

## Research Article

# Recognition of Functional Areas Based on Call Detail Records and Point of Interest Data

Guang Yuan <sup>1</sup>, Yanyan Chen <sup>1</sup>, Lishan Sun <sup>1</sup>, Jianhui Lai <sup>2</sup>, Tongfei Li <sup>1</sup>,  
and Zhuo Liu <sup>1</sup>

<sup>1</sup>Beijing Key Laboratory of Traffic Engineering, Beijing University of Technology, Beijing 100124, China

<sup>2</sup>Beijing Institute for Scientific and Engineering Computing, Beijing University of Technology, Beijing 100124, China

Correspondence should be addressed to Yanyan Chen; [cdyan@bjut.edu.cn](mailto:cdyan@bjut.edu.cn)

Received 12 July 2019; Revised 3 December 2019; Accepted 13 February 2020; Published 2 April 2020

Academic Editor: Giuseppe Musolino

Copyright © 2020 Guang Yuan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the recent emergence of big data, there has been significant progress in the study of big data mining and rapid developments in urban computing. With the integration of planning and management in urban areas, there is an urgent need to focus on the identification of urban functional areas (UFAs) based on big data. This paper describes the concept of communication activity intensity, which is more meaningful than the number of communication activities or the user density in identifying UFAs. The impact of diverse geographical area subdivisions on the accuracy of UFA recognition is discussed, and a  $k$ -means clustering method for dynamic call detail record data and kernel density estimation technique for static point of interest data are established at the traffic analysis zone level. A case study on the region within Beijing's 3rd Ring Road is conducted, and the results of UFA identification are qualitatively and quantitatively verified. The causes of large passenger flows on certain metro lines in Beijing are also analyzed. The highest identification accuracy is obtained for park and scenery areas, followed by residential areas and office areas. In conclusion, the proposed method offers a significant improvement over the identification accuracy of previous techniques, which verifies the reliability of the method.

## 1. Introduction

In the process of urban planning and management [1], the division of urban functional areas (UFAs) is a fundamental step. The distribution of UFAs is directly related to decision-making regarding urban transportation, resource management, and factory relocation [2]. As a city develops, the requirements for the integration of urban planning and management change, requiring some dynamic adjustment to the urban planning procedure. At the same time, as urban traffic congestion increases, it is important to alleviate this congestion to prevent an imbalance between urban traffic supply and demand caused by an unreasonable layout of urban functions. However, there is a certain deviation between the existing urban planning and real-world urban development. Therefore, the precise and timely identification of UFAs is urgently required. Furthermore, the identification of UFAs has positive significance for

policy formulation, resource allocation, transportation, and enterprise development [3]. Of course, it also has great significance for refining future traffic demand management.

Traditional urban land use classification is largely based on questionnaire surveys, which are time-consuming, labor-intensive, and nonexhaustive and do not reflect the structure of the city in real time [4]. However, some researchers believe that the arrival of the big data era signifies a change in our mode of thinking [5–7], and so the application of big data in planning is currently a hot topic of research [8]. There is also a recognition that constructing UFAs based on big data is essentially self-fulfilling.

In recent years, many studies have made full use of big data for urban land use classification or UFA detection [9, 10]. For example, the number of regional mobile phone calls has been used to represent the characteristics of urban functions [11], and points of interest (POIs) data have been collected to demonstrate the land use of an area [12, 13].

However, three challenging problems must be solved before mapping the functional area to very-high-resolution images [14], namely, the spatial units, features used for the analysis, and category criteria.

There have been many studies on UFAs using massive mobile phone data, including Call Detail Record (CDR) databases and Location-Based Service (LBS) databases. Several studies have also focused on the division and selection of Geographical Area Subdivisions (GASs) when using big data. Previous studies considering CDR volumes did not take the GAS size into account. Additionally, there has been a lack of data such as POIs, which include the attributes of land use at the application of CDR data, and little application of combined qualitative and quantitative methods in the verification of results.

This paper describes a set of data-driven methods for UFA identification. We consider the abovementioned factors comprehensively, including the influence of different GAS sizes, statistical indicators of CDR data, data sources containing land use features, and verification methods that are both qualitative and quantitative. The purpose of this study is to develop a practicable method for UFA identification, thus enabling reliable decision-making for urban planning and traffic planning and improving the utilization rate of existing big data applications in the engineering field.

Based on CDR data and POI data, the proposed approach makes the following contributions. First, their large-scale and long-period properties mean that CDR data can be used to record citizens' daily activities. A novel data-driven method of UFA recognition is proposed, and the intensity of daily activities, which depends on land use, makes a great contribution to identifying the function of a district within a city. Second, this study demonstrates that both calculation indicators and GASs must be considered before constructing the CDR model. The size of GASs is shown to have a significant impact on the results of numerical experiments, which further affect the indicators of CDR data such as the Number of Communication Activities (NCA), CAI, and the User Density (UD). Third, POI data overcome the shortcomings of CDR data in the analysis of land use characteristics. Combining POI data and CDR data can improve the accuracy of UFA identification. Finally, although many previous methods have employed qualitative verification, few have also adopted quantitative verification.

The structure of this paper is organized as follows. In the next section, relevant research related to UFA identification is reviewed. Novel region classification methods (GASs, CDR data model, and POI data model) and the required data sources are then presented in Section 3. Section 4 describes a case study of Beijing along with qualitative and quantitative verification and presents the results and a detailed discussion. Finally, Section 5 provides some conclusions and recommendations for future studies.

## 2. Literature Review

Research on land use classification and UFA recognition has been the subject of considerable effort in the field of geographic information [15–17]. However, satellite remote

sensing data and other traditional detection methods have some shortcomings, such as a long collection cycle, high cost, and poor representation of the difference between intrinsic functions. Several scholars [18, 19] have studied land use classification models, but their accuracy varies greatly depending on the input data. To overcome these limitations, mobile phone data have been used to explore the spatial structure of cities [20]. The results of spatiotemporal changes in urban activities based on mobile phone data can be displayed using thermodynamic diagrams [21]. This opened the way to a wide range of big data applications in urban computing [22–26]. For example, Croce et al. [27] used Floating Car Data (FCD) for zoning and graph building, while Alonso et al. [28] used a great quantity of observed traffic data to estimate the effects of traffic control regulation on the macroscopic fundamental diagram of the traffic network. Croce et al. [29] integrated transport models with big data on transport and energy in an attempt to design transport services with electrical vehicles.

CDR data use the auxiliary positioning function of the Global Positioning System (GPS) [30], allowing the analysis of crowd activities or human activity patterns. A literature review has investigated the use of mobile phone data to track travel behavior [31]. Population activities and human activity patterns are closely related to urban land use and UFAs [32], allowing urban functional types to be distinguished from the perspective of “humans” by CDR data. Of course, there has been much research on the application of big data for land use, for example, traffic data from loop sensors [28], Smart Card Data (SCD) [24], FCD data [27], and GPS data [33].

The employment space and commuting scope of the urban population in the suburbs of New York were analyzed by using CDR data in different periods [34]. Urban activities have also been analyzed dynamically in Monza and Brianza province, Italy, using the amount of mobile phone conversations, messages, and the number of mobile switching center users in different time intervals [11]. However, some experts have mentioned the greater influence of density than volume for CDR data applications [35]. Iounousse has identified the land use of a city using unsupervised clustering based on satellite data [36].

