Discovering Urban Functional Zones Using Latent Activity Trajectories

Nicholas Jing Yuan, *Member, IEEE,* Yu Zheng, *Senior Memeber, IEEE,* Xing Xie, *Senior Member, IEEE,* Yingzi Wang, Kai Zheng, Hui Xiong, *Senior Member, IEEE*

Abstract—The step of urbanization and modern civilization fosters different functional zones in a city, such as residential areas, business districts, and educational areas. In a metropolis, people commute between these functional zones every day to engage in different socioeconomic activities, e.g., working, shopping, and entertaining. In this paper, we propose a data-driven framework to discover functional zones in a city. Specifically, we introduce the concept of Latent Activity Trajectory (LAT), which captures socioeconomic activities conducted by citizens at different locations in a chronological order. Later, we segment an urban area into disjointed regions according to major roads, such as highways and urban expressways. We have developed a topic-modeling-based approach to cluster the segmented regions into functional zones leveraging mobility and location semantics mined from LAT. Furthermore, we identify the intensity of each functional zone using Kernel Density Estimation. Extensive experiments are conducted with several urban scale datasets to show that the proposed framework offers a powerful ability to capture city dynamics and provides valuable calibrations to urban planners in terms of functional zones.

Index Terms—Functional zones, latent activity trajectories, human mobility, points of interest

1 Introduction

Modern cities develop with the gestation, formation and maturity of different *functional zones*. These functional zones provide people with various *urban functions* to meet their different needs of *socioeconomic activities* (hereinafter interchangeably referred to as "activities"), e.g., Wall Street is a well-known financial district in New York City, and Silicon Valley is a high-technology business region of the San Francisco Bay Area. These functional zones can either be artificially designed by urban planners (termed as zoning [2]), or naturally formulated according to people's actual lifestyles. Meanwhile, both the territories and functions of these zones can be reshaped during the evolution of a city. Discovering functional zones is crucial for uncovering the physical and social characters of a city, and can enable a variety of valuable applications, such as tourism recommendation, business site selection, and calibration for urban planning.

The recent proliferation of ubiquitous sensing technologies, intelligent transportation systems, and location based services increases the availability of human trajectories. For example, in big cities like New York, Munich and Beijing, most taxis are equipped with GPS devices for dispatching and security management. These taxis regularly report their locations to the data center at a certain frequency. Hence, a large number of taxi trajectories are cumulated every day. Another good example is smart cards and integrated ticketing, which are provided by public transit operators in many cities. Customers can swipe the purchased cards to check-in and check-out when using public transport like subways or buses. Examples include London's Oyster Card, Dublin's Leap Card, Hong Kong's Octopus Card, and Beijing's BMAC Card [3].

In addition to revealing human mobility, these trajectories imply the socioeconomic activities of people at different locations at different times, since the activities are actually the essential reason that mobilizes people to commute between different places. We term such a trajectory as a *Latent Activity Trajectory* (LAT), where sequential locations visited by the users are observable while socioeconomic activities implied by the sequence of locations are latent. For instance, a taxi trajectory can be segmented into multiple trips pertaining to different customers, where each customer travels from an origin to a destination for a certain activity, e.g., going to work from home on a weekday morning, or going shopping on a weekend evening.

In this paper, we aim to discover functional zones in urban areas leveraging latent activity trajectories. Typically, a city is naturally partitioned into individual regions by major roads, like expressways and ring roads (refer to the white lines in Figure 1(a)). *A functional zone* is comprised of *a number*

- *Nicholas Jing Yuan, Yu Zheng, Xing Xie are with Microsoft Research, Email:{nicholas.yuan, yuzheng, xingx}@microsoft.com*
- *Yingzi Wang is with University of Science and Technology of China, Email: yingzi@mail.ustc.edu.cn*
- *Kai Zheng is with the University of Queensland, Email: kevinz@itee.uq.edu.au*
- *Hui Xiong is with State University of New Jersey, Email: hxiong@rutgers.edu*

This article is an expanded version of [1], which appeared in Proceedings of the 18th ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD 2012).

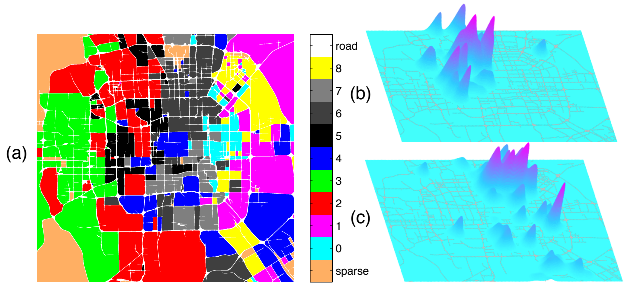


Fig. 1. Territory and intensity of functional zones. (a) functional zones identified in Beijing (indicated by different colors); (b) intensity of the developed commercial functional zone (labeled 5 in (a)); (c) intensity of the developed residential functional zone (labeled 6 in (a)).

of regions (not necessarily connected) with similar urban *functions*, where the function of a region is represented by the *distribution of socioeconomic activities*. For example, figure 1(a) shows the functional zones we have identified in the urban area of Beijing, where different colors indicate different functional zones. Furthermore, we analyze the functionality intensity in different locations of a functional zone. For instance, Figure 1 (b) and (c) show the functionality intensity of developed commercial/entertainment areas and residential areas respectively, where the higher hills suggest a higher intensity.

To identify the function of a region (the unit of a functional zone), we need to take into account two underlying signals from LAT, which reveal the socio-economic activities of citizens, thus reflecting urban functions:

1) **Mobility Semantics**. The activities conducted in a region are strongly associated with the spatiotemporal patterns of the people who visit that region. The knowledge that human mobility contributes to reveal the urban function of a region mainly is two fold. One is when people arrive at and leave a region, and the other is where people come from and leave for. Intuitively, in a workday people usually leave a residential area in the morning and return in the evening. The major time when people visit an entertainment area, however, is the evening of workdays or the whole day of non-workdays. Furthermore, different functional zones are correlated in the context of human mobility. For instance, there is a high probability that people reaching an entertainment area are originating from a working area (on a workday) and a residential area (on non-workdays). As a result, two zones are more likely to have similar functions, if people traveling to the two zones come from similar functional zones or leave for similar ones.

2) **Location Semantics**. The urban road network is leveraged as a kind of location semantics for segmenting the urban area into regions, since different regions are geo-spatially connected with each other through the road network. Another form of location semantics, the allocation of Points of Interest (POI), which are typically associated with a coordinate and a category like restaurants or shopping malls, uncovers the potential socioeconomic activities of a region. For example, a region containing a number of universities and schools has a high probability of being an educational area. A region that usually contains a variety of POIs is probably serving multiple socioeconomic activities instead of a single one. Some regions may serve as both business districts and entertainment areas in a city. In addition, the information from POI data cannot differentiate the quality of different venues and reflect the interactions between functional zones. For instance, restaurants are everywhere in a city, but they could denote different functions. Some small restaurants were built just for satisfying local residences' daily needs, while a few famous restaurants attracting many people might be regarded as a feature location of an entertainment area. As a result, sometimes two regions sharing a similar distribution of POIs could still have different functions.

