International Journal of
*Geo-Information*

*Article*

# Understanding the Functionality of Human Activity Hotspots from Their Scaling Pattern Using Trajectory Data

**Tao Jia [1,2] and Zheng Ji [1,*]**

[1] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430072, China; tao.jia@whu.edu.cn

[2] Department of Land Surveying and Geo-informatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China

* Correspondence: jz07@whu.edu.cn; Tel.: +86-276-877-8546

Received: 2 September 2017; Accepted: 2 November 2017; Published: 5 November 2017

**Abstract:** Human activity hotspots are the clusters of activity locations in space and time, and a better understanding of their functionality would be useful for urban land use planning and transportation. In this article, using trajectory data, we aim to infer the functionality of human activity hotspots from their scaling pattern in a reliable way. Specifically, a large number of stopping locations are extracted from trajectory data, which are then aggregated into activity hotspots. Activity hotspots are found to display scaling patterns in terms of the sublinear scaling relationships between the number of stopping locations and the number of points of interest (POIs), which indicates economies of scale of human interactions with urban land use. Importantly, this scaling pattern remains stable over time. This finding inspires us to devise an allometric ruler to identify the activity hotspots, whose functionality could be reliably estimated using the stopping locations. Thereafter, a novel Bayesian inference model is proposed to infer their urban functionality, which examines the spatial and temporal information of stopping locations covering 75 days. Experimental results suggest that the functionality of identified activity hotspots are reliably inferred by stopping locations, such as the railway station.

**Keywords:** trajectory data; human activity hotspots; scaling; urban functionality; Bayesian inference model

## 1. Introduction

In recent years, trajectory data have been a hot topic in the emerging domain of data science [1], of interest to data scientists in diverse fields including geography, statistics, computer science, physics, and biology. This can be attributed to the availability of massive geo-tagged data [2,3], thanks to advancements in technologies such as global positioning systems, Web 2.0, and telecommunications. A typical example of trajectory data is floating car data, which are actively collected by GPS receivers installed in vehicles for navigation or monitoring. To uncover novel patterns, much work has involved floating car data. For instance, some studies tried to segment trajectories using an arbitrary specified speed threshold [4] or elapsed time threshold [5]. Some studies estimated $CO_2$ emissions from trajectories and applied them to sustainable location planning [6] and market analysis for the retail sector [7]. A recent study explored the relationship between human activities and landscape patterns [8], and studies more relevant to our work focused on mining human activity patterns [9–13].

Among the studies on human activity patterns, some developed methods to extract activity hotspots [14–16]. These activity hotspots are the clusters of activity locations in space and time. Not only do they display dynamics in terms of their structure and lifetime [13,17,18], but they also contain rich semantic information in terms of urban functionality, which refers to the actual land use information such as residential area, commercial area, recreation, etc. However, it is not straightforward to infer their urban functionality using the available literature. On the one hand, POIs data have been directly used to reveal the functionality of urban land use [19–22], but the uncovered urban functionality cannot reflect the reality well. This is because some POIs may be visited many more times than the others, and hence the distribution of POIs may lead to an unreliable estimation of urban functionality. Besides, no activity hotspots were detected in those studies. On the other hand, POIs data have been indirectly used to infer the activity type of a stopping location or the purpose of a trip using the simple distance matching method [23,24], the decision tree model [25], the random forest method [26], the probability model [27–29], or even the visual analytic technique [30]. The functionality of the activity hotspot was then derived by merging the activity types of all stopping locations within its spatial range [31–33]. In this way, the revealed urban functionality considered the importance of different POIs, but the functionality of activity hotspots with very few stopping locations cannot be estimated reliably. In other words, the relationship between the POIs and the stopping locations is overlooked in the literature. Hence, this study aims to examine this issue by employing the underlying scaling pattern.

The scaling pattern can be understood from two perspectives. First, in physics, the scaling pattern of an entity may refer to a power law distribution of its size or quantity [34–36], which indicates invariance under contraction or dilation and is often thought to be a signature of hierarchies. Second, the scaling of an entity could be regarded as an allometric scaling relationship among its properties [37]. This pattern has been observed in the biological world [38] and in human society [39]. For instance, a recent study suggested that human interactions in terms of communication activities scale superlinearly with the size of administrative divisions such as statistical cities, urban zones, or municipalities [40]. However, in urban studies, very few were aware of the importance of this scaling pattern. Sutton [41] had observed the allometric scaling pattern between urban area and urban population for cities in the US, which helped him to identify sprawling cities; Jiang et al. [42] had investigated the scaling pattern of geographic space and further applied this finding to the process of map generalization for many geographical entities including street network, coastline, and drainage network. To our knowledge, applying the scaling pattern of human activity hotspots to understand their urban functionality is still unexamined.

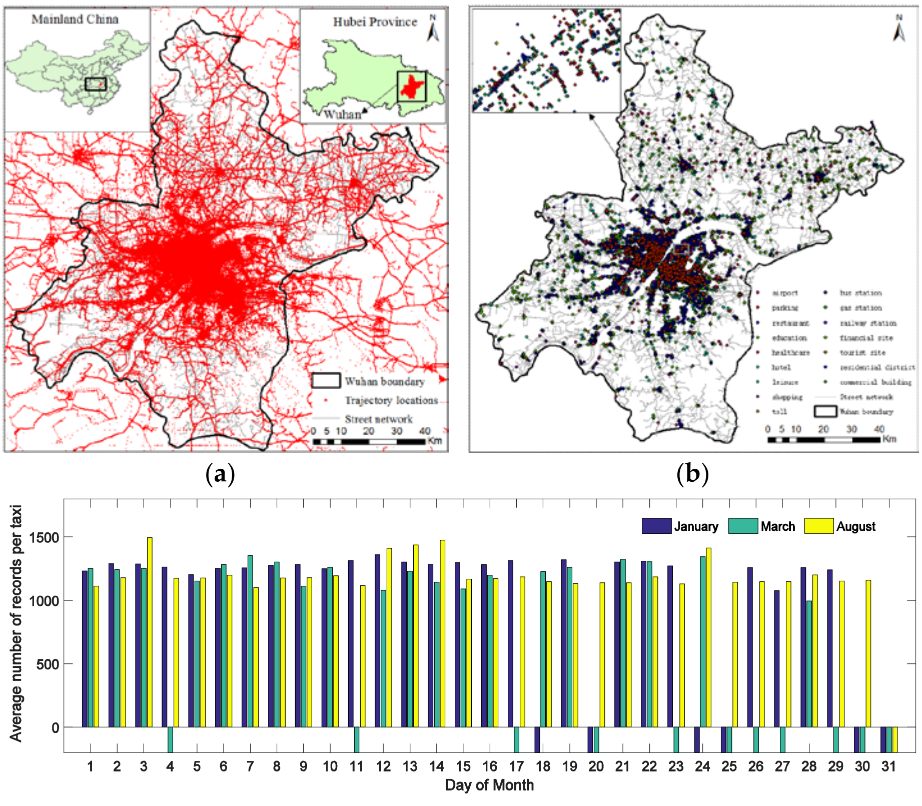
Therefore, this study is dedicated to investigate the scaling pattern of activity hotspots and further apply this rule to understand their urban functionality. Hence, it differs from the previous studies in three main aspects. Firstly, we use a head/tail break rule [43] to extract a large number of stopping locations from trajectory data, which are regarded as the activity locations for living, working, or shopping. These locations are further aggregated into activity hotspots using a newly designed temporal city clustering algorithm (TCCA). Secondly, we investigate the allometric scaling relationship between the number of POIs and the number of stopping locations, which helps to design an allometric ruler to identify the activity hotspots whose functionality could be reliably estimated using the specified number of stopping locations. This comes from the idea that humans can collectively interact with urban land use in an efficient way known as economies of scale [37]. Thirdly, a Bayesian learning model is developed to infer the urban functionality of the identified activity hotspots, where it utilizes temporal and spatial information on stopping locations covering 75 days. As a result, the benefits of this work are two-fold. First, it expands the application of scaling patterns, which contributes to a quantitative understanding of the relationship between human and urban environments. Second, it contributes a novel method to infer the urban functionality of activity hotspots, and the results will be useful in many urban applications. For example, urban planners could use them as decision supports for verifying, updating, and compiling city land use plans.

