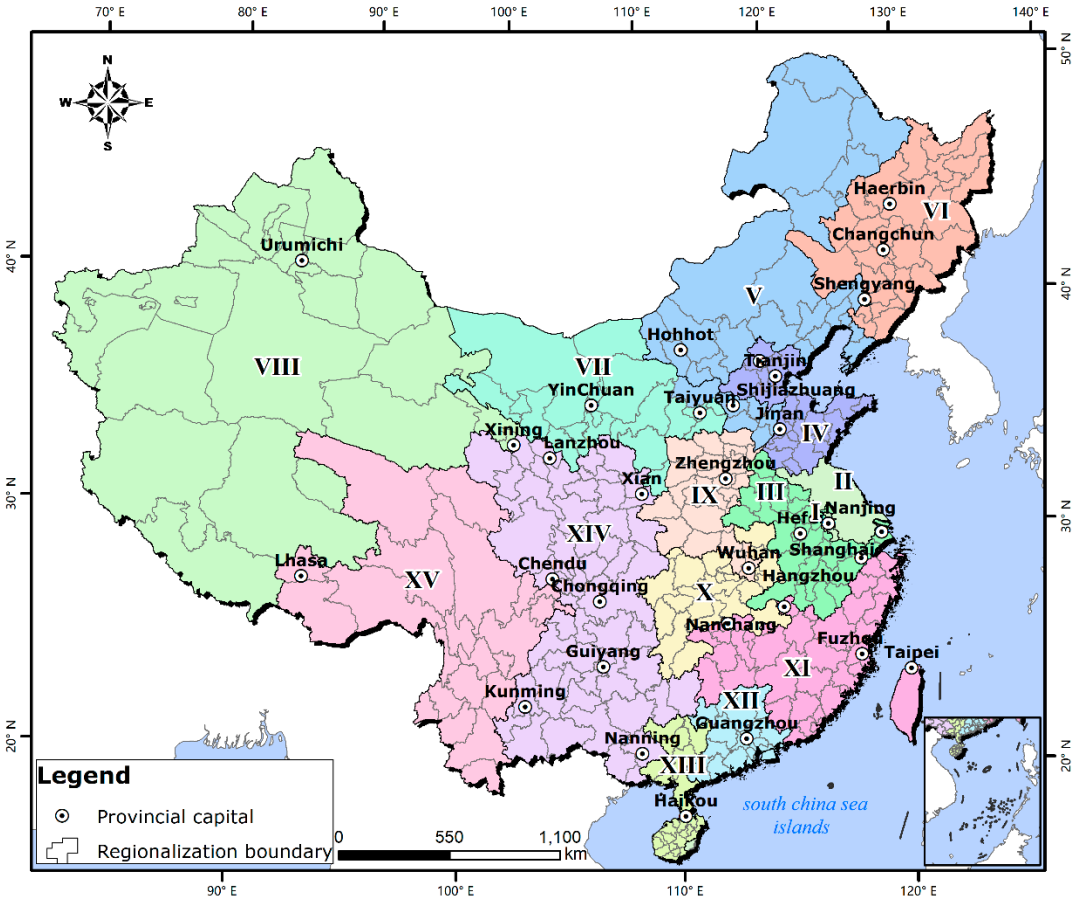
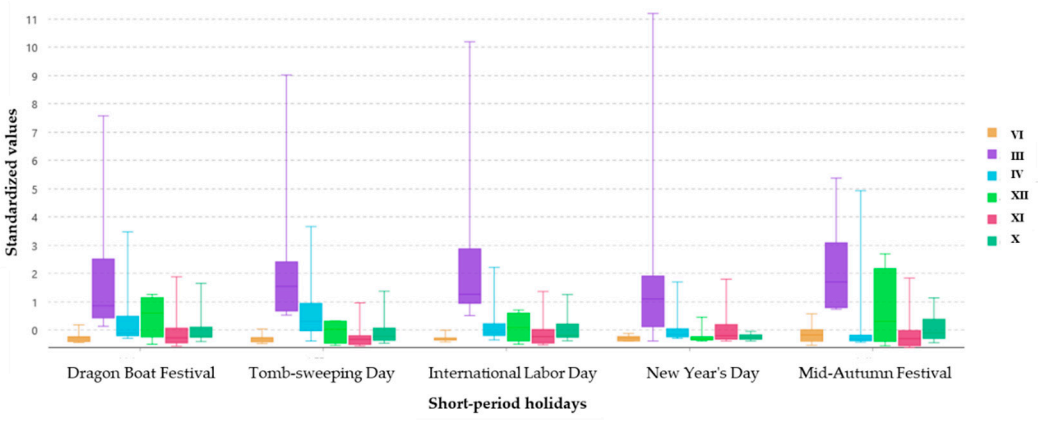


