Where You Really Are: User Trip Based City Functional Zone Ascertainment

Jiayu Gan, Jianhui Zhang*, Siwen Zheng
College of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018 China
*Corresponding author Email: jhzhang@ieee.org

Abstract—As one convenient and green transportation mode, Public Bike System (PBS) offers increasing service. It has generated a great mass of transit records as a famous application around the world. To mine the trip information in the records is an effective way to understand the city from some aspects such as the city functional zone distribution and the crowd mobility. According to the observation from daily life, it’s quite often found that some users are willing to shift their bikes at some intermedia stations so as to decrease charge. The phenomenon indicates that the drop-off stations in the system records may not be their real destinations. Focusing on this phenomenon, this paper proposes a new method named User trip based city functional Zone Ascertainment (UZA) to redefine the city functional zone. This paper divides the city into adjacent and non-overlapping regions by the Voronoi Diagram method, and discovers the topic feature distribution of each region based on the user trips. By clustering, we then integrate regions into zones and finally ascertain their functions. Extensive experiments are conducted on the real dataset collected from Hangzhou PBS, consisting of more than 23 million transit records. Compared with the Hangzhou city planning graph, the result indicates our method is reliable and effective.

Index Terms—Public Bike System; City Functional Zone; Voronoi Diagram

I. INTRODUCTION

With the raising awareness of environmental protection, lots of users choose public bikes as their daily transportation tool. Public Bike System (PBS) offers a convenient and additional option to users in the increasingly crowded city. It provides a powerful way to the first-mile and last-mile problem on the city traffic and is a vivid example of the combination of low-carbon and convenient transportation. Nowadays, many PBSs are deployed in some famous cities, such as Hangzhou and Chicago [1]. As the amount of public bike utilization increases, the wide bike rentals generate a mass of transit records. It also incurs lots of research on this interesting topic. Some existing works mainly concentrate on the prediction [2] [3], trip planning [4] [5] or redistribution problems [6] in PBS. However, the user behaviors inside the original transit records are usually ignored by the previous works [7] [8]. It is universally accepted that users are always at the core position throughout the service of PBS. This paper thus pays more attention to the user behavior rather than the records themselves.

In real transportation, it does not necessarily mean the end of a trip when the user returns the bike to a drop-off bike station. He/She may change to another bike in a short time to avoid the over-time payment during the trip because of the limitation of the free ride time in the most of PBSs. It is considered to be the transfer behavior. By observing the dataset generated by PBS in Hangzhou and the daily life of residents, we find that such behavior does exist a lot. The transfer quantity of some bike stations in a month are listed in Table I. Notice that one transit record corresponds to one bike trip, so it may not represent the user behavior correctly under the transfer behavior pattern.

As the example shown in Figure 1, the user aims to travel from the station A to C, and changes to another bike at the station B in 5 minutes. The trip from A to B and B to C are two bike trips, while the trip from A to C is a user trip. In this case, B is an end of a bike trip rather than the user’s final destination. It is a transfer station. The other case is that the user travels from the station D to E without the transfer behavior. Both of these two cases are the user trip in this paper. Thus, the difference in the final destinations between user trips and bike trips leads to a new perspective for the mobility pattern. From the macro point of view, these phenomena may cause a new layout of the city function.

In recent years, city functional zone ascertainment is a hot research area and an integral part of urban computing
TABLE I
TRANSFER QUANTITY OF SOME STATIONS IN A MONTH

<table>
<thead>
<tr>
<th>Station ID</th>
<th>Transfer Quantity 1</th>
<th>Station ID</th>
<th>Transfer Quantity 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1269</td>
<td>1745</td>
<td>8012</td>
<td>3101</td>
</tr>
<tr>
<td>1351</td>
<td>2567</td>
<td>8013</td>
<td>2612</td>
</tr>
<tr>
<td>4237</td>
<td>1539</td>
<td>8016</td>
<td>3509</td>
</tr>
</tbody>
</table>

To understand the entire city deeply and to ascertain its actual functional zones, we propose a new method named User trip based city functional Zone Ascertainment (UZA). By representing bike/user trips with vectors, UZA discovers the feature distribution based on the Latent Dirichlet Allocation (LDA) model [10] and solves the ascertainment problem.

