Discovering Transit-Oriented Development Regions of Megacities Using Heterogeneous Urban Data

Xiangjie Kong, *Senior Member, IEEE*, Feng Xia, *Senior Member, IEEE*, Kai Ma, Jianxin Li, and Qiuyuan Yang

***Abstract*— Public transport is of great significance in megacities. Transit-oriented development (TOD) has become a reliable solution to urban sustainable development, which can reshape the urban form and improve its quality. This paper focuses on leveraging heterogeneous mega urban data to answer three critical questions in TOD: what region looks like under TOD concept, which regions have the potential to be TOD regions, and how to construct these TOD regions. For region partition, we propose a connected component-based clustering algorithm, which merges the large amount of public transport stops into representative cluster ones as region centers, and then apply the Voronoi algorithm to locate the region boundaries according to the cluster centers. For TOD region identification, we present a link importance-based random walk method that considers the importance of various transits and further identifies the most valuable regions to be TOD. For discovering functions of TOD regions, we introduce a multifactor-based function characterization approach that combines both the static linguistic factor and human mobility factor together and then derives the actual function distributions. The experiments, which are conducted on three real data sets, show the superiority of the proposed methods to solve the problems of region partition, TOD region identification, and function characterization for the megacities. In the meantime, the results provide support for the government to formulate public policy to construct a TOD city.**

***Index Terms*— Function characterization, region identification, region partition, transit-oriented development (TOD), urban data.**

I. Introduction

THE process of urbanization leads to significant growth in urban population and rapid sprawl of urban space. In this process, more and more citizens spread to live in suburb, while most infrastructure (i.e., employment, commerce, and so on) still concentrate in old towns. Such imbalanced development engenders increasing social costs, including traffic congestion, energy shortage, and environmental deterioration.

Manuscript received July 1, 2018; revised May 7, 2019; accepted May 23, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61572106, in part by the Dalian Science and Technology Innovation Fund under Grant 2018J12GX048, and in part by the Fundamental Research Funds for the Central Universities under Grant DUT18JC09. *(Corresponding author: Feng Xia.)*

X. Kong, F. Xia, K. Ma, and Q. Yang are with the Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian 116620, China (e-mail: xjkong@ieee.org; f.xia@ieee.org; makai1163801756@outlook.com; qiuyuan.yang0808@gmail.com).

J. Li is with the School of Information Technology, Deakin University, Burwood, VIC 3125, Australia (e-mail: jianxin.li@deakin.edu.au).

Digital Object Identifier 10.1109/TCSS.2019.2919960

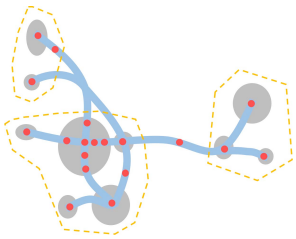


Fig. 1. Example of TOD regions.

Facing these tough issues, governments have recognized that the idea of using transit-oriented development (TOD) in reducing automobile dependence and improving the sustainability of transportation activities, which aims to design an urban form in a relatively high-density, compact and mixed form and to provide high-quality, efficient mass transportation services, together with a pedestrian-friendly environment [1].

Universally, a TOD region is a walkable neighborhood centered around public transit stops (such as metro stations, bus stops, and so on) and typically with a mix of land uses (i.e., residential and commercial) [2], as shown in Fig. 1 where the red dots represent the public transit stops, the gray colored areas represent the residential or commercial sites, the blue lines describe the public transit routes, and the dotted lines describe the TOD regions. These pedestrian-friendly and mixed-function arrangements are able to facilitate activity participation within TODs. Although some specific services are provided in other parts of a city (i.e., in other TODs), citizens can take public transport available in TOD to access these services. Hence, TOD policy maximizes transit ridership and nonmotorized mobility, which is an answer to the unsustainable, car-dependent, and transit-poor urban form that nearly characterizes the growth of modern cities. Nowadays, the TOD concept has spread worldwide and acquired fruitful achievement in practice, such as San Francisco, Copenhagen, and Hong Kong. However, contemporary works on TOD remain at the stage of theoretical introduction and statistical analysis [3], which cannot catch up with the actual need for urban development. In this paper, we will consider the significant role of different types of real-time urban data on discovering effective TOD regions.

Benefiting from the advance in wireless technique and popularity of ubiquitous terminals, the information of urban elements, such as individuals, vehicles, and surroundings, can be captured systematically, forming urban big data, including

geographical data, traffic data, commuting data, and so forth [4]–[6]. Spatial structure analysis and region function discovery with urban data have drawn extensive attention among academics. For instance, a polycentric structure is verified in several metropolia with the evidence extracted from taxi trajectories, passenger flows, or telecommunication records [7], [8]; a labeling region function is implemented via well-equipped ubiquitous sensors on cars and human beings [9], [10]. These works confirmed the effectiveness of multisource urban data in characterizing city morphology. However, there is no work to explore the potential of such heterogeneous urban data for investigating the TOD region discovery. In this paper, we aim to leverage heterogeneous urban data to answer three critical questions in TOD research.

1) What region will look like under TOD concept?
2) Which regions have the potential to be TODs?
3) How to construct these potential regions to meet TOD's need of mixed function?

To achieve the aforementioned goals, we put forward a series of data-driven algorithms. For the first question, the region partition, we propose a maximal clique and cluster combination algorithm and a connected component-based cluster algorithm to merge public transit stations within a certain range into representative ones as region centers, and then, the Voronoi algorithm [11] is applied to locate region boundaries according to cluster centers. The second question can be answered by our proposed transportation importance-based random walk method, which makes it easier to randomly walk to nodes with higher link importance and further to identify the most valuable regions to be TODs. To answer the last question, we introduce a multifactor-based function characterization approach. It defines a cost function to combine the land uses influenced by both static linguistic factor and human mobility factor and then derives the actual distribution of region functions with gradient descent. We apply the solutions to one of the megacities in China named Hangzhou with real data sets, where the TOD policy has just been officially supported recently. The structure of our methods is shown in Fig. 2.

The major contributions of this paper are summarized as follows.

1) We tackle three vital issues (region partition, region identification, and function characterization) in TOD study utilizing multisource urban data and advanced data mining techniques. To the best of our knowledge, this is the first work to explore TODs with large-scale and real-world data sets in a scientific and systematic way.
2) For region partition, we propose a maximal clique and cluster combination algorithm and a connected component-based cluster algorithm that are able to cluster region centers from various public transit stations, and then, we employ the Voronoi algorithm to locate region boundaries according to these cluster centers.
3) For region identification, we present a transportation importance-based random walk method. This method considers the link importance of diverse public transportations in regional networks of megacities, thereby identifying the most valuable regions to form TODs.
4) For function discovery, we introduce a multifactor-based function characterization approach, which defines a cost function to represent what the region is actually functioning and what the region reflects in both the static and dynamic levels and then infers the actual distribution of region functions by gradient descent.

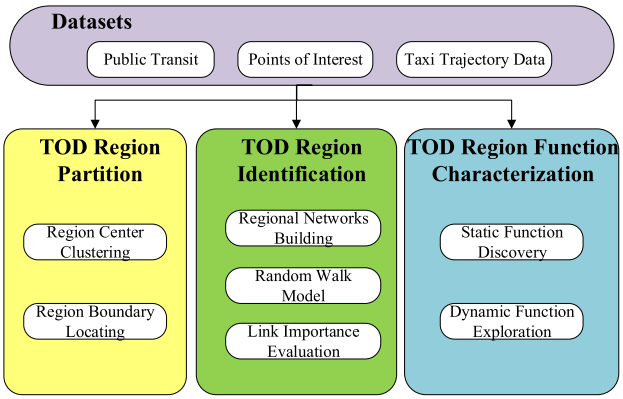


Fig. 2. Framework of this paper.

The rest of this paper is organized as follows. An overview of TOD investigation, spatial structure analysis, as well as region function discovery is presented in Section II. Section III introduces the connected component-based cluster algorithm and Voronoi algorithm for region partition. Section IV gives a description of regional networks and transportation importance-based random walk method for TOD identification. In Section V, the multifactor-based function characterization approach is presented. Section VI describes the data sets and experiments results. Finally, Section VII concludes this paper.

II. Related Work

A. TOD Investigation

In TOD study, researchers mainly focus on utilizing survey data to investigate traffic, land use, or population situations near stops [12]. Kong *et al.* [13] analyzed the resident travel behaviors to obtain five predictive features, such as flow, time, week, location, and bus, and utilized them to predict travel requirements accurately based on a machine learning model. After that, they combined the prediction results and station properties to gain shared bus optimal routes resulted in the increase in public transportation ridership.

