IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 9, 2019, accepted November 20, 2019, date of publication November 25, 2019,
date of current version December 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2955755

Urban Functional Area Division Based on Cell Tower Classification

YUEWEI WU, YUANYUAN QIAO, (Member, IEEE), AND JIE YANG
Research Center of Network Monitoring and Analysis, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding author: Yuanyuan Qiao (yyqiao@bupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61671078 and Grant 61701031, in part by the Beijing Laboratory of Advanced Information Networks of BUPT, in part by the Beijing Key Laboratory of Network System Architecture and Convergence of BUPT, in part by the National Key Research and Development Program of China under Grant 2017YFB0802701, in part by the Fundamental Research Funds for the Central Universities under Grant 2018XKJC04, in part by the Science and Technology on Space Intelligent Control Laboratory under Grant 6142208180205, in part by the 111 Project of China under Grant B17007, and in part by the Data Intelligence Innovation and Entrepreneurship Platform conducted in the Cooperation between the Beijing University of the Post and Telecommunications and China Telecom Corporation Limited Cloud Computing Branch Corporation.

ABSTRACT Each city can be divided into different functional areas for different purposes, such as commercial areas, residential areas, leisure areas, office areas, etc. How to correctly divide these areas is very helpful for urban self-cognition and urban planning. Mobile phones have become almost a must-have for everyday life nowadays in big cities. People carry their mobile phones all the time, making phone calls and surfing on the Internet. For this reason, the Open Information Dynamic Data (OIDD) is a more representative indicator of the mobility characteristics of the population. In this study, we use the user, time and location information extracted from OIDD data to analyze and divide the urban functional areas for the purpose of understanding the regional composition of the city. We use the Latent Dirichlet Allocation (LDA) model and incorporate time and space information into the Dirichlet distribution to participate in the model hypothesis. This allows for the model's high-level analysis capabilities to exploit potential urban area functions through human mobility patterns.

INDEX TERMS Open information dynamic data, human mobility, mode mining, region divide, temporal and spatial analysis.

I. INTRODUCTION

In order to make the planning and construction of the city more conducive to its citizens, the division of urban functional areas is a key point for urban planners to take into consideration [1]. Since human mobility has high regularities [2]–[5], and it can well reflect the functional areas of the city to some extent [6], [7], the study of the human mobility patterns has become a hot topic [8]–[13]. In the past several years, surveys of urban functional areas are mostly relied on field trips. However, with the development of technologies and the increasing popularity of Mobile Internet, usage of mobile phones has become much more widespread. The Open Information Dynamic Data (OIDD) [14] collected in large cities can cover almost all citizens' movement patterns since nearly everyone is using a mobile phone to make phone calls or surf the internet, which makes it a convincing data source to conduct this research. Therefore, in this paper, we use OIDD data to study human mobility patterns, and based on this, we further study the method for dividing urban functional areas. This will provide urban planners with an easy, convenient way to understand the city's structure.

Recent studies have shown that the second, third and fourth generation (2G/3G/4G) mobile Internet's network traffic data is very helpful for studying human mobility [15]–[17]. When users access to the mobile Internet, their locations will be collected. This kind of passive acquisition of trajectory has a lot of advantages. Firstly, it has low cost and high efficiency, and it takes relatively less time and effort. Secondly, it has a wide collection range, capable of covering a large number of people. Finally, it has good temporal characteristics, such as it contains time information, including the time of occurrence and the length of time. This can help us understand the impact of time on crowd behavior. Although the distribution of cellular towers is sparse (average position error is 175 meters [18]), and the density of cellular towers in urban areas is much

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Asif.

larger than that in suburban or rural areas, this margin of error is tolerable for researches in human mobility patterns and division of urban functional areas [14]. In addition, studies have shown that the results based on CDR data analysis are credible [8], [10]. Individual mobility models based on CDR data [19] can even measure the state of dynamic changes in population, activities and environment to some extent [20]–[23].

Since the OIDD data contains time and space information, and we need to learn the user, time and space information in the model training process. In view of this, the two inputs in the traditional LDA model are not suitable. So, in this article, we will evolve our algorithm based on LDA model. The temporal and spatial attributes are added to the original LDA algorithm to obtain our TS-LDA (Temporal and Spatial Latent Dirichlet Allocation algorithm), which can be applied to time and space classification. At the same time, the TCV (Temporal Coherence Value) evaluation algorithm [6] is used to provide an evaluation standard for TS-LDA. In this paper, we use OIDD data from Beijing and Zhuhai as the basic dataset. More specifically, Beijing dataset is used as the main research object and Zhuhai dataset is used to verify whether the algorithm is transferable. After statistical analysis and integration of the OIDD data, we extract time, space and user information required by TS-LDA to test its effectiveness. The algorithm can effectively divide the urban functional areas according to time and space attributes. Therefore, experiments in follow sections prove that the algorithm we propose in this paper is beneficial for urban development planning. The contributions of our work are summarized as follows.

- Our model aims to find urban functional area division method using OIDD-based TS-LDA model. From the perspective of individual number, our data covers the whole 2G/3G/4G mobile networks and nearly 10 million users. On the other hand of geographical extensiveness, the area we analyzed covers nearly 50 thousand cellular towers across the city.
- Our model creatively takes both time and space factors into consideration to divide urban functional areas. On the basis of the original LDA model, we add the analysis of the human movement trajectories [2] as well as the consideration of time factors. In addition, a matching evaluation system [6], [24] is used to modify the LDA model into one that can be applied to OIDD data.
- After successfully completing the urban functional area division in Beijing, we also used the OIDD data of Zhuhai to conduct this experiment again to verify its effectiveness. Beijing and Zhuhai differ greatly from each other in terms of geographical distance, cultural diversity and development level. The success of this model's application in Zhuhai proves that it is not only applicable to a certain region, but also transferable to others.

The remainder of this paper is organized as follows. In Section II, we review some methods about analyzing human mobility from mobile phone data, as well as how LDA algorithm can be applied. In Section III, we introduce the characteristic of the study area we choose and the data we use. In section IV, we propose a new method based on classical LDA algorithm. The detailed hypothetical scenario and model structures are illustrated in Section IV. Furthermore, in Section V, we apply the TSLDA method to actual analysis based on Beijing OIDD data. After discussing about its performance, we use the Zhuhai OIDD data to verify the transferability of the method. Finally, we discuss the operational effects of the method and our future work.

II. RELATED WORK

A. HUMAN MOBILITY

With the rapid development of the Internet, mobile phones are increasingly becoming a necessity in people's everyday lives. Everyone uses mobile phones to make calls, chat, shop, and even work [1], [25], [26]. Therefore, OIDD data covers a lot of crowd movement information as well as spatio-temporal information. Although the spatial information provided by OIDD and CDR data does not incorporate sufficient continuity, many papers still applied CDR data to conduct researches on human mobility characteristics in recent years [2], [8]–[10], [22], [27].

Jiang et al. first extracted and modeled the trajectory features of crowd movement, and obtained six human mobility patterns. She also utilized the position information in the CDR data to extract the trajectory information of individual user. By corresponding them to human mobility patterns, urban functional area distribution of Singapore is finally obtained [2]. Marta et al. observed the human movement trajectories from a mathematical point of view and learned that they have a high degree of temporal and spatial regularity. Each individual's mobility characteristics are temporally repetitive in days or weeks. It also shows a spatial repetitiveness which is independent of time, indicating by the fact that people always visit several constant places [8]. Wang et al. found that there are two types of human mobility problems: Network Proximity and Mobile Homophily. Using mobile phone data, they found that human's online behaviors (network behaviors) and offline behaviors (space trajectories) are correlated, and proposed a method of using people's online behaviors to predict their spatial trajectories [9].

It can be seen that in recent years, the usage of CDR data to research human mobility has become more and more in-depth, and the results are of great help to better implement urban designing and study human mobility more exactly.