Actually, we face a challenge to mine the records of PBS. The real records collected from Hangzhou PBS are incomplete. Therefore, we need to do a great deal of preprocessing work before our data analysis. In addition, the huge dataset also brings heavy computing burden to UZA. Even so, the experiment results confirm that UZA performs better than the approaches in the existing works. The contributions of this paper are as follows:

- Trip identification. By considering the transfer behavior, we clearly distinguish the user trip from the bike trip. The former makes it easy for us to understand the user behavior and to clarify user’s actual destination.
- City partition. We divide the city based on the Voronoi Diagram method and public bike stations. The method ensures the reasonability of the partition result.
- New method on the novel perspective. From the perspective of users, we propose UZA and redefine the functional zones of the city based on user trips.
- Experiments on masses of real data. Extensive experiments are conducted under the real-world dataset in Hangzhou PBS. By comparing with Hangzhou city planning graph and the ascertainment result based on bike trips, UZA can accurately discover the function of most regions in Hangzhou.

The remaining of this paper is organized as follows. Section II discusses the difference between the bike trip and the user trip and describes the feature discovery problem. In Section III, this paper introduces UZA in detail. Based on the real dataset from Hangzhou PBS, the experiment setting and ascertainment result are described in Section IV. Then this paper discusses some related works in Section V and briefly draws the conclusion in Section VI.

II. PROBLEM DESCRIPTION

Before ascertaining city functional zones, this section discusses the distinction between the bike trip and the user trip at first and then describes the feature discovery problem, which is the key issue in the process of ascertainment.

A. Preliminary

The real dataset utilized in this paper is collected from Hangzhou PBS, which contains large-scale transit records. When users borrow/return bikes by their smart cards, the system records the information of the card ID, station ID, timestamp, etc. With the information, we define a bike trip as follows.

**Definition 1 (Bike Trip):** A bike trip represents that user $u_i$ borrows a bike from the pick-up station $s^p_i$ and then returns it to the drop-off station $s^d_i$. Let $bt_i$ denote the bike trip and we have:

$$bt_i = (s^p_i, s^d_i) \quad (1)$$

Bike trips are usually utilized to represent user behavior in the existing works [7]. But it is not always the case in real life. To some extent, the bike trip cannot necessarily fully express the user behavior. There are numbers of reasons why users may change to another bike at the intermedia bike stations in a short time. For instance, they may change bikes to avoid the overtime payments. Most PBSs are free of charge in a certain time (e.g. the first 30 minutes of each bike trip) and require user fees for the additional time. Such as the Velib’ system in Paris and the Bicing in Barcelona [1]. Analogously, Hangzhou PBS offers a free ride for the first one hour of each bike trip. If a user takes a ride longer than one hour, it can save charge to change to another bike at one intermedia station. Occasionally, the bike may have a mechanical fault such as being uncomfortable to ride or broken. So the user has to change to another one. Due to the transfer behavior, there may cause a great deal of variation in user actual destinations. Considering this situation, a user trip is defined as follows.

**Definition 2 (User Trip):** A user trip represents that user $u_i$ borrows a bike from the pick-up station $s^p_i$, changes to another bike at the transfer station $s^t_i$ in a short consecutive time and finally returns it to the drop-off station $s^d_i$. Let $ut_i$ denote the user trip and we have:

$$ut_i = \begin{cases} (s^p_i, s^t_i, s^d_i), & \text{transfer} \\ (s^p_i, s^d_i), & \text{otherwise} \end{cases} \quad (2)$$

According to the Definition 2, the user trip is equivalent to the bike trip in the case of no bike transfer. The latter thus is a special case of the former.

B. Feature Discovery Problem

Only by knowing the features of things can we identify them clearly. Thus, the key issue in the functional zone ascertainment problem is to discover the feature distribution.
We firstly divide the city into regions and utilize the LDA model to discover the feature distribution of each region. Based on that, the functional zone ascertainment problem can be solved along.