For rail transit-based TOD investigation, Sung and Oh [14] illustrated the characteristics of TOD planning factors, such as land use, transit supply service, street network, and urban design at rail station areas in Seoul. They identified that these factors have a positive impact on constructing a transit-oriented city, and some of them need to be carefully applied to high-density cities. However, their research is only limited to Seoul, and there is not much discussion about other cities. To make the study more general, Papa and Bertolini [15] discussed

the relationship between TOD-form urban spatial structure and rail-based accessibility in six metropolitan areas. The comparison demonstrates that rail-based accessibility is higher where residents and jobs are more concentrated and lower in areas with higher network connectivity. A limitation of this paper is the use of an aggregate accessibility indicator without taking into account the subjective dimension of individual mobility choices.

For the public transport-based investigation, Cervero and Dai [16] compared bus rapid transit (BRT) density and ridership performance in forming TOD based on a survey of 27 cities. As an investigation, the authors simply used statistical correlates to list the phenomena. In contrast to the above-mentioned survey, Kamruzzaman *et al.* [17] applied the clustering methods and logistic regression models on the basis of statistics, treated each census collection district as a unit to cluster four types of TODs, including residential TODs, activity center TODs, potential TODs, and TOD nonsuitability, and they also proved that people living in TODs are significantly more likely to use public transport.

B. Spatial Structure Analysis

Spatial structure analysis helps to better explore urban TODs. The hot topics in this research direction are to identify centers, boundaries, and their characteristics in cities. For spatial structure analysis, there are many perspectives, including statistics, spatial–temporal graphs, mobile phone data, traffic data, and so on.

At first, the quantitative comparison of the urban structure was based on statistics surveys, which is coarse-grained in evolution. For example, Zhong *et al.* [18], [19] introduced a centrality index and its corresponding attractiveness indices for detecting centers and their spatial impacts using Singapore travel surveys.

Some researchers use spatial–temporal graphs of human mobility to analyze city structure. Wang *et al.* [20] proposed a collective embedding framework to learn the community structure from multiple periodic spatial–temporal graphs of human mobility. Specifically, they first exploit a probabilistic propagation-based approach to create a set of mobility graphs from periodic human mobility records. A collective deep autoencoder method is then developed to collaboratively learn the embeddings of points of interest (POIs) from multiple spatial–temporal mobility graphs. In addition, they develop an unsupervised graph-based weighted aggregation method to align and aggregate the POI embeddings into the representation of the community structures.

Mobile phone data, implying human behavior and interactions, are widely used in analyzing spatial structure at present. It is of vital importance for businesses, government, and institutes to understand how peoples' behaviors in the online cyberspace can affect the underlying computer network, or their offline behaviors at large [21]. Louail *et al.* [7], [8] had an exhaustive analysis of hotspots in 31 Spanish cities. They argued that the spatial structure of hotspots can distinguish monocentric and polycentric cities, and the essential difference between these cities is the proportion of flows between these hotspots. Likewise, Chen *et al.* [22] focused on hot lines between districts identified by two indexes, density and diversity. As for regional boundary, Ratti *et al.* [23] demonstrated that the geographically cohesive regions inferred from a telecommunications database in Great Britain correspond well with administrative regions.

Making use of traffic data to indicate urban structure also makes a series of valuable progress. Kong *et al.* [24] proposed a human mobility pattern of functional regions through analyzing the quantitative relationship between passengers getting on and off taxis in every period. Rinzivillo *et al.* [25] found a good match between the clusters formed by GPS tracks and the existing administrative borders in Pisa. Dissimilar to the above-mentioned efforts, the socioeconomic borders generated by smart card records are different from the existing administrative ones, and some new communities emerge leading to a polycentric urban form in Singapore [26]. The same polycentric phenomenon appears in London with evidence of large subway flows organizing around a limited number of activity centers according to [27]. However, it is traditionally challenging to model large-scale heterogeneous human mobility data (HHMD), since the data are collected from different sources to reflect distinct mobility patterns. Fu *et al.* [28] developed a general collective learning approach to model the HHMD at an individual level toward identifying and quantifying the urban forms of residential communities. Specifically, their proposed method exploits two geographic regularities among HHMD.

C. Region Function Discovery

Earlier region function assessment relied on the on-site investigation and questionnaire [10]. Apart from the consumption of manpower and time, the reliability is seriously influenced by subjective factors, such as the personal experience of investigators. Later, using remote-sensing image data to classify regions got noticed, and the comparison of various processing algorithms is made in [39]. Chen *et al.* [29] introduced two novel fine-tuned community detection algorithms to divide different regions and evaluate the community quality. By these two algorithms, they can assess the region functions and quality measurements by splitting and merging the network structure. Another kind of location semantics data, POI, can also be applied to cluster similar regions as well [30], [31]. Besides, Meng *et al.* [32] studied the problem of clustering moving objects in a spatial network, and they introduced two trajectory clustering algorithms and proposed a framework based on cluster block.

Recently, movement trajectory data, such as taxi trace, smart card payment, and user-generated content, provide alternative solutions to understand regions, and the connection between human mobility pattern and land uses has been pointed in some works. Like what Wang *et al.* [40] thought, how to dynamically observe and predict movement trajectory to ensure the low resource usage is a great challenge. For instance, Qi *et al.* [33] observed that get-on/off value of taxi passengers can depict the social activity dynamics in regions. Similarly, Peng *et al.* [34] found that people travel for three

TABLE I
MULTIFACTOR APPROACHES OF REGION FUNCTION DISCOVERY MENTIONED IN THE RELATED WORK

Paper ID	Core Approach	Object of Study	Advantage	Disadvantage
[9]	Urban Areas with Latent Activity Trajectories	Major Roads	Effectively Finding Functional Zones	Ignoring the Actual Information of the Road Network
[10]	Buildings in Cities	Public Transportation	Good at Depicting Dynamic Urban Space	Affected by Border Effects, Scale Issues, Pattern Insufficiency and Transportation Data Insufficiency
[29]	Community Network	Clique	Multiple Community Detection Algorithms Can Be Used to Improve Performance	The Actual Geographic Information Is Not Taken into Account
[30]	Continuous Spatial Regions	Regions on the Spatial Map	Saving Resource and Responding Quickly	Overlook Other Types of Spatial Features for Region Similarity Definition
[31]	Urban Areas	Open Geospatial Data	The Datasets Are Easy to Obtain, the Partitioning Result Is Accurate	Small Scope, Large Granularity of Regional Division
[32]	Spatial Networks	Road Networks	Avoiding the Redundant Computation of Random Expanding	Too Abstract and Not Closely Related to the Actual Situation
[33]	Urban Areas	Get-on/off Characteristics of Taxi Passengers	High Accuracy of Urban Areas Division	Narrow Application Scope, Very Simple Division Types of Areas
[34]	Urban Areas	Traffic Flow	Transforming Human Mobility into the Study of Traffic Flows between Different Destinations	Only Applicable to Datasets Containing the Beginning and Ending Information
[35]	Urban Areas	Human Mobility Data, POIs	Realizing the Functional Division of The Region	Processing Data Is Small in Size and Latitude
[36]	Urban Areas	Human Mobility Data	Examine the Spatial Patterns of These Classifications and Their Association with Different Types of Land Uses	Using Only Mobility Data, Lacking of Dynamics
[37]	Urban Areas	Taxi GPS Traces	High Accuracy for Land-use Classification	The Effect Is Worse When the Data Is Scarce
[38]	Urban Areas	SCD (Smart Card Data), POIs (Points of Interest)	Helping People Understand the Spatial Structure of A Complex City	Ignoring the Possibility of Mixed Use

purposes on workdays, including commuting between home and workplace, traveling from workplace to workplace, and others such as leisure.

Counting on the detailed location in taxi traces, Yuan *et al.* [35] extracted origin–destination (OD) information, combined with POIs, to discover functional zones with Dirichlet-multinomial regression. Liu *et al.* [36] adopted the concept of "source-sink" in ecology for identifying how different types of land uses influence trip generation at different times. Pan *et al.* [37] designed six features to characterize OD pattern, and the combination of these features achieved a promising recognition accuracy using support vector machine as a classifier. The idea of OD pattern extraction can also be used to dealing with smart card records [10], [38]. Especially, building on the achievement in [35], Yuan *et al.* [9] added payment records and presented a collaborative-filtering-based approach to further enhance the performance. Additionally, user-generated contents not only point out the location but also contain user property, which is also a good medium to study topic distributions, such as in [41] and [42].

