



Article

# Identifying Urban Functional Zones Using Public Bicycle Rental Records and Point-of-Interest Data

Xiaoyi Zhang <sup>1</sup>, Wenwen Li <sup>2</sup>, Feng Zhang <sup>1,3</sup>, Renyi Liu <sup>3</sup> and Zhenhong Du <sup>1,3,\*</sup>

<sup>1</sup> School of Earth Sciences, Zhejiang University, Hangzhou 310027, China; zhangxixi110@hotmail.com (X.Z.); zfcarnation@zju.edu.cn (F.Z.)

<sup>2</sup> School of Geographical Sciences & Urban Planning, Arizona State University, Tempe, AZ 85287-5302, USA; wenwen@asu.edu

<sup>3</sup> Zhejiang Provincial Key Laboratory of Geographic Information Science, Department of Earth Sciences, Zhejiang University, Hangzhou 310028, China; liurenyi@zju.edu.cn

\* Correspondence: duzhenhong@zju.edu.cn

Received: 25 September 2018; Accepted: 22 November 2018; Published: 27 November 2018



**Abstract:** Human mobility data have become an essential means to study travel behavior and trip purpose to identify urban functional zones, which portray land use at a finer granularity and offer insights for problems such as business site selection, urban design, and planning. However, very few works have leveraged public bicycle-sharing data, which provides a useful feature in depicting people's short-trip transportation within a city, in the studies of urban functions and structure. Because of its convenience, bicycle usage tends to be close to point-of-interest (POI) features, the combination of which will no doubt enhance the understanding of the trip purpose for characterizing different functional zones. In our study, we propose a data-driven approach that uses station-based public bicycle rental records together with POI data in Hangzhou, China to identify urban functional zones. Topic modelling, unsupervised clustering, and visual analytics are employed to delineate the function matrix, aggregate functional zones, and present mixed land uses. Our result shows that business areas, industrial areas, and residential areas can be well detected, which validates the effectiveness of data generated from this new transportation mode. The word cloud of function labels reveals the mixed land use of different types of urban functions and improves the understanding of city structures.

**Keywords:** human mobility; traffic analysis zones; topic modelling; k-means; land use

## 1. Introduction

Urban functional zones are the areas assigned to different social and economic activities. It can be divided into several types (i.e., residential, commercial, educational, and industrial zones) [1], which are firmly related to human mobility and social behavior [2–4] and are essential for urban planning applications such as transportation design, air pollution prevention and other social studies [5,6]. Therefore, the effective and efficient detection of urban functional zones has been an essential issue in recent studies.

The challenge of this work is capturing high-level latent semantic concepts that reflect human activities other than physical characteristics (such as reflectivity and texture). The majority of scientists take advantage of human mobility data such as taxi trajectories [7], mobile trace data [8,9] and check-in data on social media [10] to gather information about place and model urban structure. These studies perform well in understanding the human activities that occur in space and demonstrate the correlation between travel patterns and urban functions [6].

Meanwhile, public bicycle-sharing systems (BSS), aiming at providing the missing links (also called the last-mile connection) in public transportation systems and promoting short-distance green transportation [11], has a promising position among urban cities in China and a fundamental role in residents' daily travel [12]. A mixture of land use or urban functions could offer residents the opportunity to live, work, shop and enjoy recreation facilities within their community. Such mixed land use and urban functions lead to shorter trip length and a possible transportation mode shift from cars to bicycles [13]. Thus, compared to taxis or floating cars, the public bicycle-sharing system may provide richer information for understanding a complex urban structure. A public bicycle-sharing system also has broader station coverage than other public transit/commuting tools [14], making it an ideal choice for a seamless city-scale analysis.

Some place-related questions have been discussed in the literature by BSS researchers. Froehlich [15] tries to use the distance between station location and the city center to explain the interrelationship between place and the temporal usage pattern. The same considerations have appeared in Faghih-Imani's [14] bicycle flow prediction model. These studies focus more on explaining travel behavior by applying urban functions/land use information instead of using this short-trip human mobility for urban functional zone categorization. In Froehlich's following work [16], he pointed out that trends in station usage could be used to infer attributes about neighborhoods. However, he did not give a complete explanation on how to perform the prediction. O'Brien [17] applies a clustering method to bicycle stations to understand the demographics, but no local characteristic is addressed. Thus, we intend to perform urban functional zone identification using BSS data and develop an analysis framework using state-of-the-art topic modelling and machine learning techniques in human mobility studies.

In this paper, we propose to use public BSS data with POI information for the automatic identification of urban functional zones. The analysis framework is tailored for the station-based sample data and extends existing topic modelling approaches by adding a visual expression of land function word clouds for each functional zone. The function tags are retrieved from POI information to help the manual interpretation of land use and human behavior. Our model is implemented to sense the urban functional zones in Hangzhou City, China, which has the most extensive public bicycle system with substantial and stable daily usage. As the first attempt to integrate BSS station-based rental records with POI information to detect urban functional zones, we contribute a new perspective to use this new station-based data to depict urban structures and mixed land use patterns.

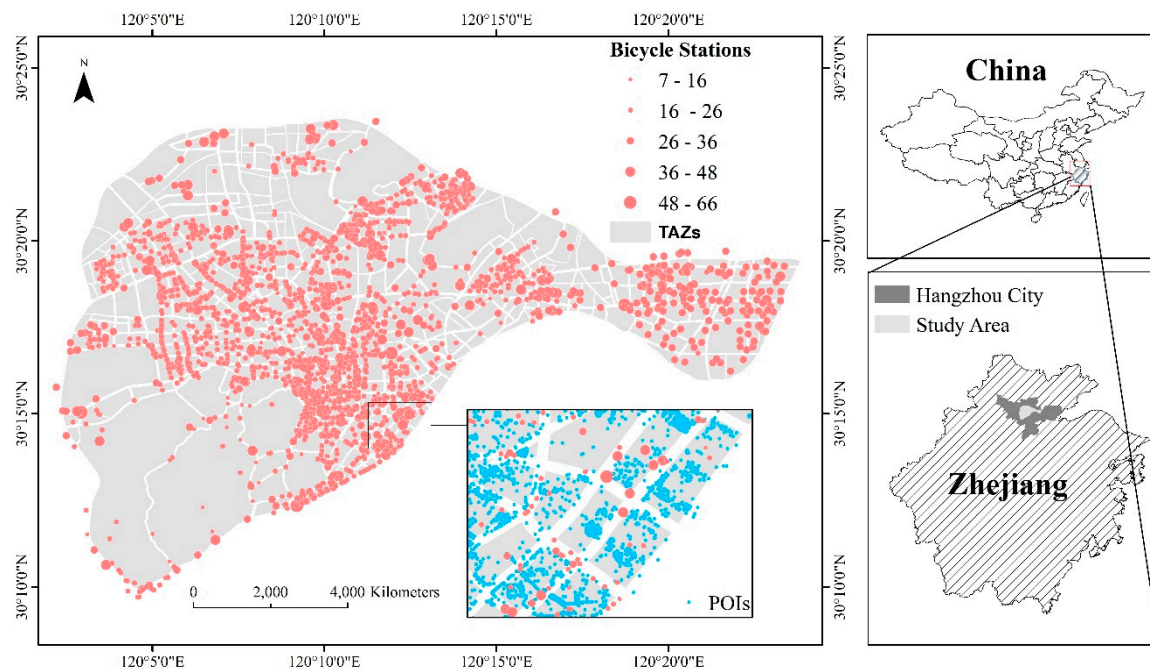
## 2. Study Area and Data

### 2.1. Study Area

Being the host city of the Group of 20 Summit in 2016 and the Asian Games in 2022, Hangzhou City attracts world attention. It is a first-tier city in China and is moving towards becoming a technology hub, contributing 1131.4 billion yuan in gross domestic product in 2016 [18] or a 1.52% of the whole country. The population of its urban area (Hangzhou City in Figure 1) has exceeded 5.44 million by 2016 [18].