B. LDA MODEL

LDA uses the typical “bag-of-words" hypothesis, which is often used for article analysis, and text generation in recent years. Since the model has a good application in the NLP field, it has been widely used in recent years [24], [28]–[32].

Chen et al. innovatively proposed to add a third type of input-time information, to the LDA model [31], and proved that it has better performance over the original one.

Zhou et al. applied this method to the WeChat dataset and proposed the TCV algorithm to evaluate this model [6]. They used WeChat data to analyze the supply and demand situation of urban cultural venues in Beijing and came up with the result that the cultural demand in Beijing's central areas is larger than that in the peripheral areas while cultural patterns change regularly with the time, month and season.

It can be seen that the LDA model has been widely used in various fields and its performance is excellent.

C. URBAN FUNCTIONAL AREA DIVISION

In earlier studies, researchers used cell phone data or mobile trajectory data to directly analyze the results of the division of urban functional areas [33]–[38]. The study of Jameson Toole *et al.* [33] and Zhao *et al.* [34] considered the space and time factors, and they used random forest and VGI algorithms respectively. Furno *et al.* [35] find that mobile traffic signature can provide an evident association of prototypical mobile communication dynamics to precise urban fabrics and verified in ten cities.

Later, researchers began to refer to the topic model for functional area division [39], [40]. The study by Xing *et al.* [39] and the study of Yuan *et al.* [40] explicitly use the LDA algorithm for urban functional area division. Yuan et al. used POI data as the basis for topic partitioning. Hanfa Xing et al. used the LDA model in the process of clustering POI data.

However, original POI data are obtained through traditional methods such as user submission and field survey, which requires a lot of manpower and material resources, and there is still some noise [41]. It can be seen from the Baidu POI data that the amount of POI data in underdeveloped cities is small and inaccurate. But every city must hope that the urban functional area information can be obtained simply, quickly and accurate [37]. So, we want to come up with an algorithm that works for any city.

Meanwhile, now more and more cities have complex functional areas, not only related to the location of the trajectory, but also related to the length of stay and the time of appearance [42]. For example, residential areas and educational areas, people are staying for a long time, but the residential area is night and morning, but the education area is daytime. And the fact that load peaks undergo geographic shifts between precise urban areas throughout the day [36] and during weekday-to-weekend transitions [38]. Therefore, in the study of the division of urban functional areas, the consideration of time factors has become increasingly important.

Thus, in our research, we have tried innovatively to add time information **directly** to the LDA algorithm and tried to propose an algorithm which does not require any data other than mobile data as a support. At the same time, we use POI (Point of Interesting) data as the ground truth [40], and we make an exact show on Baidu Map to verify the correctness of the model. Furthermore, we use multiple datasets to validate our results in order to provide stronger support for our conclusions.

III. PRELIMINARY

In this section, we introduce the two cities in China we studied, the basic information and characteristics of massive OIDD data used in our experiments.

We use Beijing and Zhuhai as two study cases in this paper. Beijing, the capital of China, is a bustling and advanced metropolis with a range of 176 kilometers in the north-south direction and 160 kilometers in west-east direction. It has a population of 21.7 million in 2017, of which 13.6 million are residents and 8.1 million non-residents. It has one of the world's highest mobile penetration rates – above 178 percent. We use the data of Zhuhai to verify the reliability based on Beijing data experiments.

In our study, the functional area division granularity is the cell tower coverage area. For Beijing data, the coverage area of the base station is 0.3282 square kilometers on average, and the density of base stations in urban areas is greater. Therefore, our partition granularity is acceptable in engineering.

STATEMENT Here, we solemnly declare that all data we use is encrypted data. The data does not reveal any user privacy, and we do not research for a specific user.

A. OPEN INFORMATION DYNAMIC DATA

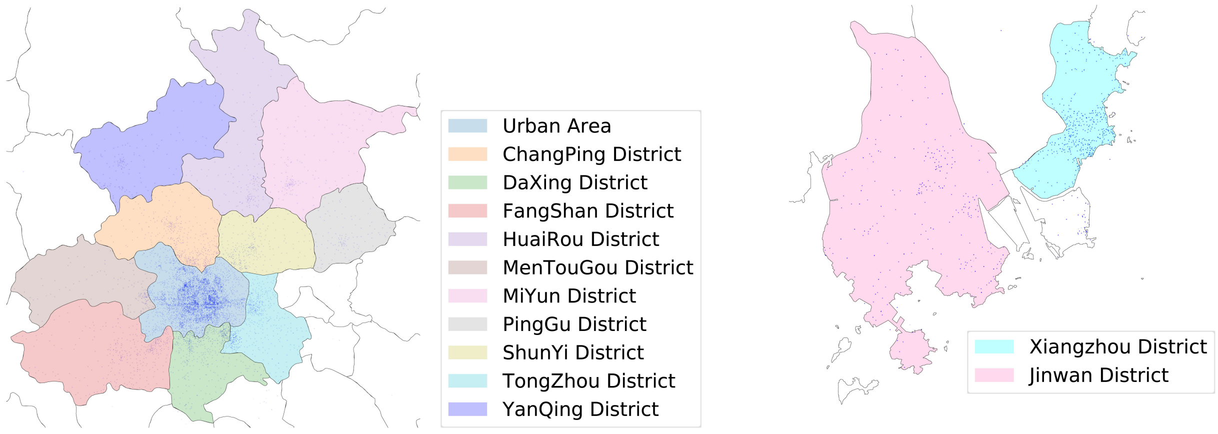
We use three consecutive weeks of mobile phone OIDD data (in March of 2018) from one carrier in China to examine the mobility patterns of anonymous individuals in the metropolitan area. There are nearly 50 thousand cellular towers in Beijing, with a spacing gap of around 100 meters in the urban area to a few kilometers in suburban region. In a word, the cellular tower network has a very high density covering the whole urban area. Therefore, we can use the OIDD data to study human mobility andregion functions. The distribution of cellular towers in all Beijing is shown as Figure 1. In which, the urban area represents the six central districts of Haidian, Xicheng, Dongcheng, Chaoyang, Fengtai and Shijingshan district. All urban areas mentioned below in this paperrepresent these six administrative regions.

1) BEIJING

The data set of the Beijing contains 7.87 million anonymous mobile phone users, and a total of 7.232 billion records of phone usages. An average of approximately 300 million records will be produced and 4 million users will be online within one day period. In which, 8.13 percent users make phone calls, 7.62 percent users send messages and 84.25 percent users merely update their locations. The statistics of the data set are shown in Figure 2.

2) ZHUHAI

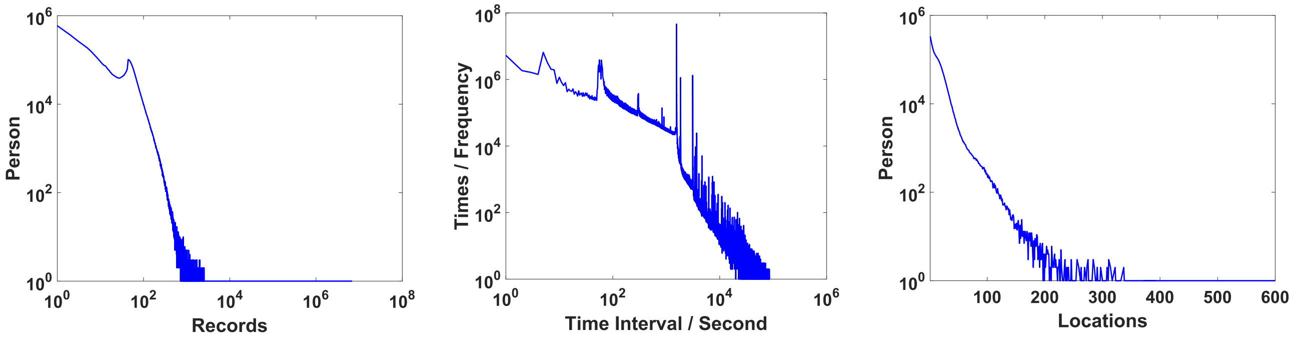
The data set of the Zhuhai contains an average of approximately 30 million records per day, and 600 thousand users will be online within one day period. In which, 7.00 percent users make phone calls, 1.99 percent users send messages and 91.01 percent users merely update their locations.