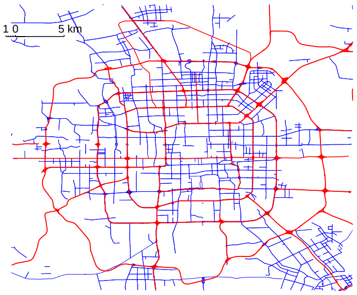


Fig. 2. Beijing road network. red: level-0/1; blue: level-2

This paper is an extension of our previous paper [1], in which we presented a topic-modeling-based approach for discovering region functions and intensity of functionality using POIs and human mobility. In this paper, we further offer the following contributions:

- We have introduced the concept of latent activity trajectory, and generalized the problem of identifying functional zones using both location and mobility semantics mined from latent activity trajectories.
- We have developed and detailed a morphological approach to segment a city into individual regions and presented a collaborative-filtering-based approach to learn the location semantics from POI configurations of a region, which outperforms the TF-IDF vectors (as metadata for topic modeling) used in our previous paper based on experimental results.
- We performed exploratory study with extensive experiments using large-scale and real-world datasets with regard to Beijing. In addition to the taxi trajectory data used in our previous work, we utilized pubic transit records of 1.5M trips from 0.3M card holders. The results suggest that the performance of our model is improved by integrating heterogeneous mobility datasets and considering collaborative location semantics of different regions.

2 Map Segmentation

A road network is usually comprised of some major roads like highways and ring roads, which naturally partition a city into regions. For example, as shown in Figure 2, the red segments denote freeways and city expressways in Beijing, and blue segments represent urban arterial roads. The three kinds of roads are associated with a road level 0, 1, and 2 respectively (in a road network database), forming a natural segmentation of the urban area of Beijing. Intuitively, we consider each segmented region a basic unit carrying urban functions since POIs often fall inside regions and people perform socioeconomic activities (such as staying home and working) inside regions, which is also the root cause of human mobility.

Typically, in a Geographical Information System (GIS), there are two models to represent spatial data: a *vector*-based model and a *raster*-based model. The vector-based model uses geometric primitives such as points, lines and polygons to represent spatial objects referenced by Cartesian coordinates, while the raster-based model quantizes an area into small discrete grid-cells. Both models have advantages and disadvantages depending on the specific applications. For instance, on one hand, the vector-based method is more powerful for precisely finding the shortest-paths, but requires intensive computation when performing topological analysis, such as map simplification[4], which is proven to be NP-complete [4]. On the other hand, the raster-based model is more computationally efficient and succinct for territorial analysis, but the accuracy is limited by the number of cells used for discretizing road networks.

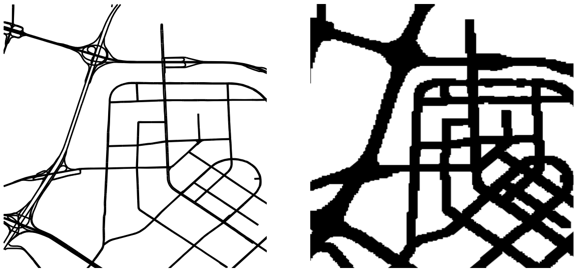
We display the vector-based road network on a plane by performing map projection [5], which transforms the surface of a shpere (i.e., the Earth) into a 2D plane (we use Mercator projection in our implementation [6]). Then we convert the vector-based road network into the raster model by gridding the projected map. [1] Intuitively, each pixel of the projected map image can be regarded as a grid-cell of the raster map. Consequently, the road network is converted to a binary image, e.g., 1 stands for the road segments (termed as foreground) and 0 stands for the blank areas (termed as background).

This section introduces an image processing approach for segmenting the raster-based road network into regions through morphological operators.[2]

2.1 Dilation

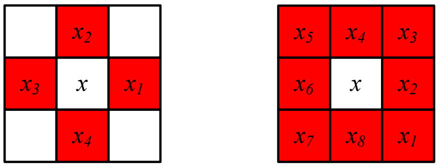
In general, a morphological operator calculates the output image given the input binary image and a *structure element*, whose size and shape are pre-defined. *Dilation* is a basic morphological operator. Let A be a binary image and B be the structure element, the dilation of A by B is defined as:

$$A \oplus B = \bigcup_{b \in B} A_b, \tag{1}$$



(a) Before dilation (b) After dilation

Fig. 3. Dilation operator



(a) 4-neighbours (b) 8-neighbours

Fig. 4. The 4-neighbours and 8-neighbours of x

where $A_b = \{a + b | a \in A\}$, i.e., the translation of A by vector b. The dilation operator is commutative.

For any p in A, after the dilation, $p = 1$ iff the intersection between A and B, centred at p, is not empty.

The purpose of the dilation operation is to remove the unnecessary details for map segmentation, avoiding the small connected areas induced by these unnecessary details such as bridges and lanes. Figure 3(a) plots a portion of road network before the dilation operator. As is shown, the small holes between the lanes and viaducts are filled, where we use a 3×3 matrix with all values set to 1 as the structure element B.

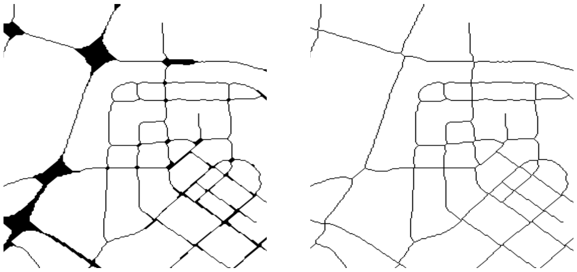
2.2 Thinning

As a consequence of the previous dilation operator, the road segments are turgidly thickened. In this step, we aim to extract the skeleton of the road segments while keeping the topology structure (such as the Euler number) of the original binary image. The *thinning operator* is performed to remove certain foreground pixels from the input binary image. For a given pixel in the input image, whether it should be removed depends on its neighbouring pixels. For a given pixel x, the neighbouring 4 pixels shown in Figure 4(a) are called the *4-neighbours* of x. Similarly, the 8 neighbouring pixels shown in Figure 4(b) are called the *8-neighbours* of x. Here, we employ the subfields-based parallel thinning algorithm proposed in [7]. This method first divides the binary image space into two disjointed subfields in a checkerboard pattern, then iterations are performed to remove foreground pixels. Each iteration consists of two sub-iterations in these two subfields:

- In the first sub-iteration, we check every pixel p in the first subfield, delete p iff Condition 1, 2 and 3 are all satisfied.

[1] We used a 2400× 2400 grid to rasterize the map of Beijing with left-top geo-coordinates (40.09, 116.17), right-bottom geo-coordinates (39.77, 116.56), which covers the main area of downtown Beijing.

[2] Sample dataset and code can be downloaded at http://1drv.ms/1lhQ4xn.



(a) After 7 iterations (b) Until convergence

Fig. 5. Thinning operator

- In the second sub-iteration, we check every pixel p in the second subfield, delete p iff Condition 1, 2 and 4 are all satisfied.

Condition 1: $X_H(p) = 1$, where

$$X_H(p) = \sum_{i=1}^{4} b_i$$

$$b_i = \begin{cases} 1 & \text{if } x_{2i-1} = 0 \text{ and } (x_{2i} = 1 \text{ or } x_{2i+1} = 1) \\ 0 & \text{otherwise} \end{cases}$$

Condition 2: $2 \leq \min\{n_1(p), n_2(p)\} \leq 3$, where

$$n_1(p) = \sum_{k=1}^{4} x_{2k-1} \vee x_{2k}$$

$$n_2(p) = \sum_{k=1}^{4} x_{2k} \vee x_{2k+1}$$

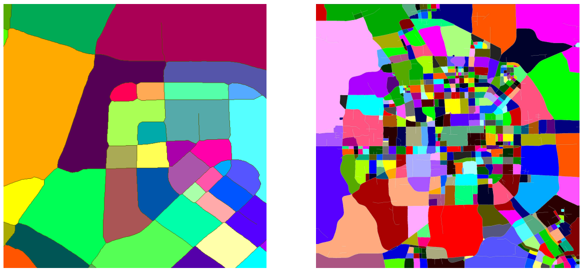
Condition 3: $(x_2 \vee x_3 \vee \bar{x}_8) \wedge x_1 = 0$

Condition 4: $(x_6 \vee x_7 \vee \bar{x}_4) \wedge x_5 = 0$

The above conditions ensure that the connectivity of the pixels is preserved when a certain pixel is deleted. Note that in this operation, connectivity paradox may be induced if we keep the same type of connectivity (4-connected or 8-connected) for both the foreground and the background[8]. Since it is desired for the road segments (foreground) to have unit width, typically, we preserve the 8-connectivity of the foreground (i.e., the 8-connectivity does not change before and after the thinning process for the road segments) and the 4-connectivity of the background [9]. Figure 5(a) and Figure 5(b) are the results after 7 iterations and until convergence (no pixel will be deleted any more) respectively.