LDA is a kind of document topic generation model and is also known as the hierarchical Bayesian model, which is proposed by Blei et al. [10]. It contains the three-layer structure of words, topics, and documents. Based on the unsupervised machine learning technique, LDA identifies latent topic information in the large document collections by treating each document as a word frequency vector. It transforms the textual information into the easy-to-model digital information. After the training and the iteration, the model generates a topic distribution of each document and the word distributions of all topics. The process of LDA can be stated as follows.

1. Choose \( N \sim \text{Poisson}(\xi) \).
2. Choose \( \theta \sim \text{Dir}(\alpha) \).
3. For each of the \( N \) words \( w_n \):
   (a) Choose a topic \( z_n \sim \text{Multinomial}(\theta) \).
   (b) Choose a word \( w_n \) from \( p(w_n|z_n, \beta) \) and a multinomial probability conditioned on the topic \( z_n \).

As shown in Figure 2, we utilize the plate notation to display the LDA model. The circle with shadow is an observable variable and other circles are the latent variables. The link between two circles indicates the conditional dependence. \( N \) and \( M \) at the top of the plates indicate the sampling repeated times of variables inside [11]. \( \alpha \) and \( \beta \) are hyper-parameters. To be specific, \( \alpha \) is a \( T \)-vector of the Dirichlet distribution on topic and \( \beta \) is a \( T \times V \) matrix of topic-word probabilities. \( \theta \) in this model is the probability of the topic \( z \) appearing in each given document and \( \varphi \) is the probability of the word \( w \) appearing in each given topic.

In this paper, each region is mapped to a document, which will be detailedly described in Section III. Then we respectively formulate words based on bike/user trips and discover the document-topic distribution which we call the feature distribution of each region.

Most symbols utilized are displayed in Table II.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( u )</td>
<td>Each user</td>
<td>( p )</td>
<td>Mobility preference</td>
</tr>
<tr>
<td>( s )</td>
<td>Bike station</td>
<td>( w )</td>
<td>Word topic</td>
</tr>
<tr>
<td>( bt )</td>
<td>Bike trip</td>
<td>( z )</td>
<td>Document topic</td>
</tr>
<tr>
<td>( ut )</td>
<td>User trip</td>
<td>( N )</td>
<td># of words in a document</td>
</tr>
<tr>
<td>( r )</td>
<td>City region</td>
<td>( M )</td>
<td># of documents</td>
</tr>
<tr>
<td>( t )</td>
<td>Time slot</td>
<td>( T )</td>
<td># of topics</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>Status tag</td>
<td>( V )</td>
<td>Vocabulary of all words</td>
</tr>
</tbody>
</table>

III. DESIGN OF UZA

In this section, we describe UZA in detail. To achieve the purpose of ascertaining city actual functional zones, we propose UZA with three stages. In the first stage, we divide the entire city into small regions. After that, we utilize the mobility preference deduced from bike/user trips to discover the feature distribution. At last, we ascertain functional zones and compare those with the result based on bike trips. The framework of UZA is shown in Figure 3.

A. City Partition

As we all know, the regionalization plays an important guiding role in understanding the rational planning of cities [12] [13]. Thus, we firstly divide the entire city into regions. It is necessary for several reasons. Chiefly, the destination of each user is within a region around the public bike station. It is reasonable to guess that the user will go to a nearby place after returning his/her bike, which means the user behavior is regional and the city partition is meaningful. Besides, we divide the city into regions based on the divide-and-conquer strategy. The strategy divides a complex problem into two or more of the similar sub-problems. In this way, we effectively simplify the functional zone ascertainment problem. But there comes an issue that how to reasonably divide the whole city.

Some state-of-the-art methods divide the city depending on the spatial granularity of the collected data [14] or Point of Interest (POI) [15]. J Yuan et al. choose the raster-based model to represent the road network and utilize morphological image processing technique to address the task of city partition [9]. However, in real life, some drawbacks may exist in these methods. For example, there are numerous bike stations being set around POIs in most areas of the city, such as large shopping plazas and restaurants. But for some areas, there may not be as many emblematic buildings around one bike station. Due to the different supply and demand in different areas, the station distribution is not balanced. Thus, dividing the city with POIs may lead to the imbalance of data. Besides, it is very likely that some of the regions divided by the road network may not contain any bike stations, which makes it impossible for us to infer the mobility preference of those regions.