(a) Beijing

(b) Zhuhai

FIGURE 1. The distribution of cellular towers in two cities. (a) Beijing. (b) Zhuhai.



(a)

(b)

(c)

FIGURE 2. (a) The distribution of number of records for each user in a day. (b) The distribution of length of time interval for each user in a day. (c) The distribution of number of distinct locations that each user visited in a day.

B. POINT OF INTEREST DATA

We use POI information data obtained from Baidu to verify the correctness of our region classification. POI data can show the type of supply service for a location. For example, POI data can indicate whether a corresponding to a latitude and longitude is a residential building or a shopping mall, or a park. It determines the cultural type of a region from a geographical perspective. It can help us to confirm our experiments.

C. REGIONAL ATTRIBUTE DIVISION

In this part, we explore how to divide urban functional areas by human mobility characteristics. Before discussing the classification method any further, we first introduce the basic definitions used in this method.

1) FUNCTIONAL REGIONS

Functional regions based on POI information data can be divide into 18 primary classifications and 139 secondary classifications, as shown in the Table1. And we choose 126 different functional region labels in our research.

2) CROWD AND REGIONAL FEATURE DEFINITION

Here we give a few definitions that need to be used in subsequent models:

- **Functional Modes:** We use OIDD data to divide urban areas. Each category is called a functional mode.
- **User Appears:** Users appear in one of the functional regions are called user appears. It means that this user appears at this location, no matter what action he did, such as making a phone call, sending a message or just refreshing his location.
- **Reliable Crowd:** The number of appears that were screened was sufficient for the experiment. It will be described in detail in Section 5.
- **Passionate Crowd:** To classify some crowd of users. Each crowd has the same habit. For example, one crowd of people like hanging up in shopping malls during afternoon, while another crowd usually work in companies during the day. According to different activities of different groups of people, we can judge the function of the region where a certain group appears at a certain time.

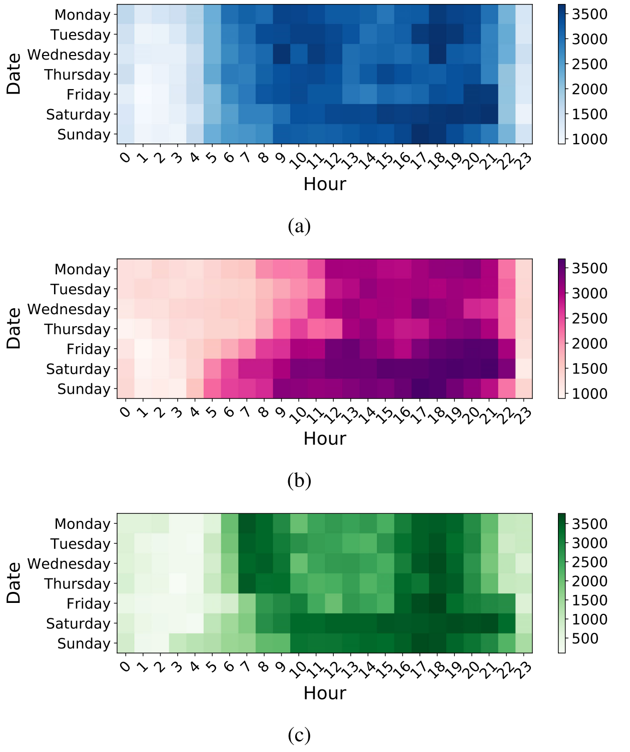
TABLE 1. POI information classification.

Primary Classification	Secondary Classification
Food	Chinese restaurant, snack bar, dessert shop, cafe, etc.
Hotel	Star hotel, express hotel, apartment hotel, etc.
Domestic Service	Communication business hall, post office, ticket office, laundry, graphic fast printing shop, photo studio, pet service, etc.
Beauty	Hairdressing, manicure, etc.
Tourist Attraction	Parks, zoos, botanical gardens, amusement parks, museums, aquariums, bathing beaches, heritage sites, scenic areas, etc.
Leisure and Entertainment	Resorts, farmyards, cinemas, KTV, theaters, dance halls, internet cafes, gaming venues, bathing massages, etc.
Work Out	Sports venues, extreme sports venues, fitness centers, etc.
Education	Institutions of higher learning, middle schools, primary schools, kindergartens, adult education, parent-child education, etc.
Media	Press and publication, radio and television, art groups, art galleries, exhibition halls, etc.
Medical	General hospitals, specialist hospitals, clinics, pharmacies, nursing homes, emergency centers, disease control centers, etc.
Car Service	Car sales, car repair, car beauty, car rental, etc.
Transportation Facilities	Airport, railway station, subway station, bus station, parking lot, service area, toll station, roadside parking space, etc.
financial	Banks, ATMs, credit unions, pawn shops, etc.
real estate	Office building, residential area, dormitory, etc.
Corporate	Company, agriculture, factory, etc.
Government	Central institutions, public security agencies, foreign-related institutions, welfare agencies, etc.
Entrance and Exit	Highway entrance and exit, airport entrance and exit, station entrance and exit, parking lot entrance and exit, etc.
Nature	Island, mountain, water system, etc.

D. TIME PERIODICITY OF DATA

If human movement has a certain time regularity, then we will have the opportunity to use this law to study the laws of human movement to determine the urban functional area. Thus, in this part, we introduce the time regularity of human movement.

As we usually know about human activities, the activities of the crowd are time-related and have a certain periodicity. We counted the number of users appear at different locations in an administrative week on an hourly basis. In order to have the best display effect, we have unified the normalization of the number of people on the line to ensure that the threshold of each picture is between 0-3500. From the overall picture of these three figures, the number of people going online in the early morning is very small, and the user's online time is mainly concentrated from 6 to 22 o'clock. From the perspective of each picture, figure 3 (a) shows the number of users appear in the urban area of Beijing. Figure 3 (b) shows the number of users appear in shopping, food and entertainment region like Table 1 shows. Figure 3 (c) shows the number of users appear in Transportation facilities and entrance and exit region like Table 1 shows. In Figure 3 we can see that, from 7 am to 9 am and from 17 pm to 19 pm, there are a large number of users on the line near the transportation facilities. This is because during this time, everyone is on the way to or off work, which is the morning and evening peak hours. During the weekends (Friday and Saturday), the number of users appear in the shopping and entertainment region significantly exceeds that in working days, and the number of people appearing near the restaurants at noon and night on weekdays is more than work time. These show that the OIDD data is time-recurring and the data structure depends on the time. Therefore, taking advantage of the time factor when using the OIDD data will help improve the result.



(a)

(b)

(c)

FIGURE 3. The number of users appear in a certain region. (a) Urban area of Beijing. (b) Shopping, food and entertainment region. (c) Transportation facilities and entrance & exit region.

E. SPATIAL-TEMPORAL REGULAR HUMAN MOBILITY

If human movement also has spatial fixation, we will have the same opportunity to start from the law of space and find the law of human movement. So, in this part, we introduce the spatial-temporal regularity of human movement.

In our life, although each individual has differences, the activities of most people have certain regularity [2].