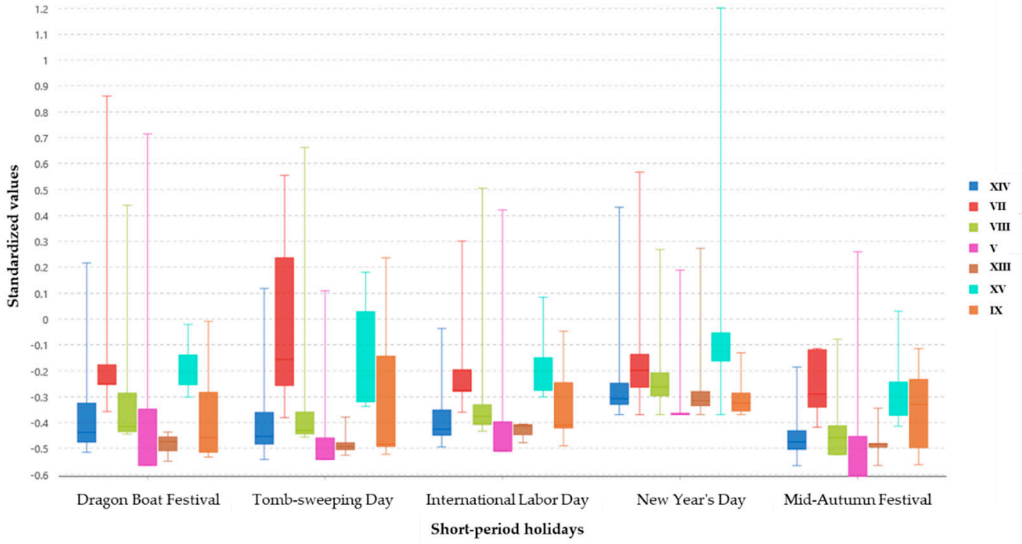
Figure 2. Regionalization results of tourist flow intensity. The whole of China is divided into 15 regions, and each division is coded in Roman numerals.

It is obvious that the new divided geographic units are not consistent with the provincial administrative units shown in Figure 1, but they have certain similarities. For example, the III district has similarities with Anhui Province in terms of geographic location, and the IV district has geographical similarities with the four provincial administrative regions including Beijing and Tianjin, and Hebei and Shandong provinces. The VI district has geographical similarities with the two provinces of Heilongjiang and Jilin.

Because it varied greatly in the normalized values among different regions, it is difficult to use the same box plot to present all of the values. Therefore, the 13 units obtained by the regionalization method are divided into two groups according to the standardized values. Units with larger normalized values are grouped and presented with side-by-side boxplots, as shown in Figure 3a; units with smaller normalized values are grouped and presented with side-by-side boxplots, as shown in Figure 3b.



(**a**) Parts of boxplots with high standardized values.



(**b**) Parts of boxplots with low standardized values.

Figure 3. Side-by-side boxplots for regionalization results of cities in China.

As shown in Figure 3a, the average tourist rate of division III during all short holidays is always the highest, whereas the volatility of the travel rate is also apparent. In Figures 2 and 3a, it can be seen that division III, with the highest volatility, is adjacent to Nanjing, and division III shared strong similarities with the administrative boundary of Anhui Province. The VI, IV, XI, and X areas have similar average travel rates and comparatively less fluctuation. Compared with other units, the average travel rate of division XII is higher, and the volatility is larger except for New Year's Day. Overall, except for divisions VI and XII, the above units are almost close to Nanjing. This result shows that those units with higher travel rates are mainly distributed in the surrounding areas of Nanjing, demonstrating a significant proximity effect.

As shown in Figure 3b, the travel rate of division XV in all five holidays is high, and it enjoys a comparatively stable travel rate except for Qingming Festival. Divisions VI, X, VIII, and XIII show similar travel rates, and their travel rates fluctuate moderately. The average travel rate of divisions V and XIII is generally low in these short-period national holidays, except for New Year's Day, whereas division V experiences more volatility and XIII enjoys the smallest fluctuation. In general, the mean and lower margin values of the all the units presented in Figure 3b are generally higher, and the lower margin values tend to be consistent, on New Year's Day.