2.3 Connected Component Labeling

The *connected component labeling* operation finds the connected 0 pixels (the blank area) in the binary image, after the thinning operation. We call the sequence $y_1, y_2, \ldots, y_n$ an *8-path* (*4-path*), if $\forall i = 1, 2, \ldots, n-1$, y_{i+1} is an 8-neighbour (4-neighbour) of y_i. We say a region Q in a binary image is *8-connected* (*4-connected*) iff all the pixels in Q have the same value and for any two pixels in Q, there exists an 8-path (4-path) connecting the two pixels. There exist many algorithms for connected component labeling. Here, we apply the classical two-pass algorithm introduced in [10] to the binary image 5(b), and obtain the segmented regions as shown in Figure 6(a). Figure 6(b) presents the result for Beijing's entire road network. We note that the computational complexity of all the morphological operations in the map segmentation method is linear in terms of number of pixels.



(a) Local result (b) Global result

Fig. 6. Segmented regions after connected component labeling

3 Discovery of Activities in a Region

In this section, we infer the distribution of activities in each region unit using a topic-model-based method.

3.1 Preliminary

Definition 1 (Transition): A *transition* Tr is a quadruple containing the following four items: origin region ($Tr.r_O$), leaving time ($Tr.t_L$), destination region ($Tr.r_D$) and arrival time ($Tr.t_A$). Here, $Tr.r_O$ and $Tr.r_D$ are spatial features while the others are temporal features.

Definition 2 (Mobility Pattern): A *mobility pattern* M is a triple extracted from a transition. Given a transiton $Tr = (Tr.r_O, Tr.r_D, Tr.t_L, Tr.t_A)$, we obtain two mobility patterns: the *leaving mobility pattern* $M_L = (Tr.r_O, Tr.r_D, Tr.t_L)$, and the *arriving mobility pattern* $M_A = (Tr.r_O, Tr.r_D, Tr.t_A)$.

Definition 3 (Transition Cuboids): A *transition cuboid* C is an $R \times R \times T$ cuboid, where R is the number of regions and T is the number of time bins. Since there exist two types of mobility patterns, we define two types of transition cuboids: *leaving cuboid* C_L and *arriving cuboid* C_A. The cell with index (i, j, k) of the leaving cuboid records the number of mobility patterns that leave r_i for r_j at time t_k, i.e.,

$$C_L(i,j,k) = \|\{M_L = (x,y,z) | x = r_i, y = r_j, z = t_k\}\|.$$

Similarly,

$$C_A(i,j,k) = \|\{M_A = (x,y,z) | x = r_i, y = r_j, z = t_k\}\|.$$

In order to derive mobility semantics (represented by the transition cuboids defined above) from latent activity trajectories, we project each trajectory on the segmented region units, turning a trajectory into a transition (note that for both taxi trajectories and public transit records, the transitions obtained pertain to a certain individual). Then, we discretize time of day into time bins in each of which we deposit the transitions and formulate mobility patterns. Here, we do not differentiate different weekdays but differ the time bins in weekdays from those in weekends. For example, setting 2 hours as a bin, we will have 24 bins (12 for weekdays and 12 for weekends) in total. Later, two transition cuboids are built using LAT.

3.2 Collaborative POI Feature Vector

To learn the location semantics, we calculate the distribution of POIs for each region. A POI is recorded with a tuple (in a POI database) consisting of a POI category (as listed in Table 2), name and a geo-position (latitude, longitude). For each region $r_i, i = 1, 2, \ldots, R$, the number of POIs in each POI category can be counted. Later, we calculate the *term frequency-inverse document frequency* (TF-IDF) to measure the importance of a POI in a region. Specifically, for a given region r_i, we formulate a POI vector, $f_i = (v_{i1}, v_{i2}, \ldots, v_{iC})$ where v_{ij} is the TF-IDF value of the j-th POI category and C is the number of POI categories. The TF-IDF value v_{ij} is given by:

$$v_{ij} = \frac{n_j}{N_i} \times \log \frac{R}{\|\{r_i|\text{the } j\text{-th POI category} \in r_i\}\|}, \quad (2)$$

where n_j is the number of POIs belonging to the j-th category and N_i is the number of POIs located in region r_i. The idf term is calculated by computing the quotient of the number of regions R divided by the number of regions which have the j-th POI category, and taking the logarithm of that quotient.

However, the TF-IDF vector is still not a good representation of a region's location semantics, which mainly suffers from the following limitations: 1) *Missing values*. There might exist some POIs in a region, which are not in the current POI database, while featuring the location semantics of that region. 2) *Latent structure*. The frequency of POI is not an intrinsic representation of the latent structure of the POI configuration for each region [11], thus it is not suitable for measuring the similarity of location semantics between regions.

Motivated by the collaborative filtering techniques in recommender systems [12], we employ the Singular Value Decomposition (SVD) method to obtain the latent semantics of each region in terms of POI configuration, which inherently tackles the above limitations. Specifically, let $\mathbf{F} = (f_1, f_2, \ldots, f_R)^\mathsf{T}$ be the matrix containing the TF-IDF vectors for all regions, with dimension $R \times C$. As is shown in Figure 7, $\mathbf{F}$ is a sparse matrix, since for many regions, there may be no certain categories of POIs. We employ SVD to decompose $\mathbf{F}$ by

$$\mathbf{F} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}, \quad (3)$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices with dimension $R \times R$ and $C \times C$ respectively, and $\mathbf{\Sigma}$ is a diagonal matrix with singular values of $\mathbf{F}$. Then we can approximate $\mathbf{F}$ with $\hat{\mathbf{F}} = \mathbf{U}_l\mathbf{\Sigma}_l\mathbf{V}_l^\mathsf{T}$, where $\mathbf{\Sigma}_l$ is a $l \times l$ low rank matrix containing only the largest l singular values of $\mathbf{\Sigma}$, and $\mathbf{U}_l$, $\mathbf{V}_l$ are the reduced matrices with corresponding l columns and C rows, respectively. Now, the location semantics space is represented by the $R \times l$ matrix $\mathbf{U}_l = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_R)^\mathsf{T}$, where row $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{il})$ is termed as the *collaborative POI feature vector* for region r_i. This representation can be regarded as the coordinates of each region in the location semantics space.

$$\begin{array}{c} \mathbf{F} \\ \begin{array}{cc} & \begin{array}{cccc} c_1 & c_2 & \cdots & c_C \end{array} \\ \begin{array}{c} r_1 \\ r_2 \\ \vdots \\ r_R \end{array} & \begin{pmatrix} \times & & & \\ & & \times & \\ & & & \\ \times & & & \times \end{pmatrix} \end{array} \\ R \times C \end{array} \approx \hat{\mathbf{F}} = \begin{array}{c} \mathbf{U}_l \\ \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \cdots \\ \mathbf{x}_R \end{pmatrix} \\ R \times l \end{array} \begin{array}{c} \mathbf{\Sigma}_l \\ \begin{pmatrix} \quad \end{pmatrix} \\ l \times l \end{array} \begin{array}{c} \mathbf{V}_l^\mathsf{T} \\ \begin{pmatrix} \qquad\qquad \end{pmatrix} \\ l \times C \end{array}$$

Fig. 7. Computing collaborative POI feature vector

As a result, the collaborative POI feature vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_R$ are incorporated as metadata in our model introduced below.