The local government initiated the Hangzhou Public Bicycle System to foster seamless public transportation in 2008 [19]. The maximum daily usage so far has reached 473,000 times, showing that it has been a mainstay in citizens' regular travel choice in ten years' development.

Like other BSS around the world, except for Paris [20], the spatial distribution of bicycle stations was first limited to the central urban district and some high population density areas around major metro stations, residential areas, commercial centers, and tourist attractions and then gradually expanded to the peripheral regions. There were 50,000 bicycles in 2000 stations at the end of 2009, increasing to 85,573 bicycles in 3403 stations in November 2017 [21]. The local government is still investing in the public bicycle infrastructure to improve and enlarge the network coverage.



**Figure 1.** The spatial distribution of bicycle stations and collection of POIs in the study area (2017).

## 2.2. Data

The public bicycle-sharing system is classified as a third-generation BSS. It enables citizens to use smart cards and kiosks to check in and check out and to pick up bicycles at any bicycle station in 24 h. However, it lacks real-time control for each bicycle and dynamic access to other transport modes. The capacity of a station/hub ranges in size from 7 docks (temporal station or less dense area) to up to 66 docks (city center).

In our study, we sampled the rental records of all bicycle stations hourly from 2017-11-11 06:00 to 2017-11-21 05:00 and observed the number of rented and restored bicycles for each station. We do not use any personally identifiable information.

The POI data are provided by Gaode Map, also known as AutoNavi Maps, which is a Chinese web mapping, navigation, and location-based service provider. A typical Gaode POI record for Beijing Tiananmen Square is {"id": "B000A60DA1", "name": "Tiananmen Square", "type": "Tourist Attraction; Scenery Spot; National View Spot", "typecode": "110202", "biz\_type": "tour", "address": "East Chang'an Road", "location": "116.397477, 39.908692", ...}. Gaode POI data have a three-level category system (see Figure 2) with 23, 264, and 871 POI types, respectively. The first two digits of *typecode* represent the first-level category of the *type* field, the following two digits represent the second level category of the *type* field, and the last two digits represent the third level category of the *type* field. A higher-level category has more detailed *type* information. In our study, we focus on the second level *type* field of POI data.

```

11----- Tourist Attraction
  1101----- Park & Square
  ...
  1102----- Scenery Spot
    110202----- National View Spot
    ...

```

**Figure 2.** The three-level category system of Gaode POI.

By sending Hypertext Markup Language (HTML) requests to the POI Search Application Program Interface (API) with a polygon parameter, we obtain all POIs within the polygon. For each query,

the API returns POIs within the polygon. We generated fishnets over the incorporated study area and use each of them as a query polygon. A total of 292,885 POIs were obtained and drawn in Figure 1.

We use Open Street Map (OSM) to assist with our studies [22]. Its Planet OSM project provides the city boundary data and road network data. We select the primary, secondary and tertiary roads as the main road network of Hangzhou City, and construct our transportation analysis zone (TAZ) as the land unit through the method proposed by Yuan [23] and Liu [24]. The strategy is described as follows:

Step 1: Use buffering operation to delineate the road space from the road network. The buffer distance in the TAZ spatial join operation is chosen by the Pareto principle of using the average nearest neighbor station distance. The 80% service distance is between 240 meters to 250 meters, which is slightly larger than the allocation distance of 200 meters published by the local transportation agency.

Step 2: Edit the disconnected areas, which are caused by dangle roads in which the endpoint of the road lines does not touch another road line.

Step 3: Merge all administrative boundaries into one polygon and remove the road space from that polygon.

We obtained 503 TAZs, covering approximately 13% of the eight districts of Hangzhou City (Figure 1).

### 3. Methodology

Text mining technologies, especially topic modelling are introduced to extract high-level latent semantic information; to mine the land functions and gain human understanding of city knowledge [10,25]. Latent Dirichlet allocation (LDA) is the most popular generative topic model for collections of discrete data [23,25,26] and has been widely applied in the urban functional zone identification process. Different mobility metrics are built and assembled as word/term to capture semantic topics. For example, Yuan illustrates a transition cuboid [1] to aggregate taxi pick-up/drop-off behavior. He also explores a more advanced topic model, Dirichlet Multinomial Regression (DMR) [27], to couple taxi pick-up/drop-off behavior with POI information. Gao [28] introduces the visiting frequency for different Foursquare venues to derive urban functional regions. A more relevant research paper comes from Bao [26] that uses numbers of BSS check-in events and a combination of temporal characteristics and station type to form a semantic word. However, we cannot directly transform and convert their models to solve our problem because the bicycle rental data are station sampled instead of trip information, which means neither check-in frequency nor origin/destination stations can be used for a topic model. Thus, we introduce the Normalized Available Bicycles (NAB) index in our topic modelling process. This index has been applied in Froehlich [16] and O'Brien [17]'s work as a sign of the usage status for a BSS station.

The analysis framework we propose for the identification of urban functional zones from bicycle rental records and POI data have two steps (see Figure 3): (1) topic modelling and (2) semantic annotation. The topic model extracts semantic topics from terms formed by station character and takes POI information as an enriched local character. Function matrices representing the distribution of different semantic topics for each TAZ is generated. Next, a clustering method is applied to delineate the functional zones, aggregating TAZs with similar topic distribution into a zone. Additionally, word clouds for each zone are drawn to help identify urban functions. Please note that the function types are not defined beforehand but are identified by our clustering and word cloud results.

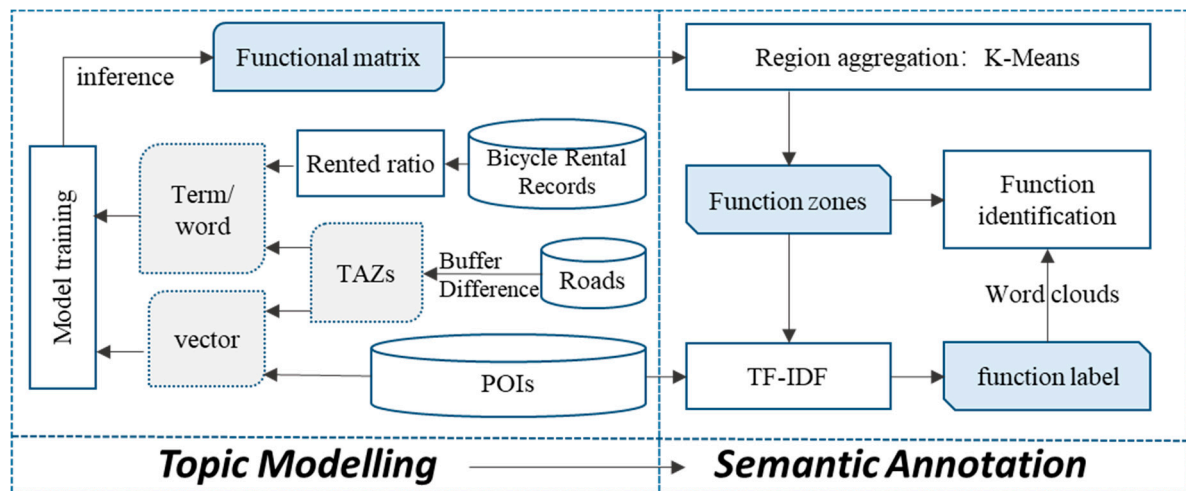


Figure 3. The analysis framework.