4.3. Regionalization of Tourist Flow Sink

The mapping of regionalized geographical divisions based on the geographical concentration index in each scenic spot is shown in Figure 4. The pseudo-*F* test value shows that the peak value of the pseudo-*F* test is 6, so the threshold 6 here was suitable for geographical division. The Yangtze River, the largest river in China, passes through the city, as shown in Figure 4. Nanjing's major urban area is situated on the south side of the Yangtze River, and is covered mostly by divisions I and II. Thus, the downtown areas are mainly located in division I, and the other downtown area is located on the north side of division II. Also, most of the 51 monitored attractions in Nanjing are mainly distributed in urban areas (see Figure 1).

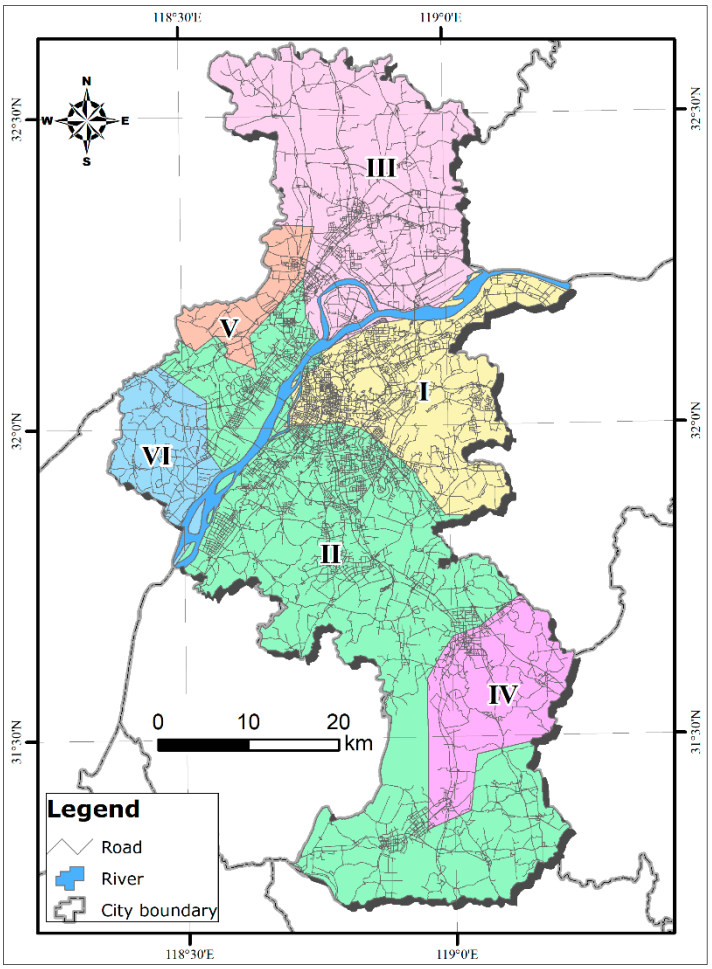


Figure 4. Regionalization result of the sink. (Note: The whole of Nanjing is divided into six regions, and each region is coded in Roman numerals.).

To interpret the regionalization features and their connotations of the six divisions in Figure 4, the mean line boxplots are drawn in Figure 5. The abscissa in Figure 5 represents five different short holidays, and the ordinate represents the standardized variance values of geographic concentration indices. As seen from Figure 5, the values of the geographic concentration index of division IV are higher than those of the other divisions, and all of them are higher than the upper quartile threshold. The standardized values of the geographic concentration index of division VI are generally high during each short holiday and exceed the upper quartile threshold of each short holiday. The standardized values of division III during each short holiday are all near the quartile threshold. In division II, the standardized values during each short holiday show a moderate value, which is between the lower median and the upper quartile threshold. Division I and division V are the lowest, and all are below

the lower quartile threshold. In general, if the location of the division is close to the downtown area of Nanjing, the geographical concentration index tends to be low and the volatility during each short holiday also tends to be small, and vice versa.