To solve the problem, we utilize the Voronoi Diagram method to divide the entire city. The method divides a plane into different cells based on a set of specific points [16]. We call the specific points seeds. The Voronoi Diagram method has been widely utilized in many areas. Based on the bus trajectory data, Jiwon Kim et al. utilize the Voronoi Diagram method to divide the urban network into spatial cells [17] and
analyse network-wide traffic patterns. Besides, based on the Voronoi Diagram method, Junming Liu et al. divide the bike station in the service area [7] and Feng Liang et al. segment road network [18].

With the Voronoi Diagram method, we divide the city with public bike stations. Each bike station is a seed in the region. Figure 4 shows a part of the partition result. The black points represent the bike stations and different cells represent different regions. The method ensures one bike station in each region and avoids the aforementioned problem. It makes the city partition result more natural and reasonable.

B. Feature Discovery

After dividing the city, we infer the mobility preference among regions and discover the feature distribution of each region. As we discussed above, both bike trip and user trip are directional (from $s^p_i$ to $s^d_i$). By taking the characteristic of bike/user trips and transit time into account, we utilize them to represent the mobility preference. In simple terms, each bike/user trip in a certain time slot is a kind of bike/user mobility preference. So we give the definition as follows.

Definition 3 (Mobility Preference): The mobility preference $p$ of the current region is the travel behavior in a given time slot, which consists of three elements: transit region ID $r_m$, time slot $t_i$, and status tag $\lambda$.

Then we have the following expression:

$$p = (r_m, t_i, \lambda)$$  \hspace{1cm} (3)

Since the bike station of each region is unique, $r_m$ can be obtained by the pick-up/drop-off station ID in bike/user trips. $\lambda$ is a status tag, which indicates the moving direction of a bike/user trip. There are two cases. If $\lambda = L$, the expression means that the user leaves the current region to region $r_m$ during the $i$th time slot. Similarly, if $\lambda = A$, it means that the user comes from region $r_m$ and arrives at the current region during the $i$th time slot. In this paper, we set two hours as an interval and divide a whole day into 12 time slots. The first time slot starts from 7:00 A.M. to 9:00 A.M.

For a region, we separate the mobility preference of bike/user trips into two tables: the arriving table and the leaving table. The arriving table shows the mobility preference from other regions to the current region in different time slots and the leaving table is on the contrary. Figure 5 takes two tables of the user trips as an example. Each column is a transit region ID and each row is a time slot. Specifically, the number of the first grid in arriving table means there are 10 trips from $r_1$ to the current region during the time slot $t_5$. If the number in the grid is zero, it means there has no trip during this time slot. By combining the two tables together, we call it a document in the LDA model. The $p$ mentioned above represents a word in each document. In other words, each region is a unique document, and each grid in the document is a word. The number in each grid is the word frequency. Accordingly, we turn the feature discovery problem to the topic discovery problem with the LDA model.

We train the LDA model with the bike/user mobility preference of each region and then obtain its document-topic distribution. The document-topic distribution is a $T$-dimension vector, and $T$ is the quantity of topics. We call each vector a feature distribution.

C. Functional Zone Ascertainment

Due to the development of urbanization, the functions of the city are actually mixed and complex. We cannot characterize functional zones solely on the basis of the largest proportion topic of the feature distribution in each region. Considering that, we further cluster regions based on the feature distributions. There are always some regions where they have the similar feature distributions. The purpose of clustering is to bring them together and form a functional zone. In this regard, we give the definition of a functional zone.

Definition 4 (Functional Zone): A functional zone consists of one or more regions, which have the similar feature distributions.

We utilize K-means to cluster the feature distribution. In order to select an appropriate $K$ and to evaluate the clustering result, the average silhouette coefficient is set as the evaluation parameter.