We do the summary about the method of region function discovery mentioned in the related work, which shown in Table I. From the previous description, we can see that TOD study still stays on the phase of statistics and analysis of survey data, whose reliability is profoundly affected by the time, place, and investigator. However, a megacity is a complex and dynamic system, which is hardly clarified by a single and static data set. Moreover, the current progress made in the related directions, such as spatial structure analysis and region function discovery, cannot reveal the inherent property of TOD, such as public transport centered (the TOD region can be regarded as the center of public transport, such as subway station and bus stop) and mixed functions (a TOD region may have the mixed land use, such as residential, commercial, and business). Therefore, utilizing multisource and real-time data with advanced research techniques is desperately needed in TOD research, which differs this paper from the previous ones.

III. TOD REGION PARTITION

TOD policy emphasizes the core status of public transit, and the coverage is also influenced by field conditions. However, most works on TOD only focus on a circle area, of which the center is subway station and the radius is a few hundred meters. Such coarse-grained partition falls behind the policy requirements. To overcome the disadvantage of traditional approaches, we conduct a more meticulous and data-driven study following the TOD concept in this section.

Generally, public transit not only refers to subway but also includes bus and BRT, which jointly form the urban public transportation system and all play an essential role in daily traveling. Hence, we consider these three public travel modes in determining region centers. Such consideration brings a direct problem: the huge number and redundancy of stations. The solution to this problem is our proposed connected component-based cluster algorithm in Section III-A, since stations often gather near buildings with specific features and residents share a similar human mobility pattern around there.

As for the border of TOD, there is no exact definition, as long as TOD covers a pedestrian-friendly environment, such as the walking distance within 5–10 min. In reality, we prefer the nearest stops for traveling and the nearer commercial centers for leisure, even the workplaces as near to our homes as possible. This idea coincides with the nearest neighbor principle of the Voronoi algorithm. Thus, we apply the Voronoi algorithm to locate the region boundary in Section III-B.

A. Region Center Clustering

Merging similar stations in a certain area as the region center is the target of this part. Clustering in machine learning is to group a collection of objects in such a way that the objects in the same cluster are more similar than those in different clusters [43]. Among cluster algorithms, K-means is the most popular one. Nonetheless, there exist two problems in embedding K-means in our research: 1) K-means requires user to give the cluster number K in advance, but at present, we cannot easily determine how many aggregated stops are appropriate; 2) K-means randomly selects K objects as centroids in the initialization process, of which the randomness leads to the instability of cluster results, so it demands repeated randomization in centroids to get better results.

To overcome the above-mentioned shortcomings and better solve the practical problems, we proposed a connected component-based cluster algorithm, which is also an improvement of K-means. Algorithm 1 presents the pseudocode of the proposed cluster method. First, we add edges between two nodes (stops) m and n of which the distance is less than d (Lines 2–6). Through such processing, several connected components are formed in the network, and we denote the nodes in each connected component as $T\{S_1, S_2, \ldots, S_n\}$ (Line 7). Next, for any connected component (cluster) S in T (Line 8), we find the farthest two nodes m and n (Lines 9–12). If the distance of m and n is larger than $2d$, which implies that this cluster S can be divided into two clusters at least, we choose m and n as initialized centroids to execute K-means and generate two clusters S' and S'' (Lines 13–17). Otherwise, we can think that the distance between the node and its centroid is no more than d, and hence, the loop will end (Line 18). At last, we calculate the average value of nodes p_i in each cluster S_i (Lines 19–21), and this is the position of region center.

From the procedure of the proposed connected component-based cluster algorithm, we can see that the uncertainty of K value and the instability of cluster results caused by random selection are solved by choosing farthest two nodes in connected components as centroids in the initialization process of K-means. The presented method introduces a graph theory into clustering, which has universality in solving the clustering problem related to map, graph, network, and so on. Note that a distance parameter d is brought in the new approach, whereas it is relatively easy to be identified according to the actual requirement. For instance, following the concept, a TOD is suitable to be built within 5–10 min walk, about 400–800 m, which can be regarded as d.

B. Region Boundary Locating

After clustering region centers, the next step is to locate region borders. The Voronoi algorithm is applied here, which avoids the imperfect coverage and overlapping phenomena in the previous studies. Specially, the Voronoi algorithm partitions a plane X into regions r (called Voronoi cell) based on a set of points p (called seed). For each seed p_i, there exists a corresponding region r_i containing all the nodes x closer to that seed p_i than to any others p_j, as shown in the following equation:

$$r_i = \{x \in X | D(x, p_i) \leq D(x, p_j) \text{ for all } j \neq i\}. \quad (1)$$

Take Fig. 3 as an example. The blue dots are the seeds, and the convex polygons are Voronoi cells. Each cell consists of the nodes whose distance to the corresponding seed is less than its distance to any other seeds. The edges of cells are the points equidistant to the nearest seeds.

In this paper, we treat all clustering stations generated from connected component-based cluster algorithm in Section III-A as the seeds in the Voronoi algorithm, and then, the study area will be divided into Voronoi cells that are all TOD candidates. After the partition, we can assume that citizens in a region are more likely to take public transportation at the corresponding cluster stations, while they not easily crossover to another region for a remote station.

Algorithm 1 Pseudocode of Connected Component-Based Cluster Algorithm

```
assume all stops as nodes;
for all node pair u, v do
    if Distance(u, v) < d then
        AddEdge(u, v);
    end if
end for
record nodes in each connected component T{S_1, S_2, ..., S_n};
for all S in T do
    for all node pair u, v do
        calculate Distance(u, v);
    end for
    find farthest two nodes m, n in S;
    if Distance(m, n) > 2d then
        execute K-means(m, n) and generate S', S'';
        delete S from T;
        add S', S'' to T;
    end if
end for
for all S_i in T do
    calculate average value p_i;
end for
```

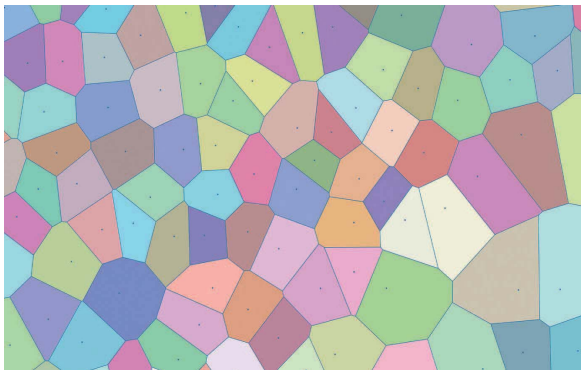


Fig. 3. Example for Voronoi.

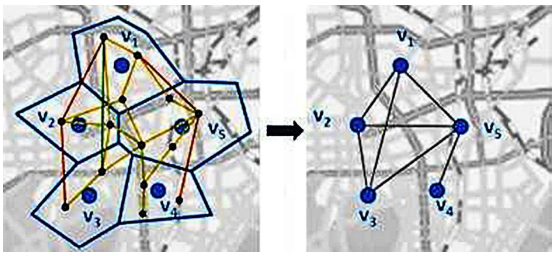


Fig. 4. Regional networks.

IV. TOD Region Identification

In Section III, we partition the study area into regions that are all TOD candidates. As we know, the random walk model is remarkable for its characteristic of integrating rich information of nodes and links simultaneously, which is widely used in a network analysis. Nonetheless, basic random walk ignores the importance of link, which cannot reflect the different kinds of transits playing a different importance in the TOD policy. To this end, we introduce a transportation importance-based random walk method for identifying TODs.

A. Regional Networks

In order to identify the most valuable regions from these candidates, the first task is to build regional networks shown in Fig. 4. As described in Section III, we have acquired a few cluster stations and their corresponding cells that constitute the node set V in regional networks. Afterward, if two regions are connected by a public transit with two adjacent cluster stations, we add an edge between these two regions, all of which compose the edge set E. More than that, we take three kinds of public transports into account, i.e., bus, BRT, and subway, represented by different colors in Fig. 4. Different kinds of transports impact the construction of TODs to different extents, which drives us to endow different weights to different connections. The detailed description of weights is in Section IV-C. In consequence, the regional networks can be noted as a weighted graph represented by $G = (V, E)$.

B. Random Walk Model

The task here is to identify the most valuable regions to be TODs, which is to find the most valuable nodes in regional networks. A random walk model is embedded as the foundation model for region identification. It evaluates each node with a rank score, which is determined by two factors: 1) the number of nodes that this node connected to; 2) the importance of these nodes. The formal definition is presented in the following:

$$\overrightarrow{RW}[v_i] = \frac{1-\alpha}{N} + \alpha \sum_{v_j \in M(v_i)} \frac{\overrightarrow{RW}[v_j]}{L(v_j)} \tag{2}$$

where $\overrightarrow{RW}$ is the rank score vector and $\overrightarrow{RW}[v_i]$ is the rank score of node v_i. $M(v_i)$ represents a set containing all the neighbors of node v_i, and $L(v_j)$ denotes the number of neighbors node v_j has. Besides, N is the total number of nodes in the networks, and α is the probability that the walker will continue to walk to the next neighbor, which is generally set as 0.85 [44].