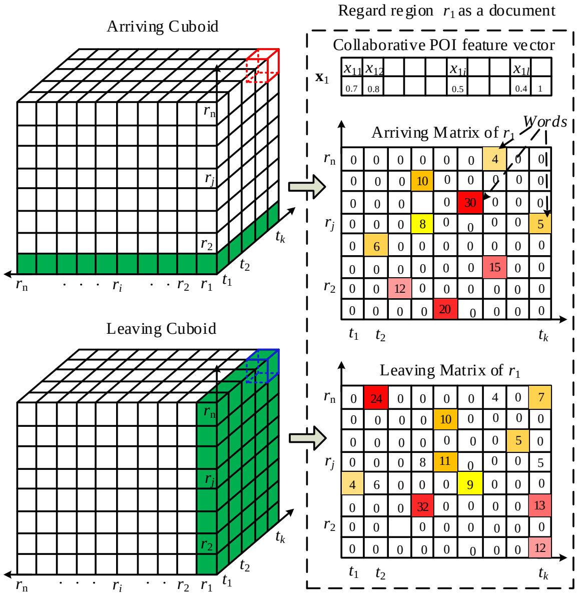
3.3 Topic Modeling

In text mining, probabilistic topic models have been successfully used for extracting the hidden semantic structure in large archives of documents[13]. In this model, each *document* of a *corpus* exhibits multiple *topics* and each *word* of a document supports a certain topic. Given all the words of each document in a corpus as observations, a topic model is trained to infer the hidden semantic structure behind the observations.

The problem of identifying the latent activities in a region can be analogized to the problem of discovering the latent topics of a document. As shown in Table 1, we regard a region as a document and an activity as a topic. In other words, a region having multiple activities is just like a document containing a variety of topics. Meanwhile, we deem the mobility patterns (representing mobility semantics) associated with a region as words and collaborative POI feature vectors (representing location semantics) as metadata of a document. Since a functional zone is characterized by its agglomeration of activities, its intraregional transport infrastructure, mobility of people, and inputs are within its interaction borders [14].

TABLE 1
Analogy from region-activities to document-topics

transition cuboids	⟶	vocabulary
regions	⟶	documents
activities of a region	⟶	topics of a document
mobility patterns	⟶	words
collaborative POI feature vector	⟶	metadata of a document


Arriving Cuboid
r_n
r_j
r_2
t_k
t_2
t_1
r_n · · · r_i · · r_2 r_1
Leaving Cuboid
r_n
r_j
r_2
t_k
t_2
t_1
r_n · · · r_i · · r_2 r_1
Regard region r_1 as a document
Collaborative POI feature vector
Words

$\mathbf{x}_1$	x_{11}	x_{12}				x_{1i}			x_{1l}	
	0.7	0.8				0.5			0.4	1

Arriving Matrix of r_1

r_n	0	0	0	0	0	0	4	0	0
	0	0	0	10	0	0	0	0	0
	0	0	0		0	30	0	0	0
r_j	0	0	0	8	0	0	0	0	5
	0	6	0	0	0	0	0	0	0
	0	0	0	0	0	0	15	0	0
r_2	0	0	12	0	0	0	0	0	0
	0	0	0	0	20	0	0	0	0
	t_1	t_2							t_k

Leaving Matrix of r_1

r_n	0	24	0	0	0	0	4	0	7
	0	0	0	0	10	0	0	0	0
	0	0	0	0	0	0	0	5	0
r_j	0	0	0	8	11	0	0	0	5
	4	6	0	0	0	9	0	0	0
	0	0	0	32	0	0	0	0	13
r_2	0	0		0	0	0	0	0	0
	0	0	0	0	0	0	0	0	12
	t_1	t_2							t_k

Fig. 8. Analogy between mobility patterns and words based on transition cuboids

Figure 8 further details the analogy using an example. In our method, given the mobility dataset, we build the arriving and leaving cuboids respectively according to Definition 3. For a specific region r_i, the mobility patterns associated with r_i are counted by $C_A(1{:}R, i, 1{:}T)$ and $C_L(i, 1{:}R, 1{:}T)$, which are two "slices" extracted from the arriving cuboid and the leaving cuboid (termed as arriving matrix and leaving matrix respectively). The right part of Figure 8 shows a "document" we compose for region r_1, where a cell (in the matrices) represents a specific mobility pattern and the numbers in the cell denote the occurrences of the pattern. For example, in the right most column of the arriving matrix, the cell containing "5" means on average the mobility that went to r_1 from r_j in time bin t_k occurred 5 times per day.

Latent Dirichlet Allocation (LDA) is a generative model that includes hidden variables. The intuition behind this model is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words [15]. Let α and η be the prior parameters for the Dirichlet document-topic distribution and topic-word distribution respectively. Assume there are K topics and β is a $K \times M$ matrix where M is the number of words in the vocabulary (all the words in the corpus D). Each β_k is a distribution over the vocabulary. The topic proportions for the dth document are θ_d, where $\theta_{d,k}$ is the topic proportion for topic k in the dth document. The topic assignments for the dth document are z_d, where $z_{d,n}$ is the topic assignment for the nth word in the dth document. Finally, the observed words for document d are w_d, where $w_{d,n}$ is the nth word in document d, which is an element from the fixed vocabulary.

Using the above notations, the generative process can be described as follows:

1) For each topic k, draw $\beta_k \sim Dir(\eta)$.
2) Given the dth document d in corpus D, draw $\theta_d \sim Dir(\alpha)$.
3) For the nth word in the dth document $w_{d,n}$,
   a) draw $z_{d,n} \sim Mult(\theta_d)$;
   b) draw $w_{d,n} \sim Mult(\beta_{z_{d,n}})$.

Here, $Dir(\cdot)$ is the Dirichlet distribution and $Mult(\cdot)$ is the multinomial distribution. The central problem of topic modeling is to estimate the posterior distribution $P(\theta, z, \beta | w, \alpha, \eta)$, which can be accomplished by different approaches, such as Gibbs sampling and variational inference [15].

Using the basic LDA model, region topics can be discovered using mobility patterns. However, as stated in Section 1, the region topics (i.e., activities) are products of both mobility semantics and location semantics. In order to combine the information from both of them, we utilize a more advanced topic model based on LDA and Dirichlet Multinomial Regression (DMR) [16].

Specifically, we incorporated the learned collaborative POI feature vectors (introduced in Section 3.2) into our model. The collaborative POI feature vector of r_i is denoted by $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{il}, 1)$ where the last "1" is a default feature (as shown in Figure 8 for region r_1) to account for the mean value of each topic, as explained in [16]. This vector is regarded as the metadata of each region, which is an analogue of the observed features such as author/email/institution of a document. Such information is used as a prior knowledge to generate the "topics" of a document.

The DMR-based topic model (for simplicity, DMR in the rest of the paper) takes into account the influence of the observable metadata in a document by using a flexible framework, which supports arbitrary features [16]. Compared to other models designed for specific data such as Author-Topic model and Topic-Over-Time model (a member in the supervised-LDA family of topic models), DMR achieves similar or improved performance while is more computationally efficient and succinct in implementation [16].