In terms of GASs, their size differs from buildings to administrative regions [37]. Moreover, the GASs may not represent a complete region in the city [38]. Additionally, researchers have conducted experiments that indicated the significance of traffic analysis zones (TAZs) in CDR applications and provided useful suggestions for urban transportation planning agencies [39].

UFA recognition typically uses a clustering method [24, 36, 38, 40, 41] or a semantic model [9, 15, 31, 42]. Semantic models can realize hierarchical recognition, but ignore the shape and size of objects, which have a great impact on the results. In addition, erroneous classification objects can also lead to incorrect results, and the correlations between the UFAs are known to have a strong influence on the overall classification. Clustering methods can overcome these shortcomings; furthermore, the clustering approach is adaptive to individuals and obtains results quickly and precisely. The lack of discussion on GASs and quantitative

verification in previous studies has led to inaccurate recognition results, and the combination of static data in existing methods is inadequate when using CDR data. In this study, to identify UFAs, the  $k$ -means clustering model is applied to dynamic CDR data that have been translated into the CAIs of GASs, and kernel density estimation (KDE) is used for static POI data based on TAZs. Additionally, in the verification of UFA recognition, qualitative and quantitative analyses are used based on static Baidu high-resolution image map data and field survey data.

### 3. Methodology

#### 3.1. Data Sources

**3.1.1. CDR Data.** The case study covers the region within Beijing's 3rd Ring Road, an area of 159 km<sup>2</sup>. The study described in this paper was conducted using Beijing CDR data from June 1–30, 2015. These data were obtained from strategic cooperation projects undertaken by our research team and the Beijing branch of China Mobile Communications Group Co., Ltd. As a result, the data have strict privacy protection (with private information removed) and right of use protection. The research included 3198 mobile communication base stations, with 880 macrocellular stations and 2318 microcellular stations. This covered an average of about 4.94 million daily users and 100.73 million daily records. The CDR data format and examples are presented in Table 1.

**3.1.2. POI Data.** The POI data refer to all geographic entities that can be abstracted as points. The POI data were extracted from the Beijing electronic map in 2015 (see Table 2).

**3.1.3. Baidu High-Resolution Image (BHRI) Map Data.** The BHRI map data used in this paper can be found at <https://map.baidu.com> and are publicly available.

**3.2. Discussions of GASs.** Five different GASs were collected from previous studies, namely, the raster layer [43], Voronoi layer [41, 44], road network segmentation layer [33], TAZ layer [24], and administrative layer [6]. The influence of the different GASs on the identification of UFAs is discussed in Table 3.

From the discussion in Table 3, the TAZ layer appears to have several advantages in terms of UFA identification. There are 235 TAZs within Beijing's 3rd Ring Road.

**3.3. CDR Data Model.** Compared with other methods, clustering has many advantages, such as easy operation, rapid output of results, and the ability to focus on individuals. The  $k$ -means clustering method is widely used in clustering analysis of UFA recognition based on human activity data. Hence, the  $k$ -means clustering is used in this study to deal with CDR data. As there is some difference between human travel characteristics on workdays and at the weekend [45], these periods are analyzed separately, which is very helpful for the recognition of UFAs. In addition, NCA, CAI, and UD are also considered in the CDR data model.

**3.3.1. Several Definitions.** The following items are used in our model of CDR data (see Table 4):

- (1) CAIs of GAS: the ratio of the number of calls made or received in a certain GAS at a fixed time interval of the day to the area of the GAS coverage
- (2) UDs of GAS: the ratio of the number of users in a certain GAS at a fixed time interval of the day to the area of the GAS coverage
- (3) Matrix of NCAs, CAIs, and UDs: the distributions of NCA, CAI, and UD in each GAS at a 5 min time slot of the day, expressed as  $V_n(\tau)$ ,  $\overline{v}_n^w(\tau)$ , and  $\overline{u}_n^w(\tau)$ , where  $\tau$  denotes the 5 min time slot,  $\tau \in \{0, 1, \dots, 287\}$
- (4) Signature of each GAS: the aggregation result of the NCA matrix, CAI matrix, and UD matrix, which indicates a certain UFA, expressed as  $S_{n, \Omega_w}(\tau)$ ,  $s_{n, \Omega_w}(\tau)$ ,  $t_{n, \Omega_w}(\tau)$ ,  $\Omega_1 = \{\text{Monday, Tuesday, Wednesday, Thursday, Friday}\}$ ,  $\Omega_2 = \{\text{Saturday, Sunday}\}$ , and  $\Omega_1, \Omega_2 \in \Delta$ ,  $\Omega_1 \cap \Omega_2 = \emptyset$ ,  $\Omega_1 \cup \Omega_2 = \Delta$ .

#### 3.3.2. Index Calculation

- (1) Matrix of user numbers:

$$U_n(\delta, \tau) = \sum_{c=1}^C P_n^c(\delta, \tau), \quad (1)$$

where  $U_n(\delta, \tau)$  is the matrix of user numbers in the  $n$ th GAS;  $n \in \{1, 2, \dots, N\}$  is the number of GASs;  $\delta$  represents each day of a month ( $\delta \in \{1, 2, \dots, 30\}$  in this paper);  $C$  is the number of mobile communication base stations in a certain GAS;  $P_n^c(\delta, \tau)$  represents the number of users in the  $n$ th GAS connected to the  $c$ th mobile communication base station in the  $\tau$ th 5 min interval of day  $\delta$ .

- (2) Matrix of communication numbers:

$$V_n(\delta, \tau) = \sum_{c=1}^C R_n^c(\delta, \tau), \quad (2)$$

where  $V_n(\delta, \tau)$  is the matrix of communication numbers in the  $n$ th GAS;  $R_n^c(\delta, \tau)$  represents the number of communications in the  $n$ th GAS connected to the  $c$ th mobile communication base station in the  $\tau$ th 5 min interval of day  $\delta$ .

- (3) Area of GASs: area statistics are mainly determined using ArcGIS, and the  $n$ th GAS area is referred to as  $A_n$ . The specific statistical operations are not discussed in this article.
- (4) Average value calculation:
  - (a) The average user numbers in GASs:

$$\overline{U}_n^w(\tau) = \frac{1}{N_w} \sum_{\delta \in \Omega_w} U_n(\delta, \tau), \quad (3)$$

TABLE 1: CDR data format and examples.

Field name	Description	Examples
IMSI	Unique identification of a cell phone, also the user ID	15*****756
Time	Time of the successful signaling trigger, exact to the second	2015/06/1 1:05:00
LAC	ID number of position area	4128
E	Longitude	116.327407
N	Latitude	39.887484
Event ID	Event type 8: calling or called 9: position update 10: sending and receiving SMS	8
Type	Type of base station 0: macrocellular station 1: microcellular station	0

TABLE 2: POIs within 3rd Ring Road of Beijing.

No.	Type	Amount
1	Hotel	2083
2	Restaurant	6882
3	Supermarket and shopping mall	5152
4	Exits and entrances	499
5	Mansion	1111
6	Bus station	4406
7	Park	121
8	Others	23787
9	Parking lot	893
10	School	1417
11	Drug store	632
12	Medical	1182
13	Bank	2759
14	Government agency	1118

where  $\overline{U}_n^w(\tau)$  represents the matrix of average user numbers in GASs on weekday or weekends;  $w \in \{1, 2\}$ , with 1 denoting weekday and 2 denoting weekend.

(b) Average communication numbers of GASs:

$$\overline{V}_n^w(\tau) = \frac{1}{N_w} \sum_{\delta \in \Omega_w} V_n(\delta, \tau), \quad (4)$$

where  $\overline{V}_n^w(\tau)$  represents the matrix of average communication numbers in GASs on a weekday or weekend.