Equation (2) shows the score of a node in one step. For all nodes in the whole networks, the random walk process that is an iterative process is defined as

$$\overrightarrow{RW}^{(t+1)} = \alpha \tilde{S} \overrightarrow{RW}^{(t)} + (1-\alpha)q \tag{3}$$

where $\overrightarrow{RW}^{(t)}$ is the rank score vector in the tth step and q is a row vector of which form is $(0, \ldots, 1, \ldots, 0)$. $\tilde{S}$ is the transform matrix representing the probability for each node skipping to other nodes. The iteration process will end when the model assigns each node v_i with a stable rank score $\overrightarrow{RW}[v_i]$. Then, we sort nodes in accordance to their scores and select TopN nodes as most valuable regions.

C. Link Importance

Basic random walk assumes that the weights of edges are the same. Hence, it defines the element $s_{i,j}$ in $\tilde{S}$ as $1/L(p_j)$, which means that the walker transmits to the node's neighborhoods with the same probability. Such setting cannot reflect different relationship strengths between the nodes in the networks. Therefore, we assign the edges with related weights according to the quantity and quality of connections formed by various transport modes and denote the weight as link importance LI.

To be specific, the edges in regional networks are produced by transportation lines, and the number of lines connecting two regions reflects the relationship strength between them to a great extent. Thus, we assume the numbers of bus, BRT, and subway lines connecting node v_i and v_j is $m_{i,j}$, $n_{i,j}$, and $k_{i,j}$, respectively. Furthermore, the quality of transportation modes should be taken into account as well. For instance, rail transit provides large capacity and high-speed services, which attracts more citizens to take, and further motivates an increasing number of researchers to construct TODs around rail transit stations. Therefore, the quality of these three modes is assigned by w_m, w_n, and w_k in this paper. Combining the two factors mentioned earlier, the link importance between nodes v_i and v_j is calculated by the following equation:

$$LI(v_i, v_j) = m_{i,j} w_m + n_{i,j} w_n + k_{i,j} w_k. \tag{4}$$

Thereafter, the transform probability is in proportion to the link importance, which is described as

$$s_{i,j} = \frac{LI(v_i, v_j)}{\sum_{v_k \in M(v_i)} LI(v_i, v_k)}. \tag{5}$$

As presented earlier, we introduce edge features into the network structure and present a transportation importance-based random walk model. It is easy for this model to randomly walk to nodes with higher link importance and further to identify the most valuable regions to be TODs.

V. TOD Region Function Characterization

Since TODs generally contain diverse functions to support diverse life needs, what actual functions do the regions have? The actual distribution of region functions is not only

formulated by urban planners but also evolves as people's activities. In other words, the actual distribution is influenced by static semantic factor and human mobility factor. Thus, a multifactor-based function characterization approach is needed.

Discovering the distribution of POIs, which typically contain the coordinate and category of building (see Section VI-A2 for details), is a way to preliminarily understand the functions under static factor. For instance, if a region consists of massive shopping plazas and restaurants, it has a high probability to support commercial function. A term frequency–inverse document frequency (TF-IDF) approach [45] from information retrieval is applied here to characterize the static function. Human mobility is another factor to affect region function. For instance, along with the increase in staff, new infrastructure emerges in the workplace, which brings new features to this area. The topic model from natural language processing has been proven to discover region function under the view of human mobility from taxi trajectories [5], [35], which is employed in this paper.

The distribution information we acquire so far can be regarded as the actual function reflecting in a static and dynamic way, that is, the actual function has these two appearances at the same time. Inspired from machine learning, we define a cost function to represent what the region is actually functioning and what the region function reflects in both the static and dynamic levels and then infer the actual region function by gradient descent.

A. Static Function Discovery

We can consider the relationship between POI data and static function from two aspects, i.e., the absolute number and the relative number, and this idea can be realized by the TF-IDF method. To be specific, if the absolute number of a specific kind of POI is high, the corresponding function should possess a larger proportion in this area. For instance, a region with a lot of shopping centers and restaurants should undertake higher proportion of commercial function. The term frequency (tf) term in TF-IDF can present the absolute number, as shown in the following equation:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_{k=1}^{C} n_{i,k}} \tag{6}$$

where $tf_{i,j}$ denotes the tf value of the jth POI category in region r_i, $n_{i,j}$ is the number of POIs belonging to jth category in region r_i, and C is the number of POI categories.

Furthermore, a special kind of POI rarely appears in other regions, even if the absolute number is not high in this region, but it still can feature this region. For example, a university town is surrounded by abundant restaurants, even if the absolute number of restaurants is high, but the region where these universities locate still should give priority to education. This can be described by inverse document frequency (idf) term, as shown in the following equation:

$$idf_j = \log \frac{R}{|\{i|n_{i,j} \neq 0, i = \{1, 2, \ldots, R\}\}| + 1} \tag{7}$$

TABLE II
ANALOGY BETWEEN REGION-FUNCTION EXPLORATION AND DOCUMENT-TOPIC DISCOVERY

Region-Function Exploration	Document-Topic Discovery
study area	document collection
region	document
human mobility pattern	word
function	topic

where idf_j is the idf value of the jth POI category and R is the number of regions.

Afterward, the $tf\text{-}idf$ value of the jth POI category in region r_i can be obtained by multiplying two variables mentioned earlier, as shown in the following equation:

$$tf\text{-}idf_{i,j} = tf_{i,j} \times idf_j. \tag{8}$$

At last, we formulate a vector $\overrightarrow{Y_i}$ for each region r_i to denote the distribution of POIs in the following equation:

$$\overrightarrow{Y_i} = (tf\text{-}idf_{i,1}, tf\text{-}idf_{i,2}, \ldots, tf\text{-}idf_{i,C}). \tag{9}$$

Through the earlier calculation, we can acquire the distribution of POIs in each region, which represent the region function from the perspective of static semantic.

B. Dynamic Function Exploration

Taxi trajectories produced by human activities reflect region functions from the dynamic angle, since people usually travel from regions with similar functions to another similar region at a similar time, such as from residential areas to workplaces at workday morning and from residential to entertainment areas at weekends. The mobility data set of social vehicles cannot be acquired because of the limitations of privacy and security, while the data set of taxis can be easily obtained through various methods. Besides, both taxis and social vehicles can represent the peoples' urban mobility pattern, so we can use taxi trajectories data to replace the public vehicle [46]. The relationship between human mobility pattern and region function can be uncovered by topic models. In this regard, we make an analogy between exploring functions of a region and discovering topics of a document, as shown in Table II. Given all the words of each document in document collection, latent Dirichlet allocation (LDA), a topic model, can infer the distribution of topics for each document. Accordingly, given all the human mobility patterns of each region in study area, LDA can infer the distribution of functions for each region.

We formalized the analogy as follows. For region r_i, we define an $R \times T$ leaving matrix L^i, where R is the number of regions and T is the number of time slots. The element $L^i[j,k]$ denotes the taxi trajectory from region r_i to region r_j in the time slot t_k, and the value of element represents the number of corresponding trajectories. Similarly, we define an $R \times T$ arriving matrix A^i, and the element $A^i[j,k]$ denotes the taxi trajectory from region r_j to region r_i in the time slot t_k. Thereafter, we regard region i as a document, and all the regions form the document collection. Moreover, the elements $L^i[j,k]$ and $A^i[j,k]$ are all specific

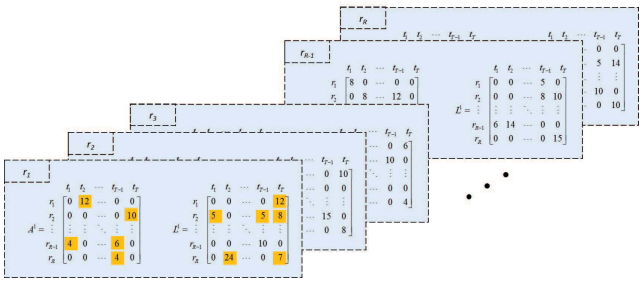


Fig. 5. Topic modeling in region-function exploration.