As presented in Figure 9, the generative process of the DMR model is:

1) For each activity k,
   a) draw $\lambda_k \sim \mathcal{N}(0, \sigma^2 I)$;
   b) draw $\beta_k \sim Dir(\eta)$.
2) Given the ith region r_i,
   a) for each activity k, let $\alpha_{i,k} = \exp(\mathbf{x}_i^T \lambda_k)$;
   b) draw $\theta_i \sim Dir(\alpha_i)$;
   c) for the nth mobility pattern in the ith region $m_{i,n}$,
      i) draw $z_{i,n} \sim Mult(\theta_i)$;
      ii) draw $m_{i,n} \sim Mult(\beta_{z_{i,n}})$.

Here, $\mathcal{N}$ is the Gaussian distribution with σ as a hyper parameter, and λ_k is a vector with the same length as the collaborative POI feature vector. The nth observed mobility pattern of region r_i is denoted as $m_{i,n}$. Other notations are similar to the previous LDA model. In our implementation, the parameters of this model are trained using Gibbs sampling following the method provided in [16].

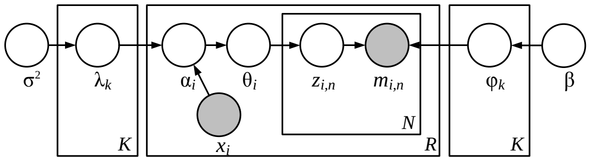


Fig. 9. DMR-based topic model

TABLE 2
POI category taxonomy

code	POI category	code	POI category
1	car service	16	banking and insurance service
2	car sales	17	corporate business
3	car repair	18	street furniture
4	motorcycle service	19	entrance/bridge
5	café/tea bar	20	public utilities
6	sports/stationery shop	21	Chinese restaurant
7	living service	22	foreign restaurant
8	sports	23	fastfood restaurant
9	hospital	24	shopping mall
10	hotel	25	convenience store
11	scenic spot	26	electronic products store
12	residence	27	supermarket
13	governmental agencies and public organizations	28	furniture building materials market
14	science and education	29	pub/bar
15	transportation facilities	30	theaters

Unlike the basic LDA model, here, the Dirichlet prior α is now specified to individual regions (α_i) based on the observed collaborative POI feature vector of each region, i.e., $\alpha_{i,k} = \exp(\mathbf{x}_i^T \lambda_k)$. Therefore, for different combinations of POI category distributions, the resulting α values are distinct. Thus the activity distributions extracted from the data are induced by both the collaborative POI features and mobility patterns. As a result, by applying DMR, given the mobility patterns and collaborative POI feature vectors, we obtain the activity assignment for each region and the mobility pattern distribution of each activity.

4 Territory Identification

4.1 Region Aggregation

This step aggregates similar regions in terms of activity (topic) distributions by performing a clustering algorithm. Regions from the same cluster have similar functions, and different clusters represent different functions. For region r_i, after parameter estimations based on the DMR model, the topic distribution is a K dimensional vector $\theta_i = (\theta_{i,1}, \theta_{i,2}, \dots, \theta_{i,K})$, where $\theta_{i,k}$ is the proportion of topic k for region r_i. We perform the k-means clustering method on the K-dimensional points $\theta_i,\ i \in 1, 2, \dots, R$. The number of clusters can be predefined according to the needs of an application or determined using the average *silhouette* value as the criterion [17]. The silhouette value of a point i in the dataset, denoted by $s(i)$ is in the range of $[-1, 1]$, where $s(i)$ close to 1 means that the point is appropriately clustered and very distant from its neighboring clusters; $s(i)$ close to 0 indicates that the point is not distinctly in one cluster or another; $s(i)$ close to -1 means the point is probably assigned to the wrong cluster. The average silhouette value of a cluster measures how tightly the data in this cluster is grouped, and the average silhouette of the entire dataset reflects how appropriately all the data has been clustered. In practice, we perform cross validation on the dataset for different k multiple times and choose an appropriate k with the maximum overall silhouette value. Consequently, we aggregate the regions into k clusters, each of which is termed as a *functional zone*.

4.2 Functionality Intensity Estimation

On one hand, the functionality of a functional zone is generally not uniformly distributed within the entire region. On the other hand, sometimes, the core functional area may span multiple regions and may have an irregular shape, e.g., a hot shopping street crossing several regions. In order to reveal the degree of functionality and glean the essential territory of a functional zone, we estimate the *functionality intensity* for each aggregated functional zone (a cluster of regions).

Intuitively, the number of visits implicitly reflects the popularity of a certain functional zone. In other words, people's mobility patterns imply the functionality intensity. As a result, we feed the origin and destination of each mobility (represented by latitude and longitude) into a Kernel Density Estimation (KDE) model to infer the functionality intensity in a functional zone. Note that the real place that an individual visited may not be the destination that we can obtain from a mobility dataset. For example, the drop-off points of taxi trajectories may not be people's final destinations like a shopping mall. However, the pick-up/drop-off points should not be too far from the really-visited locations according to commonsense knowledge. The farther distance a location to the drop-off point, the lower probability that people would visit the location.

Given n points $x_1, x_2, \dots, x_n$ located in a 2D spatial space, we estimate the intensity at location s using a kernel density estimator, defined as:

$$\lambda(s) = \sum_{i=1}^{n} \frac{1}{nr^2} K(\frac{d_{i,s}}{b}), \tag{4}$$

where $d_{i,s}$ is the distance from x_i to s, b is the bandwidth and $K(\cdot)$ is the kernel function whose value decays with the increasing of $d_{i,b}$, such as the Gaussian function, Quartic function, Conic and negative exponential. The choice of the bandwidth usually determines the smoothness of the estimated density – a large b achieves smoother estimation while a small b reveals more detailed peaks and valleys. In our case, we choose the Gaussian function as the kernel function, i.e.,

$$K(\frac{d_{i,s}}{b}) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{d_{i,s}^2}{2b^2}), \tag{5}$$

and the bandwidth b is determined according to MISE criterion [18].

4.3 Region Annotation

In this step, given the results we have obtained, we try to annotate each cluster of regions with some semantic terms, which can contribute to the understanding of its real functions. Note that region annotation is a very challenging problem in both traditional urban planning and document processing. Essentially, the issue is the visualization of the topic model, which is listed as a future direction of topic modeling in the recent survey paper by Blei [13]. A compromised method used thus far is to utilize the most frequent words in a discovered topic to annotate a document. But in our case, listing the frequent mobility patterns (analogue to words) is far from enough to name a functional zone.

In our method, we annotate a functional zone by considering the following 4 aspects: 1) The POI configuration in a functional zone. We compute an average POI density vector across the regions in functional zone, where the *density* ρ_j of the jth POI category in region r_i is calculated by:

$$\rho_j = \frac{\text{Number of POIs of the } j\text{th POI category}}{\text{Area of region } r_i \text{ (measured by grid-cells)}}. \quad (6)$$

According to density value of each POI category in the calculated POI density vector, we rank POI categories in a functional zone (termed as internal ranking) and rank all functional zones for each POI category (referred to as the external ranking). We will give an example in the experiment as shown in Table 5. 2) The most frequent mobility patterns of each functional zone. 3) The functionality intensity. We study the representative POIs located in each functionality kernel, e.g., a function region could be an educational area if its kernel is full of universities and schools. 4) The human-labeled regions. People may know the functions of a few well-known regions, e.g., the region contains the Forbidden City is an area of historic interests. After clustering, the human labeled regions will help us understand other regions in a cluster. Refer to the experiments for the detailed results and analysis.

5 Experiments

5.1 Settings

5.1.1 Datasets

We use the following datasets for the evaluation:

1) Data representing location semantics.

- **Points of Interest (POI):** The Beijing POI dataset covers 328,669 POIs from the year 2011, where each POI is associated with the information of its latitude, longitude and the category (see Table 2 for a complete list of categories).
- **Road Networks:** The road network of Beijing is used to segment the urban area into regions, with statistics shown in Table 3.