### 3.1. Topic Modelling

#### 3.1.1. Term Formulation

The mobility metric we applied for each TAZ is an extension of the Normalized Available Bicycles. There is observation data for a bicycle station every hour that reports the number of rented bicycles and restored bicycles. First, all stations within one TAZ are selected from the spatial join result of the bicycle station layer with the TAZ layer. Then, Equation (1) is used to integrate each station's NAB value [20] of a given hour  $t$  and calculate the rental ratio for  $i$ -th TAZ. The total bicycle number is the sum of rented bicycles and restored bicycles.

$$\text{rental\_ratio}_{i,t} = \frac{\sum_{\text{stations}} \text{NAB}_t \cdot \text{bikenum}}{\sum_{\text{stations}} \text{total}_t} \quad \forall \text{stations} \in i \quad (1)$$

Next, the Natural Breaks classification method transforms the rental ratio to a five-level categorized value: *empty*, *low*, *moderate*, *high*, and *full*, representing the different rental status of a TAZ. The Natural Breaks, also known as Jenks, can identify groups with similar values and maximizes the differences between groups [29,30]. Finally, appending the status code with temporal characters of day and hour, we construct the term/word depicting the TAZ. For example, TAZ #1 may have one term of "Mon08empty", which means the rented bicycles within the region on Monday at 8 a.m. is nearly zero.

#### 3.1.2. Vector Calculation

Except for the term formulation and mobility metrics choice in the identification work, it is tricky to give a function definition when researchers use these human mobility data because (1) they are high in volume and full of noise; and (2) different types of media give different semantic information. Auxiliary information such as POI data [23,31,32] or prior knowledge [33,34] are often introduced to help define the land function after the clustering process. POI shows basic information related to specific coordination, i.e., address, name, and category, which has the potential to unveil the urban land use characteristics [23,35].

In our paper, we use the first-level categories of POI data and calculate all POIs within a TAZ (with a buffer distance of  $0.001^\circ$ , equal to approximately 100 meters) and then construct a meta vector



$x = (v_1, v_2, \dots, v_n)$ , where  $n$  denotes the number of POI categories and  $v_i$  is the frequency of the  $i$ -th POI category in a TAZ and can be calculated by Equation (2).

$$v_i = \left\lfloor \frac{\text{number of POIs of } i\text{-th category}}{\text{sum of POIs}} \right\rfloor \quad (2)$$

### 3.1.3. Model Training and Inference

The fundamental assumption of our approach is that the bicycle rental characteristic, represented through an observable indicator, is related to a place's land use and social functions [6]. In turn, we can yield the land function matrix through a latent topic detection process. Additionally, we hypothesize that regions with similar POI distribution have a more significant possibility to share similar bicycle rental characteristics. Thus, the Dirichlet Multinomial Regression model is chosen because it takes metadata into the generative process while maintaining efficient and robust performance.

Accordingly, we define an analogy to explore the urban functions using the DMR model. We take the bicycle rental status of a particular time of a land unit as a word/term and a continuous ten-day hourly rental status series of a TAZ as a document. The corpus is comprised of such documents of all TAZs. Additionally, the POI information within a land unit is regarded as the metadata. The data generating process of the land functional matrix can be summarized as:

1. For each latent land function type  $f \in \{1, \dots, F\}$ ,
  - a. Draw  $\lambda_f \sim \mathcal{N}(\mu, \delta^2)$ .
  - b. Draw  $\varphi_f \sim \text{Dir}(\beta)$ .
2. For each TAZ  $i \in \{1, \dots, I\}$ ,
  - a. For each latent land function type  $f$ , let  $\alpha_{i,f} = \exp(x_i^T \lambda_f)$ , which is where the assumption of POI distribution may affect land use and land function works.
  - b. Draw  $\theta_i \sim \text{Dir}(\alpha_i)$ .
3. For the  $t$ -th hour in the  $i$ -th land unit,
  - a. Draw  $z_{i,t} \sim \text{Mult}(\theta_i)$ .
  - b. Draw  $w_{i,t} \sim \text{Mult}(\varphi_{z_{i,t}})$ .

Where  $\lambda_f$  is the  $n$ -length, normally distributed random vector. The mean value is  $\mu$  and the standard deviation is  $\delta$ .  $n$  is decided by the number of POI categories.  $\varphi_f$  represents the mixture weights of hourly rental statuses.  $\text{Dir}$  denotes the Dirichlet distribution with  $\beta$  as a hyper parameter.  $x_i^T$  encodes the  $n$ -length POI vector depicted in Section 3.1.2, and  $\theta_i$  represents a  $f$ -length vector representing the mixture weights for each land function.  $z_{i,t}$  is the latent land function type. From  $\varphi_{z_{i,t}}$ , we finally get the terms  $w_{i,t}$  formed in Section 3.1.1.

## 3.2. Semantic Annotation

### 3.2.1. Region Aggregation

After deriving the latent thematic topics by running the DMR model, each region can be represented as a vector of the  $T$ -dimensional topics. Regions that are similar in the topic space share the same urban function and can be aggregated into the same cluster as a functional zone. K-means is an unsupervised clustering approach that is widely used in aggregating functional zones and activity pattern [23,26,28,36]. In Zhan's [36] comparative studies, after testing the standard hard partitioning methods, the fuzzy partitioning methods, and some more advanced clustering algorithms, he concludes that the K-means algorithm demonstrated the best performance in the land use inference.

During the region aggregation process, the K-means clustering method is implemented and spatial units are assigned to  $k$  clusters where the number  $k$  needs to be predefined. We use cluster validation indices to measure if a structure found with the cluster analysis is adequate and how well an object is appropriately clustered. Among the current indices, the silhouette criterion [37–39] is the most widely used index for determining an appropriate value of a cluster number. The range of the silhouette value is between 0 and 1. A high value (close to 1) indicates that an object is appropriately clustered and is highly unique from other clusters.

### 3.2.2. Word Cloud: TF-IDF

Word clouds have been widely used in the text mining domain to provide a content overview [31,40]. In our proposed approach, we generate the word cloud to give a compact visual form of land functions of different clusters. Moreover, a group of colors is predefined to distinguish different urban functions.

Land function labels are extracted and broken down from POI category information. We introduce a series of filters to remove special characters (i.e., “/” and “&”) and stop words. The stop words include not only traditional stop words for common text mining tasks (i.e., “it”, “and” and “a”) but also some high-frequency spatial referenced words (i.e., “related”, “building” and “places”). Then, we calculate the term frequency-inverse document frequency (TF-IDF) value of each term as the font size, as in Equation (3). Term frequency (TF) counts the frequency of category  $t$  in the label collection of a cluster  $i$ . The inverse document frequency (IDF) is the probability that term  $t$  will occur in a given document  $d$  in the corpus and can be calculated as Equation (4).  $N$  is the total documents of the corpus.

$$tf - idf(t, d, D) = tf(t, d) \times idf(t, D) \quad (3)$$

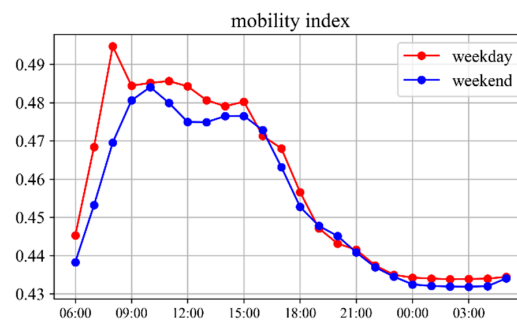
$$idf(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|} \quad (4)$$

## 4. Results and Analysis

### 4.1. Exploratory Analysis

To have a first impression of the data, we present the mobility index to highlight the temporal demand pattern in Figure 4. The curve lines of weekdays and weekends both have a higher value in the daytime and a lower demand at night. Early on weekday mornings, there is a high rented bicycle ratio at approximately 8 a.m. A local minimum occurs at 1 p.m. to 2 p.m. as people have lunch and take a break, and then the curve goes up slightly approximately 3 p.m. and continues to fall off afterward. A possible explanation may be that people tend to use the bicycles to commute to the office in the morning instead of going home, since the time limitation is much looser at night and allows for more travel choices such as walking. If so, a bicycle re-balancing task shall be done every day to ensure there are enough bicycles near residents' home.