The remainder of this paper is organized as follows. In Section 2, we describe the datasets and the procedures to derive stopping locations and human activity hotspots. In Section 3, we report the scaling pattern of human activity hotspots and the allometric ruler to identify those whose urban functionality could be reliably estimated. In Section 4, we propose the Bayesian learning model to infer the urban functionality of the identified activity hotspots. The limitations are discussed in Section 5. Conclusions are drawn in Section 6.

## 2. Datasets and Preprocessing

### *2.1. GPS Trajectory Dataset*

The GPS trajectory dataset is acquired from the transportation agency of Wuhan, which shares its dataset with the research community for noncommercial use. It is composed of 830,062,777 records with the size of 46 GB, contributed by 16,787 taxis collecting for almost three months including January, March, and August of 2013. Each record in the dataset contains the ID number of the taxi and the information when the GPS signal is received at an interval of 60 s or less. The information includes the longitude (x) and latitude (y), the time (t), and the velocity (v) of the taxi. It should be noted that the longitude and latitude are referenced to the World Geodetic System 84 and measured with a horizontal accuracy of 5 m. Records with missing or incorrect information in terms of time or position are removed from the dataset, resulting in 93.1% of the records being adopted in this study. Spatially, it not only has trips to nearby cities but also covers the downtown area and suburbs of our study area, (Figure 1a); temporally, it covers 81% of the collection days due to business security concerns, which shows a good temporal coverage although each month contributes a different percentage, say 81% in January, 65% in March, or 97% in August. Besides, each taxi on average contributes approximately 1200 records per day, which indicates the consistency of data collection of each taxi in each day across the three months (Figure 1c). Hence, this trajectory dataset with such a large volume is chosen to ensure the robustness of our study.



(**a**) (**b**)

(**c**)

**Figure 1.** Datasets used in this study: (**a**) trajectories in March overlaid on the urban street network; (**b**) Points of Interest (POI) dataset overlaid on the urban street network; (**c**) average number of records per taxi during each day in January, March, and August, where *x*-axis is the day of month, *y*-axis labels the average number of records per taxi, and the day with negative number implies the absence of records.

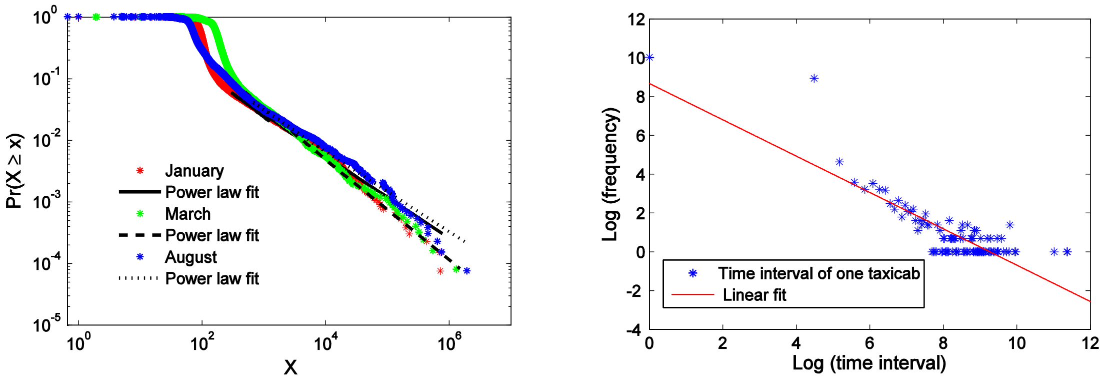
## *2.2. POI Dataset*

The POI dataset is obtained from the company NavInfo, a provider of navigation services in Mainland China with the most detailed POI dataset. It has a total number of 59,876 POIs comprised of 16 categories relating to most of our daily activities. The names of these categories are: airport, commercial building, bus station, residential district, education, financial site, gas station, healthcare, hotel, leisure, parking, railway station, restaurant, shopping, toll, and tourist site (Figure 1b). Among the POIs, shopping sites seem to have the highest total number—21,630—followed by restaurants and leisure sites which total 8382 and 6586 respectively. There are three airports and 27 railway stations. POI categories can reflect land use types, and hence they can be used to estimate the functionality of urban regions. It is POIs that attract the human activities in neighborhoods and the whole city. Therefore, the POI dataset is chosen for understanding its relationship with human activities and further supplying a prior distribution of urban functionality.

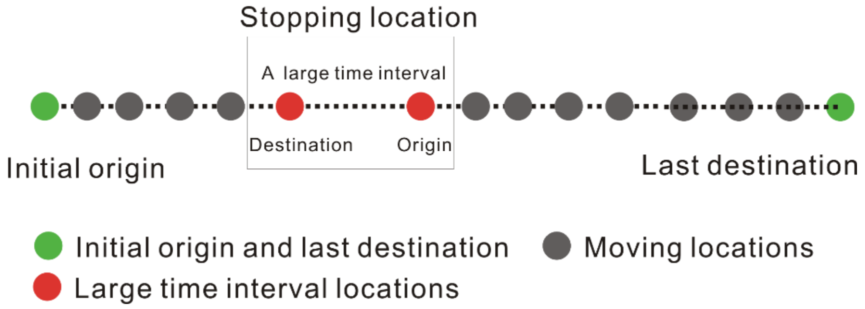
## *2.3. Stopping Locations*

Stopping locations denote the origins and destinations of the taxi trips. They are likely to contain implicit semantic information related to human activities [29,31–33] and are regarded as proxies of human activity locations for working, living, shopping, etc. Hence, in a continuous trajectory, stopping locations tend to display significant dissimilarities with other moving locations. From the perspective of time, stopping locations refer to adjacent locations with large time intervals between them, and they are used to demarcate the continuous trajectories into individual trips. They can be seen in Figure 2c, where the red dots indicate stopping locations with large time intervals.

Some trajectory datasets, like the New York taxi travel records, are accompanied by tag information that indicates whether the taxi is picking up or dropping off a fare. This information can be directly used to extract the stopping locations; however, there are many other trajectory datasets, like ours, which do not have such valuable information. Hence, it is not a trivial task to derive stopping locations, because of the ambiguity of setting threshold values for the time intervals. In this study, we take the arithmetic mean value of the time intervals of each taxi in each month as its own time threshold value, which is different from a previous study [36] where only one mean value is used. This strategy can be justified by the great differences among taxis in their mean values of time intervals, which obey remarkable power law distributions with respect to the three months as shown in Figure 2a. Besides, the reason for adopting the mean value as the time threshold for each taxi is the power-law-like distribution of the time interval of each taxi (Figure 2b), which incurs the usage of the head/tail break method [43–47] to demarcate the trajectory locations into two parts: the head (minority) of stopping locations and the tail (majority) of moving locations. Bearing the threshold values in mind, we extract 52,633,829 stopping locations within our study area (Figure 2c), which means each taxi generates around 42 stopping locations per day on average. This is reasonable for the normal business of taxi service.



(**a**) (**b**)



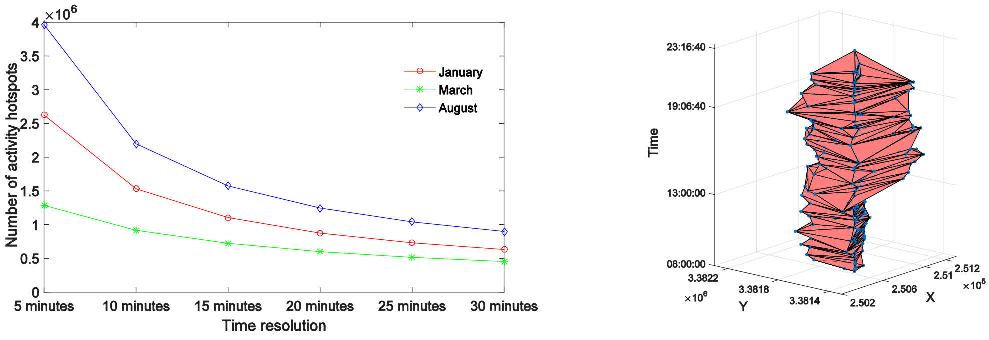
(**c**)

**Figure 2.** Extracting stopping locations: (**a**) a log-log plot for the power law distributions of the mean values of time intervals of individual taxis with respect to January, March, and August, where the significance tests are conducted with *p*-values as 0.5, 0.5, and 0.25 for the three months respectively (please refer to [48] for the details on calculating the *p*-value); (**b**) a log-log plot for the power-law-like distribution of the time interval values of one taxi; (**c**) an illustration on extracting stopping locations from a continuous trajectory.