(5) Intensity calculations:

$$\begin{aligned} \overline{v}_n^w(\tau) &= \frac{\overline{V}_n^w(\tau)}{A_n}, \\ \overline{u}_n^w(\tau) &= \frac{\overline{U}_n^w(\tau)}{A_n}, \end{aligned} \quad (5)$$

where  $\overline{v}_n^w(\tau)$ ,  $\overline{u}_n^w(\tau)$  denote the matrixes of CAIs and UD, respectively and  $A_n$  is the area of the  $n$ th GAS ( $\text{km}^2$ ).

(6) Signature calculations:

(a) Signature of NCA:

$$S_{n,\Omega_w}(\tau) = \overline{V}_n^w(\tau). \quad (6)$$

(b) Signature of CAI:

$$s_{n,\Omega_w}(\tau) = \overline{v}_n^w(\tau). \quad (7)$$

(c) Signature of UD:

$$t_{n,\Omega_w}(\tau) = \overline{u}_n^w(\tau), \quad (8)$$

where  $S_{n,\Omega_w}(\tau)$ ,  $s_{n,\Omega_w}(\tau)$ ,  $t_{n,\Omega_w}(\tau)$  are the signatures of GASs based on the NCA, CAI, and UD on a weekday or a weekend. The signatures are calculated by SPSS.

3.3.3. *Clustering Analysis.* Unsupervised clustering technology requires the number of clusters to be known beforehand. In the case of  $k$ -means, the optimal number of clusters is determined by whether close clustering or good separation is required. A validation method [46] can be used to select a better value of  $k$ .

The cluster validity index is the ratio of the intracluster distance to the intercluster distance. The ideal classification will minimize the intracluster distance and maximize the intercluster distance, so a smaller value of the validity index indicates better classification. The cluster validity index is calculated as follows:

$$\begin{aligned} C_{\text{ita}} &= \frac{1}{N} \sum_{p=1}^k \sum_{Z_n \in C_p} \|Z_n - c_p\|^2, \\ C_{\text{ite}} &= \min_{p \neq q} \|c_p - c_q\|^2, \\ \text{VI} &= \frac{C_{\text{ita}}}{C_{\text{ite}}}, \end{aligned} \quad (9)$$

where  $C_{\text{ita}}$  and  $C_{\text{ite}}$  denote the intracluster and intercluster distances; VI is the validity index;  $C_p$  is the set of signatures belonging to the cluster defined by centroid  $c_p$ ;  $Z_n$  represents a signature, such as  $S_n$ ,  $s_n$ , and  $t_n$ ; and  $p, q \in k$ .

TABLE 3: Discussion of five GASs.

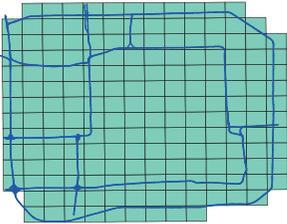
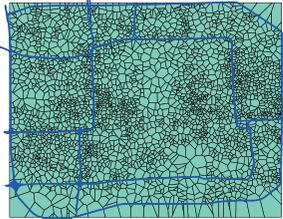
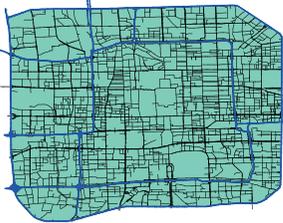
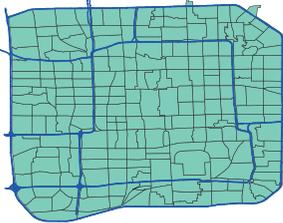
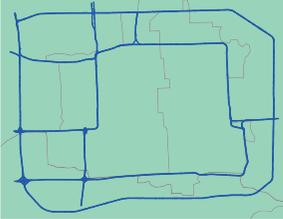
Name	Layer diagram	Discussion
Raster layer		This kind of layer has no difference between cells. In reality, small segmentation makes little significance, but the theoretical basis is insufficient with large divisions. Moreover, this layer crosses the traffic corridor in the city, which may lead to incorrect identification results
Voronoi layer		All units in this layer have irregular sizes, and the accuracy of recognition obviously decreases in areas where the density of mobile base stations is high, making incorrect results more likely in these units. The phenomenon of crossing traffic corridors also exists
Road network segmentation layer		Although this layer does not cross traffic corridors, there is a huge discrepancy in the road network distribution between urban centers and suburbs, which strongly affects the results. Furthermore, the results from this layer lack practical significance
TAZ layer		There is no crossing of traffic corridors in this layer, and the division takes multiple factors into account, such as landform, administration, human history, road network, and urban function. Thus, the results are much more meaningful in terms of traffic field, land use, and urban planning
Administrative layer		The granularity of the division is somewhat coarse, making correct identification difficult. Furthermore, the recognition results are of little significance for urban planning and transportation planning

Figure 1 shows the clustering results obtained with different values of  $k$ . The following can be inferred: (1) The validity value of the CAI data is smaller than that of the NCA and UD data, which indicates that clustering analysis based on CAIs results in a large intracluster distance and small intercluster distance. This suggests a better clustering result and demonstrates that the size of the GAS has a significant impact on the recognition of urban functions. (2) The UD and NCA data do not provide good results, indicating that the NCAs or UDs may not have as great an impact on the clustering results as the CAIs, which can broadly distinguish the mechanisms of

CDR data. There are many situations in which the communication base stations are triggered, including active triggering and passive triggering, cross-region triggering, and switching on-off. In practice, there may be great deviations in results if the communication activity is ignored. (3) The different  $k$  values produce different values of VI. The smallest value is given by the CAI data with  $k=6$ . Thus, combined with some relevant research about the types of urban functions and the Chinese standard [47], five single UFAs (residential, commercial, park and scenery, office, and education areas) plus mixed areas are considered in this paper.

TABLE 4: Parameter representation.