Silhouette Coefficient: Silhouette coefficient takes full account of the tightness within the cluster and the separation between the clusters [19], which contributes to the interpreta-

![Fig. 4. Part partition result of Hangzhou. The black point in each cell is the bike station in each region.](image1)

![Fig. 5. An example of the arriving table and the leaving table based on the user trips. They respectively represent the mobility preference of users who arrive at/leave the current region in different time slots.](image2)
tion of the cluster analysis result. The tightness \( a(i) \) is denoted by the average distance of \( i \) to all other samples in the same cluster. And the separation \( b(i) \) is denoted by the average distance of \( i \) to other clusters’ samples. Thus, we have the equation of the silhouette coefficient \( s(i) \):

\[
s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}
\]

(4)

If the cluster \( A \) contains only one object \( i \), the \( s(i) \) will be set to zero. From Equation 4, it can be easily observed that \(-1 \leq s(i) \leq 1\). If \( s(i) \) is close to 1, then the clustering result of sample \( s(i) \) is reasonable. If \( s(i) \) is close to -1, then the sample \( i \) is wrongly clustered and should be classified into another cluster. We utilize the average silhouette coefficient of all samples as the parameter of clustering result and select an appropriate quantity of clusters based on that.

IV. EXPERIMENT

In order to verify the effectiveness of UZA, the extensive experiments are conducted based on the real dataset of Hangzhou PBS. We provide an analysis of the dataset at first and then show the experiment result.

A. Dataset

In this paper, the experiment data collected from Hangzhou PBS cover three months (from late March to late June in the year of 2014). We have a total of 3123 bike stations with over 23 million transit records. Table III shows a brief schema of the dataset.

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>card_no</td>
<td>Card ID of each user</td>
</tr>
<tr>
<td>lease_time</td>
<td>Timestamp when user borrows the bike</td>
</tr>
<tr>
<td>lease_station</td>
<td>ID of pick-up station</td>
</tr>
<tr>
<td>return_time</td>
<td>Timestamp when user returns the bike</td>
</tr>
<tr>
<td>return_station</td>
<td>ID of drop-off station</td>
</tr>
</tbody>
</table>

The dataset is so large that it brings us a challenge. As we know, the transit records are collected from millions of users, but the value density of the data is inversely proportional to the data size. Even some of the data are incomplete. So before the experiments, we carry some preprocessing work and make an analysis of the data.

By observing the transit records from the database, we find that most of those whose trip time lasts less than 1 minute usually have the same pick-up station and drop-off station. The reason for this phenomenon may be due to the user is not familiar with PBS. After the user swipe his smart card, the bike may not be ridden away timely or the user may do some wrong operations. It makes him swipe again in a short time. Besides, some records that last for an extended period may be abnormally recorded. Thus we remove the records with too long or too short trip time and keep those whose trip time is from 1 minute to 5 hours. Moreover, we correct some error records in the database and delete some incomplete records.

After that, we further analyse the transit records. As shown in Figure 6, the graphs effectively demonstrate the characteristics of Hangzhou PBS. From the Cumulative Distribution Function (CDF) in Figure 6(a), it can be observed that the duration for most trips ranges from 1 minute to 60 minutes. The peak of the trip duration approximately appears at 10 minutes, which is displayed in Figure 6(b). As mentioned earlier, we have a one-hour-free ride time, and users usually have public bikes for short distance trips. Such as solving the last-mile problem from the bus station or the metro station to the workplace. After the preprocessing, we have over 17 million transit records for bike trips. Following the consideration that everyone has up to one transfer station in this paper, we choose 5 minutes as the transfer time interval and get over 15 million user trips.

![Image](image_url)

Fig. 6. Transit records analysis

B. Experiment Result

Based on the Voronoi Diagram method and the location of public bike stations, we divide Hangzhou into 3123 regions. To compare the ascertainment results of user trips and bike trips, we respectively sort out the mobility preference for each of these two cases. The model is trained with \( T \) topics for 3000 iterations, and the parameters are optimized every 10 iterations. To find a reasonable result, we test \( T \) from 1 to 10. After that, we utilize K-means to cluster and compare the silhouette coefficient under different quantity of topics. Respectively based on bike trips and user trips, the comparison results are given from 7 to 10 topics.