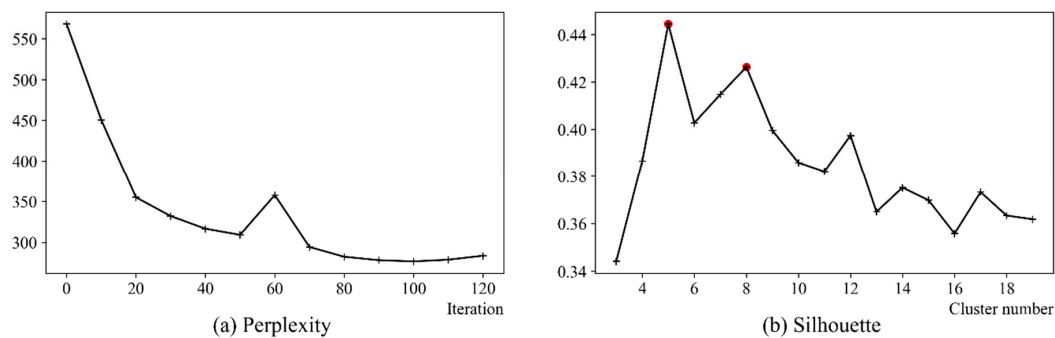
In contrast to the weekday activity, on the weekends there is no sign of the 8 a.m. commute. Instead, the morning peak hour shifts from 8 a.m. to 10 a.m., indicating that the function activity transforms from commute into recreation and daily activities. The trend coincides with the previous daily peak pattern of Vogel and Froehlich [15,41]. Moreover, the bicycle demand on weekdays is relatively higher than on weekends during the daytime, showing that the public bicycle system in Hangzhou City functions more as a commuter transportation system.



**Figure 4.** The daily temporal demand pattern between weekdays (red) and weekends (blue).

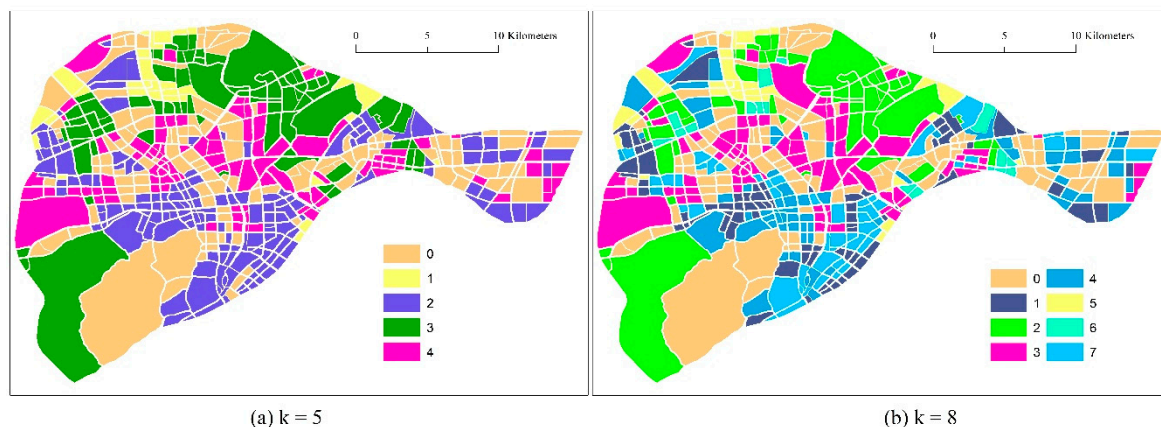
#### 4.2. Clustering Results

Considering the number of urban functional types in the official land use map, we set the topic number as 8. Figure 5a shows the perplexity during our topic modelling process, and it achieves convergence after 100 iterations.



**Figure 5.** The change of perplexity value in the topic modelling process (a) and the choice of different cluster numbers in the clustering process (b).

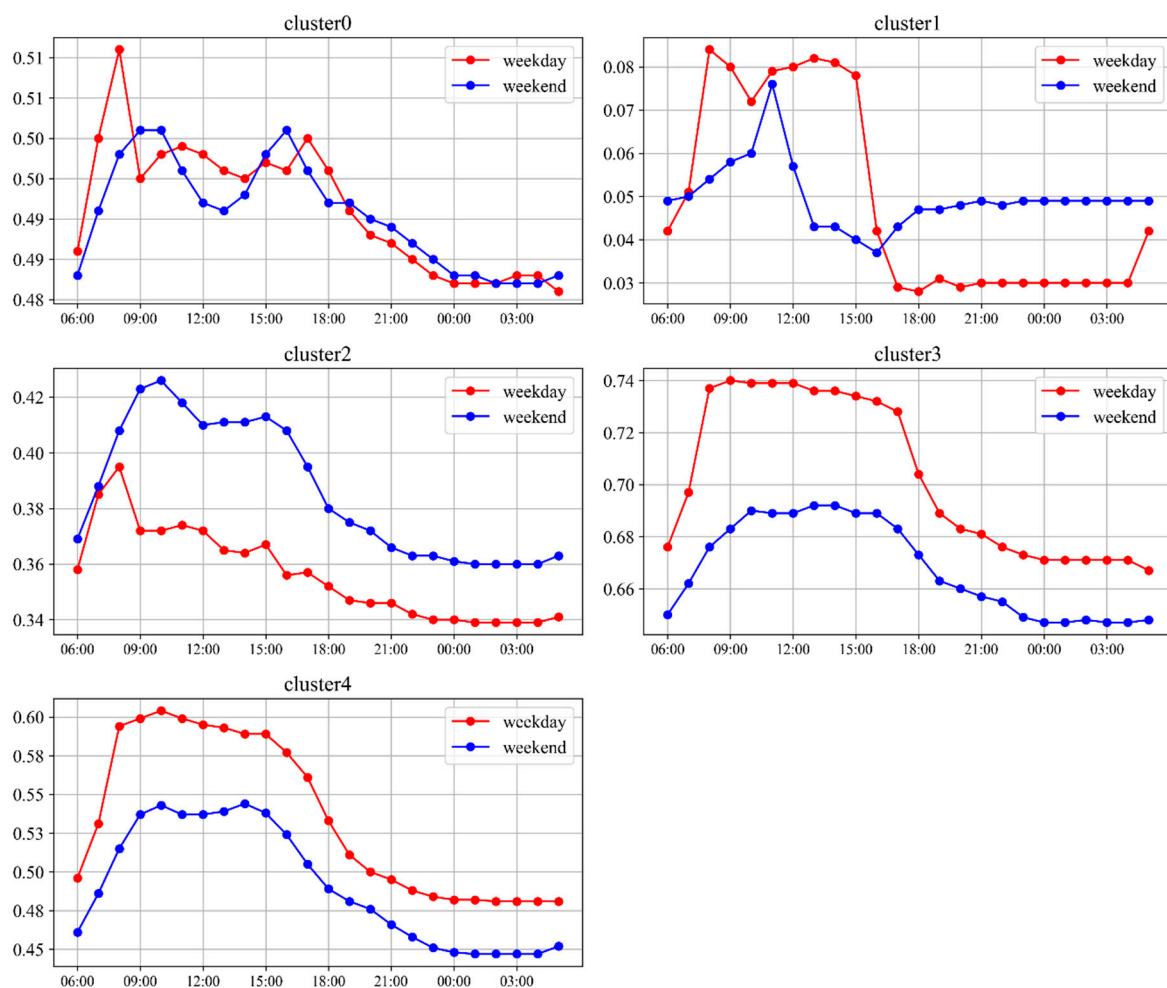
In the clustering process, we tried different  $k$  values ranging from 3 to 19 and calculated the silhouette value in Figure 5b. The silhouette value first gets a high score when the cluster number is 5, then sharply decreases and get a second local maximum at 8. Thus, we chose 5 and 8 to be the candidate cluster number, and the corresponding cluster results are drawn in Figure 6. We find that the results are quite similar. Some categories on the left ( $k = 5$ ) are separated into sub-categories on the right ( $k = 8$ ): cluster 2 ( $k = 5$ ) is separated into clusters 1, 4 and 7 ( $k = 8$ ); cluster 3 ( $k = 5$ ) is separated into clusters 2 and 6 ( $k = 8$ ). The others remained the same, i.e., cluster 0 ( $k = 5$ ) and cluster 0 ( $k = 8$ ), cluster 1 ( $k = 5$ ) and cluster 5 ( $k = 8$ ), and cluster 4 ( $k = 5$ ) and cluster 3 ( $k = 8$ ).