Name	Meaning
$\tau$	Denotes a 5 min time slot throughout the day, $\tau \in \{0, 1, \dots, 287\}$
$n$	$n \in \{1, 2, \dots, N\}$ is the number of GASs
$\delta$	Denotes the day of a month, $\delta \in \{1, 2, \dots, 30\}$ in this paper
$w$	Denotes weekday or weekend, $w \in \{1, 2\}$ , 1 for weekday and 2 for weekend
$\Omega_w$	$\Omega_1 = \{\text{Monday, Tuesday, Wednesday, Thursday, Friday}\}$ , $\Omega_2 = \{\text{Saturday, Sunday}\}$ , $\Omega_1, \Omega_2 \in \Delta$ , $\Omega_1 \cap \Omega_2 = \emptyset$ , $\Omega_1 \cup \Omega_2 = \Delta$
$\Delta$	Seven days of the week, $\Delta \in \{\text{Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday}\}$
$U_n(\delta, \tau)$	Matrix of user numbers in the $n$ th GAS
$V_n(\delta, \tau)$	Matrix of communication numbers in the $n$ th GAS
$C$	Number of mobile communication base stations in a certain GAS
$P_n^c(\delta, \tau)$	Number of users in the $n$ th GAS
$R_n^c(\delta, \tau)$	Number of communications in the $n$ th GAS, which is the $c$ th mobile communication base station in the $\tau$ th 5 min interval of day $\delta$
$A_n$	Area of $n$ th GAS
$N_w$	Number of weekdays or weekends in a month, $N_1 = 22$ , $N_2 = 8$ in this paper
$\bar{U}_n^w(\tau)$	Matrixes of average user numbers of GASs on a weekday or weekend
$\bar{V}_n^w(\tau)$	Matrixes of average communication numbers of GASs on a weekday or weekend
$\bar{\gamma}_n^w(\tau)$	Matrixes of CAIs
$\bar{u}_n^w(\tau)$	Matrixes of UDIs
$S_{n, \Omega_w}(\tau)$	Signature of GASs based on the NCAs on weekday or weekend
$s_{n, \Omega_w}(\tau)$	Signature of GASs based on the CAIs on weekday or weekend
$t_{n, \Omega_w}(\tau)$	Signature of GASs based on the UDIs on weekday or weekend
$C_{ita}$	Intracluster distances
$C_{ite}$	Intercluster distances
VI	Validity index
$C_p$	Set of signatures that belong to the cluster defined by a centroid $c_p$
$Z_n$	Signature, $\{S_n, s_n, t_n\} \in Z_n$ , $p, q \in k$
$f(s)$	KDE function at spatial position $s$
$h$	Distance attenuation threshold
$m$	Number of elements for which the distance is less than or equal to $h$ from location $s$
$c_j$	Sample element
$g(x)$	Spatial weighting function
$i$	Type of cluster; $i \in (1, 2, 3, 4, 5)$
$a_i$	Identification index of type $i$
$S_A^i$	Actual function area of a cluster $i$ , $\text{km}^2$
$S_B^i$	Area of detecting function, which is the area of the GAS containing $i$ , $\text{km}^2$

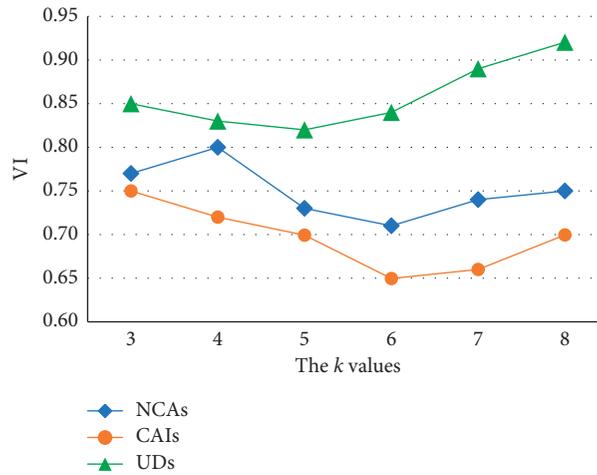


FIGURE 1: Calculation results.

### 3.4. POI Model

**3.4.1. POI Data Processing.** The POI data were classified for modeling. First, any POI data unrelated to functional identification were removed, for instance, bus station data, exit and entrance data, and other POI data. Several POIs were then reclassified according to the Chinese standard [47]; in this study, the school POI data were divided into university, high school and middle school, and primary school and kindergarten. Residential areas were permitted to include some public facilities, such as primary schools, kindergartens, and convenience stores. In addition, office buildings, government agencies, and parking lots of office buildings were classified as office areas. Commercial areas were distinguished by supermarkets and shopping malls, hotels, and restaurants. The processing results of the POI data are presented in Table 5.

**3.4.2. POI Model.** In general, nonparametric estimation, which is not affected by the overall parameters, is the most widely used method for determining the probability density. Moreover, it can be applied to any sample analysis. KDE is a nonparametric estimation method for the unknown probability density function. Thus, the POI data model uses KDE. The calculation can be expressed as follows:

$$f(s) = \sum_{j=1}^m \frac{1}{h^2} g\left(\frac{s-c_j}{h}\right), \quad (10)$$

where  $f(s)$  is the KDE function at spatial position  $s$ ;  $h$  is a distance attenuation threshold;  $m$  is the number of elements for which the distance is less than or equal to  $h$  from location  $s$ ;  $c_j$  is the sample element; and  $g(x)$  is the spatial weighting function.

The two key parameters in the KDE function are  $g(x)$  and  $h$ . Different average weights must be selected when choosing a certain  $g(x)$  function. The uniform function gives the same weight to all points within the scope of the study; the triangular function gives a linear decreasing trend; the Epanechnikov function is relatively slow; and the Gaussian function has no boundaries, allowing weights to be assigned to all points. This study adopts an adaptive bandwidth for the KDE of the Gaussian kernel function [48], as this ensures better convergence and smoothness than the fixed-bandwidth KDE function.

**3.5. Recognition Procedures.** The identification procedure is illustrated in Figure 2. First, based on CDR data, we calculate the parameters required for the index calculations. Second, the characteristics of the CDR clustering results are analyzed, including the weekday and weekend features, number of peak values, intensity of peak values, and distribution of peak values; this allows the travel behaviors and public cognitions to be understood. Third, based on the clustering of POI data, the UFA identification results are modified.

Fourth, verification is conducted using the BHRI data, field data, and the identification index. Finally, we obtain the final UFA results.

## 4. Case Study and Discussion

**4.1. Results and Discussion.** To recognize the UFAs, the clustering results of POI data are shown in Figures 3 and 4. The characteristics of residents' travel behavior and public cognition are now introduced to explain the signatures. UFA identification based on Figure 3 is also discussed.

Cluster 1: the main feature of this cluster is that it has a very high CAI on a weekday and an obvious double peak concentrated at 08:00–11:00 and 14:00–16:00. The CAI in the morning is greater than that in the afternoon. Furthermore, there is still a certain number of CAIs between 18:00 and 21:00, which indicates that some people are still working during this period. The CAIs of this cluster are lower on the weekend than on weekdays, probably the result of overtime being worked on weekends. However, the value of CAI begins to decrease at 16:00 and is very low after 17:00, which suggests that the work is much more flexible on weekends than on weekdays, so employees can leave their offices early. Based on this analysis, cluster 1 is considered to represent office areas.

Cluster 2: this cluster is characterized by the fact that the CAIs on the weekend are higher than those on weekdays, and there is no double peak on weekdays. In contrast, there is a peak activity from 15:00 to 17:00 on weekends, which indicates that people in these areas use their mobile phones to contact friends, fellow travelers, or drivers to arrange their journey home. Combined with the POI results in Figure 4(a), we can infer that this is the signature of park and scenery areas.

Cluster 3: this cluster has the obvious feature that the CAI values on workdays and weekends are relatively low. Additionally, there is a double peak on weekdays and higher CAIs than on weekends. However, no double peak occurs at the weekend. This can be explained by residents working at home on workdays and taking a nap after lunch. In contrast, people who are enjoying their leisure time do not need a specific period of rest. There are around 500 calls/km<sup>2</sup> on weekdays, which might be to invite friends or clients to dinner. Thus, these GASs are likely to be residential areas.

Cluster 4: in this cluster, a notable double peak occurs on the left side and has a higher value on weekdays than on weekends. However, the CAIs on both workdays and weekends are not especially high. The trends on weekdays and weekends are similar after 19:00, and the intensity values are only slightly different between day and night, which indicates that it is mainly young people living and working here. In conclusion, this kind of schedule suggests universities and high schools. With the help of the POI data in Figure 4(b), we can firmly conclude that these are education areas.

Cluster 5: the fifth cluster type features a slight difference in intensity between workdays and weekends, and there is a

TABLE 5: POI reclassification based on Chinese land use standard.