## *2.4. Human Activity Hotspots*

We propose a temporal city clustering algorithm (TCCA), which is an extension of the city clustering algorithm (CCA) [15], to aggregate the individual stopping locations "from the bottom up" into human activity hotspots. They refer to the clusters within an urban area where most of the human activities take place. Specifically, this method starts from a randomly selected location, on which a cylinder is drawn to check whether other locations are encompassed. The cylinder is constructed by drawing a circle with a specified spatial radius *r*, which is then extruded along the time dimension with a specified time resolution $\tau$. This process goes recursively until no locations are within the cylinder or the contained locations have already been checked.

How to specify the values of the two parameters requires further investigation. In this study, we intuitively set the spatial radius *r* at 200 m, which is the approximate average length of road segments in the study area. This setting is consistent with the argument that human activities are mostly constrained to city blocks delineated by road networks [35]. On the other hand, for a better understanding of the effects of time resolution on scaling pattern, we change the time resolution $\tau$ from 5 min to 30 min with 5 min increments. This results in 6 groups of activity hotspots with respect to different time resolutions in each month. As shown in Figure 3a, the number of activity hotspots is decreasing with the increment of time resolution. Besides, we show one activity hotspot in Figure 3b, which clearly depicts its spatio-temporal process.



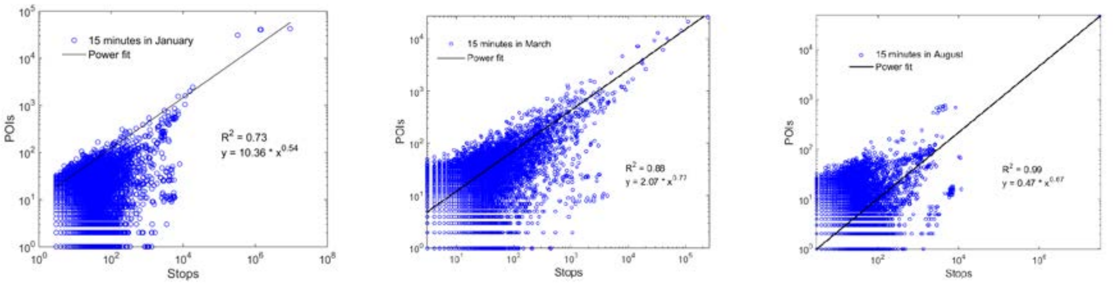
(**a**) (**b**)

**Figure 3.** Human activity hotspots for the three months: (**a**) the number of activity hotspots decreases with the increasing value of time resolution from 5 min to 30 min; (**b**) 3D boundary of an activity hotspot lasting from 1 January 2013 08:07:26 to 1 January 2013 22:19:14, where its spatial area expands at approximately 13:00.

## 3. Identify the Reliable Human Activity Hotspots from Their Scaling Pattern

### *3.1. Scaling Pattern of Human Activity Hotspots*

Human activity hotspots are good subjects to test the inherent scaling pattern. Hence, we investigate the scaling relationship between the number of stopping locations and the number of POIs, say $y = ax^b$, where $x$ is the number of stopping locations, $y$ is the number of POIs, $a$ is the constant, and $b$ is the scaling exponent. In this context, the number of stopping locations relates to the potential needs of the people visiting the activity hotspots, which reflects the importance of the activity hotspots. The number of POIs relates to the extent of urban land use that can satisfy the needs of the people, which reflects the potential functionality of the activity hotspots. As shown in Figure 4 and Table 1, the findings coincide with our initial assumption suggesting a clearly allometric sublinear scaling relationship for the three months irrespective of the time clustering resolution. This consistency implies that the universal law of economies of scale dominate human interaction with urban land use. In addition, these allometric scaling relationships display very high values of goodness of fit in term of R-square values, and they are all statistically significant with very small $p$-values approaching to 0 using F-test. Hence, the goodness of fit test and the significance test might guarantee the reliability of these scaling patterns.



(**a**) (**b**) (**c**)

**Figure 4.** Allometric scaling relationship between the number of stops and POIs for activity hotspots with time resolution of 15 min in: (**a**) January; (**b**) March; (**c**) August.

**Table 1.** Allometric scaling relationships between stops and POIs in different time resolutions for the three months (Note: All these models are statistically significant with $p$-values of 0 using F-test).

| | Time Resolution | *a* | *b* | *R*² | Time Resolution | *a* | *b* | *R*² |
|---|---|---|---|---|---|---|---|---|
| **January** | 5 min | 4.02 | 0.64 | 0.82 | 20 min | 12.00 | 0.53 | 0.73 |
| | 10 min | 7.45 | 0. 58 | 0.85 | 25 min | 13.68 | 0.52 | 0.73 |
| | 15 min | 10.36 | 0.54 | 0.73 | 30 min | 15.20 | 0.52 | 0.74 |
| **March** | 5 min | 1.54 | 0.74 | 0.60 | 20 min | 3.18 | 0.74 | 0.89 |
| | 10 min | 1.10 | 0.83 | 0.84 | 25 min | 3.43 | 0.72 | 0.92 |
| | 15 min | 2.07 | 0.77 | 0.88 | 30 min | 4.48 | 0.70 | 0.92 |
| **August** | 5 min | 1.58 | 0.60 | 0.97 | 20 min | 0.29 | 0.69 | 0.99 |
| | 10 min | 0.65 | 0.65 | 0.99 | 25 min | 0.25 | 0.70 | 0.99 |
| | 15 min | 0.47 | 0.67 | 0.99 | 30 min | 0.24 | 0.71 | 0.99 |

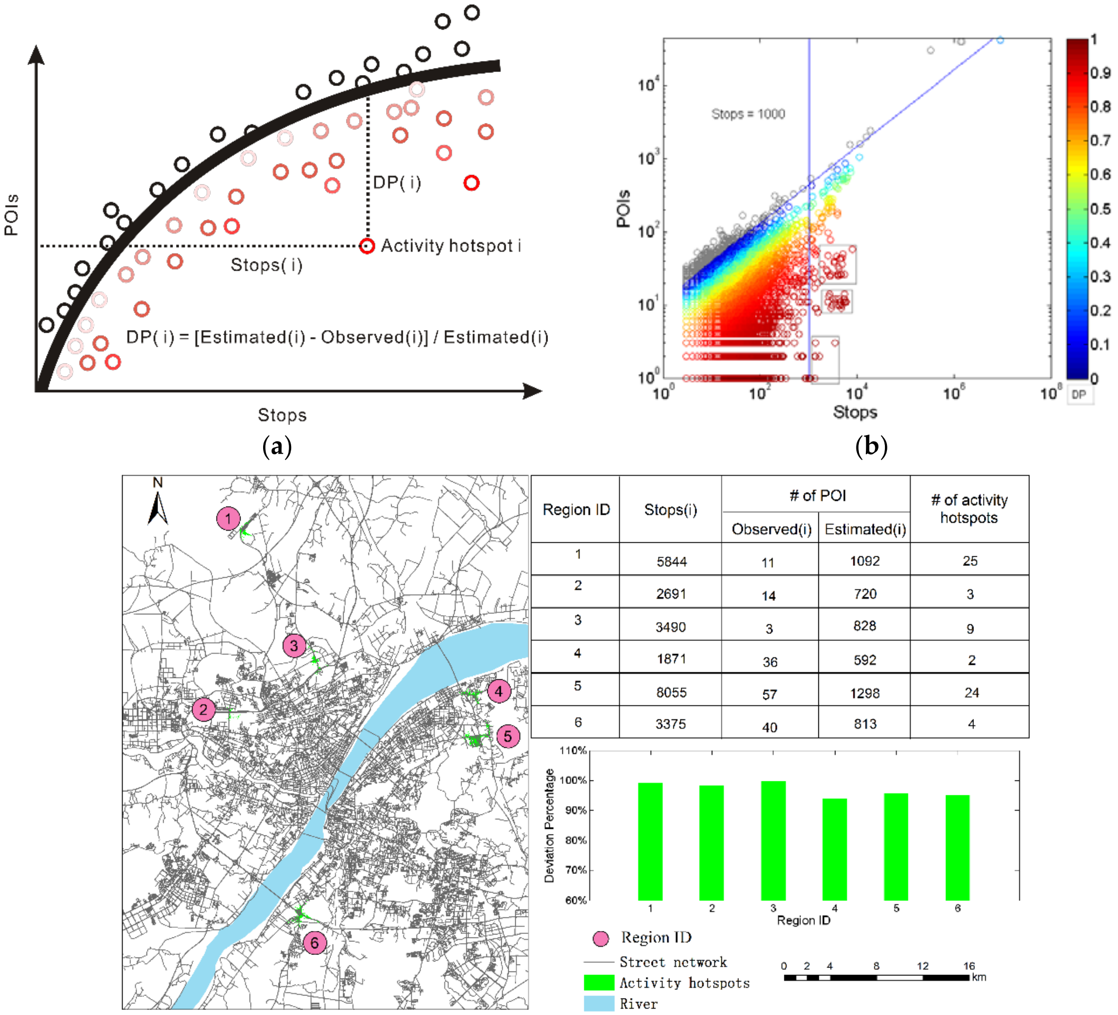
### *3.2. Identification of the Reliable Human Activity Hotspots*