**Figure 6.** The clustered results when  $k = 5$  (a) and  $k = 8$  (b).



We draw the mobility index of each cluster in Figure 7. The definition of x-axis and y-axis are the same as in Figure 4. However, the temporal tendencies are sharply different with the drift of overall stations in the study area, although the morning peak hour still occurs at 8 a.m. in the weekdays and shifts from 8 a.m. to 10 a.m. on weekends. It should be noted that the dependent values, the range of rent ratio, are distinct in each cluster, and cluster 1 has an almost zero rent ratio. For other clusters, cluster 0 is more like a standard commuting curve, cluster 2 has a morning peak, and clusters 3 and 4 both show a flat daytime curve. Explanations of such patterns will be discussed in the next section.



**Figure 7.** The daily temporal demand pattern of different clusters when  $k = 5$ .

#### 4.3. Themed Functions

In the word cloud we removed the POI data with a first-level type of transport services (*typecode* starts with “15”), road furniture (*typecode* starts with “18”), place name (*typecode* starts with “19”), pass facilities (*typecode* starts with “99”) and indoor facilities (*typecode* starts with “97”) and kept the rest, which are more related to land use. The references for removing and re-organizing the POI categories for detecting land use can be found in [10,26,32]. We produced the word cloud using the mid category system of Gaode and extracted the semantic characteristic of the clustering results.

At the beginning of the word cloud generation process, every cluster gave some high-weighted words, although we have added the traditional words and some spatially referenced words as stop words (in Table 1), which created a barrier for us to extract the themed function type. A similar situation occurred in Gao’s [28] research in distinguishing different functional zoning, where coffee shops widely exist in most regions. Therefore, we filtered the influence of these general features within

every cluster and re-calculated the TF-IDF value. These non-distinctive words are also listed in Table 1 for reference.

**Table 1.** Stop words in our study.

Category	Contents
traditional words	and, related, parts, comprehensive
spatially referenced words	building, house, center, place, places, area, name
non-distinctive words	store, food, restaurant, Chinese, service, company, organization, institution, public

The word clouds are rendered by different theme color. Orange is mapped to daily life label sets, including residential, convenience store, hairdressing, pharmacy, and beauty. Yellow is used to depict recreation and leisure activities such as bath and massage. Blue is related to public services such as schools, hospitals, public toilets, and governmental agencies. Dark blue is mapped to business and commercial activities such as the electronic hypermarket, house materials market, automobile, clothing, insurance, and finance. Finally, green is used to depict attractions and travel facilities such as scenic areas, attractions, hotels, and hostels. By looking at the color composition and high-weighted keywords, researchers have an impression of the urban functional zone clusters.

Figure 8 gives the semantic annotation results when the cluster number is 5. We have two residential clusters (cluster 0 and cluster 4) with a batch of daily life labels. Apart from some general daily life labels, typical Hangzhou-style (Leisure City) labels, such as *massage* and *bath* are highlighted in the word clouds. Compared to cluster 0, the word cloud of cluster 4 shows some business function labels, which may be why temporal curve does not behave like a standard commuting curve with a dip at 11 a.m. to 2 p.m. Cluster 1, with a combination of residential and industrial labels and a low rent ratio, belongs to the suburban area and urban fringe. Cluster 2, providing public services, leisure actives as well as attractions and travel facilities, is the only one with a higher rent ratio on weekends than weekdays. The dominant blue color in the word cloud of cluster 3 indicates primary industrial land use of land. Although the word clouds and temporal curves of clusters 3 and 4 are similar, the loss of residential labels in the cluster 3 results in a flatter daytime curve.

The definitions above are based on the primary function of each cluster even it may have other minor functions. For example, cluster 2 is mixed with residential labels. When we further draw word clouds for the clustering result of  $k = 8$ , Figure 9 clearly shows that the function label changes. Cluster 2 ( $k = 5$ ) extracts its business role to clusters 1 and 4 ( $k = 8$ ) and its daily life and recreation activities to cluster 7 ( $k = 8$ ). At the same time, the temporal curve in Figure 10 changes synchronically. Clusters 1 and 4 inherit the nature of the business function and are more active on weekends, while cluster 7 behaves a commuting curve that is more seen in the residential area. Although the same labels, i.e., governmental, school and hotel appear in both clusters 1 and 4, the former has distinctive labels such as enterprises, finance, and insurance, addressing financial activities more, while the latter has labels such as scenery, spot and foreign, indicating tourism services.