2) Data representing mobility semantics.[3]

TABLE 3
Statistics of taxi & public transit trips and road networks

Taxi	#taxis	13,597
	#occupied trips	8,202,012
	#effective days	92
	average trip distance(km)	7.47
	average trip duration(min)	16.1
	average sampling interval(sec)	70.45
PTC	#trips	1,503,101
	#card IDs	295,720
Road	#road segments	162,246
	percentage of major roads	17.1%
	#segmented unit regions	554
	size of "vocabulary" (non-0 items)	3,244,901

- **Taxi Trajectories:** We used a GPS trajectory dataset generated by Beijing taxis in the year 2011, with the statistics shown in Table 3. We only chose occupied trips (identified by the information of a taxi meter) from the data, and accordingly segmented the trajectories to individual transitions (refer to Definition 1). It is worth noticing that there are over 30 cities in the world with over 10,000 taxicabs, and Beijing has over 67,000 taxis. The taxi trips represent a significant portion of people's urban mobility. According report by the Beijing Transportation Bureau, taxi trips occupy over 12 percent of traffic flows on road surfaces[19].
- **Public Transit Data (PTC):** This dataset logs the transactions of public transit including buses and subways in 2011. By pre-processing the transactions, we obtained a total of 1.5M trips (after removing the trips that have no information of origins and destinations), which is complementary to the taxi trips for representing urban mobility.

5.1.2 Platforms and baselines

We implement our method on a 64-bit server with a Quad-Core 2.67G CPU and 16GB RAM. We train our model with 10 topics for 1000 iterations, and optimize the parameters every 50 iterations. For k-means clustering, we incorporate the average silhouette value to determine the k and use the average results based on a 5-fold cross-validation. The efficiency (on average) is presented in Table 4.

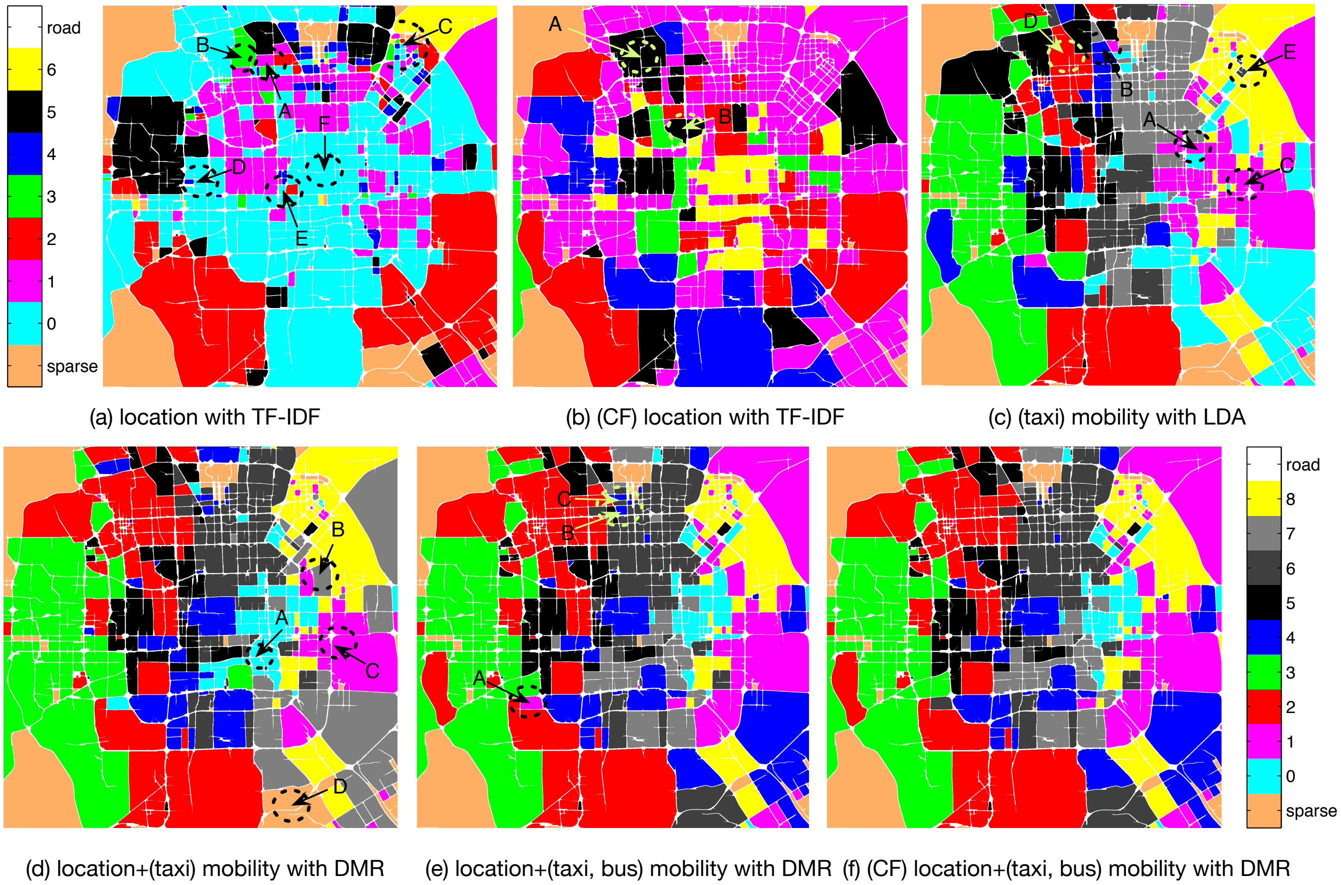
TABLE 4
Overall efficiency

operation	time(min)
map segmentation	0.325
building transition cuboids	41.3
learning location semantics using SVD	2.127
estimating topic model(1000 iterations)	1372
region aggregation	0.124
total	1394

We compare our method with several baselines:

- *TF-IDF-based Methods*, which include two approaches: 1) using POI distribution as feature vectors and 2) using collaborative POI feature vectors (introduced in Section 3.2).

[3] Although we used taxi trajectories and public transit data to evaluate our framework, we note that other mobility data such as mobile phone traces can also be directly incorporated into this framework.



(a) location with TF-IDF

(b) (CF) location with TF-IDF

(c) (taxi) mobility with LDA

(d) location+(taxi) mobility with DMR

(e) location+(taxi, bus) mobility with DMR

(f) (CF) location+(taxi, bus) mobility with DMR

Fig. 10. Functional zones discovered by different methods ((a),(b) share the left-top label legend, and (c)–(f) share the right-bottom label legend)

A k-means clustering is employed to cluster the regions into k functional zones based on their POI feature vectors.

- *LDA-based Topic Model*, which uses only the mobility semantics. Similar to our analogy from regions to documents, this method feeds the mobility patterns (the analogue to words) into an LDA model. Later, we perform a k-means clustering, similar to the method we used when grouping all regions based on their topic distributions learned from LDA. The parameters such as number of iterations, number of topics are set in accordance with the DMR-based method. As the number of POI categories usually has the same scale as the topics, applying the LDA model solely to POIs (as words) will not reduce the dimension of words.

We carried out the following studies to evaluate the effectiveness of our framework (though it is very difficult). 1) We invited 12 local people (who have been in Beijing for over 6 years) and asked them to label two representative regions for each kind of function (in total 24 labeled regions). We checked whether the regions having the same labels are assigned into the same functional zone and whether the regions with different labels are improperly clustered into one functional zone. 2) We matched our results against the land use planning of Beijing.