The above scaling pattern of human activity hotspots hints that we could devise an allometric ruler, the de facto power regression line between the number of stops and POIs, to identify the activity hotspots whose functionality can be reliably estimated using stopping locations. Specifically, from the allometric ruler, we can define the deviation percentage (*DP*) for an activity hotspot *i* as:

$$DP(i) = (Estimated(i) - Observed(i))/Estimated(i) \tag{1}$$

where *Observed*(*i*) is the number of POIs in activity hotspot *i* and *Estimated*(*i*) is the number of POIs estimated from the stopping locations. As shown in Figure 5a, activity hotspots are separated into two parts by the allometric ruler, where the upper ones in gray have *DP* values less than 0 and the lower ones in red have *DP* values greater than 0. Besides, the number of stopping locations (stops) is also used to identify the reliable activity hotspots. Intuitively, the larger the values of *DP* and stops, the more reliable the activity hotspots can be estimated using stopping locations.

As an example, the ruler is applied to activity hotspots with a time resolution of 15 min in January. In Figure 5b, they are colored according to the *DP* values using a jet colormap. In this study, for the purpose of illustration, an activity hotspot is selected if its *DP* value is greater than 0.92 and its stops value is greater than 1000. The former ensures a reliable estimation of urban functionality, while the latter guarantees an important activity hotspot. Using the ruler, a total number of 67 activity hotspots are identified as the reliable ones (enclosed within boxes in Figure 5b). In the city, they are spatially related to six urban regions as shown in Figure 5c. For instance, some regions are associated with a large number of activity hotspots, while others are related to a very small number. Intuitively a popular region, say, a railway station, would correspond to a large number of activity hotspots. However, for ease of analysis, we take only the activity hotspot with the largest number of stops as the representative of each region.



| Region ID | Stops(i) | # of POI | | # of activity hotspots |
|---|---|---|---|---|
| | | Observed(i) | Estimated(i) | |
| 1 | 5844 | 11 | 1092 | 25 |
| 2 | 2691 | 14 | 720 | 3 |
| 3 | 3490 | 3 | 828 | 9 |
| 4 | 1871 | 36 | 592 | 2 |
| 5 | 8055 | 57 | 1298 | 24 |
| 6 | 3375 | 40 | 813 | 4 |

(**a**) (**b**)

(**c**)

**Figure 5.** Usage of the allometric ruler: (**a**) schematic illustration of the allometric ruler; (**b**) application of the allometric ruler to human activity hotspots with a time resolution of 15 min in January; (**c**) map of six regions corresponding to the identified activity hotspots.

## 4. Inferring the Functionality of Reliable Activity Hotspots Using a Bayesian Model

### *4.1. Construct the Bayesian Inference Model*

Once the activity hotspots are identified, the main question is how to construct a model to infer their urban functionality using the stopping locations. To solve this problem, we use a Bayesian inference model, aiming to employ the Bayes' theorem to update the probability of a hypothesis with observed evidence. In this study, the distribution of POIs is considered as the hypothetic functionality distribution, whereas the stopping locations act as observed evidence to infer the real functionality distribution. Specifically, the proposed Bayesian inference model is designed to determine the most likely POI type, given a stopping location's spatial coordinate and occurrence time. The model is shown in Equation (2), where $\Pr((x, y)|type)$ is the probability of observing one stopping location at coordinate *x*, *y*, given a POI *type*, $\Pr(t|type)$ is the probability of observing one stopping location around time *t*, given a POI *type*, and $\Pr(type)$ is the probability of observing a certain type of POI in the study area. Note that it is assumed that the occurrence of a stopping location at coordinate *x*, *y* was independent of the time *t*, given a POI *type*.

$$\Pr\left(type\middle|(x, y), t\right) = \frac{\Pr\left((x, y)\middle|type\right) \bullet \Pr\left(t\middle|type\right) \bullet \Pr(type)}{\Pr\left((x, y), t\right)}, \tag{2}$$

To put it in detail, this model is composed of three steps.

The first step is to derive the empirical probability of observing one stopping location around time *t*, given a POI *type* (Equation (3)). To achieve this step, we decompose the entire region into Voronoi cells using each type of POI, which is similar to the assumption in central place theory [49], so that each POI acted as a nodal point for the distribution of goods or services to the surrounding Voronoi cell. Consumers were assumed to act as homo economicus to the nearest POI, given that travel effort was equal in all directions. Then, stopping locations falling inside each Voronoi cell are aggregated according to time of day, which is discredited into 24 h. Note that the aggregation process does not rely on simply counting the number of stopping locations, but considers the geographical weight in terms of its inverse distance to the POI. In other words, near stopping locations are more important than far ones. Finally, a further aggregation is performed for all the Voronoi cells regarding the time of day, which leads to the empirical probability of observing one stopping location at a certain time of day, given each type of POI. In Equation (3), $SL_i$ is the $i^{th}$ stopping location, $VN_j$ is the $j^{th}$ Voronoi cell, $N_{type}$ is the number of a certain *type* of POI, $n_t^j$ is the number of stopping locations within the $j^{th}$ Voronoi cell around a certain time *t* of the day, and $n^j$ is the number of stopping locations within the $j^{th}$ Voronoi cell.

$$P(\mathrm{t}|type) = \sum\nolimits_j^{N_{type}} \sum\nolimits_i^{n_t^j} (1/dist(SL_i, VN_j))^2 \Big/ \sum\nolimits_j^{N_{type}} \sum\nolimits_i^{n^j} (1/dist(SL_i, VN_j))^2, \tag{3}$$

The second step focuses on deriving the empirical probability of observing one stopping location at coordinate *x*, *y*, given a POI *type*. It is very difficult to directly derive this probability from observations, but we assume that it was proportional to the gravity force between the POI and the stopping location [33]. The gravity force can be approximated by the product of the attractiveness value of the POI and its inverse distance to the stopping location. In Equation (4), $A_{type}$ is the attractiveness value of a POI *type* specified according to its physical size or importance, and $\mu$ is the distance decay factor, in this study set as the value of 1.5 [33].