Type	Reclassification type	Potential functional area	Amount
Park	Park University	Park and scenery areas	121
School	High school and middle school Primary school and kindergarten	Education area	861
Drug store Medical	Drug store Medical	Residential area	6022
Supermarket and shopping malls	Convenience store Supermarket and shopping mall		
Hotel Restaurant Bank	Hotel Restaurant Bank	Commercial area	2645
Mansion	Commercial building Office building		
Government agency Parking lot	Government agency Parking lot of office building	Office area	3011

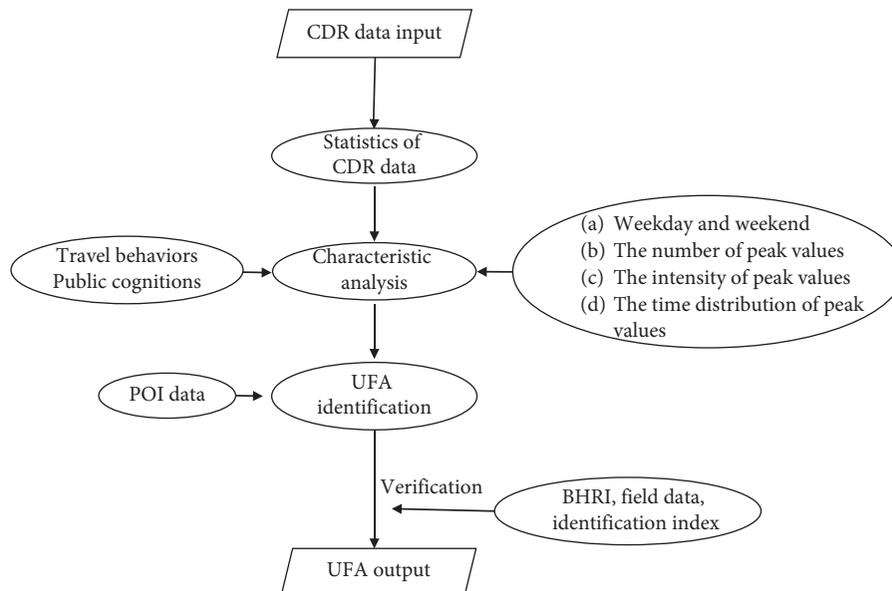


FIGURE 2: Flowchart of identification.

double peak on working days. Furthermore, high CAIs are maintained from 08:00 to 21:00 and longer into the night on weekends. Though the CAI values decline on both weekdays and weekends after 21:00, their number and duration during this period on weekends are stronger and longer than that on weekdays. All of these features are more likely to occur in commercial areas.

Cluster 6: with the lowest CAI values on weekdays and almost the highest values (albeit with significant fluctuations) on weekends, this cluster cannot be accurately summarized, especially at weekends. At the same time, no travel behaviors or human activities can fully explain this pattern. Thus, this region is tagged as a mixed area.

The cluster members of each signature, as calculated by SPSS, were displayed in GIS, and the spatial distribution of the UFA recognition results are shown in Figure 5.

According to the UFA recognition results in Figure 5, several conclusions can be drawn. First, the residential areas have a high density of occupation and are widely distributed. However, the distribution of park and scenery areas is relatively concentrated. Second, most of the GASs south of metro line 1 are residential areas; in contrast, the educational areas are largely located to the north of metro line 1, which may result in tidal traffic situations. As a result, the passenger flow on north-south subway lines (e.g., metro lines 4 and 5) is very high. Third, office areas are mainly distributed around and between metro lines 6 and 1. This places significant traffic pressure on these metro lines, with the spatial and temporal characteristics of passenger flow making for heavy daily average passenger numbers. Fourth, the concentrated distribution of park and scenery areas, especially in urban central areas, brings greater

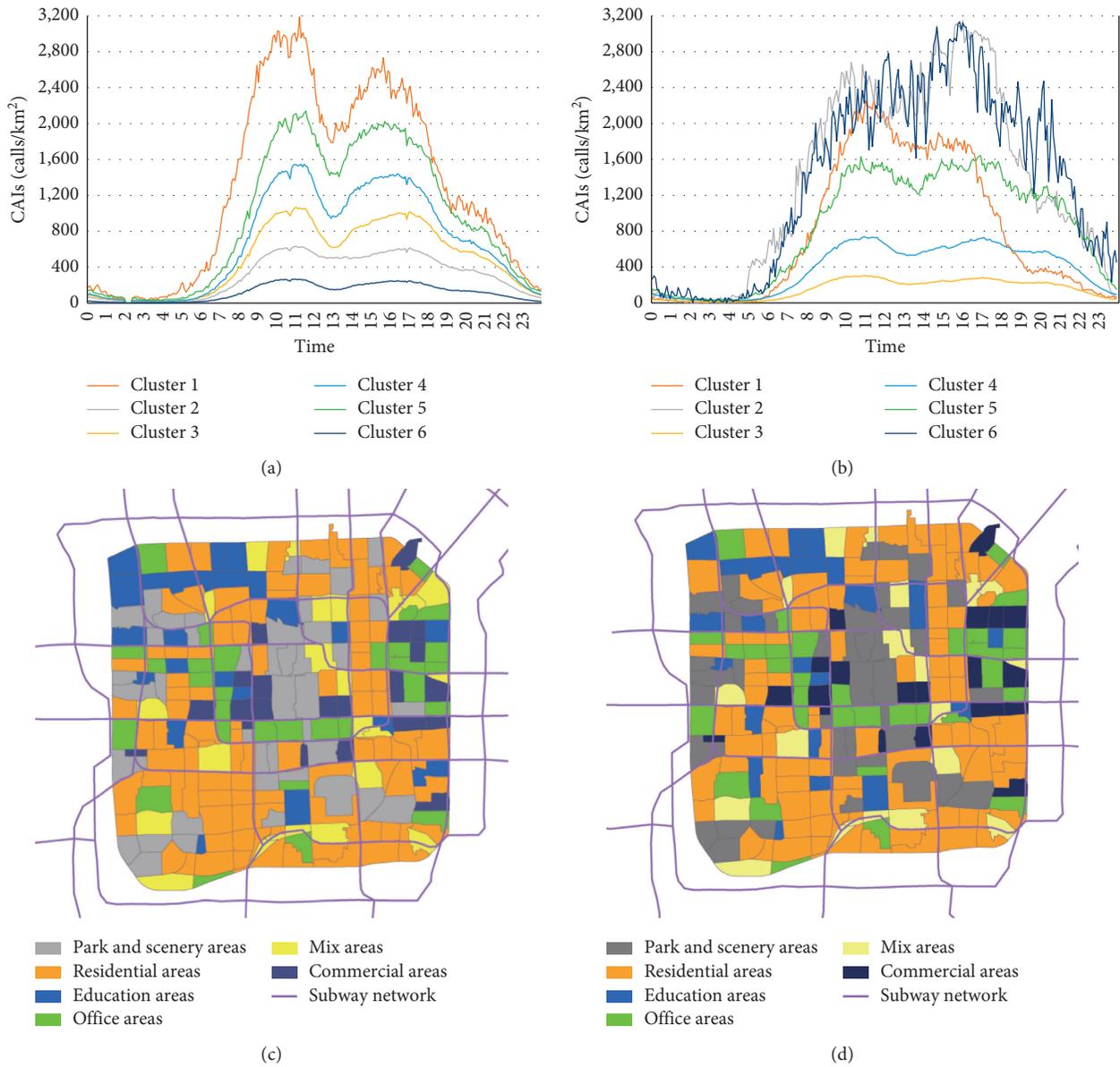


FIGURE 3: Results of CDR data clustering: (a) clustering results on weekday, (b) cluster results on weekend, (c) spatial distribution of six clusters on weekdays, and (d) spatial distribution of six clusters on weekends.

centripetal traffic pressure to the urban traffic operations and management. This phenomenon is particularly remarkable on holidays and at weekends.