TABLE III
TIME SLOTS FOR WEEKDAYS AND WEEKENDS

Slot No.	Weekday	Slot No.	Weekend
1	07:00-10:30	6	9:00-12:30
2	10:30-16:00	7	12:30-18:30
3	16:00-19:30	8	18:30-21:00
4	19:30-21:00	9	21:00-09:00
5	21:00-07:00		

mobility patterns, which are treated as words, and the element value means the occurrence number of words.

To better depict the analogy, we take Fig. 5 as an example. All the regions in the figure compose the document collection, each of them is a document, and each element representing a mobility pattern is a word. For instance, digit "7" in the bottom-right corner of region r_1 indicates that the mobility pattern traveling from r_1 to r_R in time slot t_T occurred seven times.

Note that the trajectories under consideration are all occupied taxi trips with a certain origin region and a certain destination region, only which imply human mobility. Besides, in terms of the time slot, we show the traffic conditions on weekdays and at weekends in Fig. 6. According to the figure and mobility purposes, we form nine slots ($T = 9$) in Table III. Moreover, due to the uneven division of slots, the values of each element in matrixes are changed to the average number of the corresponding trajectories.

By the analogy described earlier, we acquire all the patterns in each region, which means all the words in each document. Then, LDA can infer the distribution of functions for each region. For region r_i, the output of LDA is a K-dimensional vector, as shown in the following equation:

$$\overrightarrow{Z}_i = \{z_{i,1}, z_{i,2}, \ldots, z_{i,K}\} \tag{10}$$

where $\overrightarrow{Z}_i$ is the topic distribution of region r_i, $z_{i,k}$ is the proportion of topic k in region r_i, and K is the topic number. Thereby, we calculate the similarity of region r_i and r_j in terms of dynamic function as follows:

$$\lambda_{i,j} = \cos < \overrightarrow{Z}_i, \overrightarrow{Z}_j > . \tag{11}$$

C. Actual Function Estimation

The actual distribution of region functions is influenced by many factors, i.e., static semantic factor and human mobility factor. Although the inherent function is largely determined by the static semantic factor, the evolution of function is driven by human activities. Thus, combining these two factors is essential in function characterization.

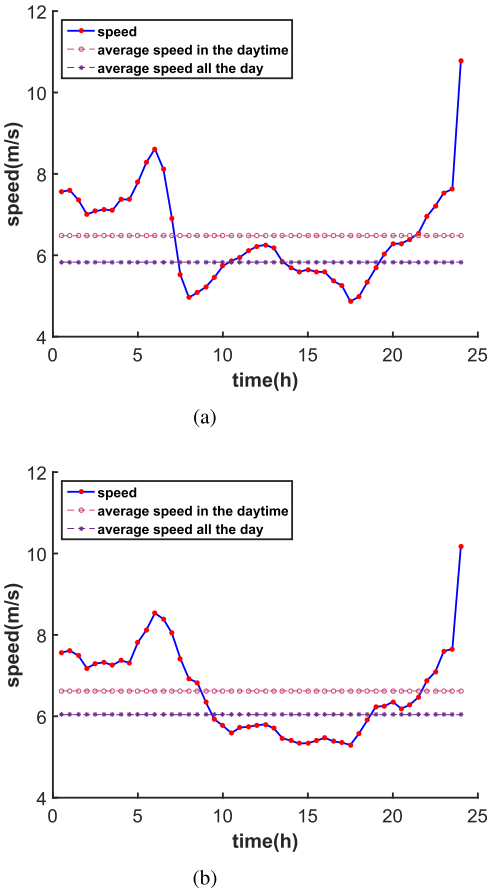


Fig. 6. Traffic conditions changing over time. (a) Weekdays. (b) Weekends.

1) Cost Function Definition: We define a cost function J to represent the differences between what the region is actually functioning and what the actual function reflects in both static way and dynamic way, as shown in (13).

In (13), R is the region number and W balances the proportion of static and dynamic functions. $\overrightarrow{X}_i$ is the distribution of actual function in region r_i that we desire, $\overrightarrow{Y}_i$ is the static function distribution that we have obtained in Section V-A, and λ_{ij} represents the similarity of region r_i and region r_j in terms of dynamic function that has been calculated in Section V-B. In addition, since the actual distribution of region functions cannot deviate from its inherent features a lot, we initialize $\overrightarrow{X}_i$ with $\overrightarrow{Y}_i$

$$J = \frac{1}{R^2} \sum_{i=1}^{R} \sum_{j=1}^{R} W(\cos\theta_{ij} - \lambda_{ij})^2 \tag{12}$$

$$\cos\theta_{ij} = \begin{cases} \dfrac{\overrightarrow{X}_i \cdot \overrightarrow{Y}_j}{\|\overrightarrow{X}_i\| \times \|\overrightarrow{Y}_j\|}, & i = j \\ \dfrac{\overrightarrow{X}_i \cdot \overrightarrow{X}_j}{\|\overrightarrow{X}_i\| \times \|\overrightarrow{X}_j\|}, & i \neq j \end{cases} \tag{13}$$

$$W = \begin{cases} R-1, & i = j \\ 1, & i \neq j. \end{cases} \tag{14}$$

As shown in (13), for $i = j$, we define the cost as the difference between the actual function and its static appearance, which is obtained by TF-IDF from POIs. For $i \neq j$, the cost is defined as the difference between the actual function and its representation in terms of human mobility, which is extracted from taxi trajectories by LDA. In this way, the total cost is the square sum of these differences, which is denoted as J.

2) Gradient Descent: After the definition, we desire the minimum point of cost function, which means that the smaller the differences between actual function and its two appearances, the better. In this paper, we utilize a gradient descent algorithm to find the local minimum point of cost function J around static function distribution as the actual distribution of region functions.

A gradient descent algorithm searches the local minimum of a cost function along the negative gradient direction iteratively. Since the gradient direction is decided by derivative function, we first take the partial derivative of J with respect to $X_{ik} (i = 1, 2, \ldots, R, k = 1, 2, \ldots, C)$ as follows:

$$\frac{\partial J}{\partial X_{ik}} = \frac{1}{R} \sum_{j=1}^{R} 2W \times \left(\frac{\overrightarrow{X}_i \cdot \overrightarrow{Z}}{\| \overrightarrow{X}_i \| \times \| \overrightarrow{Z} \|} - \lambda_{ij} \right) \times \left(\frac{Z_k \times \| \overrightarrow{X}_i \| \times \| \overrightarrow{Z} \| - \frac{\| \overrightarrow{Z} \|}{\| \overrightarrow{X}_i \|} \times \overrightarrow{X}_i \cdot \overrightarrow{Z} \times X_{ik}}{\| \overrightarrow{X}_i \|^2 \times \| \overrightarrow{Z} \|^2} \right) \tag{15}$$

$$\overrightarrow{Z} = \begin{cases} \overrightarrow{Y}_j, & i = j \\ \overrightarrow{X}_j, & i \neq j \end{cases} \tag{16}$$

$$W = \begin{cases} R - 1, & i = j \\ 1, & i \neq j. \end{cases} \tag{17}$$

Therefore, the iterative update process of X_{ik} is given as follows:

$$X'_{ik} = \begin{cases} X_{ik} - \alpha \dfrac{\partial J}{\partial X_{ik}}, & X_{ik} \neq 0 \\ X_{ik}, & X_{ik} = 0 \end{cases} \tag{18}$$

where α is the learning rate and we set it as 1 in this paper. According to the equation, if region r_i has no POI in the kth category ($X_{ik} = 0$), we assume that the region dose not undertake this kind of function and X_{ik} will not change with the iteration ($X'_{ik} = X_{ik}$). Otherwise, X_{ik} will take one step proportional to the negative of the gradient ($X'_{ik} = X_{ik} - \alpha(\partial J / \partial X_{ik})$). Note that in each iteration, the sum of components in $\overrightarrow{X}_i$ is not equal to 1 ($\sum_k X_{ik} \neq 1$), hence, we normalize $\overrightarrow{X}_i$, and it can be proved that the normalization of $\overrightarrow{X}_i$ cannot affect the value of J.

The update process will not end until the number of iterations reaches its threshold, of which we regard as the achievement of local minimum of J. At this point, X_{ik} shows the proportion of kth actual function in region r_i, and thereby, $\overrightarrow{X}_i$ records the distribution of actual functions in region i which we desire. In addition, the achieved region functions combine both the static semantic factor and the human mobility factor in function characterization, which reflects the situation more accurately.