According to the Figure 7, we can observe that each silhouette coefficient value is maximized when \( K \) equals to the topic quantity. In Figure 7(a), the maximum silhouette coefficient of bike trips and user trips are similar. But in other figures, they all have some differences. However, the somewhat weaker performance of user trips proves the diversification of urban development. City functions are no longer as simple as they used to be. Instead, many regions have a great variety of functions. With the transfer behavior, user trips can really reveal their final destinations. It is no longer mechanically based on the two bike stations in the record as the starting and ending points for a trip. User trips of this real sense bring us a more authentic feature distribution.
Equation 4, the difference between $b(i)$ and $a(i)$ of a user trip is smaller than that of a bike trip. It means the feature distribution trained from user trips are more similar than that trained from bike trips. The similarity between two feature distributions brings some difficulties to K-means. In this case, we fit the results of different topics on the real map and finally select 8 topics.

Figure 8 shows the functional zones under two types of trips. It displays the results for the core area of Hangzhou. To ascertain its functions, we follow two principles: the word distribution of each topic and the real situation in Hangzhou. As we can see, there are some blank areas in the results. After verification, we find that we do have 3123 regions. But the dataset only contains the transit records for most of them, and some of the stations have no records. Therefore, it only discovers 2442 of these regions. Based on the result in Figure 8(a), we reveal the following six functional zones, and the color disk shows the ID of each topic.

**Transportation Hub:** $z_0$. This zone contains the largest railway station in Hangzhou, which is also one of the largest transportation hubs in Asia. However, the feature of its word distribution is not as obvious as we expect. There are millions of people choosing to travel by railway every day. It is more complicated to ascertain its function. Besides, most people will travel with luggage. They may not choose public bikes as their transportation. For the above reasons, we ascertain its function mainly by comparing it with the map and the real situation in Hangzhou.

**Industrial Park:** $z_2$. We define $z_2$ as the industrial park. In real life, this zone mainly has two parts. The larger part contains a number of well-known internet companies. And the smaller part contains some places like the electric industrial park, logistics companies, etc. In addition, according to the top 100 frequent words in each topic, we draw the word-time distributions as shown in Figure 9. The $x$-axis represents the
time slot and the y-axis represents the word frequency. Figure 9(a) shows that most users in this zone go to work and get off work during the normal time slot, but some users will work overtime until late.

**Educational and Residential Area:** \(z_3\). \(z_3\) is a mixed topic of the educational park and residential area. To our knowledge, it has the largest high-tech educational park in Zhejiang Province, which contains about 15 universities [20]. Besides, due to the advantage of proximity to schools, many people choose to live around. Figure 9(b) shows a clear regularity of \(z_3\). Users in this zone often use public bikes in the morning and evening rush hours.

**Tourist Attraction:** \(z_4\), \(z_7\). It is well known that the West Lake is a famous tourist attraction in Hangzhou. In our ascertainment result, it belongs to \(z_4\). The users in this zone are mainly tourists. As Figure 9(c) shows, the word frequency distribution is more uniform. It can be understood as the situation that users may visit the scenic spot from 7:00 A.M. to 21:00 P.M. Comparing with the Figure 9(d), \(z_4\) and \(z_7\) have the similar features. In the real life, \(z_7\) also contains some tourist attractions, but they are not as famous as the West Lake. People who visit here may also include some local residents.

**Daily Living and Cultural Area:** \(z_1\), \(z_5\). According to the word frequency distributions, the bike usage time of these two topics is more well-proportioned. We thus call this zone the daily living and cultural area. The regions in topic \(z_1\) are close to the center of the city where many residents live and work nearby. The regions in topic \(z_5\) are also a large living area, which contains some museums and cultural heritage.

**Entertainment and Commuting Area:** \(z_6\). This zone contains some comprehensive shopping malls and workplaces, such as the Wanda Plaza and the INTIME City. We also find that the time for using public bikes in this zone is more concentrated.

It should be noted that the same color in Figure 8(a) and 8(b) does not necessarily mean the same function. Even so, we can see that there are some differences between two sub-figures in Figure 8. For instance, in Figure 8(b), it misclassifies a small part of the West Lake to another functional zone. It indicates that the ascertainment result based on bike trips may not be applicable to all kinds of functions. Some regions in the figure cover a large area, which is caused by the following two reasons. (i) It is caused by the nature of the Voronoi Diagram method. In Hangzhou, the distribution of bike stations is uneven. Some areas are more dense, while others are sparse. The Voronoi Diagram method will be as much as possible to make the partition result cover the entire city. For some stations, the regions will have a larger proportion of area. (ii) In our database, some bike stations have only a small amount of the transit records. The model may judge the regions of these bike stations as the same topic. Thus this kind of zones have a large area, and it is more challenging for us to ascertain their functions.