**4.2. Result Verification.** To check the accuracy of the results, qualitative and quantitative verifications are applied based on the recognition results, field survey results, and BHRI map data. The following are typical GASs considered in this analysis: Temple of Heaven Park, Financial Street, Beijing Institute of Technology (BIT), Fangzhuang Region, Qianmen Street, and Beijing South Railway Station.

**4.2.1. Qualitative Verification.** In terms of qualitative verification, we consider some typical GASs and field survey data, as well as the BHRI map. Representations of the six clusters are discussed in Table 6.

**4.2.2. Quantitative Analysis.** In the quantitative analysis of the proposed methodology, the mixed areas about 8.5% of the total area are neglected because more than one functional component is present. For those areas with a sole functional result, the identification index is defined as the ratio of the area covered by that function to the whole area of the GAS. This is schematically illustrated in Figure 6. Of

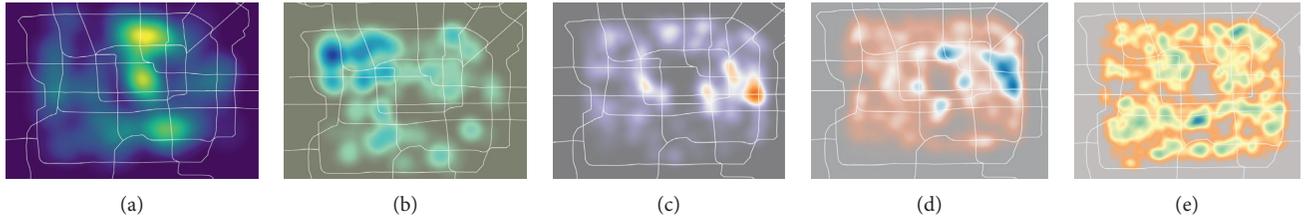


FIGURE 4: Results of POI data clustering. The color represents the density of POI numbers: (a) park and scenery areas, (b) education area, (c) office area, (d) commercial area, and (e) residential area.

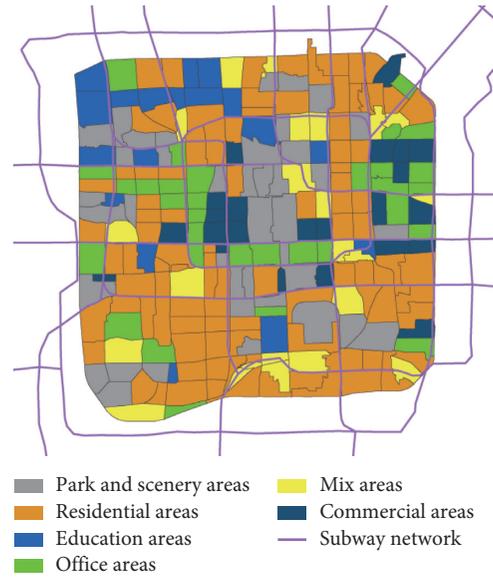


FIGURE 5: Result of identification.

course, this index can be used to represent the accuracy of recognition. The actual function area is calculated using the field survey results and the BHRI map, and the area of each GAS is computed by GIS. The identification index is computed as follows:

$$a_i = \frac{S_A^i}{S_B^i} \times 100\%, \quad (11)$$

where  $i$  is the cluster type,  $i \in (1, 2, 3, 4, 5)$ ;  $a_i$  is the identification index of type  $i$ ;  $S_A^i$  is the actual function area of cluster  $i$  ( $\text{km}^2$ ); and  $S_B^i$  is the area of the GAS in which  $i$  is located ( $\text{km}^2$ ).

The lowest identification index was found to be 63.16% for the commercial area, which is higher than the overall accuracy obtained in previous studies [49, 50]. The dynamic needs of urban planning and management can be satisfied if the identification index is above 60% [50]. Thus, the identifications have great practical significance because the results are all above 60%. The average identification index is 78.30%, far more than the mean value achieved in the

previous research, which demonstrates the great progression made by this study.

As Table 7 shows, the park and scenery areas have the highest identification index of 96.00%. This can be explained by the fact that the GASs or TAZs were considered when these functions were divided; furthermore, it shows the significance of choosing reasonable GASs before identifying the urban functions.

The next-highest identification index values are given by the residential areas and office areas. The POIs of multitype residential facilities (e.g., kindergartens, drug stores, and convenience stores) are very helpful in identifying residential areas. Moreover, there are very high CAIs in office areas, so an impressive identification index can be achieved.

The education areas and commercial areas have lower identification index values. This can be explained by the many hotels for conference attendees and departments for school staff around the education area; likewise, with complex land use close to commercial areas, people come and go, but do not stay too long, which affects the CAIs to a certain extent.

TABLE 6: Results of qualitative verification.

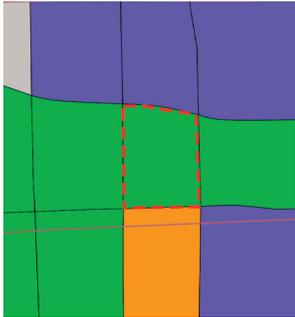
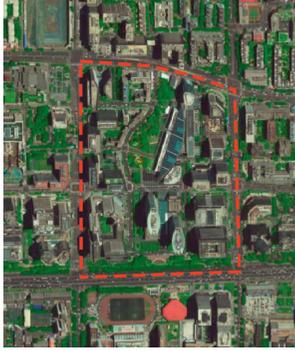
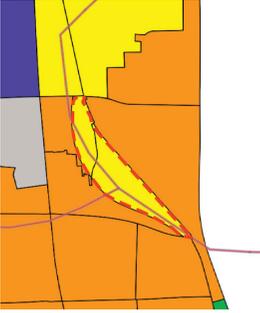
GAS of Temple of Heaven Park		
 <p>Contrast diagram</p>	 <p>Field survey</p>	 <p>BHRI map</p>
Identification results	This region agrees closely with Temple of Heaven Park, and so the identification result conforms to reality	
GAS of financial street		
 <p>Contrast diagram</p>	 <p>Field survey</p>	 <p>BHRI map</p>
Identification results	This region includes Bank of Beijing Mansion, China Unicom building, Yuhang building, Kaifu building, and others, which means the office area has been correctly identified	
GAS of BIT		
GAS of BIT		

TABLE 6: Continued.

Contrast diagram		Identification result		Field survey		BHRI map
Identification results	BJT covers most of this GAS, so the identification is correct					
GAS	GAS of Fangzhuang Region					
Contrast diagram		Identification result		Field survey		BHRI map
Identification results	The residential districts of Songjiazhuang, Fangnan, Zhengxin, and others are located in this GAS, and there is no doubt that these are residential areas					
GAS	GAS of Qianmen Street					
Contrast diagram		Identification result		Field survey		BHRI map
Identification results	This region includes Madame Tussauds, Qianjude Road Duck, Dashilar, and the famous Qianmen Street, all of which have very strong commercial functions					

TABLE 6: Continued.