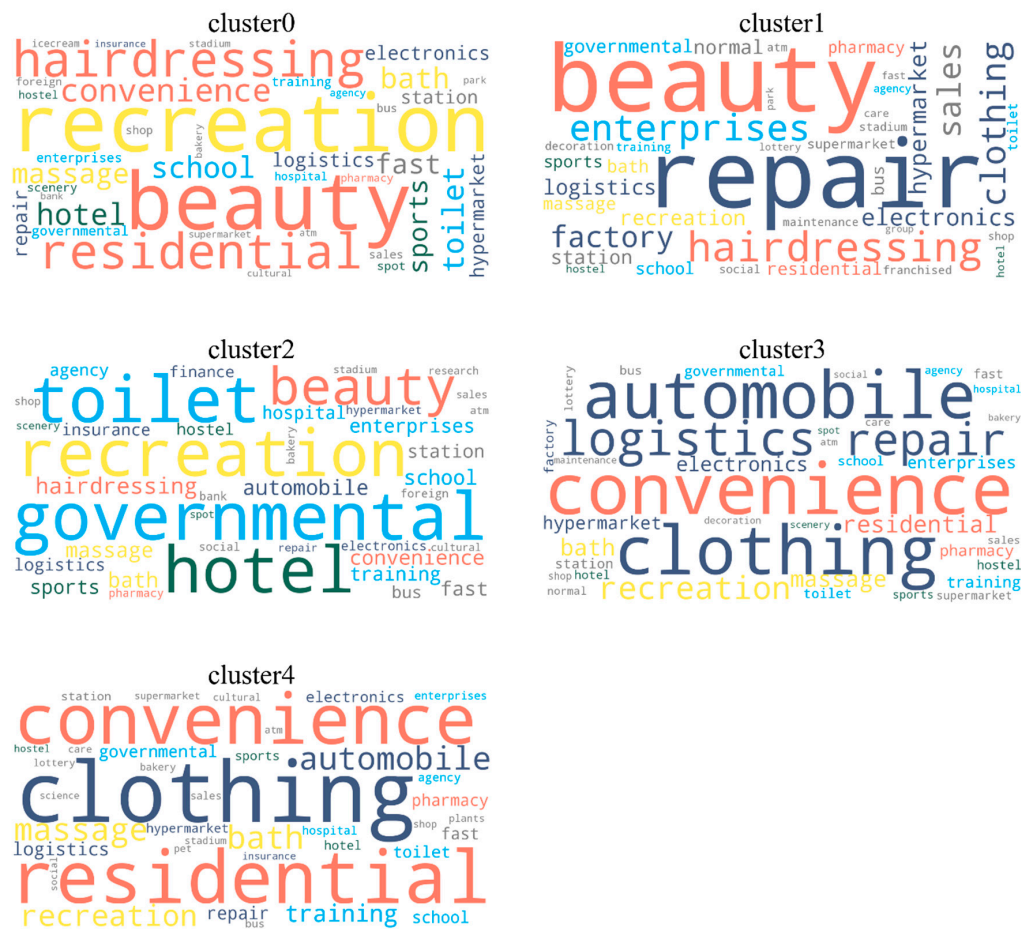


Figure 8. The semantic annotation result of different clusters when  $k = 5$ .



Figure 9. The semantic annotation result of different clusters when  $k = 8$ .

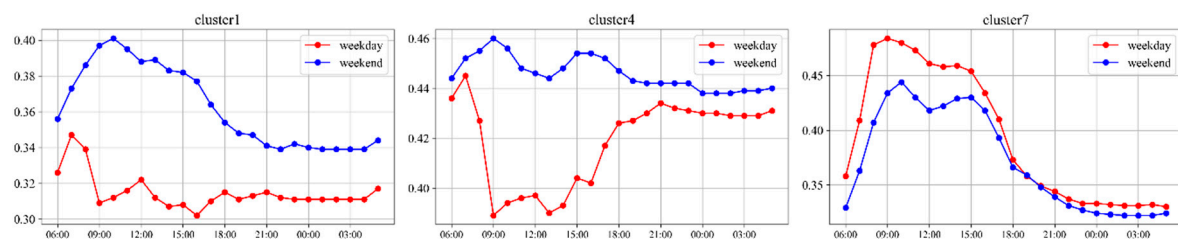
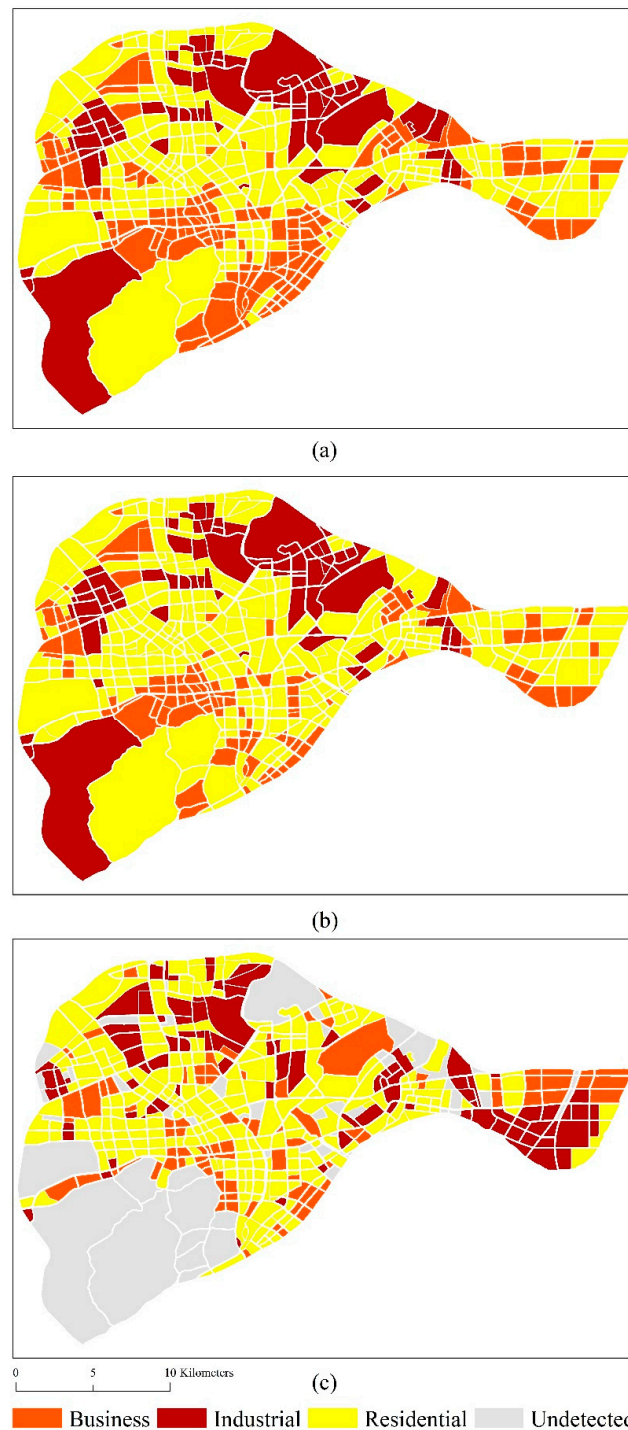


Figure 10. The daily temporal demand pattern of different clusters when  $k = 8$ .

#### 4.4. Consistency with the Urban Master Plan

To understand how well the clustering results match with urban land use, we measured the consistency by the percentage of overlapping that exists between our results with the 2016 Urban Master Plan (<http://www.hzplanning.gov.cn/index.aspx?tabid=641facd5-9004-46a0-91b2-7936584281fe>). We manually checked every TAZ using both official Urban Master Plan and Google Map to obtain up-to-date urban land use data.

Since the semantic annotation is a little different from the category in the 2016 Urban Master Plan, we use a general mapping scheme (Table 2) to test the accuracy of our clustered result. The comparison is drawn in Figure 11. Educational land, business area, and public facilities are combined to compare with cluster 2 ( $k = 5$ ) and clusters 1 and 4 ( $k = 8$ ). They all provide different public services for citizens. Industrial land is compared to cluster 3 ( $k = 5$ ) and clusters 2 and 6 ( $k = 8$ ). Additionally, residential areas are compared to clusters 0, 1 and 4 ( $k = 5$ ) and clusters 0, 3, 5 and 7 ( $k = 8$ ). The others in the 2016 Master Plan are not calculated in our experiment because the semantic annotation did not detect any of them.



**Figure 11.** Comparison between the clustering results of (a)  $k = 5$ , (b)  $k = 8$  and (c) 2016 Urban Master Plan.

**Table 2.** The mapping between 2016 Master Plan data and cluster.

General	2016 Master Plan	Cluster $k = 5$	Cluster $k = 8$
Business	Educational Land\Business Area\Public facilities	2	1,4
Industrial	Industrial Land	3	2,6
Residential	Residential Area	0,1,4	0,3,5,7
Undetected	Green-land\Rural area\Open space	-	-

Each element in Table 3 represents a specific land use general type in the 2016 Urban Master Plan that is covered by one of our land use clusters, i.e., Industrial, Business and Residential. The last columns show the overall accuracy of our land use detection. Comparing our results across different types, we observe that residential land shows the highest percentage (59.4% when  $k = 5$  and 66.8% when  $k = 8$ ), which is comparable to Martinez's [33] study using Twitter with an accuracy of 64.14% and Toole's [8] research using mobile data with an accuracy of 54%. However, green land, rural area and open space have not been detected. We believe that the main reason for that is these land use types are inferior to other types concerning POI, which will be discussed in detail later.