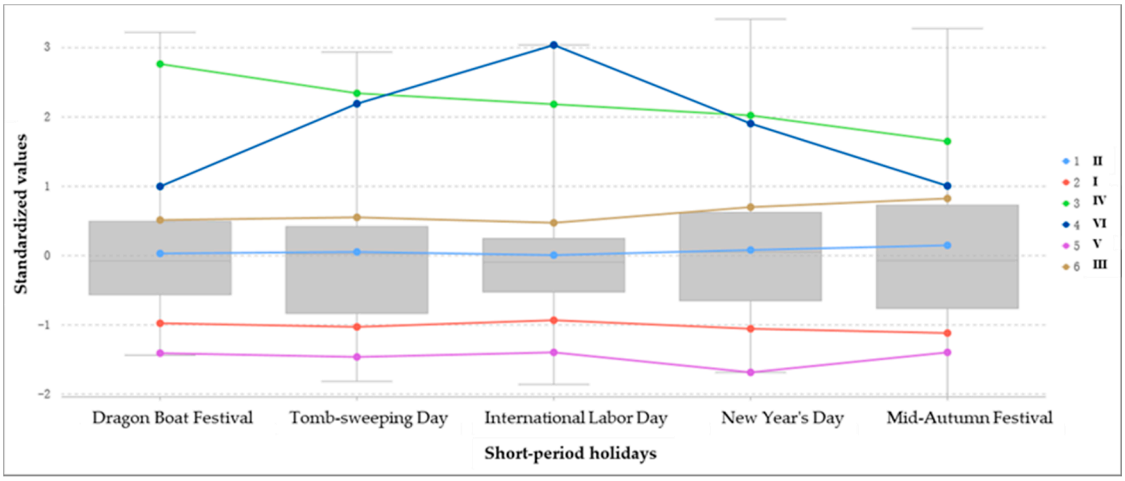


Figure 5. Mean line boxplots for regionalization results of scenic spots in Nanjing.

5. Discussion

5.1. Reliability of Data and Recognition of Tourist Flow

The tourist visitation statistics used in this paper come from the tourism big data monitoring platform of the Nanjing Culture and Tourism Bureau. The data come from China Mobile, which is currently the largest mobile communication provider in China. As of December 2018, there are 925 million users of China Mobile, 315 million China Unicom users, and 303 million China Telecom users. China Mobile users accounted for 59.95% of the total. This means that the data samples used in this study can be considered to account for approximately 59.95% of the total sample size. This proportion of users is sufficient to complete the research work of this paper and to ensure the reliability of the research data.

Currently, mobile phone users in China have essentially adopted the real name system. The real name system means that each mobile phone number corresponds to the individual origin place information, so it can be determined which province and prefecture-level city the tourists come from. The scenic spot monitors and locates visitors through the base station. Monitoring takes place at regular intervals. When users with the same mobile phone number are monitored on multiple occasions over a period of time, the repeated users will be eliminated during data processing, which allows tourists to be uniquely monitored at a particular scenic spot on each day. Therefore, the whole dataset records all the individual tourists who constitute the source, and the whole tourist statistics monitored in 51 scenic spots constitute the sink. This approach ensures the reliability of the identified tourist flow data.

5.2. Relationship between Regionalization and Geospatial Cognition

There are various complex geographical objects and phenomena in geo-space. These objects and phenomena may be natural things such as the environment or social phenomena and events generated by human activities and behavior. Some of these things are specific, and others are abstract. These objects or phenomena represent the real world. The GIS abstracts, synthetically selects, and simplifies these phenomena or objects to digitalize, informatize, and intellectualize objects or phenomena in the real world. In this process, a map has always served as a carrier of important geographic information and knowledge. Whether using traditional GIS or the current big data and intelligent GIS under the influence of artificial intelligence (AI), the use of maps is inseparable from analyzing and expressing the abstract real world.

An important issue now is how to extract rules and knowledge from massive spatial-temporal data and social flow data in a more intelligent and efficient way. This is a problem concerning technology, but it is also problem of how to better understand spatial-temporal data and how to express social flow data. Research on the abstraction, synthesis, and simplification of traditional data continues. However, what is currently in front of us is how to enhance the synthesis, abstraction and simplification of social flow data and even real-time flow data. The analysis of tourist flow as a kind of large data flow, and how to use it effectively, better, and more intelligently to show the hidden rules and knowledge through the region is very meaningful exploratory work.

The five short-period national holidays in this study occur in different seasons, and the study area is the entire mainland China. The study area is vast, and the region differs greatly in terms of climate, customs, and economy. For different short-period national holidays, people in different regions have distinctive preferences about different holidays, but there also exists the possibility of preference similarity in some places. This assumption provides tourist source geographical division possibilities. Therefore, it is particularly important to first determine whether these travel rates have a spatial agglomeration effect.

The sink area in this study contains 51 major scenic spots in Nanjing. Those scenic spots vary in terms of the distance to the city hub, and different types of attractions lead to varying levels of popularity of some scenic spots in different seasons. Moreover, the recognition of scenic spots and the varied cultural backgrounds of the visitors may affect the geographic concentration degree of a particular scenic spot. Therefore, great importance should be attached to whether the tourist arrivals in a certain scenic spot have a spatial agglomeration effect. This study used the geography concentration index to measure whether the source of tourist arrivals in different scenic spots during different short-period national holidays is agglomerated.