GAS	GAS of Beijing South Railway Station	
Contrast diagram		Identification result
Identification results		Field survey
		BHRJ map
	No regular features can be found to identify this area, so it is considered a mixed area	

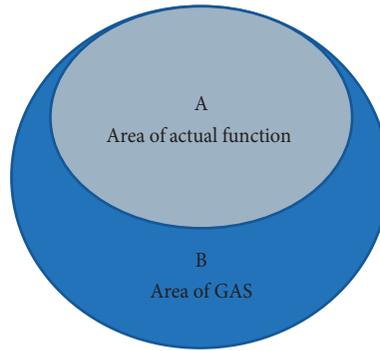


FIGURE 6: Schematic of the identification index.

TABLE 7: Results of quantitative verification.

Functional areas	Sample areas	Cluster ( $i$ )	$S_B^i$ (km <sup>2</sup> )	$S_A^i$ (km <sup>2</sup> )	Identification index (%)
Education area	BIT	4	1.06	0.71	66.98
	Beijing Normal University	4	0.89	0.64	71.91
	Beijing University of Posts and Telecommunications	4	0.56	0.40	71.43
Residential area	Fangzhuang Region	3	1.20	0.99	82.50
	Pener alley Region	3	0.49	0.43	87.76
	Zhushikou West Street	3	1.01	0.94	93.07
Office area	Financial Street	1	0.33	0.25	75.76
	Hanhai Culture Building	1	0.54	0.39	72.22
	Ganjiakou Region	1	0.38	0.29	76.32
Commercial area	Qianmen Street	5	0.25	0.16	64.00
	Jianwai SOHO	5	0.45	0.29	64.44
	Dongxinglong Street	5	0.57	0.36	63.16
Park and scenery areas	Temple of Heaven Park	2	2.00	1.92	96.00
	Yuyuantan Park	2	0.96	0.91	94.79
	Beijing Zoo	2	0.66	0.63	95.45
Avg.					78.30

## 5. Conclusions and Future Work

The tendency toward integrated urban planning and management requires dynamic recognition of UFAs. However, the selection of GASs in the previous research has nonnegligible effects on the identification of UFAs. In this study, three indexes of CDR were presented, and the concept of CAI was selected as the main focus of the study. Moreover, POI data were found to be very helpful in identifying UFAs. Thus,  $k$ -means clustering for CDR data and the KDE method for POI data were applied to the region within Beijing's 3rd Ring Road. It is worth noting that the proposed method could be used with a combination of other information, such as SCD data or blog check-in data, which contains POI data. In the final UFA identification results, the park and scenery areas were found to be most accurate. The average identification index was about 78.30%, far higher than in previous research.

The findings of this study are conducive to dynamic urban management and planning. Note that the proposed method has not been applied to the whole city in a case study. However, urban planning theories and related planning data should be considered in future research on UFA identification. Additionally, further research may focus on the application of new technologies in big data mining,

such as deep learning and machine learning, which can provide reliable information for the integration of planning and management.

### Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

### Disclosure

The manuscript abstract was previously presented in Transportation Research Board 98th Annual Meeting.

### Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

### Acknowledgments

The authors would like to acknowledge the financial support for this study provided by the National Key R&D Program of China (2017YFC0803903 and 2016YFE0206800),

Transportation Technology Project of Henan Province (2017Z8), Beijing Natural Science Foundation (no. L181001), Key Program of Beijing Natural Science Foundation (no. 4181002), Science and Technology Project of Beijing Municipal Transportation Commission (201825-HNBJ2), and Project of Beijing Municipal Science & Technology Commission (Z191100002519002).

## References

- [1] A. P. Montanges, G. Moser, H. Taubenböck, M. Wurm, and D. Tuia, "Classification of urban structural types with multisource data and structured models," in *Proceedings of the Joint Urban Remote Sensing Event (JURSE)*, March 2015.
- [2] U. Heiden, W. Heldens, S. Roessner, K. Segl, T. Esch, and A. Mueller, "Urban structure type characterization using hyperspectral remote sensing and height information," *Landscape and Urban Planning*, vol. 105, no. 4, pp. 361–375, 2012.
- [3] H. Assem, L. Xu, T. S. Buda, and D. O'Sullivan, "Spatio-temporal clustering approach for detecting functional regions in cities," in *Proceedings of the 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 370–377, IEEE, San Jose, CA, USA, November 2016.
- [4] K. Maat, B. van Wee, and D. Stead, "Land use and travel behaviour: expected effects from the perspective of utility theory and activity-based theories," *Environment and Planning B: Planning and Design*, vol. 32, no. 1, pp. 33–46, 2016.
- [5] B. Jiang, "Editorial: spatial heterogeneity, scale, data character, and sustainable transport in the big data era," *International Journal of Geo-Information*, vol. 7, p. 5, 2018.
- [6] B. Jiang, "Geospatial analysis requires a different way of thinking: the problem of spatial heterogeneity," *GeoJournal*, vol. 80, no. 1, pp. 1–13, 2015.
- [7] P. Bonnel and M. A. Munizaga, "Transport survey methods—in the era of big data facing new and old challenges," *Transportation Research Procedia*, vol. 32, pp. 1–15, 2018.
- [8] G. D. D'Angeac, *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey & Company, New York, NY, USA, 2011.
- [9] S. A. Ríos and R. Muñoz, "Land Use detection with cell phone data using topic models: case Santiago, Chile," *Computers, Environment and Urban Systems*, vol. 61, pp. 39–48, 2017.
- [10] G. Re Calegari, E. Carlino, D. Peroni, and I. Celino, "Filtering and windowing mobile traffic time series for territorial land use classification," *Computer Communications*, vol. 95, pp. 15–28, 2016.
- [11] F. Manfredini, P. Pucci, and P. Tagliolato, "Toward a systemic use of manifold cell phone network data for urban analysis and planning," *Journal of Urban Technology*, vol. 21, no. 2, pp. 39–59, 2014.
- [12] F. Scioscia, M. Binetti, M. Ruta, S. Ieva, and E. Di Sciascio, "A framework and a tool for semantic annotation of POIs in OpenStreetMap," *Procedia—Social and Behavioral Sciences*, vol. 111, pp. 1092–1101, 2014.
- [13] X. Zhou, J. Liu, A. G. O. Yeh, Y. Yang, and W. Li, "The uncertain geographic context problem in identifying activity centers using mobile phone positioning data and point of interest data," in *Advances in Geographic Information Science*, Springer, Cham, Switzerland, 2015.
- [14] X. Zhang and S. Du, "A linear dirichlet mixture model for decomposing scenes: application to analyzing urban functional zonings," *Remote Sensing of Environment*, vol. 169, pp. 37–49, 2015.
- [15] X. Zhang, S. Du, and Q. Wang, "Hierarchical semantic cognition for urban functional zones with VHR satellite images and POI data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 132, pp. 170–184, 2017.
- [16] X. Zhang and S. Du, "Learning selfhood scales for urban land cover mapping with very-high-resolution satellite images," *Remote Sensing of Environment*, vol. 178, pp. 172–190, 2016.
- [17] Q. Wang, S. Tang, X. Chen, and L. Wang, "Multinomial logistic regression for land use classification with remote sensing," in *Proceedings of the Conference Transportation Research Board 94th Annual Meeting*, Washington, DC, USA, January 2015.
- [18] T. Li, J. Wu, H. Sun, and Z. Gao, "Integrated co-evolution model of land use and traffic network design," *Networks & Spatial Economics*, vol. 16, pp. 579–603, 2016.
- [19] T. Li, H. Sun, J. Wu, Z. Gao, Y.-e. Ge, and R. Ding, "Optimal urban expressway system in a transportation and land use interaction equilibrium framework," *Transportmetrica A: Transport Science*, vol. 15, no. 2, pp. 1247–1277, 2019.
- [20] R. Ahas and Ü. Mark, "Location based services—new challenges for planning and public administration?" *Futures*, vol. 37, no. 6, pp. 547–561, 2005.
- [21] C. Ratti, D. Frenchman, R. M. Pulselli, and S. Williams, "Mobile Landscapes: using location data from cell phones for urban analysis," *Environment and Planning B: Planning and Design*, vol. 33, no. 5, pp. 727–748, 2006.
- [22] B. N. Sudarmanto, M. Chikaraishi, A. Fujiwara, J. Zhang, and L. Gang, "Exploring variation of trip fares by taxi-like paratransit services in Jakarta city: a multilevel analysis," in *Proceedings of the Conference Transportation Research Board 91st Annual Meeting*, Washington, DC, USA, January 2012.
- [23] T. Hu, J. Yang, L. Xuecao, and P. Gong, "Mapping urban land use by using landsat images and open social data," *Remote Sensing*, vol. 8, p. 2, 2016.
- [24] H. Han, X. Yu, and Y. Long, "Discovering functional zones using bus smart card data and points of interest in Beijing," 2015, <https://arxiv.org/abs/1503.03131>.
- [25] X. Cheng, W. Li, F. Jia, D. Yang, and Z. Duan, "Analyzing human activity patterns using cellular phone data: case study of jinhe new town in Shanghai, China," in *Proceedings of the Conference Transportation Research Board 92nd Annual Meeting*, Washington, DC, USA, January 2013.
- [26] A. Blei, K. Kawamura, M. Javanmardi, and A. Mohammadian, "Evaluation methods for estimating vehicle miles traveled with GPS travel survey data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2495, no. 1, pp. 112–120, 2015.
- [27] Croce, Musolino, Rindone, and Vitetta, "Transport system models and big data: zoning and graph building with traditional surveys, FCD and GIS," *ISPRS International Journal of Geo-Information*, vol. 8, no. 4, p. 187, 2019.
- [28] B. Alonso, Á. Ibeas, G. Musolino, C. Rindone, and A. Vitetta, "Effects of traffic control regulation on network macroscopic fundamental diagram: a statistical analysis of real data," *Transportation Research Part A: Policy and Practice*, vol. 126, pp. 136–151, 2019.
- [29] A. I. Croce, G. Musolino, C. Rindone, and A. Vitetta, "Sustainable mobility and energy resources: a quantitative assessment of transport services with electrical vehicles," *Renewable and Sustainable Energy Reviews*, vol. 113, Article ID 109236, 2019.