TABLE IV
STATISTICS OF THE PUBLIC TRANSIT DATA SET

Modes	Line Number	Stop Number
Bus	641	8255
BRT	14	446
Subway	5	127

TABLE V
STATISTICS OF POIS

Code	Category	Subcategory	Number
1	residence	apartment, house	24595
2	workplace	company, factory	49891
3	education	school, training, science	11005
4	commerce	restaurant, mall, plaza, theatre, cafe, club, bar, supermarket	41951
5	service	car service, gym, bank, living service, institution, hospital, clinic, pharmacy	56089
6	scenic	park, square, museum	6354
Total			189885

TABLE VI
STATISTICS OF TAXI TRAJECTORIES

Properties	Values
Dataset Size (G)	54.4
Taxi Number	2540
Effective Days	30
Occupied Trips	5522177
Average Trip Distance (km)	5.83
Average Trip Duration (min)	17.37

VI. EXPERIMENT

To validate the effectiveness of our proposed methods, we utilize real data sets in Hangzhou to conduct the experiments with MATLAB and Python. In the following, we introduce three data sets utilized in the experiment and then show the results and analysis.

A. Data Sets

We use three data sets in the evaluation as follows.

1) Public Transit: The data set consists of three kinds of public transit including bus, BRT, and subway in February 2015. We mainly focus on the information related to lines and stops in these modes, such as latitude and longitude. The corresponding numbers are summarized in Table IV.

2) Points of Interest: The POI data set contains 189 885 records in 2015 with the name, longitude, latitude, and category. The detailed category information and its corresponding number are shown in Table V.

3) Taxi Trajectories: We utilize a GPS data set generated by Hangzhou taxis in March 2014, and some properties of the data set are shown in Table VI. Note that we use Manhattan distance [47] to show the average trip distance. Moreover, we find that the data on March 11st only contain partial information which cannot imply human mobility, and thus, we eliminate the GPS records on that day and use the rest of the occupied trips to conduct the experiments.

Although the collecting intervals of the above-mentioned three data sets are not in the same time dimension, the urban form cannot significantly change within one year.

(a) (b) (c)

Fig. 7. Comparison among region partition methods. (a) Nonclustering method. (b) Connected component-based partition method. (c) Connected component-based clustering method (proposed method).

Therefore, we think that those data sets describe the same state of the city.

B. Evaluation on Region Partition

We compare region partition results generated by two traditional methods [method 1) and 2)] with our proposed method [method 3)]: 1) nonclustering method, which regards all stops as seeds in the Voronoi algorithm to partition the study area; 2) connected component-based partition method, which forms connected components of all stops with the threshold of d and regards the average value in each component as seeds in Voronoi; and 3) connected component-based clustering method, which generates the cluster centers and divides the region division based on these cluster center. The difference between method 2) and 3) is whether it gets the cluster centers or not. The data set associated with public transit stops, which is introduced in Section VI-A1, is used here, and d is set as 800, which is in line with the pedestrian-friendly distance in TOD concept.

Fig. 7 shows the region partition results of three approaches. All the transit stops are treated as region centers in Fig. 7(a), which causes the problem of producing regions that are too small and too dense, and it is unrealistic to develop a single stop as a TOD unit as well. Thus, merging stops effectively must be first solved. From Fig. 7(b), the connected component-based partition method can generate different sized regions, and however, the sizes are generally large. In addition, the regions in downtown are larger than those in suburb. This is because prosperous areas have numerous and dense stops which results in the wide coverage of connected component, while the situation in the remote district is opposite. Such results are not beneficial for studying TOD. The results of our proposed method are presented in Fig. 7(c), which generates 645 regions with moderate sizes in both downtown and suburb. Moreover, the center and the border of regions also accord with actual situations, such as regional boundaries, which imitate the trend of the Qiantang River. From the experiment results, our proposed method, connected component-based clustering, can solve the problem of redundant stops and divide the study area into reasonable regions; meanwhile, it avoid the determination of K value and instability of clustering results in K-means algorithms.

C. Evaluation on Region Identification

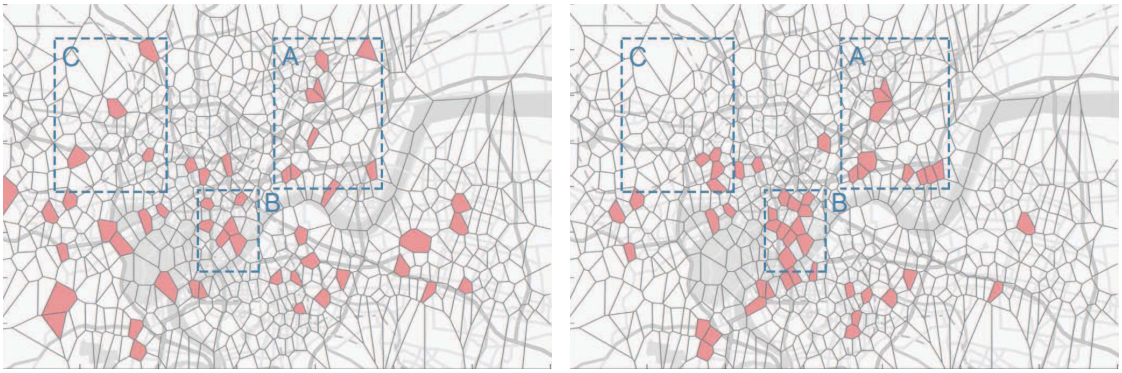
The comparison is the random walk model, which deems that the relationship between nodes is equally important. The data utilized here are the public transit data set, as introduced in Section VI-A1. In terms of parameter setting, the total number of nodes in the networks N is 645 as the region number; the iteration ends when the difference of all nodes' RW values is less than 10^{-16}; the ratio of transit modes' quality is $w_m : w_n : w_k = 1 : 2 : 3$, which takes capacity, frequency, and speed into account. In the experiment, RW values achieve convergence via about 100 iterations, and we choose TOP 50 nodes as TOD regions. Although new towns at different levels from Hangzhou urban planning[1] cannot be exactly equivalent to TOD regions, they also reflect the direction of city development to a great extent. Hence, we use these new towns as a kind of label in region identification.

The experiment results are presented in Fig. 8. The comparing method can identify a certain number of new towns in urban planning in Fig. 8(a), whereas those identified regions are very scattered, which cannot reflect the fact that the development of a central region often brings prosperity to the neighborhood. From Fig. 8(b), we can see that the proposed method identifies most new towns and achieves the effect of overall decentralization and local concentrations simultaneously.

Specifically, we take Blocks A–C in Fig. 8 as an example. For Block A, our proposed method can identify the new towns of Linping (middle), Jiuqiao (bottom left), and Xiasha (bottom right) accurately in Fig. 8(b), while the comparing algorithm is insufficient to recognize the surrounding of three centers and label some regions which are not in urban planning instead in Fig. 8(a). Block B covers the area near West Lake and Qianjiang CBD, which is the center of Hangzhou with the highest development level and the largest coverage. Compared with the random walk [see Fig. 8(a)], our method can mirror this fact more precisely [see Fig. 8(b)]. Similarly, with regard to North Town in the bottom-right corner of Block C, our proposed transportation importance-based method is still outstanding.

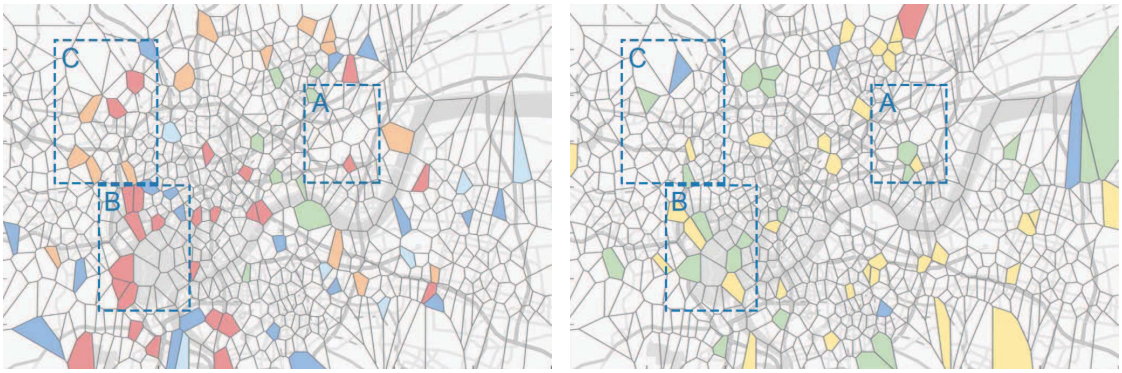
From the earlier discussion, the proposed method can identify the TOD regions effectively. This is due to the

[1] Hangzhou Urban Planning Files, http://www.hzplanning.gov.cn/index.aspx?tabid=903d39b5-a4d8-41eb-bec0-13b068e0bf54



(a) (b)

Fig. 8. Comparison between region identification methods. (a) Random walk method. (b) Transportation importance-based random walk method (proposed method).