If the geographic concentration index of these scenic spots in different short-period national holidays is also agglomerated, it indicates a necessity to conduct a regional analysis and map the geographical concentration index based on different short-period national holidays. If the agglomeration is statistically significant, then the geographical division and mapping for both the source and the sink will be performed, and the regionalization characteristics and regulations will be described and explained.

6. Conclusions and Future Directions

6.1. Conclusions

This study provides a new analytical approach and methodology for the analysis of tourism flows from a regional perspective. At present, flow data have become the main form of data in the field of geography research, including human flow, logistics, information flow, capital flow, and technology flow [58–60]. Among these forms, tourist flow is one of the most important types of various people flow data. The expression and analysis of tourist flow is different from traditional GIS modeling such as point, line, and surface [61,62]. Individual flow data include the source region and destination of tourists, which together constitute the tourist flow unit. As the core content of geographic regional analysis, regionalization is also one of the main objectives of GIS spatial analysis and problem modeling. Regionalization of flow data must reconsider the applicability and scalability of traditional regionalization methods. Based on tourist flow data with time series characteristics, this paper uses a machine learning method to regionalize source and sink tourist flow data, which provides a feasible concept for the regional structure analysis of flow data.

In the analysis of tourist flow data with Nanjing as the tourist destination, the tourist flow source area in this paper takes the prefecture-level city as the basic unit, and the tourist flow sink area takes the scenic spot as the basic object. After regionalization analysis, the tourist flow source areas are divided into 15 divisions, and the tourist flow destinations are divided into six divisions. Among the source divisions, many have similarities with existing divisions or more provincial administrative divisions.

Many provinces in China have maintained self-similarity of the internal culture of the province and the cultural differences between the different provinces to a certain extent [63,64]. In terms of the physical geographical environment, regions with the same physical geographical features accord with the boundaries of adjacent provinces to some extent. Differences in the geographical environment bring about differences in the economy, transportation, and climate [65]. Accordingly, the formation of these divisions with travel rates as regionalization variables has a certain correlation with the social and the natural environment.

For the districts that include the main urban areas of Nanjing, the geographic concentration indices of the source are lower in all the short holidays, and the geographic concentration indices of the different short holidays are more stable. Many well-known scenic spots mainly located in the downtown area are more developed and convenient in terms of transportation and other supporting facilities. Therefore, nonlocal tourists tend to travel to such scenic spots, whereas remote and unpopular scenic spots are more likely to be visited by those with a strong sense of focused purpose. This study shows that the regionalization and cartographic analysis methods of tourist flow can also be extended to the regional analysis and application of other OD flow data.

6.2. Future Directions

In this study, the travel rate and geographic concentration index of different short holiday tourists are used as regionalization source and sink variables. The main goal is to regionalize and map tourist flow sources and sinks based on these regionalized variables. However, although this study takes the source and sink as a whole to explore the regional characteristics of the flow itself, the possible relevance of regionalization results between the source and the sink were not explored. Future analyses will conduct further exploration, and a few patterns will be abstracted. It will be further interesting and meaningful work to explore this regional difference of tourist sources in different types of scenic spots, which will also be taken into account in follow-up studies. In addition, this study used tourist frequency statistics to carry out the geographical division of the source. Future research might need to consider the usage of the number of tourists from the respective city and create geographical division. More insights will need to be obtained to fully reveal the feature of regionalization of the tourist source.

Author Contributions: Qiushi Gu and Haiping Zhang contributed paper writing and method implementation. Min Chen and Chongcheng Chen contributed ideas of the paper and system.

Funding: This work was funded by the Key Laboratory of Spatial Data Mining & Information Sharing of Ministry of Education, Fuzhou University (No. 2019LSDMIS03), and the Priority Academic Program Development of Jiangsu Higher Education Institutions under grant 164320H116.