- [30] J. C. Herrera, D. B. Work, R. Herring, X. Ban, Q. Jacobson, and A. M. Bayen, "Evaluation of traffic data obtained via GPS-enabled mobile phones: the mobile century field experiment," *Transportation Research Part C: Emerging Technologies*, vol. 18, no. 4, pp. 568–583, 2010.
- [31] J. Wang, X. Kong, A. Rahim, F. Xia, A. Tolba, and Z. Al-Makhadmeh, "IS2Fun: identification of subway station functions using massive urban data," *IEEE Access*, vol. 5, pp. 27103–27113, 2017.
- [32] S. Phithakkitnukoon, T. Horanont, G. D. Lorenzo, R. Shibasaki, and C. Ratti, "Activity-aware map: identifying human daily activity pattern using mobile phone data," in *Human Behavior Understanding*, Springer, Berlin, Germany, 2010.
- [33] F. Oloo, "Mapping rural road networks from global positioning system (GPS) trajectories of motorcycle taxis in sigomre area, Siaya county, Kenya," *ISPRS International Journal of Geo-Information*, vol. 7, no. 8, p. 309, 2018.
- [34] A. Furno, M. Fiore, R. Stanica, C. Ziemlicki, and Z. Smoreda, "A tale of ten cities: characterizing signatures of mobile traffic in urban areas," *IEEE Transactions on Mobile Computing*, vol. 16, no. 10, pp. 2692–2696, 2017.
- [35] M. R. Vieira, V. Frias-Martínez, N. Oliver, and E. Frias-Martínez, "Characterizing dense urban areas from mobile phone-call data: discovery and social dynamics," in *Proceedings of the Second International Conference on Social Computing*, August 2010.
- [36] J. Iounousse, S. Er-Raki, A. El Motassadeq, and H. Chehouani, "Using an unsupervised approach of probabilistic neural network (PNN) for land use classification from multitemporal satellite images," *Applied Soft Computing*, vol. 30, pp. 1–13, 2015.
- [37] G. Pan, G. Qi, Z. Wu, D. Zhang, and S. Li, "Land-use classification using taxi GPS traces," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 113–123, 2013.
- [38] G. Qi, X. Li, S. Li, P. Gang, Z. Wang, and D. Zhang, "Measuring social functions of city regions from large-scale taxi behaviors," in *Proceedings of the International Conference on Pervasive Computing & Communications Workshops*, March 2011.
- [39] H. Dong, M. Wu, X. Ding et al., "Traffic zone division based on big data from mobile phone base stations," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 278–291, 2015.
- [40] J. L. Toole, M. Ulm, D. Bauer, and M. C. Gonzalez, "Inferring land use from mobile phone activity," in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing—UrbComp'12*, Beijing, China, August 2012.
- [41] V. Soto and E. Frias-Martínez, "Automated land use identification using cell-phone records," in *Proceedings of the 3rd ACM International Workshop on MobiArch—HotPlanet'11*, New York, NY, USA, June 2011.
- [42] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining—KDD'12*, Beijing, China, August 2012.
- [43] M. Veloso, S. Phithakkitnukoon, and C. Bento, "Urban mobility study using taxi traces," in *Proceedings of the International Workshop on Trajectory Data Mining and Analysis (TDM) in Conjunction with the 13th International Conference on Ubiquitous Computing ACM*, Beijing, China, September 2011.
- [44] E. Graells-Garrido, O. Peredo, and J. García, "Sensing urban patterns with antenna mappings: the case of Santiago, Chile," *Sensors*, vol. 16, no. 7, p. 1098, 2016.
- [45] J. Ma, S. Wang, R. Jian, and L. Sun, "Using point of interest data from electronic map to predict transit station rideship," in *Proceedings of the Transportation Research Board 93rd Meeting*, Washington, DC, USA, January 2014.
- [46] R. H. Turi and R. Siddheswar, "Determination of number of clusters in K-means clustering and application in colour image segmentation," in *Proceedings of the 4th International Conference on Advances in Pattern Recognition And Digital Techniques*, Calcutta, India, July 1999.
- [47] China Construction Industry Press, *Code for Classification of Urban Land use and Planning Standards of Development Land in Gb 50137-2011*, China Construction Industry Press, Beijing, China, 2011.
- [48] J. D. Zhang and C. Y. Chow, "GeoSoCa: exploiting geographical, social and categorical correlations for point-of-interest recommendations," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval—SIGIR'15*, August 2015.
- [49] T. Pei, S. Sobolevsky, C. Ratti et al., "A new insight into land use classification based on aggregated mobile phone data," *International Journal of Geographical Information Science*, vol. 28, no. 9, pp. 1988–2007, 2014.
- [50] S. Hu and L. Wang, "Automated urban land-use classification with remote sensing," *International Journal of Remote Sensing*, vol. 34, no. 3, pp. 790–803, 2013.