(a) (b)

Fig. 9. Comparison between function characterization methods. (a) Region functions in which MuF discovers but TF-IDF does not. (b) Functions in which TF-IDF treats as region functions but MuF does not.

consideration of different importance of public transit in forming connections and further facilitating valuable regions with higher rankings. Admittedly, the identification in some remote areas with sparse data remains to be improved, such as the southeast of Hangzhou.

D. Evaluation on Function Characterization

POI and taxi trajectory data sets are utilized to conduct the function characterization experiment. We compare the proposed multifactor-based function characterization approach (MuF) with TF-IDF and treat the categories accounting for over 25% in function distribution vector as the region functions. In LDA, the topic number K is 7 which is consistent with POI categories; the number of iterations is 1000; the rest parameters are set as default values [48]. In addition, the iteration number of gradient descent is set as 300.

Function characterization-related results are shown in Fig. 9. To better illustration, the functions, in which MuF discovers but TF-IDF does not, are presented in Fig. 9(a); the functions, in which TF-IDF treats as region functions but MuF does not, are shown in Fig. 9(b); the functions, in which both two methods label, are not displayed. Overall, MuF can characterize the region with various kinds of functions [see Fig. 9(a)] and weaken the proportion of ubiquitous residence and commerce [see Fig. 9(b)]. Moreover, the adjustments involve the whole city. To be specific, MuF discovers the education function at the top-right corner of Block B, which matches the existence of several educational institutions such as two campuses of Zhejiang University, and the other scenic functions in Block B are located in West Lake and Xixi Wetland Park, as shown in Fig. 9(a). In contrast, TF-IDF cannot reveal the education function and mainly regards the rest as residence and commerce functions from Fig. 9(b). In addition, MuF characterizes Block A with scenic and commerce and Block B with workplace, scenic, and education, which are all consistent with the actual situations. Through the analysis, MuF can discover region functions effectively, especially for those functions that are not ubiquitous but can characterize the region features, such as scenic and education.

E. Analysis of TOD Regions

In the previous experiments, we have verified the effectiveness of the proposed methods in terms of region partition, region identification, and function characterization. Hereby, we formally give the function distributions of identified TOD regions and their surroundings, as shown in Fig. 10. The underpainting represents the function with the first share, the dot shows the second proportion function, and the circle around the dot is the third. From Fig. 10, we can see that

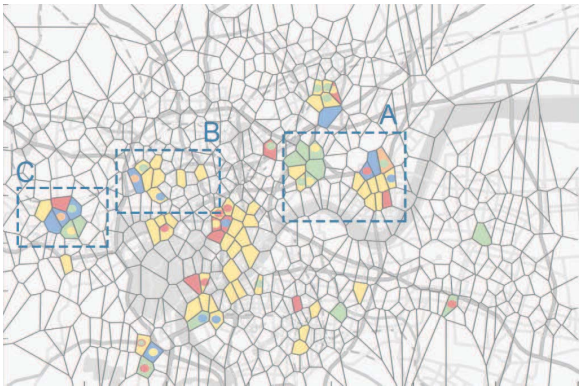


Fig. 10. Function distributions of TOD regions and their surroundings.

the functions of TOD regions are relatively independent and distinctive, which accords with the goal of function diffusion in Hangzhou urban planning. Concretely speaking, the left part of Block A is Jiuqiao Commercial Town that indeed undertakes commerce function at present; the right part is Xiasha that is planned to be an industrial and education zone, but it lacks industrial function according to the current construction status. Furthermore, Yuhan Group in Block C achieves the initial success toward the residential suburb and scientific research base, while North Town in Block B still needs further development to meet the requirement of a business and financial town.

VII. Conclusion

In this paper, we managed to answer the three critical problems in TOD study, especially for megacities leveraging heterogeneous urban data. The experiments, which were conducted on three real data sets, including public transit data, POIs, and taxi trajectories, proved the effectiveness of the proposed methods in their respective fields. Moreover, we made a careful analysis of some representative blocks, which also offers scientific data support for the government to make policy for the development of a megacity, using Hangzhou as an example.

As a first step on data-driven studying TOD, the data quality directly affects the experimental outcomes, which has been embodied by the unsatisfied results in some remote small areas and some complex situations in heartlands. Hence, high-quality data sets can facilitate the ongoing research. Furthermore, we strive to further improve the performances of introduced approaches in each research area.

Acknowledgment

The authors would like to thank M. Wang and X. Yu for their help to carry out the experiments.

References

[1] B. P. Y. Loo, C. Chen, and E. T. H. Chan, "Rail-based transit-oriented development: Lessons from New York City and Hong Kong," *Landscape Urban Planning*, vol. 97, no. 3, pp. 202–212, 2010.

[2] J. Holmes and J. van Hemert, "Transit oriented development," in *Rocky Mountain Land Use Institute*, 2008.

[3] J. L. Renne, *Transit Oriented Development: Making it Happen*. Evanston, IL, USA: Routledge, 2016.

[4] X. J. Kong, X. Song, F. Xia, H. Guo, J. Wang, and A. Tolba, "LoTAD: Long-term traffic anomaly detection based on crowdsourced bus trajectory data," *World Wide Web*, vol. 21, no. 3, pp. 825–847, 2017.

[5] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, p. 38, 2014.

[6] X. Kong, Z. Xu, G. Shen, J. Wang, Q. Yang, and B. Zhang, "Urban traffic congestion estimation and prediction based on floating car trajectory data," *Future Generat. Comput. Syst.*, vol. 61, pp. 97–107, Aug. 2016.

[7] T. Louail *et al.*, "From mobile phone data to the spatial structure of cities," *Sci. Rep.*, vol. 4, p. 5276, Jun. 2014.

[8] T. Louail *et al.*, "Uncovering the spatial structure of mobility networks," *Nature Commun.*, vol. 6, p. 6007, Jan. 2015.

[9] N. J. Yuan, Y. Zheng, X. Xie, Y. Wang, K. Zheng, and H. Xiong, "Discovering urban functional zones using latent activity trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 3, pp. 712–725, Mar. 2015.

[10] C. Zhong, X. Huang, S. M. Arisona, G. Schmitt, and M. Batty, "Inferring building functions from a probabilistic model using public transportation data," *Comput., Environ. Urban Syst.*, vol. 48, pp. 124–137, Nov. 2014.

[11] T. T. Cao, H. Edelsbrunner, and T. S. Tan, "Triangulations from topologically correct digital Voronoi diagrams," *Comput. Geometry*, vol. 48, no. 7, pp. 507–519, 2015.

[12] D. Pojani and D. Stead, "Ideas, interests, and institutions: Explaining dutch transit-oriented development challenges," *Environ. Planning A, Economy Space*, vol. 46, no. 10, pp. 2401–2418, 2014.

[13] X. Kong, M. Li, T. Tang, K. Tian, L. Moreira-Matias, and F. Xia, "Shared subway shuttle bus route planning based on transport data analytics," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 4, pp. 1507–1520, Oct. 2018.

[14] H. Sung and J.-T. Oh, "Transit-oriented development in a high-density city: Identifying its association with transit ridership in Seoul, Korea," *Cities*, vol. 28, no. 1, pp. 70–82, 2011.

[15] E. Papa and L. Bertolini, "Accessibility and transit-oriented development in European metropolitan areas," *J. Transp. Geogr.*, vol. 47, pp. 70–83, Jul. 2015.

[16] R. Cervero and D. Dai, "BRT TOD: Leveraging transit oriented development with bus rapid transit investments," *Transp. Policy*, vol. 36, pp. 127–138, Nov. 2014.

[17] M. Kamruzzaman, D. Baker, S. Washington, and G. Turrell, "Advance transit oriented development typology: Case study in Brisbane, Australia," *J. Transp. Geography*, vol. 34, pp. 54–70, Jan. 2014.

[18] C. Zhong, X. Huang, S. M. Arisona, and G. Schmitt, "Identifying spatial structure of urban functional centers using travel survey data: A case study of Singapore," in *Proc. 1st ACM SIGSPATIAL Int. Workshop Comput. Models Place (COMP)*, Orlando, FL, USA, Nov. 2013, pp. 28:28–28:33.

[19] C. Zhong, M. Schläpfer, S. M. Arisona, M. Batty, C. Ratti, and G. Schmitt, "Revealing centrality in the spatial structure of cities from human activity patterns," *Urban Stud.*, vol. 54, no. 2, pp. 437–455, 2017.