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

References

1. Bao, J.G.; Ma, L.J.C. Tourism geography in china, 1978–2008: Whence, what and whither? *Prog. Hum. Geog.* **2011**, *35*, 3–20.
2. Pandy, W.R.; Rogerson, C.M. The timeshare industry of Africa: A study in tourism geography. *Bull. Geogr. Socio-Econ. Ser.* **2013**, *21*, 97–109. [CrossRef]
3. Youyin, Z.; Jing, G.; Yaofeng, M. Tourist flow research progress, evaluation and outlook. *Tour. Trib./Lvyou Xuekan* **2013**, *28*, 38–46.
4. Degen, W. The influence of Beijing-Shanghai high-speed railway on tourist flow and time-space distribution. *Tour. Trib./Lvyou Xuekan* **2014**, *29*, 75–82.
5. Feng, N.; Li, J. A couple analysis of the extraversion online tourism information, inbound tourist flow: A case of the American, Canadian inbound tourist flow. *Tour. Trib.* **2014**, *29*, 79–86.
6. Xi, Y.; Wang, Z.F.; Department of Culture Industry Management, Hunan University of Commerce; Business College of Jishou University. Exploratory analysis of Hunan tourist flow agglomeration and diffusion based on traffic network. *Areal Res. Dev.* **2014**, *3*, 1–28.

7. Fuchs, M.; Hopken, W.; Lexhagen, M. Big data analytics for knowledge generation in tourism destinations—A case from Sweden. *J. Destin. Mark. Manag.* **2014**, *3*, 198–209. [CrossRef]
8. Irudeen, R.; Samaraweera, S. Big data solution for Sri Lankan development: A case study from travel and tourism. In Proceedings of the 2013 International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 11–15 December 2013; pp. 207–216.
9. Xiang, Z.; Fesenmaier, D.R. Big data analytics, tourism design and smart tourism. In *Analytics in Smart Tourism Design*; Springer: Cham, Switzerland, 2017; pp. 299–307.
10. Li, J.J.; Xu, L.Z.; Tang, L.; Wang, S.Y.; Li, L. Big data in tourism research: A literature review. *Tour. Manag.* **2018**, *68*, 301–323. [CrossRef]
11. March, R.; Woodside, A.G. Testing theory of planned versus realized tourism behavior. *Ann. Tour. Res.* **2005**, *32*, 905–924. [CrossRef]
12. Miller, D.; Merrilees, B.; Coghlan, A. Sustainable urban tourism: Understanding and developing visitor pro-environmental behaviours. *J. Sustain. Tour.* **2015**, *23*, 26–46. [CrossRef]
13. Woodside, A.G.; Hsu, S.Y.; Marshall, R. General theory of cultures' consequences on international tourism behavior. *J. Bus. Res.* **2011**, *64*, 785–799. [CrossRef]
14. Buhalis, D.; Law, R. Progress in information technology and tourism management: 20 years on and 10 years after the internet—The state of Etourism research. *Tour. Manag.* **2008**, *29*, 609–623. [CrossRef]
15. Martin, B.G. Weather, climate and tourism—A geographical perspective. *Ann. Tour. Res.* **2005**, *32*, 571–591. [CrossRef]
16. Khan, S.A.R.; Dong, Q.L.; Wei, S.B.; Zaman, K.; Zhang, Y. Travel and tourism competitiveness index: The impact of air transportation, railways transportation, travel and transport services on international inbound and outbound tourism. *J. Air. Transp. Manag.* **2017**, *58*, 125–134. [CrossRef]
17. Louca, C. Income and expenditure in the tourism industry: Time series evidence from Cyprus. *Tour. Econ.* **2006**, *12*, 603–617. [CrossRef]
18. Min, L. Research on income difference between urban and rural inhabitants during tourism development in zhangjiajie. *J. Jishou Univ. Nat. Sci. Ed.* **2011**, *3*, 103–106.
19. Iwashita, C. Media representation of the UK as a destination for Japanese tourists: Popular culture and tourism. *Tour. Stud.* **2006**, *6*, 59–77. [CrossRef]
20. Smith, S. A sense of place: Place, culture and tourism. *Tour. Recreat. Res.* **2015**, *40*, 220–233. [CrossRef]
21. Astawa, I.P.; Triyuni, N.N.; Santosa, I.D.M.C. Sustainable tourism and harmonious culture: A case study of cultic model at village tourism. *J. Phys. Conf. Ser.* **2018**, *953*, 012057. [CrossRef]
22. Hongsong, P.; Lin, L.; Xingfu, L. The network structure of cross-border tourism flow based on the social network method: A case of Lugu lake region. *Sci. Geogr. Sin.* **2014**, *34*, 1041–1050.
23. Crampon, L.; Tan, K. A model of tourism flow into the pacific. *Tour. Rev.* **1973**, *28*, 98–104. [CrossRef]
24. Hemson, G.; Maclennan, S.; Mills, G.; Johnson, P.; Macdonald, D. Community, lions, livestock and money: A spatial and social analysis of attitudes to wildlife and the conservation value of tourism in a human-carnivore conflict in botswana. *Biol. Conserv.* **2009**, *142*, 2718–2725. [CrossRef]
25. Tsai, P.W.; Chen, Z.S.; Xue, X.S.; Chang, J.F. Studying the influence of tourism flow on foreign exchange rate by iabc and time-series models. In *Advances in Intelligent Information Hiding and Multimedia Signal Processing, Pt I*; Springer: Cham, Switzerland, 2018; Volume 81, pp. 225–232.
26. Yang, X.Z.; Wang, Q. Exploratory space-time analysis of inbound tourism flows to china cities. *Int. J. Tour. Res.* **2014**, *16*, 303–312.
27. Kelly, J.; Haider, W.; Williams, P.W. A behavioral assessment of tourism transportation options for reducing energy consumption and greenhouse gases. *J. Travel Res.* **2007**, *45*, 297–309. [CrossRef]
28. Bieger, T.; Wittmer, A. Air transport and tourism—Perspectives and challenges for destinations, airlines and governments. *J. Air Transp. Manag.* **2006**, *12*, 40–46. [CrossRef]
29. Papatheodorou, A.; Rosselló, J.; Xiao, H. Global economic crisis and tourism: Consequences and perspectives. *J. Travel Res.* **2010**, *49*, 39–45. [CrossRef]
30. Klabbers, M.; Timmermans, H. Measuring tourism consumer behaviour using escape: A multimedia interview engine for stated choice and preference experiments. In *Information and Communication Technologies in Tourism 1999*; Springer: Vienna, Austria, 1999; pp. 179–190.
31. Strang, E.; Peterson, Z.D. Unintentional misreporting on self-report measures of sexually aggressive behavior: An interview study. *J. Sex Res.* **2017**, *54*, 971–983. [CrossRef]