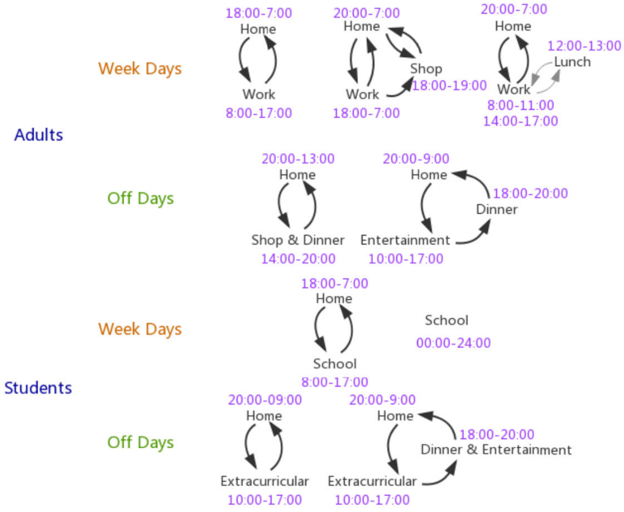


FIGURE 4. Common human movement.

We found that most of the crowd's flow patterns are shown in Figure 4. For adults who already work, during weekdays, people usually travel between home and office units. Most of them generally stay at home in the evening, work at the company during the day, and visit stores or shops during off-hours. During weekends, they choose to stay at home or go shopping and dining at the mall. However, for students in primary, middle, or high school, they travel regularly from school and home every weekday. Even some students living in school, they spend nearly all their time in school. On weekends, students will attend some extracurricular activities and class, or do some entertainment. Therefore, the time and place of concentration of the crowd can reflect the function of the area to a certain extent.

IV. METHODS

A. TEMPORAL AND SPATIAL LATENT DIRICHLET ALLOCATION

The traditional LDA algorithm is based on the 'bag-of-words' assumption, which is mostly applied to NLP. In recent years, the LDA algorithm has proven to be a very effective unsupervised learning model in text learning. When we migrate the LDA algorithm to the problem in this paper, we need to consider the time-space-dependent nature of the OIDD data. Therefore, we add time factor and space factor to the LDA model. Our TSLDA model inherits from the traditional LDA model, and the difference between them is shown in Figure 5, and the structure of TSLDA is shown as Figure 6. In which, α and θ_u represent modes-user layer, η and φ_k represent region-modes layer, δ and λ_t represent modes-time layer. The specific distribution will be introduced in the next subsection.

B. DISTRIBUTION ASSUMPTHON BASED ON LDA

In classical LDA algorithm, the binary problem uses a binomial conjugate distribution, that is β distribution. While in the multi-classification problem, the Dirichlet distribution hypothesis is used. Therefore, as an algorithm for solving multi-classification problems, the TSLDA algorithm continues to use the Dirichlet distribution. K-dimensional multi and Dirichlet distribution in the general sense are as follow.

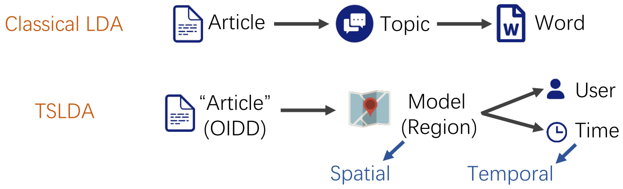


FIGURE 5. Model inheritance. The difference between classical LDA and TSLDA.

$$multi(\vec{m}|n,\vec{p}) = \frac{n!}{\prod_{k=1}^{K} m_k!}\cdot\prod_{k=1}^{K} p_k^{m_k} \tag{1}$$

$$Dirichlet(\vec{p}|\vec{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\cdot\prod_{k=1}^{K} p_k^{\alpha_k-1} \tag{2}$$

In which, m represents the number of samples in each category, p is the corresponding probability.

We confirm the distribution like figure 7. Suppose the mode's prior distribution based on each user is a Dirichlet distribution.

$$\theta_u = Dirichlet(\vec{\alpha}) \tag{3}$$

In which, $\vec{\alpha}$ is Dirichlet prior distribution of modes based on each user.

As the same, suppose the time's prior distribution based on mode is a Dirichlet distribution.

$$\lambda_t = Dirichlet(\vec{\delta}) \tag{4}$$

In which, $\vec{\delta}$ is Dirichlet prior distribution of time based on each mode.

Suppose the user's prior distribution based on mode is a Dirichlet distribution.

$$\phi_k = Dirichlet(\vec{\eta}) \tag{5}$$

In which, $\vec{\eta}$ is Dirichlet prior distribution of time based on each mode.

For any user at a certain time, $z_{u,t}$ is obeys the $multi(\theta_u, \lambda_t)$ distribution.

$$z_{u,t} = multi(\vec{\theta}_u, \vec{\lambda}_t) \tag{6}$$

$$p(z_{u,t}|\alpha,\delta) = \sum_{\vec{\theta}_u} p(z_{u,t}|\theta_u)p(\theta_u|\alpha)\cdot\sum_{\vec{\lambda}_t} p(z_{u,t}|\lambda_t)p(\lambda_u|\delta) \tag{7}$$

And then, we can know the relationship between $w_{u,t}$ and other parameters like following equation.

$$\begin{aligned} p(w_{u,t}|z_{u,t},\eta) &= \sum_{\vec{\phi}_k} p(w_{u,t}|\phi_k)p(\phi_k|\eta) \\ &\quad\cdot\sum_{\vec{\theta}_u} p(z_{u,t}|\theta_u)p(\theta_u|\alpha)\cdot\sum_{\vec{\lambda}_t} p(z_{u,t}|\lambda_t)p(\lambda_u|\delta) \end{aligned} \tag{8}$$

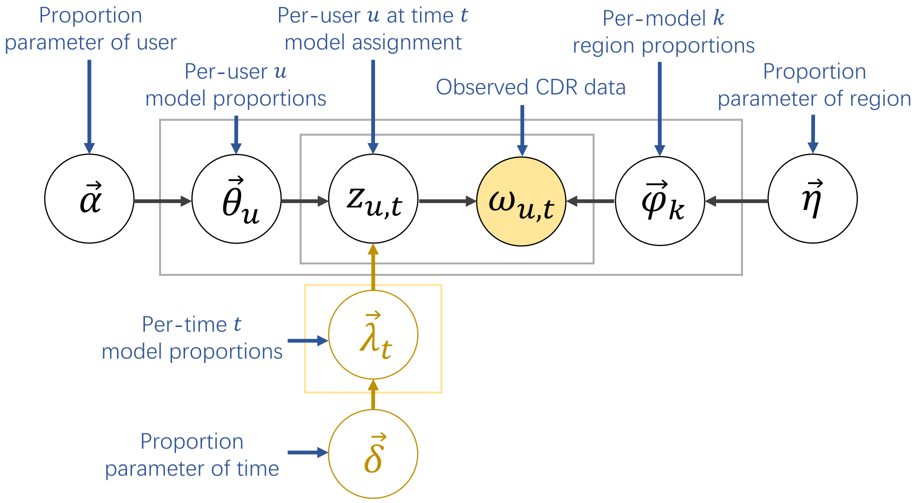

Proportion parameter of user
Per-user u at time t model assignment
Per-model k region proportions
Per-user u model proportions
Observed CDR data
Proportion parameter of region
$\vec{\alpha}$
$\vec{\theta}_u$
$z_{u,t}$
$\omega_{u,t}$
$\vec{\varphi}_k$
$\vec{\eta}$
Per-time t model proportions
$\vec{\lambda}_t$
Proportion parameter of time
$\vec{\delta}$

FIGURE 6. TSLDA model structure.

$$\vec{\alpha} \xrightarrow{\text{Dirichlet}} \vec{\theta}_u \xrightarrow{\text{Multinomial}} z_{u,t}$$
$$\vec{\delta} \xrightarrow{\text{Dirichlet}} \vec{\lambda}_t \xrightarrow{\text{Multinomial}} z_{u,t}$$
$$\vec{\eta} \xrightarrow{\text{Dirichlet}} \vec{\varphi}_k \xrightarrow{\text{Multinomial}} \omega_{u,t}$$

FIGURE 7. Distribution relationship between model parameters.