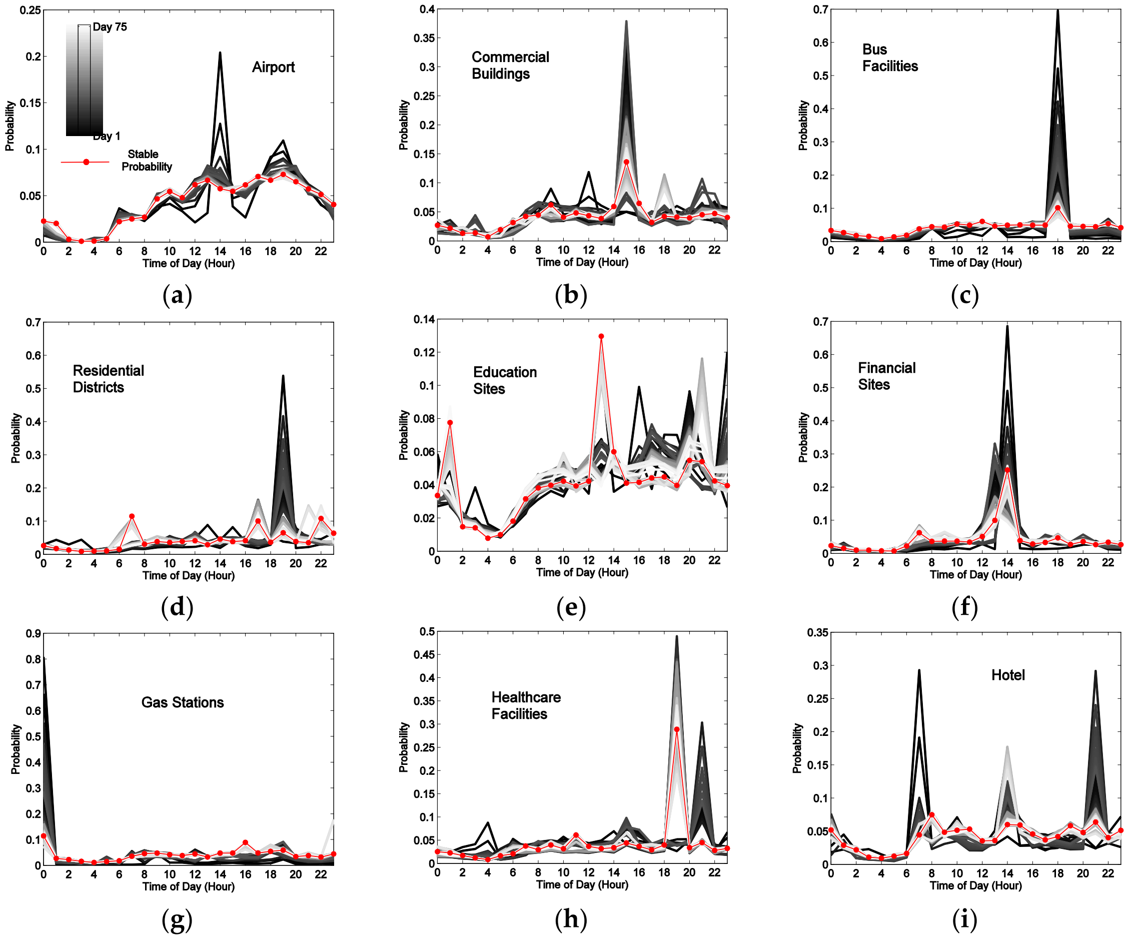
$$\Pr\left((\mathrm{x}, \mathrm{y})\middle|\mathrm{type}\right) \propto A_{\mathrm{type}} \bullet (\mathrm{dist}(\mathrm{SL_i}, \mathrm{POI_{type}}))^{-\mu}, \tag{4}$$

The third step is to infer the activity type of a stopping location with the maximum probability according to the naïve Bayesian theory. To further mimic the real urban functionality of activity hotspots, we add the damping factor $\beta$ to the inference model. It was originally used in the PageRank [50] algorithm to model the probability that a web surfer would continue clicking on the current web page without jumping to a randomly chosen one. In this context, it allows the stopping location with a probability to be matched with a POI *type* in the entire study area to supplement the urban functionality of an activity hotspot. The formula can be seen in Formula (5), where $\Phi_{\text{global}}$ is the set of POI types in the study area, $\Phi_{\text{local}}$ is the set of POI types in the activity hotspot, and $\beta_{\text{threshold}}$ is the threshold value of the damping factor.

$$\begin{cases} Max_{type\in\Phi_{local}}\left\{A_{type}\bullet \text{dist}(\text{SL}, \text{POI}_{type})^{-\mu}\bullet \Pr\left(t\middle|type\right)\bullet \Pr(type)\right\}, & \beta<\beta_{threshold} \\ Max_{type\in\Phi_{global}}\left\{\Pr\left(t\middle|type\right)\bullet \Pr(type)\right\}, & \beta\geq\beta_{threshold} \end{cases}, \tag{5}$$

## 4.2. Train the Model and Infer the Urban Functionality

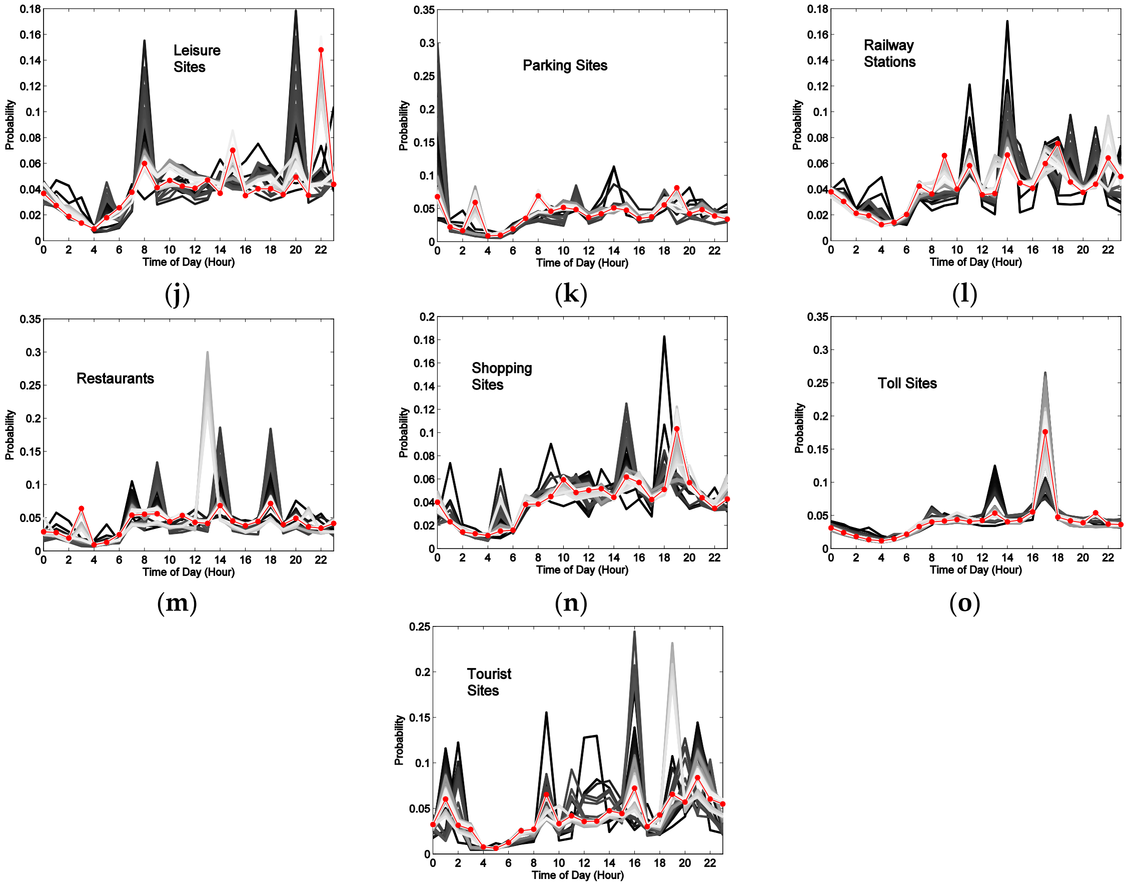
Once the model is constructed, we train it with the stopping locations. The training process is carried out in two steps. First, we calculate the probability of observing a stopping location given a POI *type* (likelihood function), which is actually to train Equation (3). However, for Equation (4), there is no need to use the empirical stopping locations, because it can be calculated on the fly. Specifically, the training process adopts all the stopping locations covering three months. To test stability, stopping locations are supplied to the model gradually in daily increments for a total of 75 days, which eventually leads to a saturated likelihood function curve, shown as the red dotted line in Figure 6a. Figure 6 shows how each type of POI functions at different times of day; generally, the probability of human activities decreases from midnight to morning and displays different patterns in the daytime for different types of POI. For example, in Figure 6d, we can observe roughly two peaks in human activities; going to work from 06:00 to 08:00 and returning home from 16:00 to 18:00. Second, we calculate the probability of each POI type before stopping locations are available, which is actually to train the prior probability, say $\Pr(type)$. Hence, this training process is simply to compute the percentage of POIs with respect to different types in the study area.