32. Kawase, J.; Kurata, Y.; Yabe, N. When and where tourists are viewing exhibitions: Toward sophistication of GPS-assisted tourist activity surveys. In Proceedings of the Information and Communication Technologies in Tourism 2012, Helsingborg, Sweden, 25–27 January 2012; pp. 415–425.
33. Zheng, W.; Huang, X.T.; Li, Y. Understanding the tourist mobility using GPS: Where is the next place? *Tour. Manag.* **2017**, *59*, 267–280. [CrossRef]
34. Dijk, J.V.; Jong, T.D. Post-processing GPS-tracks in reconstructing travelled routes in a GIS-environment: Network subset selection and attribute adjustment. *Ann. GIS* **2017**, *23*, 203–217. [CrossRef]
35. Li, H.; Gao, W. Study on region difference of tourism development in Guangdong province based on spatial statistical analysis. *J. Xinyang Norm. Univ. Nat. Sci. Ed.* **2016**, *29*, 71–74.
36. Li, M.M.; Fang, L.; Huang, X.T.; Goh, C. A spatial-temporal analysis of hotels in urban tourism destination. *Int. J. Hosp. Manag.* **2015**, *45*, 34–43. [CrossRef]
37. Kline, C.; Hao, H.; Alderman, D.; Kleckley, J.W.; Gray, S. A spatial analysis of tourism, entrepreneurship and the entrepreneurial ecosystem in North Carolina, USA. *Tour. Plan. Dev.* **2014**, *11*, 305–316. [CrossRef]
38. Morley, C.; Rossello, J.; Santana-Gallego, M. Gravity models for tourism demand: Theory and use. *Ann. Tour. Res.* **2014**, *48*, 1–10. [CrossRef]
39. Baggio, R.; Scott, N.; Cooper, C. Network science a review focused on tourism. *Ann. Tour. Res.* **2010**, *37*, 802–827. [CrossRef]
40. Zhang, H.P.; Zhou, X.X.; Gu, X.; Zhou, L.; Ji, G.L.; Tang, G.A. Method for the analysis and visualization of similar flow hotspot patterns between different regional groups. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 328. [CrossRef]
41. Wang, H.; Huang, H.; Ni, X.; Zeng, W. Revealing spatial-temporal characteristics and patterns of urban travel: A large-scale analysis and visualization study with taxi GPS data. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 257. [CrossRef]
42. Zhou, X.; Zhang, H.-P.; Ji, G.; Tang, G.-A. A method to mine movement patterns between zones: A case study of subway commuters in Shanghai. *IEEE Access* **2019**, *7*, 67795–67806. [CrossRef]
43. Agiomirgianakis, G.; Serenis, D.; Tsounis, N. Short-and long-run determinants of tourist flows: The case of South Korea. In *Advances in Applied Economic Research*; Springer: Cham, Switzerland, 2017; pp. 861–872.
44. Han, Y.; Wang, S.; Ren, Y.; Wang, C.; Gao, P.; Chen, G. Predicting station-level short-term passenger flow in a citywide metro network using spatiotemporal graph convolutional neural networks. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 243. [CrossRef]
45. Lu, Z.B.; Rao, W.M.; Wu, Y.J.; Guo, L.; Xia, J.X. A kalman filter approach to dynamic od flow estimation for urban road networks using multi-sensor data. *J. Adv. Transport.* **2015**, *49*, 210–227. [CrossRef]
46. Li, X.; Huang, G.; Tang, J. Passenger flow forecasting based on od-matrix estimation model. *J. China Railw. Soc.* **2008**, *30*, 7–12.
47. Alexander, L.; Jiang, S.; Murga, M.; Gonzalez, M.C. Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transport. Res. C Emerg. Technol.* **2015**, *58*, 240–250. [CrossRef]
48. Miah, S.J.; Vu, H.Q.; Gammack, J.; McGrath, M. A big data analytics method for tourist behaviour analysis. *Inf. Manag.* **2017**, *54*, 771–785. [CrossRef]
49. Nie, Y.M.; Zhang, H.M. A relaxation approach for estimating origin-destination trip tables. *Netw. Spat. Econ.* **2010**, *10*, 147–172. [CrossRef]
50. Lankford, P.M. Regionalization: Theory and alternative algorithms. *Geogr Anal.* **1969**, *1*, 196–212. [CrossRef]
51. Getis, A.; Aldstadt, J. Constructing the spatial weights matrix using a local statistic. *Geogr. Anal.* **2004**, *36*, 90–104. [CrossRef]
52. Marcon, E.; Puech, F. Measures of the geographic concentration of industries: Improving distance-based methods. *J. Econ. Geogr.* **2010**, *10*, 745–762. [CrossRef]
53. Lewis, D.B. *Elementary statistics for geographers*, 3rd ed.; Guilford Press: New York, NY, USA, 2010; p. 464.
54. Maurel, F.; Sedillot, B. A measure of the geographic concentration in French manufacturing industries. *Reg. Sci. Urban Econ.* **2004**, *29*, 575–604. [CrossRef]
55. Wu, H.; Hayes, M.J.; Weiss, A.; Hu, Q. An evaluation of the standardized precipitation index, the china-z index and the statistical z-score. *Int. J. Climatol.* **2001**, *21*, 745–758. [CrossRef]
56. Assuncao, R.M.; Neves, M.C.; Camara, G.; Da Costa Freitas, C. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *Int. J. Geogr. Inf. Sci.* **2006**, *20*, 797–811. [CrossRef]