Afterwards, we use Gibbs sampling as following equation.

$$\hat{\theta}_{uk} = \frac{n_{u,\neg i}^{(k)} + \alpha_k}{\sum_{k=1}^{K}(n_{u,\neg i}^{(k)} + \alpha_k)} \tag{9}$$

$$\hat{\phi}_{kl} = \frac{n_{k,\neg i}^{(l)} + \beta_l}{\sum_{l=1}^{L}(n_{k,\neg i}^{(l)} + \beta_l)} \tag{10}$$

$$\hat{\lambda}_{uk} = \frac{n_{t,\neg i}^{(m)} + \beta_m}{\sum_{m=1}^{M}(n_{t,\neg i}^{(m)} + \beta_m)} \tag{11}$$

Finally, we can get the probability distribution of the mode-user, mode-time and region-mode, like the three layers shows.

C. TEMPORAL AND SPATIAL CORRELATION CALCULATION

After we add time and space factors into the LDA model, we will find that the original LDA model evaluation parameters have failed. Therefore, in order to find the appropriate number of classification modes K, we need to design an evaluation system based on the output of TSLDA to measure its effect. Firstly, we should make clear that the output of TSLDA has three parts, that is the distribution of mode-user (top user U* in each modes), mode-time (top time period T* in each modes) and the top region categories V*. Secondly, our goal is to find a best classification mode K, so that in the case of K, the distribution of modes and time has the greatest degree of similarity to the distribution of modes and users. We use Temporal Coherence Value (TCV) algorithm and TCV score which is proposed in [24], [43]. The input data used in TCV is constructed by a sliding window which moves over the original appear of all users.

$$S_{set} = \{(v^*, V^*, T^*) | v^* \in V^*\} \tag{12}$$

$$\vec{w}(j) = NPMI(v^*, t_j^*)^{\tau} \tag{13}$$

$$\vec{W}(j) = \sum_{i=1}^{I} NPMI(v^*, t_j^*)^{\tau} \tag{14}$$

The TCV score is like the following equitation.

$$m_q = cos(\vec{w}_q, \vec{W}_q) \tag{15}$$

$$\hat{m} = \frac{\sum_{q=1}^{Q} m_q}{Q} \tag{16}$$

V. EXPERIMENTS

In this section, we will use the OIDD data for experiments to divide Functional Modes. Firstly, we use Beijing urban area OIDD data as a case to do this experiment and use the POI data of Beijing and map visualization to verify the accuracy of the experiment. The total amount of data reaches 12TB. After that, we will use Zhuhai OIDD data to do the same experiment, to demonstrate the transferability of experimental methods.

A. DATA PREPROCESSING

Due to the presence of foreigners in Beijing (such as travel, medical treatment, etc.), the number of User Appears for some individual users in the OIDD data is very low, so that their behavior cannot support the experiment to reach a conclusion. The existence of these kind of data will increase the burden of the algorithm, so we first filter out these unreliable data, and the users obtained after filtering become Reliable Crowds. As same, some cellular towers receive very little

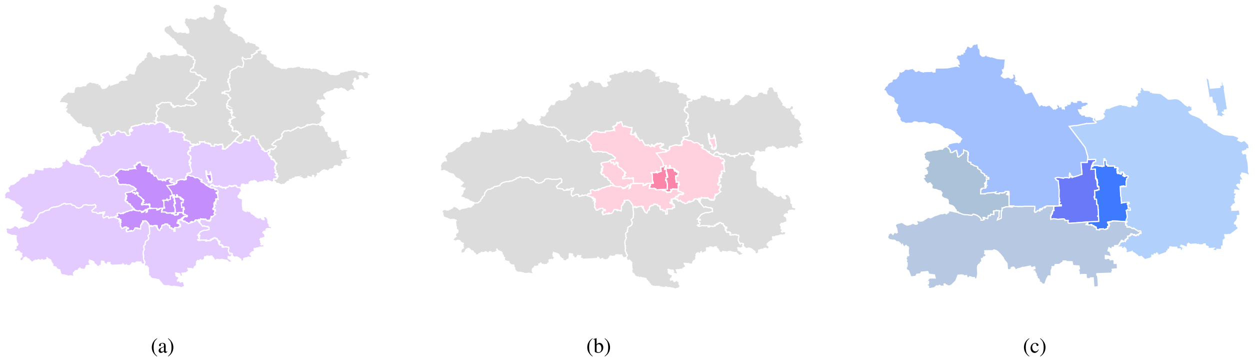


FIGURE 8. (a) The number of Reliable Cellular Towers. (b) The number of Reliable Crowds. (c) The final region we choose for follow experiments.

information due to regional remoteness or functional problems. These cellular towers are not enough to support the experiment too. So, the unqualified cellular towers are also filtered out. However, the data is densified, which means we cannot know the information of a user, so we need to develop a rule to judge the Reliable Crowds. Our rules for screening Reliable Crowds and Cellular Towers are as follows.

1) REMOVE OSCILLATION

When the user crosses the coverage area of the two cell towers, the user-generated data record will continuously jump between the two cell towers. We call this phenomenon ''Oscillation''. Oscillation can cause data redundancy and can have a negative impact on subsequent experiments. So, we first use the SOL algorithm to remove Oscillation. We used the SOL algorithm proposed in the research of Ling *et al.* [44] for data preprocessing. After removing Oscillaion, 10.2TB of valid data was obtained, which is about 85.6% of the total data.

2) FILTER RELIABLE CROWDS

Based on section 4.3, we divide every day into four time periods, 7:00-12:00, 12:00-17:00, 18:00-22:00, 22:00-7:00(+1). A user can be classified as Reliable Crowds if and only if a user appears once in each time period of the day and meets the requirements for at least one week in a row.

3) FILTER RELIABLE CELLULAR TOWERS

A cellular tower can be classified as Reliable Cellular Towers if and only if it receives more than 300 records send from Reliable Crowds in each time period of the day and meets the requirements for at least 14 days.

4) RELIABLE REGION

We use the OIDD data from whole Beijing to screen for the Reliable Crowds and Cellular Towers. We count the number of Reliable Cellular Towers firstly, as Figure 8 (a) shows. The number of Reliable Cellular Towers in the urban area is the largest, and more than half of the cellular towers in the surrounding urban areas are Reliable Cellular Towers. However, the number of Reliable Cellular Towers in Daxing, Changping, Shunyi and Tongzhou District is less than 30 percent. So, we no longer use the data for these four regions in subsequent experiments. Secondly, we examine the number of Reliable Crowds. Like Figure 8 (b) shows that, the number of reliable people per square kilometer in Dongcheng and Xicheng District is more than 300. The number of reliable people per square kilometer in HaiDian, ChaoYang, FengTai and ShiJingShan District is more than 100. But other districts are really low, even just 10 per square kilometer. Thus, we choose to use the OIDD data from urban area for the subsequent experiments.

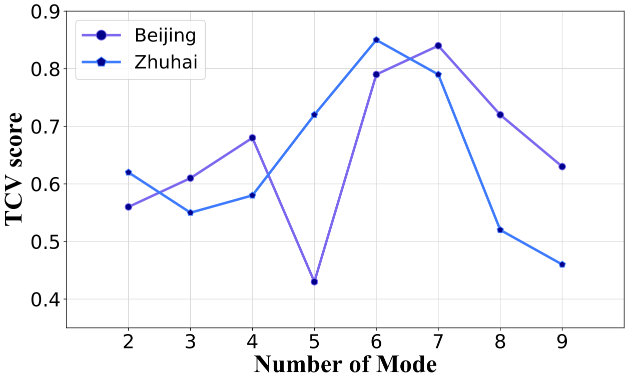


FIGURE 9. TCV scores under different Mode numbers.