**Table 3.** The coverage ratio between 2016 Master Plan and cluster.

2016 Master Plan	Business	Industrial	Residential	Undetected	Overall
$k = 5$	34.3%	23.7%	59.4%	0%	39.1%
$k = 8$	28.8%	15.1%	66.8%	0%	36.9%

## 5. Discussion

### 5.1. The Role of POI Data

To show the impact of the DMR model when incorporating the POI information during the functional matrix generation process, we use the original LDA model as a comparative exploration. We use a similarity measurement between the 2016 Master Plan and two clustering results to evaluate the clustering performance.

Specifically, we use an information theoretic index called V-measure (Andrew and Hirschberg, 2007; Niesterowicz et al., 2016). Two desirable objectives define it for any cluster assignment. By making a pairwise comparison, we know the extent to which each cluster in A contains only members of a single class in B (homogeneity) and whether all members of a given cluster in A are assigned to the same cluster in B (completeness). All these measures should be bounded in [0,1]. The closer to 1 they are, the better the clustering performance is.

From Table 4 we concluded that DMR modelling has better performance in both homogeneity and completeness indices, resulting in a better V-score. It shows the great benefit of incorporating the POI information into our functional matrix generation process.

**Table 4.** The comparison of POI included model (DMR) and excluded model (LDA).

	Homogeneity	Completeness	V-Score
DMR	0.040	0.031	0.035
LDA (bicycle data only)	0.021	0.019	0.020

### 5.2. Limitations and Possible Solutions

In the following section, we discuss in detail the possible reasons that cause inconsistency between the urban master plan and our urban functional zones. We conclude three typical errors and explain possible reasons.



- **Mixed land function.** Some land functions may inferior to other land functions. For example, middle schools and elementary schools may be surrounded by residential communities within a TAZ. Restaurants and stores are also part of the space. However, we may interpret this region as a residential area instead of education land or business land. Using a finer division (i.e., block level or building level) may help with the problem.
- **Semantic difference.** The semantic annotation process has a significant influence on the functional zone cluster results. TAZ's function varies when considering different perspectives. For example, the mountain around West Lake includes biological services (green land), tourism services (business area) and an office building for high-tech companies (industrial land). This error is different from the above because it cannot be solved by using a smaller unit. This kind of mixed function will increase with the development of the complexity of the urban system. We believe that this is quite useful for future urban planners and city managers to understand real multi-functionality.
- **Inefficiency in open space/rural area/green-land identification.** This kind of defect is caused by our data-driven method. Bicycle rental activities occur less in those areas. Additionally, the POI information comes from the commercial database assigning more weight to human commercial activities. The solution to this error is to introduce multi-source data or remote sensing imagery as the input of our functional matrix generation process.

## 6. Conclusions

In this paper, we apply topic modelling techniques to extract a city functional matrix from bicycle rental records and POI information of Hangzhou City. We use the K-means clustering method with a word cloud visualization expression to delineate the urban functional zones. Business areas, industrial areas, and residential areas are well detected. The word cloud of function labels reveals the mixed land use of different types of urban functions. It also illustrates how diverse urban functions split into different TAZ clusters when we use a larger cluster number and define urban function zones at a more detailed level.

Our work contributes in two ways. First, an analysis framework designed for station-based public rental data are proposed. The effectiveness of term formulation via mobility metrics of bicycle rental records and POI vector can enlarge our data source of human mobility in urban studies. Second, our visual expression of high-weighted labels provides a new way to help annotate the semantic function of clustering results other than looking its geographical distribution over the city map. Additionally, it improves the understanding of the mixed land functions under a well-developed urban system. We expect that our visualization method of word cloud of land functions can also be applied to other sources of similar urban digital traces, such as cellular and floating cars, to help explain the complexity of the urban structure.

Our next research question may involve the relationship between multi-modal transportation data and urban functions. Other data sources, such as Very High Resolution (VHR) images and social media data should also be incorporated to solve the limitations in finding open space, rural area, and green land and to further improve the identification efficiency. Another problem is the complexity of semantic expression and the mixed land use representation, and we recommend a function matrix with a possibility distribution of different function activities instead of a primary function definition to be used in some modelling work such as the prediction of bicycle usage.

**Author Contributions:** Conceptualization, X.Z. and Z.D.; Data curation, F.Z.; Formal analysis, X.Z.; Funding acquisition, R.L.; Methodology, X.Z. and W.L.; Project administration, Z.D.; Resources, F.Z. and R.L.; Supervision, R.L.; Validation, X.Z., W.L. and Z.D.; Writing—original draft, X.Z.; Writing—review & editing, W.L. and Z.D.

**Funding:** This research was funded by the National key Research and Development Program of China, grant number 2018YFB0505000.

**Acknowledgments:** The authors would like to thank the four anonymous reviewers of IJGI for their constructive comments that better shaped the paper. Xiaoyi Zhang would like to thank Zhejiang University for providing the scholarship for the visiting Cyber Infrastructure and Computational Intelligence (CICI) Lab in Arizona State University, and is grateful for the inputs to the early stage of this research received from the CICI group members.