(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

(**g**) (**h**) (**i**)



(j) (k) (l)

(m) (n) (o)

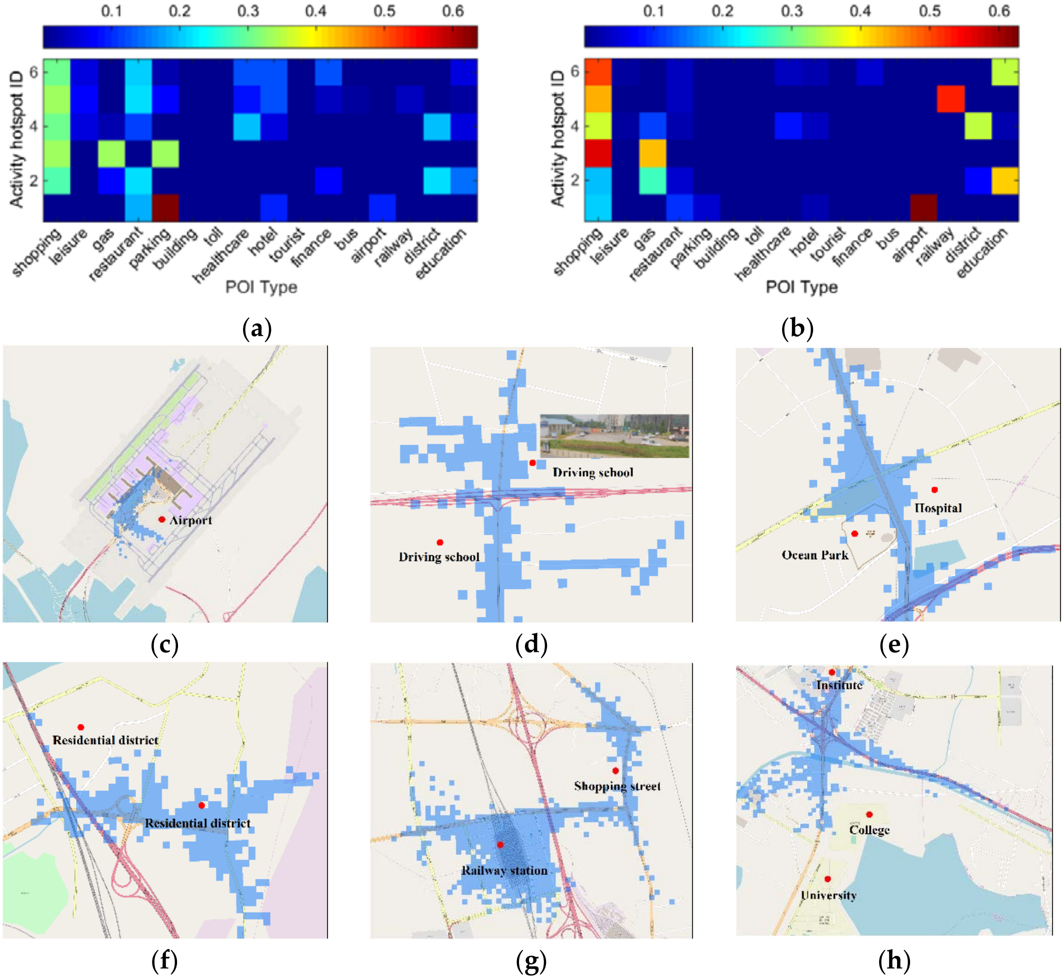
(p)

**Figure 6.** The probability of observing one stopping location at times of day for (a) airport, (b) commercial buildings, (c) bus facilities, (d) residential districts, (e) education sites, (f) financial sites, (g) gas station, (h) healthcare facilities, (i) hotel, (j) leisure sites, (k) parking sites, (l) railway stations, (m) restaurants, (n) shopping sites, (o) toll sites and (p) tourist sites, where *x*-axis is in hour and *y*-axis is the probability. (Note: In this figure, the curves with color changing from black to white are calculated by using the stopping locations from one day to 75 days respectively, and the curve with the red dotted line indicates the saturated one calculated by using all stopping locations of 75 days; this figure suggests that different types of POI might function differently in the time of day, for instance, airport, bus station and railway station present a relatively stable visiting pattern during daytime, residential district or commercial building displays roughly two peaks in the morning and in the afternoon, and leisure sites show a clear peak around 22:00 at night).

Then, the urban functionality of the identified activity hotspots is inferred using stopping locations. Specifically, for each activity hotspot *i*, we randomly select stopping locations with the same number as the estimated POIs, and these stopping locations are supplied to Formula (5) to determine the most matched POI types. Hence, these newly derived POI types are aggregated to infer their functionality by either reweighting the importance among POIs or adding new types of POIs. It should be noted that we set the damping factor at the value of 0.85, which has been generally used in many studies on human mobility patterns and models [36,51]. In this context, it means that approximately 85% of stopping locations would be matched with POIs in the local activity hotspot and approximately 15% of them would be assigned to POI types in the global study area. To avoid sampling bias, this procedure is repeated 100 times for each activity hotspot, and the final average value is used. Compared with the prior probability of functionality distribution (Figure 7a), we can enhance the reliability of activity hotspots for their urban functionality, as shown in Figure 7b.

To put the results in detail:

- Activity hotspot 1: It is enhanced by improving the importance of airport, namely many stopping locations are matched with the POI of airport. Hence, it is considered to provide aviation service, which coincides with the real situation by visual check in Figure 7c.
- Activity hotspot 2: It is enhanced by improving the importance of education and gas station. The two major functionalities indicate that it mainly provides the service of driver training, which can be visually checked in Figure 7d with two driving schools.
- Activity hotspot 3: It is not enhanced by our model, which is considered as a shopping region. But, as shown in Figure 7e, it mainly provides healthcare and recreation services. The reason can be probably attributed to the long walking distance from the stopping locations to the two POIs due to the current city planning.
- Activity hotspot 4: It is slightly enhanced by improving the importance of residential district and shopping. Hence, it is considered as a residential area with abundant shopping sites, few gas stations and healthcare sites by visual check in Figure 7f.
- Activity hotspot 5: It is obviously enhanced by improving the importance of railway stations, namely many stopping locations are matched with the POI of railway station. Hence, it is considered as a railway service area, which agrees well the real land use pattern (Figure 7g).
- Activity hotspot 6: It is clearly enhanced by emphasizing the importance of education. Hence, it is regarded as the place mainly providing education services, which is compatible with the three institutions or universities there by visual check in Figure 7h.



(**a**) (**b**)

(**c**) (**d**) (**e**)

(**f**) (**g**) (**h**)

**Figure 7.** Urban functionality of the activity hotspots and their validations on OpenStreetMap (available at: www.openstreetmap.org): (**a**) prior probability of functionality distribution; (**b**) inferred probability of functionality distribution; (**c**) activity hotspot 1, providing aviation services with an international airport; (**d**) activity hotspot 2, providing educational services with driving schools; (**e**) activity hotspot 3, providing recreation and healthcare services with an ocean park and an infectious disease hospital, which cannot be inferred by our model; (**f**) activity hotspot 4, serving as a residential

area with many shopping sites; (**g**) activity hotspot 5, providing railway services with a train station; (**h**) activity hotspot 6, serving as an educational area with several institutions or colleges.

## 5. Discussion

The underlying scaling pattern of geographic entities might be examined from two perspectives, namely the power law distribution of one quantity and the allometric relationship between two quantities. The allometric relationship could be further examined in situations of a sublinear or superlinear scaling relationship. For the sublinear case, it is typically observed that a city resembles an organism and its population size scales sublinearly with the consumed energies or resources for economies of scale. For the superlinear case, it found that the population size of a city scaled superlinearly with the amount of patents/innovation for the requirement of wealth creation [37]. The scaling pattern of geographical entities might have many potential implications, but only very few urban studies in the literature have realized its importance, such as the studies on urban sprawl [41] and map generalization [42]. Hence, the novelty of our study is to infer the urban functionality of human activity hotspots using their allometric scaling relationship with urban land use, which to our knowledge has rarely been investigated in the literature.

Nonetheless, this study has several limitations:

First, to estimate the urban functionality of activity hotspots, we use taxi trajectories spanning 75 days over three months. A major concern is the reliability of the reported results, because taxis represent only one travel mode among many chosen by urban residents, which limits the observed types of human activities. In other words, it is very difficult to estimate the statistical bias owing to the absence of transportation survey data at Wuhan, which is a common problem of investigations on human activity or mobility patterns using taxi trajectories. Therefore, how to resolve this issue is still an open question. Nonetheless, a possible solution is to integrate other types of human activity or movement data, such as the integrated circuit card data collected by a bus or metro system, the movement data of private vehicles or bicycles, or even pedestrian movement data. This requires further studies.