We finally compare our result with Hangzhou city planning graph, which proposed by Hangzhou City Planning and Design Academy [21]. The result shows that most of the city functional zones based on user trips are consistent with the graph. For a few zones which are different from the planning graph, we find the following reasons. Taking \(z_3\) as an example, we ascertained it as the educational and residential area, while the planning graph considers it as the industrial area. It is because they have the similar regularity of using public bikes, as shown in Figure 9(a) and Figure 9(b). We can hardly distinguish them completely. It seems that teachers may have the similar work time as commuters. Besides, due to the singleness of public bike data, it also leads to some differences compared with the city planning graph. Even so, we still have most of the correct ascertainment results.

V. RELATED WORK

A. Trajectory Data and Mobility Pattern

In some of the existing works, many scholars have done the research based on the mobility pattern and the urban traffic trajectory data.

[22] identifies tourists by mining mobility patterns of the crowd trajectories in public transportation, such as subways and buses. Based on the data provided by Singapore Land Transport Authority (LTA), they build an interactive web-based system to gain tourists travelling behavior. Tong Xu et al. propose a social-driven two-stage framework to reveal how the social propagation affects for better prediction of cab drivers future behavior [23].

[24] discovers the transit commute pattern of bus passengers by mining a large number of Beijing traffic card data. The mPat (Mobility PATterns) [14] combines the data from public transportation and mobile networks to explore the multi-data source-based mobility pattern.

It can be seen that the trajectory data are valuable and multiple. But in this paper, we focus on the data of PBS.

B. Public Bike System

Andreas Kaltenbrunner et al. analyse the mobility pattern of people in the city with the number of available bikes at stations in the Bicing community bike system of Barcelona [25]. They collect relevant data from the website to discover the use and geographic movement patterns of people in the city.

[26] proposes a statistical model to automate the analysis of transit data for PBS. The model identifies some potential factors that affect the geographical location of a bike trip, and provides suggestion into the relationship between the type of area in which the station is located and their mobility patterns.

Longbiao Chen et al. utilize the published station status data to deduce bike travel patterns and help users find nearby stations and bikes [8]. In the same year, the team proposes a two-stage framework aimed at sensing the social activities of city centers with the open data of PBS [27].

C. LDA Model

[28] designs a data mining approach and utilizes the LDA model to extract the main features of the spatio-temporal behavior of the PBS. In [29], they utilize the LDA model
to identify the key dimensions of customer service voiced by hotel visitors. The dimensions they find are key for hotels to manage their interactions with visitors. In [30], they analyse the large-scale geo-location data from social media to infer individual activity patterns with the LDA model.

In summary, the LDA model is a practical model. There are plenty of successful works based on that. Different from the above work, this paper trains the public bike transit records with the LDA model and ascertains city functional zones. Particularly, we compare the difference between bike trips and user trips on the ascertainment result.

VI. CONCLUSION

In this paper, we ascertain city functional zones based on the real transit records collected from Hangzhou PBS in China. Especially, we differentiate between bike trips and user trips. To solve the functional zone ascertainment problem, we propose a new method named UZA. In UZA, we firstly divide the entire city with the Voronoi Diagram method and discover the feature distribution from each region based on the LDA model. By clustering the feature distribution, the city functional zone ascertainment problem can finally be addressed. Extensive experiments are conducted to verify the performance of UZA. We also compare the result of user trips with that of bike trips and reveal the actual functional zones. The experiments indicate that we have most of the correct functional zones by comparing them with Hangzhou city planning graph. The more accurate result we gained from UZA may contribute to the management and the construction of the city.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (NSFCC) under Grant No.61473109, the Graduate Scientific Research Foundation of Hangzhou Dianzi University under Grant No.CXJJ2018053, and the General Research Foundation of the Education Department of Zhejiang Province under Grant No.Y201840731.

REFERENCES

[18] Liang Feng, Shaoyun Ge, and Hong Liu. Electric vehicle charging station planning based on weighted voronoi diagram. pages 1–5, 2011.