Compared to our experiments in [1], we further conducted the following experiments to examine the proposed DMR-based method:

- We compared the performance when using different TF-IDF-based feature vectors (i.e., non-collaborative TF-IDF feature vector / collaborative POI feature vector).
- We investigated whether the public transit data can further improve the performance of the proposed framework by comparing the results using solely the taxi data against using the combined data.

5.2 Results

5.2.1 Discovered Functional Zones

Figure 10 shows the aggregated functional zones discovered by different methods, with different colors indicating different functions. Note that in different figures, the same color may stand for different functions. As a result, TF-IDF-based methods forms 7 clusters (c_0–c_6) while LDA-based methods and DMR-based methods form 9 functional zones.

The TF-IDF-based methods considering only the location semantics perform the worst compared with other approaches. For example, as shown in Figure 10(a), region B is a university, which should be clustered with region A (another university) and region D (a high school). Meanwhile, region F (the Forbidden City) is not distinguished from other commercial areas like region E (Xidan). Another example is the Wangjing area (C), which is an emerging residential area with some companies and many living services, like apartments, shopping malls and restaurants. Unfortunately, the TF-IDF method improperly divides this area into many small functional zones as this method only considers POI distributions. The TF-IDF method using collaborative feature vectors tends to smoothen the distribution of POIs, thus merging a large portion of regions into one cluster. However, it still suffers from only considering location semantics, e.g., as shown in Figure 10(b), region A is the university campus, while region B is a developed commercial area, which should not be clustered together with A, although some universities lie in region B.

Basically, the LDA-based method and DMR-based method have a similar output of functional zones. However, there still exist several exemplary regions where using both mobility and location semantics (DMR-based) outperforms using only mobility semantics (LDA-based) obviously. For example, region F in Figure 10(c) is a developing commercial/entertainment area in the Wangjing area. But LDA aggregates it with the Forbidden City (Region E in Figure 10(a))), which is a region of historical interests; Area B (China Agricultural University) and Area D (Tsinghua University) are typical science and education areas where LDA fails to correctly cluster them together; Area A around Sanlitun is a well-known diplomatic district of Beijing, which is mixed with a developing commercial area C. The LDA-based method only using mobility semantics overlooks the location semantics implied by the POIs, thereby drops behind the DMR-based method (shown in Figure 10(d)).

Figure 10(e) presents the identified functional zones using DMR combing both the taxi trips and public transit data. Compared with solely using the taxi data (shown in Figure 10(d)), some previous sparse regions, such as region D in Figure 10(d) (where the taxi trajectories are not sufficient to train the model), can now be identified[4]. In addition, some mis-identified functional regions are further corrected by incorporating diverse types of human mobility data. For example, region A in Figure 10(d) is a residential area (with many ancient "hutongs" and old neighborhoods), which is wrongly clustered into the diplomatic area if we only use the taxi data. Region B in Figure 10(d) is the famous SOLANA Mall (an emerging commercial district in Beijing), which should be in the same cluster with the new CBD area (region C in Figure 10(d). Figure 10(f) shows the results of DMR with collaborative POI feature vectors as the metadata, which further improves the performance. As shown in Figure 10(e), region A, B and C are actually all residential areas, which failed to be identified until we fed the DMR with collaborative feature vectors. The reason behind this is probably that both the mobility data and POI configurations of these regions are biased due to the small sizes, but could be smoothed by collaborative filtering considering the semantic similarity with other regions.

TABLE 5
Overall POI density vector and ranking of functional zones. (FD: frequency density, IR: internal ranking)

	c0		c1		c2		c3		c4		c5		c6		c7		c8	
POI	FD	IR	FD	IR	FD	IR	FD	IR	FD	IR	FD	IR	FD	IR	FD	IR	FD	IR
CarServ	0.025	26	0.140	11	0.063	26	0.045	18	0.079	17	0.030	24	0.061	24	0.011	25	0.063	22
CarSale	0.016	27	0.018	25	0.067	25	0.012	25	0.010	26	0.005	28	0.019	27	0.005	28	0.008	27
CarRepa	0.014	28	0.063	17	0.073	23	0.049	17	0.064	18	0.023	26	0.062	23	0.010	26	0.041	25
MotServ	0.003	29	0.001	28	0.005	30	0.001	28	0.002	29	0.004	29	0.001	29	0.000	29	0.003	28
Caf/Tea	0.285	13	0.085	16	0.277	12	0.068	15	0.139	13	0.141	13	0.268	14	0.132	15	0.192	12
StaStor	0.116	18	0.052	18	0.141	19	0.038	20	0.057	20	0.068	18	0.139	18	0.128	16	0.088	18
LivServ	1.400	2	0.611	1	1.523	2	0.413	1	0.744	1	0.807	1	1.523	2	0.832	1	1.174	1
Sports	0.069	20	0.039	19	0.129	21	0.031	22	0.045	22	0.039	22	0.095	20	0.041	20	0.103	16
Hospital	0.198	14	0.096	14	0.268	13	0.082	11	0.166	11	0.140	14	0.299	12	0.163	13	0.223	10
Hotel	0.177	16	0.033	22	0.150	18	0.057	16	0.082	16	0.096	15	0.226	15	0.161	14	0.076	20
SceSpo	0.033	24	0.009	26	0.043	27	0.011	26	0.019	25	0.031	23	0.054	25	0.035	21	0.039	26
Residen	0.880	3	0.217	6	0.693	6	0.185	6	0.383	5	0.482	5	0.900	4	0.421	5	0.568	3
Gov/Pub	0.414	9	0.104	13	0.363	10	0.079	12	0.166	10	0.173	12	0.470	7	0.289	7	0.309	7
Sci/Edu	0.445	8	0.255	5	1.325	3	0.112	8	0.212	8	0.305	6	0.579	6	0.232	11	0.296	8
TrasFac	0.453	6	0.141	10	0.514	7	0.079	13	0.180	9	0.213	10	0.416	8	0.255	9	0.326	6
Bank/Fina	0.445	7	0.092	15	0.497	8	0.069	14	0.133	14	0.226	9	0.357	11	0.187	12	0.210	11
CopBusi	1.984	1	0.602	2	2.501	1	0.257	3	0.501	3	0.657	2	1.701	1	0.605	2	1.063	2
StrFur	0.000	30	0.001	30	0.005	29	0.001	30	0.001	30	0.001	30	0.000	30	0.000	30	0.000	30
Entr/Bri	0.365	12	0.128	12	0.240	14	0.090	10	0.155	12	0.182	11	0.268	13	0.270	8	0.138	15
PubUti	0.396	10	0.141	9	0.339	11	0.111	9	0.283	6	0.234	8	0.378	10	0.293	6	0.165	13
ChiRes	0.650	5	0.281	3	1.102	4	0.291	2	0.537	2	0.567	3	0.880	5	0.444	4	0.421	4
ForRes	0.168	17	0.027	23	0.106	22	0.007	27	0.009	27	0.042	21	0.068	22	0.034	22	0.161	14
FasRes	0.104	19	0.033	21	0.213	15	0.035	21	0.058	19	0.083	17	0.139	17	0.071	18	0.103	17
ShopMal	0.720	4	0.270	4	1.020	5	0.240	4	0.433	4	0.537	4	0.943	3	0.580	3	0.365	5
ConvStor	0.381	11	0.158	7	0.365	9	0.136	7	0.277	7	0.257	7	0.400	9	0.236	10	0.234	9
E-Stor	0.038	23	0.020	24	0.172	17	0.027	23	0.044	23	0.060	20	0.068	21	0.029	23	0.042	24
SupMar	0.051	22	0.009	27	0.068	24	0.021	24	0.041	24	0.030	25	0.049	26	0.020	24	0.051	23
FurBuil	0.066	21	0.143	8	0.184	16	0.208	5	0.099	15	0.090	16	0.136	19	0.074	17	0.076	21
Pub/Bar	0.196	15	0.037	20	0.138	20	0.040	19	0.051	21	0.066	19	0.147	16	0.069	19	0.085	19
Theater	0.029	25	0.001	29	0.007	28	0.001	29	0.006	28	0.010	27	0.008	28	0.006	27	0.002	29

Overall, the method combing location semantics and mobility semantics (including both taxi trajectories and public transit data) outperforms other approaches in terms of the accordance with the labeled functional regions.