B. APPROPRIATE NUMBER OF MODES

We analyzed the OIDD data using the TSLDA model. The optimal K value selection is discriminated by the TCV score, that is, the classification into several modes works best. For the sake of follow-up, we will explain the data of Zhuhai together. Figure 9 shows the variation of the TCV score for the number of different classification modes. As can be seen from Figure 9, for urban area of Beijing, the best results are divided into 7 modes, and a TCV score is 0.84. While for Zhuhai, the best results are divided into 6 modes, with a TCV score of 0.85.

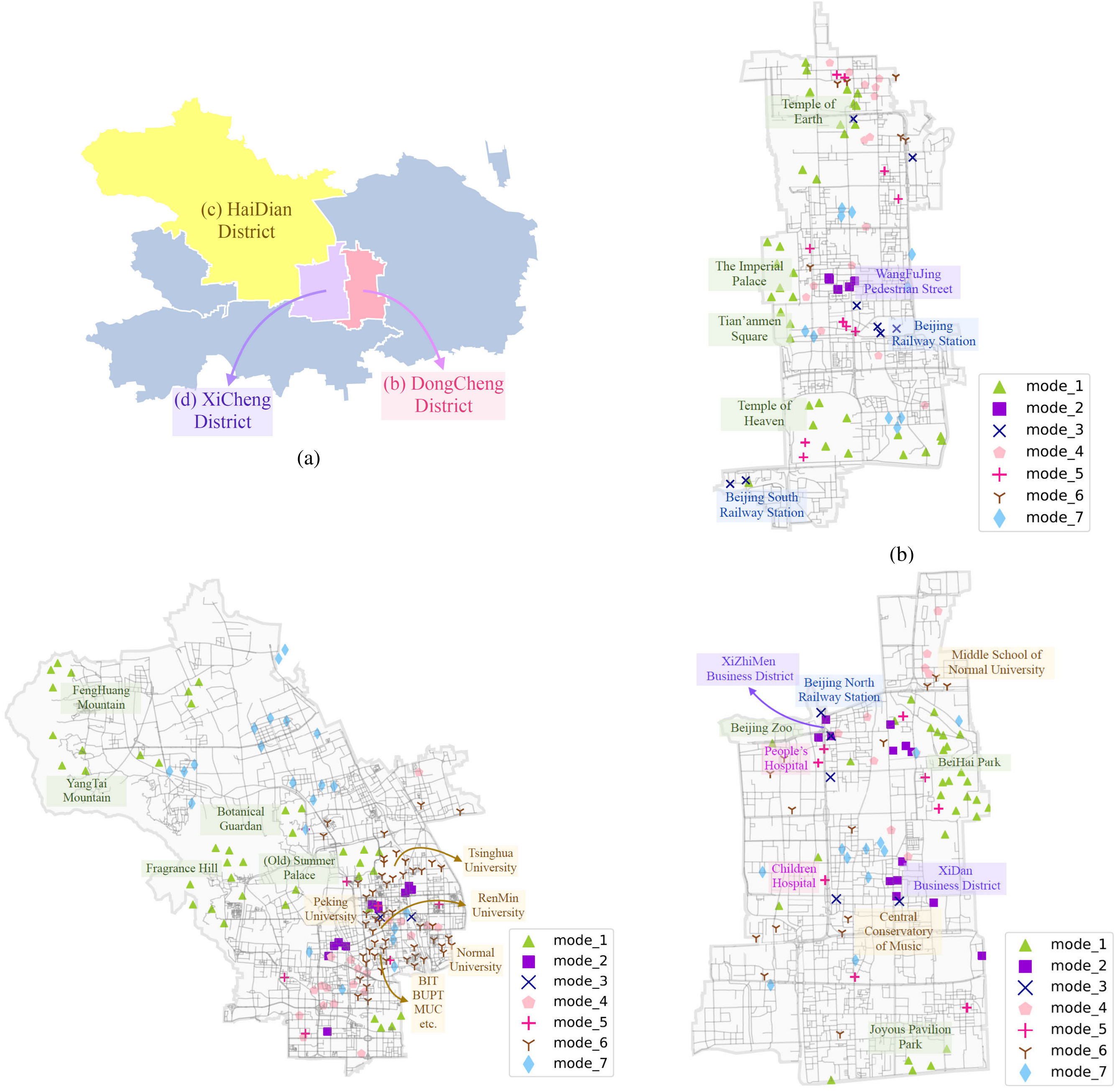


FIGURE 10. Three Cases of Map Signing. (a) The over all district relationship. (b) DongCheng District. (c) HaiDian District. (d) XiCheng District.

C. CORRECTNESS VERIFICATION

We will verify the classification effect from two angles. One is based on POI data and other is based on map visualization.

1) POI DATA VERIFY

From the output of TSLDA, we can obtain the top longitude and latitude divide and probability of each region. From Baidu POI data, we can get the POI Secondary Classification information of this coordinate through accurate latitude and longitude information. Figure 12 shows the probability of each classification based on each mode in POI Secondary Classification. We can conclude from Figure 10 that Mode_1 tend to conclude parks, zoos, scenic and museums, that is Mode_1 point to Tourist Attraction and Nature these two Primary Classifications in POI data like Table 1 shows. As same, we can analyze that Mode_2 point to Shopping, Foods, Work Out and Leisure & Entertainment,

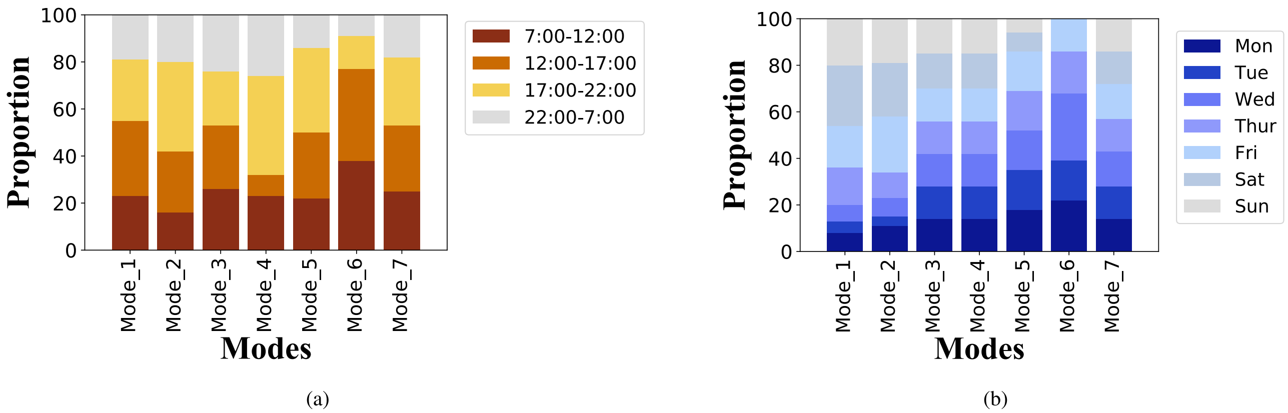


FIGURE 11. Proportion of people in different modes in different time periods in ZhuHai. (a) Daily ratio. (b) Weekly ratio.

Mode_3 point to Transportation Facilities, Mode_4 point to Hotels, Mode_5 point to Medical, Mode_6 point to Education, Mode_7 point to Corporate. So, from this aspect, TSLDA can classify different function mode from OIDD data and the classification effect is roughly consistent with the POI classification. That is, we can use this method to study the general functional division of a city.

2) MAP VISUALIZATION

We use the latitude and longitude information obtained by classification to mark different categories on the map. From this shape of map, we can see that Mode_1 covers the main parks in every district. It accurately found the location of famous scenic spots such as the Summer Palace, Forbidden City, Temple of Heaven, Tian'anmen Square, etc. Meanwhile, Mode_2 mainly find some business district. Mode_6 covers many famous universities, such as Peking University, Tsinghua University, etc. Therefore, this method can roughly give a range of typical functional areas, so that we can quickly and generally understand a city.