Secondly, we investigate the quality of GPS trajectory dataset from two aspects. Spatially, we examine the reliability of the assignment of GPS locations to the Voronoi cells of a specified type of POI. To do so, we calculate the service radius ($r$) of a Voronoi cell as the radius of a circle with the same spatial extent. As shown in Figure 8, we find that 99% of Voronoi cells have an $r$ value greater than 7 m and 90% of Voronoi cells have an $r$ value greater than 13 m. Hence, via a simple comparison with the horizontal accuracy of GPS locations (5 m), it can be generally assumed that the accuracy of GPS locations has a limited influence on their assignments to the Voronoi cells. Temporally, the major concern is that the temporal resolution of our GPS location is relatively low, and it is very difficult to estimate the stops with time duration of less than one minute. In other words, you can hardly know what happens between two consecutive GPS locations. However, we try to use a massive dataset in order to compensate for lost stopping locations. Besides, it is reasonable that each taxi contributes on average 42 stopping locations (or 41 trips) in one day (24 h), because the estimated monthly income of one taxi driver of 4613 Yuan (considering the facts of two drivers sharing a taxi, the 15 Yuan charge for a trip, and the profit rate of 0.5 in Wuhan, China) is very close to the amount (5000 Yuan) published in the income survey report [52].

Thirdly, in the Bayesian reference model we set the damping factor at 0.85, which means that approximately 85% of stopping locations are matched to the POIs in the local activity hotspot, and approximately 15% of them are matched to the POIs in the entire study area. By setting a damping factor we can mimic real situations by supplementing the urban functionality of the activity hotspot. However, it is not easy to set the value of a damping factor for a better simulation, and it requires further study.

Fourthly, to select the human activity hotspots whose functionality could be reliably estimated, we empirically determined the *DP* value of 0.92 and the stops value of 1000 for the purpose of illustration. This is a potential limitation, and it affects the number of activity hotspots to be selected.

In other words, how to select the activity hotspots is still an open question, and it should be dependent on the field of application and determined by decision makers.

Fifthly, the inferred urban functionality is validated using the OpenStreetMap (OSM) [53], an online map service aimed specifically at creating and providing free geographic data such as street maps to anyone. The validation procedure starts by visually observing the major urban functionality of the activity hotspot on the map, and then it compares them with the major inferred urban functionality. For some activity hotspots, it is difficult to identify the functionality from OSM due to the absence of geographic data, and the validation is conducted in the field. However, it suffers from subjectivity and could be applied to only a few samples, with inaccurate results. A more objective and accurate validation method is thus needed, which requires further study.

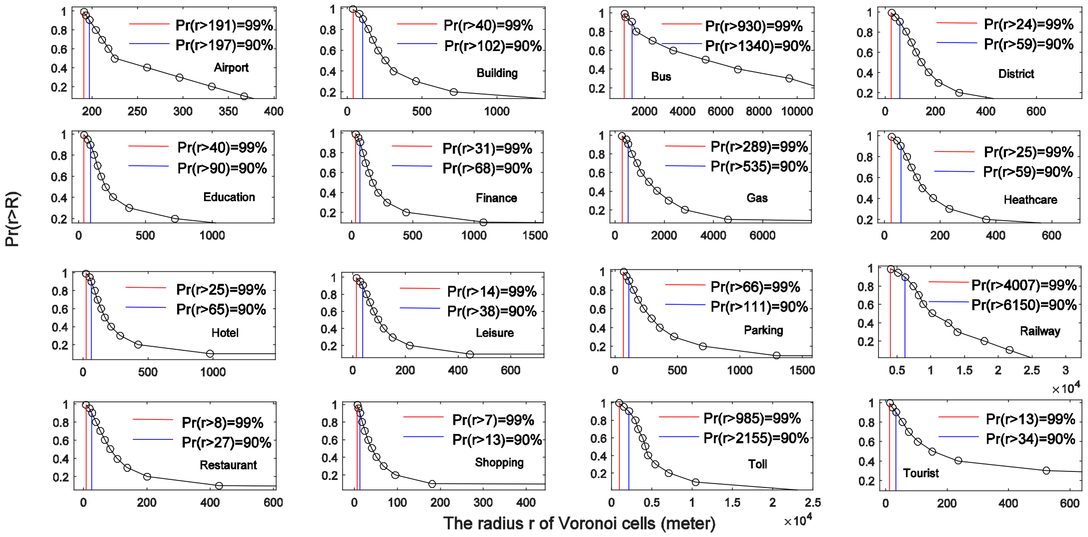


**Figure 8.** The cumulative distribution of the service radius *r* with respect to the type of POI.

## 6. Conclusions

In this paper, we aim to identify and infer the urban functionality of human activity hotspots in the light of the underlying scaling pattern. First, we derive a huge number of stopping locations from massive taxi trajectory data using the head/tail break rule, which produces what we regard as proxies of human activity locations. Second, these stopping locations are aggregated into human activity hotspots using a temporal city clustering algorithm, and human activity hotspots are reported to display the scaling pattern over time. With the scaling pattern, we devise an allometric ruler to identify the activity hotspots whose functionality can be reliably estimated using the specified number of stopping locations. Eventually, 67 activity hotspots are identified, which corresponds to six regions in space.

A Bayesian inference model is then proposed to infer their urban functionality. The model studies both temporal and spatial information in stopping locations covering 75 days, where temporal information denotes time of day and spatial information refers to spatial coordinates. Specifically, a probability curve depicting how each type of POI functions during different times of day is empirically calculated. Additionally, a prior probability of functionality distribution is calculated as the percentage of each type of POI in the study area. Thereafter, each stopping location is supplied to the model to match with the most likely POI type, which eventually leads to a reliable functionality distribution for each activity hotspot. Overall, our findings suggest that most of the activity hotspots could be understood well in terms of their functionality distribution. The results will be useful for decision-making on urban land use planning and transportation in general, and they will be beneficial for monitoring the change of urban land use and the dynamics of urban transport in particular. Future work will thus be concentrated on the application of our method to other types of human movement data such as the circuit card data collected by a bus or metro system.

**Acknowledgments:** We would like to thank the financial support from the National Natural Science Foundation of China who supported this work with Award numbers: 41401453 and 41331175. We would also like to extend our gratitude to the anonymous reviewers for their valuable comments and Angela Watson for polishing and proofreading our language.

**Author Contributions:** T.J. and Z.J. conceived and designed the experiments; T.J. performed the experiments; T.J. and Z.J. analyzed the data; T.J. wrote the paper.

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

## References

1. Cleveland, W.S. Data science: An action plan for expanding the technical areas of the field of statistics. *Stat. Anal. Data Min*. **2014**, *7*, 414–417.
2. Widener, M.J.; Li, W. Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US. *Appl. Geogr.* **2014**, *54*, 189–197.
3. Yang, W.; Mu, L. GIS analysis of depression among Twitter users. *Appl. Geogr.* **2015**, *60*, 217–223.
4. Yan, Z.X. Towards semantic trajectory data analysis: A conceptual and computational approach. In Proceedings of the 2009 Very Large Data Bases (VLDB) Conference, Lyon, France, 28 August 2009.
5. Bohte, W.; Maat, K. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transp. Res. Part C Emerg. Technol.* **2009**, *17*, 285–297.
6. Jia, T.; Carling, K.; Håkansson, J. Trips and their $CO_2$ emissions to and from a shopping center. *J. Transp. Geogr*. **2013**, *33*, 135–145.
7. Carling, K.; Håkansson, J.; Jia, T. Out-of-town shopping and its induced $CO_2$ emissions. *J. Retail. Consum. Serv*. **2013**, *20*, 382–388.
8. Li, J.; Zhang, Y.; Wang, X.; Qin, Q.; Wei, Z.; Li, J. Application of GPS Trajectory Data for Investigating the Interaction between Human Activity and Landscape Pattern: A Case Study of the Lijiang River Basin, China. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 1–17.
9. Reades, J.; Calabrese, F.; Ratti, C. Eigenplaces: Analysing cities using the space-time structure of the mobile phone network. *Environ. Plan. B Plan. Des*. **2009**, *36*, 824–836.
10. Neuhaus, F. Urban Diary-A tracking project: Capturing the beat and rhythm of the city: Using GPS devices to visualize individual and collective routines within Central London. *J. Space Syntax* **2010**, *1*, 315–336.
11. Jia, T.; Jiang, B. Exploring human activity patterns using taxicab static points. *ISPRS Int. J. Geo-Inf.* **2012**, *1*, 89–107.
12. Zhang, F.; Yuan, J.; Wilkie, D.; Zheng, Y.; Xie, X. Sensing the pulse of urban refueling behavior: A perspective from taxi mobility. *ACM Trans. Intell. Syst. Technol.* **2013**, *9*, 1–24.
13. Scholz, R.W.; Lu, Y. Detection of dynamic activity patterns at a collective level from large-volume trajectory data. *Int. J. Geogr. Inf. Sci*. **2014**, *28*, 946–963.
14. Alvares, L.O.; Bogorny, V.; Kuijpers, B.; de Macedo, J.A.F.; Moelans, B.; Vaisman, A. A model for enriching trajectories with semantic geographical information. In Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems, Seattle, WA, USA, 7–9 November 2007.
15. Rozenfeld, H.D.; Rybski, D.; Andrade, J.S., Jr.; Batty, M.; Stanley, H.E.; Makse, H.A. Laws of population growth. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 18702–18707.
16. Zhao, X.; Xu, W. A clustering-based approach for discovering interesting places in a single trajectory. In Proceedings of the 2009 Second International Conference on Intelligent Computation Technology and Automation, Zhangjiajie, China, 10–11 October 2009.
17. Louail, T.; Lenormand, M.; Cantu-Ros, O.G.; Picornell, M.; Herranz, R.; Frias-Martinez, E.; Ramasco, J.J.; Barthelemy, M. From mobile phone data to the spatial structure of cities. *Sci. Rep.* **2014**, *4*, 1–12.
18. Sun, Y.; Fan, H.; Li, M.; Zipf, A. Identifying the city center using human travel flows generated from location-based social networking data. *Environ. Plan. B Plan. Des*. **2016**, *43*, 480–498.
19. Santos, M.; Moreira, A. Automatic classification of location contexts with decision trees. In Proceedings of the CSMU-2006: Conference on Mobile and Ubiquitous Systems, Guimares, Portugal, 29–30 June 2006.
20. Jiang, S.; Alves, A.; Rodrigues, F.; Ferreira, J.; Pereira, F.C. Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Comput. Environ. Urban Syst*. **2015**, *53*, 36–46.
21. Zeng, L.R.; Lin, H.F. Analysis of land use along urban rail transit based on POI Data. In Proceedings of the 16th COTA International Conference of Transportation Professionals, Shanghai, China, 6–9 July 2016.