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

## References

1. Yuan, N.J.; Zheng, Y.; Xie, X.; Wang, Y.; Zheng, K.; Xiong, H. Discovering Urban Functional Zones Using Latent Activity Trajectories. *IEEE Trans. Knowl. Data Eng.* **2015**, *27*, 712–725. [\[CrossRef\]](#)
2. Stead, D.; Marshall, S. The Relationships between Urban Form and Travel Patterns. An International Review and Evaluation. *Eur. J. Transp. Infrastruct. Res.* **2001**, *1*, 113–141.
3. Zhang, M. The Role of Land Use in Travel Mode Choice: Evidence from Boston and Hong Kong. *J. Am. Plan. Assoc.* **2004**, *70*, 344–360. [\[CrossRef\]](#)
4. Crooks, A.; Pfoser, D.; Jenkins, A.; Croitoru, A.; Stefanidis, A.; Smith, D.; Karagiorgou, S.; Efentakis, A.; Lamprianidis, G. Crowdsourcing urban form and function. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 720–741. [\[CrossRef\]](#)
5. Gao, S.; Liu, Y.; Wang, Y.; Ma, X. Discovering Spatial Interaction Communities from Mobile Phone Data. *Trans. GIS* **2013**, *17*, 463–481. [\[CrossRef\]](#)
6. Liu, Y.; Liu, X.; Gao, S.; Gong, L.; Kang, C.; Zhi, Y.; Chi, G.; Shi, L. Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. *Ann. Assoc. Am. Geogr.* **2015**, *105*, 512–530. [\[CrossRef\]](#)
7. Liu, X.; Gong, L.; Gong, Y.; Liu, Y. Revealing travel patterns and city structure with taxi trip data. *J. Transp. Geogr.* **2015**, *43*, 78–90. [\[CrossRef\]](#)
8. Toole, J.L.; Ulm, M.; González, M.C.; Bauer, D. Inferring land use from mobile phone activity. In Proceedings of the ACM SIGKDD International Workshop on Urban Computing, Beijing, China, 12–16 August 2012.
9. Pei, T.; Sobolevsky, S.; Ratti, C.; Shaw, S.; Li, T.; Zhou, C. A new insight into land use classification based on aggregated mobile phone data. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 1988–2007. [\[CrossRef\]](#)
10. Zhou, X.; Zhang, L. Crowdsourcing functions of the living city from Twitter and Foursquare data. *Cartogr. Geogr. Inf. Sci.* **2016**, *5*, 393–404. [\[CrossRef\]](#)
11. Shaheen, S.; Guzman, S.; Zhang, H. Bikesharing in Europe, the Americas, and Asia: Past, present, and future. *Transp. Res. Rec.* **2010**, *2143*, 159–167. [\[CrossRef\]](#)
12. Zhao, J.; Deng, W.; Song, Y. Ridership and effectiveness of bikesharing: The effects of urban features and system characteristics on daily use and turnover rate of public bikes in China. *Transp. Policy* **2014**, *35*, 253–264. [\[CrossRef\]](#)
13. Dieleman, F.M.; Dijst, M.; Burghouwt, G. Urban Form and Travel Behaviour: Micro-level Household Attributes and Residential Context. *Urban Stud.* **2016**, *39*, 507–527. [\[CrossRef\]](#)
14. Faghih-Imani, A.; Anowar, S.; Miller, E.J.; Eluru, N. Hail a cab or ride a bike? A travel time comparison of taxi and bicycle-sharing systems in New York City. *Transp. Res. Part A Policy Pract.* **2017**, *101*, 11–21. [\[CrossRef\]](#)
15. Froehlich, J.; Neumann, J.; Oliver, N. Measuring the Pulse of the City through Shared Bicycle Programs. In Proceedings of the International Workshop on Urban, Community, and Social Applications of Networked Sensing Systems, Raleigh, NC, USA, 5–7 November 2008.
16. Froehlich, J.; Neumann, J.; Oliver, N. Sensing and Predicting the Pulse of the City through Shared Bicycling. In Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI), Pasadena, CA, USA, 11–17 July 2009; pp. 1420–1426.
17. Brien, O.; Cheshire, J.; Batty, M. Mining bicycle sharing data for generating insights into sustainable transport systems. *J. Transp. Geogr.* **2014**, *34*, 262–273.
18. Zhejiang Provincial Bureau of Statistics. Chapter 1: General Survey. In *Zhejiang Statistical Yearbook 2017*; China Statistics Press: Hangzhou, China, 2018.
19. Shaheen, S.; Zhang, H.; Martin, E.; Guzman, S. China's Hangzhou Public Bicycle. *Transp. Res. Rec.* **2011**, *2247*, 33–41. [\[CrossRef\]](#)
20. García-Palomares, J.C.; Gutiérrez, J.; Latorre, M. Optimizing the location of stations in bike-sharing programs: A GIS approach. *Appl. Geogr.* **2012**, *35*, 235–246. [\[CrossRef\]](#)

21. Hangzhou Public Transport Corporation. Introduction of Bicycle Service in Hangzhou. Available online: [http://www.ggzxc.cn/about.aspx?c\\_kind=521&c\\_kind2=522&c\\_kind3=531](http://www.ggzxc.cn/about.aspx?c_kind=521&c_kind2=522&c_kind3=531) (accessed on 9 October 2018).
22. Fonte, C.C.; Martinho, N. Assessing the applicability of OpenStreetMap data to assist the validation of land use/land cover maps. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 2382–2400. [CrossRef]
23. Yuan, J.; Zheng, Y.; Xie, X. Discovering regions of different functions in a city using human mobility and POIs. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012; pp. 186–194.
24. Liu, X.; Long, Y. Automated identification and characterization of parcels with OpenStreetMap and points of interest. *Environ. Plan. B Plan. Des.* **2016**, *43*, 341–360. [CrossRef]
25. Adams, B.; McKenzie, G. Inferring Thematic Places from Spatially Referenced Natural Language Descriptions. In *Crowdsourcing Geographic Knowledge*, 1st ed.; Sui, D., Elwood, S., Goodchild, M., Eds.; Springer: Dordrecht, The Netherlands, 2013; pp. 201–221.
26. Bao, J.; Xu, C.; Liu, P.; Wang, W. Exploring Bikesharing Travel Patterns and Trip Purposes Using Smart Card Data and Online Point of Interests. *Netw. Spat. Econ.* **2017**, *17*, 1231–1253. [CrossRef]
27. Mimno, D.; McCallum, A. Topic Models Conditioned on Arbitrary Features with Dirichlet-Multinomial Regression. In Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence, Helsinki, Finland, 9–12 July 2008; pp. 411–418.
28. Gao, S.; Janowicz, K.; Couclelis, H. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Trans. GIS* **2017**, *21*, 446–467. [CrossRef]
29. Zhang, Y.; Brussel, M.J.G.; Thomas, T.; van Maarseveen, M.F.A. Mining bike-sharing travel behavior data: An investigation into trip chains and transition activities. *Comput. Environ. Urban Syst.* **2018**, *69*, 39–50. [CrossRef]
30. Jahanshahi, D.; Minaei, M.; Kharazmi, O.A.; Minaei, F. Evaluation and Relocating Bicycle Sharing Stations in Mashhad City using Multi-Criteria Analysis. *Int. J. Trans. Eng.* **2019**, *6*, 265–283.
31. Wang, Y.; Wang, T.; Tsou, M.; Li, H.; Jiang, W.; Guo, F. Mapping Dynamic Urban Land Use Patterns with Crowdsourced Geo-Tagged Social Media (Sina-Weibo) and Commercial Points of Interest Collections in Beijing, China. *Sustainability* **2016**, *8*, 1202. [CrossRef]
32. Li, G.; Xi, L.; Lun, W.; Yu, L. Inferring trip purposes and uncovering travel patterns from taxi trajectory data. *Cartogr. Geogr. Inf. Sci.* **2015**, *43*, 103–114.
33. Frias-Martinez, V.; Frias-Martinez, E. Spectral clustering for sensing urban land use using Twitter activity. *Eng. Appl. Artif. Intell.* **2014**, *35*, 237–245. [CrossRef]
34. Fan, K.; Zhang, D.; Wang, Y.; Zhao, S. Discovering Urban Social Functional Regions Using Taxi Trajectories. In Proceedings of the 2015 IEEE Conference on Ubiquitous Intelligence and Computing, Oslo, Norway, 23–25 June 2016; pp. 356–359.
35. Rodrigues, F.; Pereira, F.C.; Alves, A.; Jiang, S.; Ferreira, J. Automatic Classification of Points-of-Interest for Land-use Analysis. In Proceedings of the Fourth International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing), Valencia, Spain, 30 January–4 February 2012.
36. Zhan, X.; Ukkusuri, S.V.; Zhu, F. Inferring Urban Land Use Using Large-Scale Social Media Check-in Data. *Netw. Spat. Econ.* **2014**, *14*, 647–667. [CrossRef]
37. Abonyi, J.; Feil, B. *Cluster Analysis for Data Mining and System Identification*, 1st ed.; Springer Science & Business Media: Basel, Switzerland, 2007.
38. Jain, A.K.; Dubes, R.C. Algorithms for clustering data. *Technometrics* **1988**, *32*, 227–229.
39. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *1987*, 53–65. [CrossRef]
40. Kling, F.; Pozdnoukhov, A. When a city tells a story: Urban topic analysis. In Proceedings of the 20th International Conference on Advances in Geographic Information Systems, Redondo Beach, CA, USA, 6–9 November 2012.
41. Vogel, P.; Greiser, T.; Mattfeld, D.C. Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns. *Procedia Soc. Behav. Sci.* **2011**, *20*, 514–523. [CrossRef]