[20] P. Wang, Y. Fu, J. Zhang, X. Li, and D. Lin, "Learning urban community structures: A collective embedding perspective with periodic spatial-temporal mobility graphs," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 6, p. 63, 2018.

[21] Y. Zhang *et al.*, "Mobile social big data: Wechat moments dataset, network applications, and opportunities," *IEEE Netw.*, vol. 32, no. 3, pp. 146–153, May/Jun. 2018.

[22] S. Chen, H. Wu, L. Tu, and B. Huang, "Identifying hot lines of urban spatial structure using cellphone call detail record data," in in *Proc. IEEE 11th Intl Conf Ubiquitous Intell. Comput., IEEE 11th Intl Conf Autonomic Trusted Comput., IEEE 14th Intl Conf. Scalable Comput. Commun. Associated Workshops (UTC-ATC-ScalCom)*, Bali, Indonesia, Dec. 2014, pp. 299–304.

[23] C. Ratti *et al.*, "Redrawing the map of great britain from a network of human interactions," *PLoS One*, vol. 5, no. 12, 2010, Art. no. e14248.

[24] X. Kong, F. Xia, J. Wang, A. Rahim, and S. K. Das, "Time-location-relationship combined service recommendation based on taxi trajectory data," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1202–1212, Jun. 2017.

[25] S. Rinzivillo, S. Mainardi, F. Pezzoni, M. Coscia, D. Pedreschi, and F. Giannotti, "Discovering the geographical borders of human mobility," *KI-Künstliche Intelligenz*, vol. 26, no. 3, pp. 253–260, Aug. 2012.

[26] C. Zhong, S. M. Arisona, X. Huang, M. Batty, and G. Schmitt, "Detecting the dynamics of urban structure through spatial network analysis," *Int. J. Geographical Inf. Sci.*, vol. 28, no. 11, pp. 2178–2199, Nov. 2014.

[27] C. Roth, S. M. Kang, M. Batty, and M. Barthélemy, "Structure of urban movements: Polycentric activity and entangled hierarchical flows," *PLoS One*, vol. 6, no. 1, 2011, Art. no. e15923.

[28] Y. Fu *et al.*, "Representing urban forms: A collective learning model with heterogeneous human mobility data," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 3, pp. 535–548, Mar. 2019.

[29] M. Chen, K. Kuzmin, and B. K. Szymanski, "Community detection via maximization of modularity and its variants," *IEEE Trans. Comput. Social Syst.*, vol. 1, no. 1, pp. 46–65, Mar. 2014.

[30] C. Sheng, Y. Zheng, W. Hsu, M. L. Lee, and X. Xie, "Answering top-κ similar region queries," in *Proc. 15th Int. Conf. Database Syst. Adv. Appl. (DASFAA)*, Tsukuba, Japan, Apr. 2010, pp. 186–201.

[31] G. R. Calegari, E. Carlino, D. Peroni, and I. Celino, "Extracting urban land use from linked open geospatial data," *ISPRS Int. J. Geo-Inf.*, vol. 4, no. 4, p. 2109, 2015.

[32] X. Meng, Z. Ding, and J. Xu, "Clustering analysis of moving objects," in *Moving Objects Management*, 2014, pp. 163–195.

[33] G. Qi, X. Li, S. Li, G. Pan, Z. Wang, and D. Zhang, "Measuring social functions of city regions from large-scale taxi behaviors," in *Proc. IEEE Int. Conf. Pervas. Comput. Commun. Workshops (PERCOM Workshops)*, Seattle, WA, USA, Mar. 2011, pp. 384–388.

[34] C. Peng, X. Jin, K.-C. Wong, M. Shi, and L. Pietro, "Collective human mobility pattern from taxi trips in urban area," *PLoS One*, vol. 7, no. 4, pp. 1–8, 2012.

[35] 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 (KDD)*, Beijing, China, Aug. 2012, pp. 186–194.

[36] Y. Liu, F. Wang, Y. Xiao, and S. Gao, "Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai," *Landscape Urban Planning*, vol. 106, no. 1, pp. 73–87, 2012.

[37] G. Pan, G. Qi, Z. Wu, D. Zhang, and S. Li, "Land-use classification using taxi GPS traces," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 1, pp. 113–123, Mar. 2013.

[38] Y. Long and Z. Shen, "Discovering functional zones using bus smart card data and points of interest in beijing," in *Geospatial Analysis to Support Urban Planning in Beijing*, vol. 116. Cham, Switzerland: Springer, 2015, pp. 193–217.

[39] R. R. Vatsavai, E. Bright, C. Varun, B. Budhendra, A. Cheriyadat, and J. Grasser, "Machine learning approaches for high-resolution urban land cover classification: A comparative study," in *Proc. 2nd Int. Conf. Comput. Geospatial Res. Appl. (COM.Geo)*, Washington, DC, USA, May 2011, pp. 11:1–11:10.

[40] C. Wang, X. Meng, Q. Guo, Z. Weng, and C. Yang, "Automating characterization deployment in distributed data stream management systems," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 12, pp. 2669–2681, Dec. 2017.

[41] F. Kling and A. Pozdnoukhov, "When a city tells a story: Urban topic analysis," in *Proc. 20th Int. Conf. Adv. Geographic Inf. Syst. (IGSPATIAL)*, Redondo Beach, CA, USA, Nov. 2012, pp. 482–485.

[42] T. Pei, S. Sobolevsky, C. Ratti, S.-L. Shaw, T. Li, and C. Zhou, "A new insight into land use classification based on aggregated mobile phone data," *Int. J. Geograph. Inf. Sci.*, vol. 28, no. 9, pp. 1988–2007, 2014.

[43] Z. Cao, S. Wang, G. Forestier, A. Puissant, and C. F. Eick, "Analyzing the composition of cities using spatial clustering," in *Proc. 2nd ACM SIGKDD Int. Workshop Urban Comput. (UrbComp)*, Chicago, Illinois, Aug. 2013, pp. 14:1–14:8.

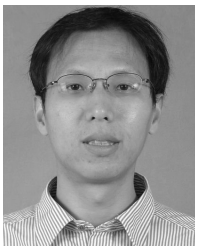
[44] V. Grolmusz, "A note on the pagerank of undirected graphs," *Inf. Process. Lett.*, vol. 115, nos. 6–8, pp. 633–634, 2015.

[45] J. H. Paik, "A novel TF-IDF weighting scheme for effective ranking," in *Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, Jul./Aug. 2013, pp. 343–352.

[46] X. J. Kong *et al.*, "Mobility dataset generation for vehicular social networks based on floating car data," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 3874–3886, May 2018.

[47] M. S. Elgamel and A. Dandoush, "A modified Manhattan distance with application for localization algorithms in ad-hoc WSNs," *Ad Hoc Netw.*, vol. 33, pp. 168–189, Oct. 2015.

[48] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.



Xiangjie Kong (M'13–SM'17) received the B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China.

He is currently an Associate Professor with the School of Software, Dalian University of Technology, Dalian, China. He has published over 100 scientific papers in international journals and conferences (with over 70 indexed by ISI SCIE). His current research interests include computational social science, mobile computing, and urban big data.

Dr. Kong is a Senior Member of China Computer Federation and a member of ACM. He has served as the (guest) editor for several international journals and the workshop chair or a PC member of a number of conferences.

Feng Xia (M'07–SM'12) received the B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China.

He was a Research Fellow with the Queensland University of Technology, Brisbane, QLD, Australia. He is currently a Full Professor with the School of Software, Dalian University of Technology, Dalian, China. He has published two books and over 200 scientific papers in international journals and conferences. His current research interests include data science, big data, knowledge management, network science, and systems engineering.

Dr. Xia is a Senior Member of ACM. He serves as the general chair, the PC chair, the workshop chair, or the publicity chair of a number of conferences. He is also the (guest) editor of several international journals.

Kai Ma received the bachelor's degree in software engineering from the Dalian University of Technology, Dalian, China, in 2017, where he is currently pursuing the M.Sc. degree with The Alpha Lab, School of Software.

His current research interests include vehicular social networks, vehicle data generation, and human behavior.

Jianxin Li received the Ph.D. degree in computer science from the Swinburne University of Technology, Melbourne, VIC, Australia, in 2009.

He is currently an Associate Professor with the School of Info Technology, Deakin University, Burwood, VIC, Australia. His current research interests include database query processing and optimization, social network analytics, and traffic network data.

Qiuyuan Yang received the bachelor's and master's degrees in software engineering from the Dalian University of Technology, Dalian, China, in 2013 and 2016, respectively.

She is currently with Baidu, Beijing, China. Her current research interests include mobile social networks, big traffic data, and socially aware networking.