22. Yue, Y.; Zhuang, Y.; Yeh, A.G.O.; Xie, J.Y.; Ma, C.L.; Li, Q.Q. Measurements of POI-based mixed use and their relationships with neighbourhood vibrancy. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 658–675.
23. Wolf, J.; Schönfelder, S.; Samaga, U.; Oliveira, M.; Axhausen, K.W. 80 weeks of GPS-traces: Approaches to enriching the trip information. *Transp. Res. Rec.* **2004**, *1870*, 46–54.
24. Xie, K.; Deng, K.; Zhou, X. From trajectories to activities: A spatio-temporal join approach. In Proceedings of the 2009 International Workshop on Location Based Social Networks, New York, NY, USA, 4–6 November 2009.
25. Griffin, T.; Huang, Y. A decision tree classification model to automate trip purpose derivation. In Proceedings of the ISCA 18th International Conference on Computer Applications in Industry and Engineering, Honolulu, HI, USA, 9–11 November 2005; pp. 44–49.
26. Montini, L.; Rieser-Schüssler, N.; Horni, A.; Axhausen, K. Trip purpose identification from gps tracks. *Transp. Res. Rec. J. Transp. Res. Board* **2014**, *2405*, 16–23.
27. Furletti, B.; Cintia, P.; Renso, C.; Spinsanti, L. Inferring human activities from gps tracks. In Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, Chicago, IL, USA, 11 August 2013.
28. Moiseeva, A.; Jessurun, J.; Timmermans, H. Semiautomatic imputation of activity-travel diaries using GPS traces, prompted recall, and context-sensitive learning algorithms. *Transp. Res. Rec. J. Transp. Res. Board* **2010**, *2183*, 60–68.
29. Zhang, W.; Qi, G.; Pan, G.; Lu, H.; Li, S.; Wu, Z. City-Scale social event detection and evaluation with taxi traces. *ACM Trans. Intell. Syst. Technol.* **2015**, *6*, 40.
30. Fuchs, G.; Stange, H.; Hecker, D.; Andrienko, N.; Andrienko, G. Constructing semantic interpretation of routine and anomalous mobility behaviors from big data. *ACM SIGSPAT. Spec.* **2015**, *7*, 27–34
31. Yuan, J.; Zheng, Y.; Xie, X. Discovering regions of different functions in a city using human mobility and POIs. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012.
32. Pan, G.; Qi, G.; Wu, Z.; Zhang, D.; Li, S. Land-Use classification using taxi GPS traces. *IEEE Trans. Intell. Transp. Syst.* **2013**, *14*, 113–123.
33. Gong, L.; Liu, X.; Wu, L.; Liu, Y. Inferring trip purposes and uncovering travel patterns from taxi trajectory data. *Cartogr. Geogr. Inf. Sci.* **2015**, *43*, 103–114.
34. Brockmann, D.; Hufnagel, L.; Geisel, T. The scaling laws of human travel. *Nature* **2006**, *439*, 462–465.
35. Jiang, B.; Jia, T. Zipf's law for all the natural cities in the United States: A geospatial perspective. *Int. J. Geogr. Inf. Sci.* **2011**, *25*, 1269–1281.
36. Jia, T.; Jiang, B.; Carling, K.; Bolin, M.; Ban, Y.F. An empirical study on human mobility and its agent-based modeling. *J. Stat. Mech. Theory Exp.* **2012**, doi:10.1088/1742-5468/2012/11/P11024.
37. Bettencourt, L.M.A.; Lobo, J.; Helbing, D.; Kühnert, C.; West, G.B. Growth, innovation, scaling, and the pace of life in cities. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 7301–7306.
38. Schmidt, N.K. *Scaling: Why Is Animal Size So Important*; Cambridge University Press: Cambridge, UK, 1984.
39. Lobo, J.; Bettencourt, L.M.A.; Strumsky, D.; West, G.B. Urban scaling and the production function for cities. *PLoS ONE* **2013**, *8*, e58407.
40. Schlapfer, M.; Bettencourt, L.M.A.; Grauwin, S.; Raschke, M.; Claxton, R.; Smoreda, Z.; West, G.B.; Ratti, C. The scaling of human interactions with city size. *J. R. Soc. Interface* **2014**, *11*, 1–9.
41. Sutton, P.C. A scale-adjusted measure of ''Urban sprawl'' using nighttime satellite imagery. *Remote Sens. Environ.* **2003**, *86*, 353–369.
42. Jiang, B.; Liu, X.; Jia, T. Scaling of geographic space as a universal rule for map generalization. *Ann. Am. Assoc. Geogr.* **2013**, *103*, 844–855.
43. Jiang, B. Head/Tail breaks: A new classification scheme for data with a heavy-tailed distribution. *Prof. Geogr.* **2013**, *65*, 482–292.
44. Dunham-Snary, K.J.; Sandel, M.W.; Westbrook, D.G.; Ballinger, S.W. A method for assessing mitochondrial bioenergetics in whole white adipose tissues. *Redox Biol.* **2014**, *2*, 656–660.
45. Long, Y.; Thill, J.C. Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. *Comput. Environ. Urban Syst.* **2015**, *53*, 19–35.
46. Long, Y. Redefining Chinese city system with emerging new data. *Appl. Geogr.* **2016**, *75*, 36–48.
47. Li, S.; Dragicevic, S.; Castro, F.A.; Sester, M.; Winter, S.; Coltekin, A.; Pettit, C.; Jiang, B.; Haworth, J.; Stein, A.; Cheng, T. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS J. Photogramm. Remote Sens.* **2016**, *115*, 119–133.

48. Clauset, A.; Shalizi, C.R.; Newman, M.E.J. Power-law distributions in empirical data. *SIAM Rev*. **2009**, *51*, 661–703.
49. Christaller, W. *Central Places in Southern Germany*, 1st ed.; Prentice-Hall: Upper Saddle River, NJ, USA, 1966.
50. Page, L.; Brin, S. The anatomy of a large-scale hypertextual web search engine. In Proceedings of the Seventh International World-Wide Web Conference, Brisbane, Australia, 14–18 April 1998
51. Mtibaa, A.; May, M.; Ammar, M. Social Forwarding in Mobile Opportunistic Networks: A Case of PeopleRank. In *Handbook of Optimization in Complex Networks*; Thai, M., Pardalos, P., Eds.; Springer: New York, NY, USA, 2012; Volume 58.
52. Ganji: the monthly income distribution of taxi drivers in Wuhan. Available online: http://wh.ganji.com/gz_zpczcsiji/ (accessed on 3 11 2017)
53. Haklay, M.; Weber, P. OpenStreetMap—User generated street map. *IEEE Pervasive Comput*. **2008**, *7*, 12–18.

© 2017 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons by Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).