D. ANALYSIS DISTRICT OF BEIJING BASED ON TSLDA

From Figure 11, we can analysis some tends in urban of Beijing.

- HaiDian district is the second largest urban area in Beijing, which is focus on education. There are many schools in HaiDian district, not only lots of famous colleges, but also many middle schools and primary schools. Meanwhile, HaiDian district has many parks and scenery areas, such as Summer Palace and Fragrance Hill. Furthermore, HaiDian district has many residential area, which can accommodate many residents. Therefore, we can conclude that HaiDian district is suitable for living and traveling.
- XiCheng district is the central of Beijing, in which exists many hospitals and business districts. Meanwhile some government institutions set in XiCheng district. Therefore, Xicheng district is a region which is suitable for shopping and working.
- DongCheng District is the heart of Beijing.Main scenery like Tian'anmen Square, Forbidden City, Temple of Heaven, as well as Temple of Earth all belongs to this district. Therefore, DongCheng District is a historical district which should be given more protection.

E. CHECK EXPERIMENT BASED ON ZHUHAI DATASET

We did the same experiment using Zhuhai OIDD data. Since Zhuhai is not a very big city, so the data we obtained cannot cover every corner well. So, we just use Xiangzhou district to try.

In this experiment, based on figure 9 we take the number of modes as 6. After confirming that the experimental results are correct, we counted the number of people in each mode included in each time period and calculated the proportion of the population in different time periods, as shown in Figure 11. At the same time, in order to verify the classification correctness of the TSLDA model, we compare the classification results with the POI data, and the verification results are shown in Figure 12. It can be seen from Fig. 12 that the region partitioning probability given by TSLDA is highly consistent with the POI data, which can prove the accuracy of the classification. Meanwhile, As the figure 12 and other experiments show, Mode_1 point to Tourist Attraction and Nature, Mode_2 point to Shopping, Foods, Work Out and Leisure & Entertainment, Mode_3 point to Transportation Facilities, Mode_4 point to Hotels, Mode_5 point to Education, Mode_6 point to Corporate. There are more people go shopping and parks in weekends, and majority of people work in office during weekdays. There are nearly no people in office during weekends. But students usually study or live in school nearly all time, so time that appears in schools is very average. From Figure 11 we can clearly obtain the temporal regular of human mobility and different function region.

Meanwhile, we use the latitude and longitude information obtained by classification to mark different categories on the map which shown in figure 13. Since our dataset mainly covers Xiangzhou District of Zhuhai City, we only use this area data for analysis. From this shape of map, we can see that

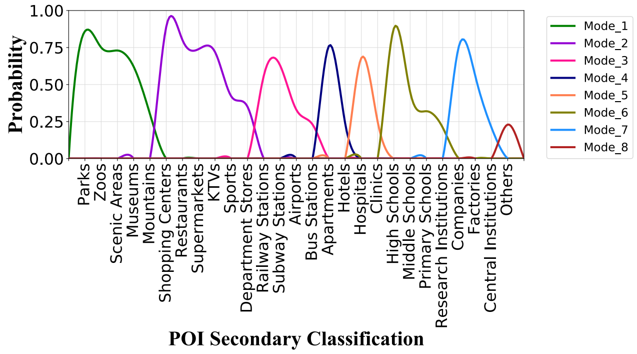


FIGURE 12. Probability of each classification based on each mode in POI Secondary Classification.

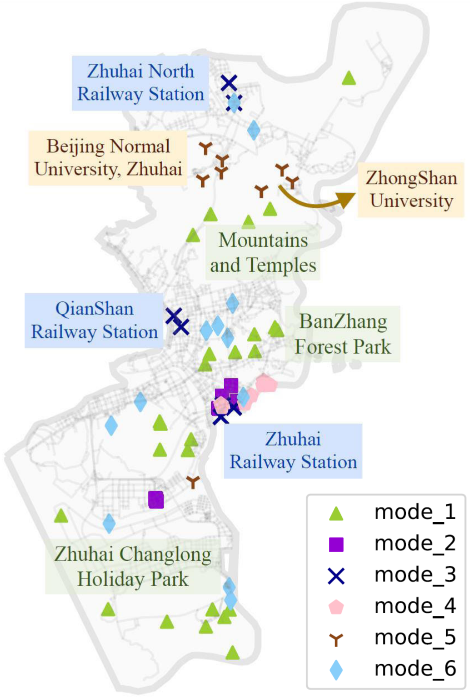


FIGURE 13. Xiangzhou District, Zhuhai Map Signing.

Mode_1 covers the main parks, Mode_2 mainly find some shops, Mode_6 covers some universities, etc. Therefore, this method can also give a range of typical functional areas based on this dataset.

VI. CONCLUSION

Urban function analysis is one of the basic tasks of urban construction. How to quickly and accurately explore urban functional domains is a useful study of urban development. In this paper, we propose the TSLDA algorithm, which improves the LDA algorithm applied to text processing to a model that can be applied to OIDD data analysis. Datasets that can be used by this method is not limited to OIDD data. As long as we can obtain a flow of information including time, place and user, we can use this method to divide and analyze the urban functional area. It is also proved in this paper that the two cities with different habits and geographical differences can have good effects, which indicates that the method is adaptable. This makes the model very scalable and can be used in a wide range of applications.

However, in this paper, it is also found that the prediction effect of this method will be significantly reduced for areas with sparse cellular towers or areas with fewer user appears. Therefore, in the future, we will continue to study this method and try to use less data to get better analysis results.

ACKNOWLEDGMENT

The authors sincerely thank China Telecom Corporation Limited Cloud Computing Branch Corporation for hosting A. Zhang and providing us with access to the datasets for this study.

REFERENCES

[1] K. Lynch, *The Image City*, vol. 11. Cambridge, MA, USA: MIT Press, 1960.

[2] S. Jiang, J. Ferreira, and M. C. Gonzalez, ‘‘Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore,’’ *IEEE Trans. Big Data*, vol. 3, no. 2, pp. 208–219, Jun. 2017.

[3] D. Brockmann, L. Hufnagel, and T. Geisel, ‘‘The scaling laws of human travel,’’ *Nature*, vol. 439, no. 7075, pp. 462–465, Jan. 2006.

[4] N. Eagle and A. S. Pentland, ‘‘Eigenbehaviors: Identifying structure in routine,’’ *Behav. Ecology Sociobiol.*, vol. 63, no. 7, pp. 1057–1066, May 2009.

[5] F. Giannotti, M. Nanni, and D. Pedreschi, ‘‘Efficient mining of temporally annotated sequences,’’ in *Proc. SIAM Int. Conf. Data Mining*, 2006, pp. 348–359.

[6] X. Zhou, A. Noulas, C. Mascolo, and Z. Zhao, ‘‘Discovering latent patterns of urban cultural interactions in wechat for modern city planning,’’ in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2018, pp. 1069–1078.

[7] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, ‘‘Trajectory pattern mining,’’ in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2007, pp. 330–339.

[8] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, ‘‘Understanding individual human mobility patterns,’’ *Nature*, vol. 453, no. 7196, p. 779, 2008.

[9] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, ‘‘Human mobility, social ties, and link prediction,’’ in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2011, pp. 1100–1108.

[10] X. Zhou, Z. Zhao, R. Li, Y. Zhou, J. Palicot, and H. Zhang, ‘‘Human mobility patterns in cellular networks,’’ *IEEE Commun. Lett.*, vol. 17, no. 10, pp. 1877–1880, Oct. 2013.

[11] M. K. Smith, *Tourism, Culture Regeneration*. Wallingford, U.K.: CABI, 2006.