57. Maulik, U.; Bandyopadhyay, S. Performance evaluation of some clustering algorithms and validity indices. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 1650–1654. [CrossRef]
58. Marty, P.F. An introduction to digital convergence: Libraries, archives, and museums in the information age. *Libr. Quart.* **2010**, *80*, 1–5. [CrossRef]
59. Andris, C.; Xi, L.; Ferreira, J., Jr. Challenges for social flows. *Comput. Environ. Urban Syst.* **2018**, *70*, 197–207. [CrossRef]
60. Andris, C. Integrating social network data into gisystems. *Int. J. Geogr. Inf. Sci.* **2016**, *30*, 2009–2031. [CrossRef]
61. Midler, J.C. Non-Euclidean geographic spaces: Mapping functional distances. *Geogr. Anal.* **1982**, *14*, 189–203. [CrossRef]
62. Wang, Y.X.; Wang, F.H.; Zhang, Y.; Liu, Y. Delineating urbanization "source-sink" regions in china: Evidence from mobile app data. *Cities* **2019**, *86*, 167–177. [CrossRef]
63. Li, Q.; Xu, M.X.; Liu, G.B.; Zhao, Y.G.; Tuo, D.F. Cumulative effects of a 17-year chemical fertilization on the soil quality of cropping system in the loess hilly region, China. *J. Plant Nutr. Soil Sci.* **2013**, *176*, 249–259. [CrossRef]
64. Lu, G.N.; Batty, M.; Strobl, J.; Lin, H.; Zhu, A.X.; Chen, M. Reflections and speculations on the progress in Geographic Information Systems (GIS): A geographic perspective. *Int. J. Geogr. Inf. Sci.* **2019**, *33*, 346–367. [CrossRef]
65. Soll, D. Feeding gotham: The political economy and geography of food in New York city, 1790–1860. *J. Interdiscipl. Hist.* **2018**, *48*, 417–418. [CrossRef]