[12] R. D. Malmgren, J. M. Hofman, L. A. Amaral, and D. J. Watts, ‘‘Characterizing individual communication patterns,’’ in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jun. 2009, pp. 607–616.

[13] T. Kurashima, T. Iwata, T. Hoshide, N. Takaya, and K. Fujimura, ‘‘Geo topic model: Joint modeling of user's activity area and interests for location recommendation,’’ in *Proc. 6th ACM Int. Conf. Web Search Data Mining*, Feb. 2013, pp. 375–384.

[14] Y. Qiao, Y. Cheng, J. Yang, J. Liu, and N. Kato, ‘‘A mobility analytical framework for big mobile data in densely populated area,’’ *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1443–1455, Feb. 2017.

[15] Y. Zhang, ‘‘User mobility from the view of cellular data networks,’’ in *Proc. INFOCOM IEEE Conf. Comput. Commun.*, Apr./May 2014, pp. 1348–1356.

[16] J. Yang, Y. Qiao, X. Zhang, H. He, F. Liu, and G. Cheng, ‘‘Characterizing user behavior in mobile Internet,’’ *IEEE Trans. Emerg. Topics Comput.*, vol. 3, no. 1, pp. 95–106, Mar. 2015.

[17] Y.-B. Lin and P.-K. Huang, ‘‘Prefetching for mobile Web album,’’ *Wireless Commun. Mobile Comput.*, vol. 16, no. 1, pp. 18–28, Jan. 2016.

[18] A. Thiagarajan, L. Ravindranath, H. Balakrishnan, S. Madden, and L. Girod, ‘‘Accurate, low-energy trajectory mapping for mobile devices,’’ *Usenix Assoc.*, 2011.

[19] C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nature Phys.*, vol. 6, no. 10, pp. 818–823, Sep. 2010.

[20] T. Louail, M. Lenormand, O. G. C. Ros, M. Picornell, R. Herranz, E. Frias-Martinez, J. J. Ramasco, and M. Barthelemy, "From mobile phone data to the spatial structure of cities," *Sci. Rep.*, vol. 4, Jun. 2014, Art. no. 5276.

[21] H. Klessig, V. Suryaprakash, O. Blume, A. Fehske, and G. Fettweis, "A framework enabling spatial analysis of mobile traffic hot spots," *IEEE Wireless Commun. Lett.*, vol. 3, no. 5, pp. 537–540, Oct. 2014.

[22] M. A. Bayir, M. Demirbas, and N. Eagle, "Mobility profiler: A framework for discovering mobility profiles of cell phone users," *Pervasive Mobile Comput.*, vol. 6, no. 4, pp. 435–454, 2010.

[23] J. Yang, X. Zhang, Y. Qiao, Z. Fadlullah, and N. Kato, "Global and individual mobility pattern discovery based on hotspots," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 5577–5582.

[24] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," in *Proc. GSCL*, Sep. 2009, pp. 31–40.

[25] Y. Zheng, "Trajectory data mining: An overview," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 3, p. 29, 2015.

[26] S. Jiang, J. Ferreira, and M. C. Gonzalez, "Clustering daily patterns of human activities in the city," *Data Mining Knowl. Discovery*, vol. 25, no. 3, pp. 478–510, Nov. 2012.

[27] X.-Y. Yan, X.-P. Han, B.-H. Wang, and T. Zhou, "Diversity of individual mobility patterns and emergence of aggregated scaling laws," *Sci. Rep.*, vol. 3, Sep. 2013, Art. no. 2678.

[28] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[29] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Trans. Neural Netw.*, vol. 20, no. 2, pp. 189–201, Feb. 2009.

[30] L.-C. Chen, "An effective LDA-based time topic model to improve blog search performance," *Inf. Process. Manage.*, vol. 53, no. 6, pp. 1299–1319, Nov. 2017.

[31] A. Anandkumar, D. P. Foster, D. J. Hsu, S. M. Kakade, and Y.-K. Liu, "A spectral algorithm for latent Dirichlet allocation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 917–925.

[32] S. Hasan and S. V. Ukkusuri, "Urban activity pattern classification using topic models from Online geo-location data," *Transp. Res. C, Emerg. Technol.*, vol. 44, pp. 363–381, Jul. 2014.

[33] J. L. Toole, M. Ulm, M. C. González, and D. Bauer, "Inferring land use from mobile phone activity," in *Proc. ACM SIGKDD Int. Workshop Urban Comput.*, 2012, pp. 1–8.

[34] Y. Zhao, X. Zhou, G. Li, and H. Xing, "A spatio-temporal VGI model considering trust-related information," *ISPRS Int. J. Geo Inf.*, vol. 5, no. 2, p. 10, Feb. 2016.

[35] 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 Trans. Mobile Comput.*, vol. 16, no. 10, pp. 2682–2696, Oct. 2017.

[36] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, "Measuring serendipity: Connecting people, locations and interests in a mobile 3G network," in *Proc. 9th ACM SIGCOMM Conf. Internet Meas.*, Nov. 2009, pp. 267–279.

[37] M. Lenormand, M. Picornell, O. G. Cantú-Ros, T. Louail, R. Herranz, M. Barthelemy, E. Frías-Martínez, M. S. Miguel, and J. J. Ramasco, "Comparing and modelling land use organization in cities," *Roy. Soc. Open Sci.*, vol. 2, no. 12, Oct. 2015 Art. no. 150449.

[38] M. R. Vieira, V. Frias-Martinez, N. Oliver, and E. Frias-Martinez, "Characterizing dense urban areas from mobile phone-call data: Discovery and social dynamics," in *Proc. IEEE 2nd Int. Conf. Social Comput.*, Aug. 2010, pp. 241–248.

[39] H. Xing, Y. Meng, D. Hou, J. Song, and H. Xu, "Employing crowdsourced geographic information to classify land cover with spatial clustering and topic model," *Remote Sens.*, vol. 9, no. 6, p. 602, 2017.

[40] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and pois," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2012, pp. 186–194.

[41] S. Kajimura, Y. Baba, H. Kajino, and H. Kashima, "Quality control for crowdsourced POI collection," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*. Cham, Switzerland: Springer, 2015, pp. 255–267.

[42] A. Fumo, M. Fiore, and R. Stanica, "Joint spatial and temporal classification of mobile traffic demands," in *Proc. IEEE IEEE Conf. Comput. Commun.*, May 2017, pp. 1–9.

[43] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proc. 8th ACM Int. Conf. Web Search Data Mining*, Feb. 2015, pp. 399–408.

[44] L. Qi, Y. Qiao, F. Ben Abdesslem, Z. Ma, and J. Yang, "Oscillation resolution for massive cell phone traffic data," in *Proc. 1st Workshop Mobile Data*, 2016, pp. 25–30.

YUEWEI WU received the B.E. degree from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2018, where she is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering. Her researches focus on machine learning algorithm, the mobile Internet traffic analysis, and big data analytics.

YUANYUAN QIAO received the B.E. degree from Xidian University, Xi'an, China, in 2009, and the Ph.D. degree from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2014. She is currently an Associate Professor with the School of Information and Communication Engineering, BUPT. Her researches focus on traffic measurement and classification, the mobile Internet traffic analysis, and big data analytics.

JIE YANG received the B.E., M.E., and Ph.D. degrees from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China in 1993, 1999, and 2007, respectively.

She is currently a Professor and the Deputy Dean of the School of Information and Communication Engineering, BUPT. She has published several articles on international magazines and conferences, including the IEEE Journal on Selected Areas in Communications, the IEEE Transactions on Wireless Communications, and the IEEE Transactions on Parallel and Distributed Systems. Her current research interests include broadband network traffic monitoring, user behavior analysis, and big data analysis in the Internet and telecommunications.

Dr. Yang was the Vice Program Committee Co-Chair of the IEEE International Conference on Network Infrastructure and Digital Content, in 2012 and 2014.