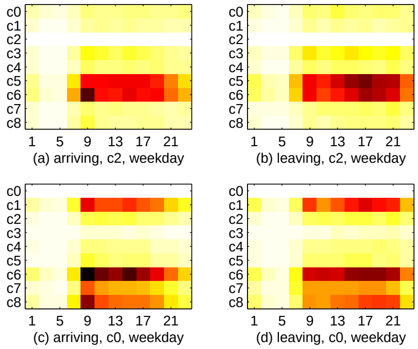
5.2.2 Annotation of Functional Zones

Table 5 shows the average POI density vector of each region cluster (c_0–c_8, remember that DMR-based method generated 9 clusters) and the corresponding internal and external rankings, where the external rank is represented by the depth of the color (1 darkest, 4 lightest). Clearly, clusters (functional zones) c_0, c_2, c_5, c_6, c_7, and c_8 are more mature and more developed areas as compared to other clusters, since they have more high ranked POI categories, which are annotated as follows:

Diplomatic/Embassy Areas[c_0]. The most characteristic POI categories in this functional zone are the international restaurants, pubs/bars, theaters, cafés and tea bars, with a significantly higher frequency density than other functional zones. Most embassies are located in these areas, which are well configured for the diplomatic function, e.g., they have the second highest external rank of residential buildings, hospitals, bank and insurance services.

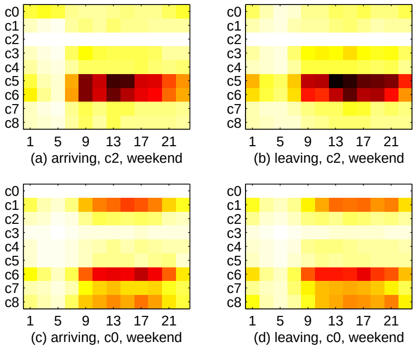
Science/Education/Technology Areas[c_2]. This functional zone contains the maximum number of science and education POIs (e.g., Tsinghua university and Beijing university), banks and corporate business POIs. In addition, the biggest electronic market in China, called "ZhongguanCun", known

[4] Note that if we solely use the public transit data as mobility semantics, many regions will be identified as "sparse" regions due to the insufficiency (only 1/5 of the taxi trips on average for each region) for learning the model.



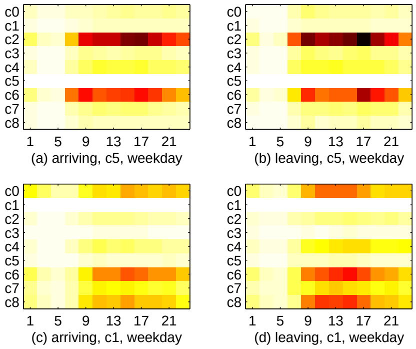
(a) arriving, c2, weekday (b) leaving, c2, weekday (c) arriving, c0, weekday (d) leaving, c0, weekday

Fig. 11. Weekday transitions of c_0, c_2



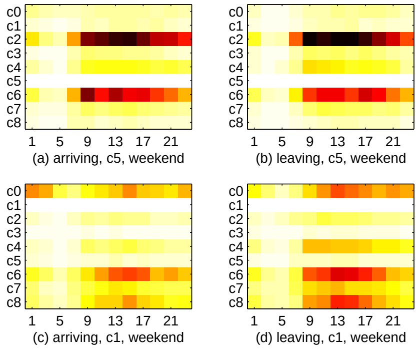
(a) arriving, c2, weekend (b) leaving, c2, weekend (c) arriving, c0, weekend (d) leaving, c0, weekend

Fig. 12. Weekend transitions of c_0, c_2



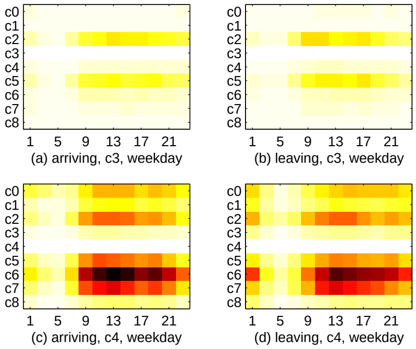
(a) arriving, c5, weekday (b) leaving, c5, weekday (c) arriving, c1, weekday (d) leaving, c1, weekday

Fig. 13. Weekday transitions of c_1, c_5



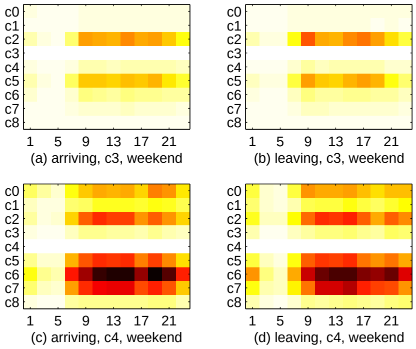
(a) arriving, c5, weekend (b) leaving, c5, weekend (c) arriving, c1, weekend (d) leaving, c1, weekend

Fig. 14. Weekend transitions of c_1, c_5



(a) arriving, c3, weekday (b) leaving, c3, weekday (c) arriving, c4, weekday (d) leaving, c4, weekday

Fig. 15. Weekday transitions of c_3, c_4



(a) arriving, c3, weekend (b) leaving, c3, weekend (c) arriving, c4, weekend (d) leaving, c4, weekend

Fig. 16. Weekend transitions of c_3, c_4

as the Silicon Valley in China, is located in this functional zone.

Developed Residential Areas[c_6]. This functional zone is clearly a mature residential area with the most residential buildings, hospitals, hotels, and convenience stores. Within this functional zone, an adequate number of services supports the people's living, such as the restaurants, shopping malls, banking services, schools, and sports centers.

Emerging Residential Areas[c_8]. This area is annotated as the emerging residential area since it has a balanced POI configuration (similar to, but less than c_6), such as living services, residential buildings, sports centers, hospitals and some companies.

Old Neighborhoods[c_7]. The areas within this functional zone are mostly residential building built before the year 1995 or even more anciently, where the old streets are known as "hutongs". The POI configuration shows that this type of zone is less developed than both c_6 and c_8.

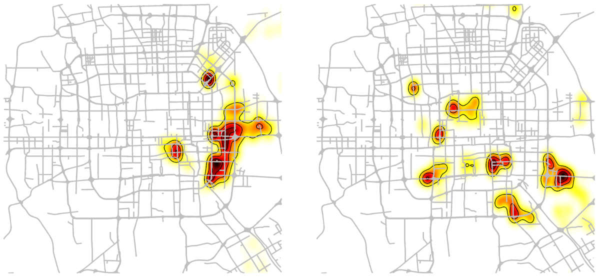
Developed Commercial/Entertainment areas[c_5]. This is a typical entertainment and commercial zone containing several mature business circles in Beijing, such as the Xidan business circle[5], Financial Street[6], and Gongzhufen business circle[7].

Emerging Commercia/Entertainment Areas [c_1]. The POI configuration (the internal rank) of this cluster is similar to cluster c_5, but in terms of the absolute quantity, c_1 is less than c_5. A certain number of shopping malls, restaurants and banking services feature this cluster as a developing commercial/ business/ entertainment functional zone (either of them is possible). In the meantime, the functionality intensity provides another corroboration for this annotation. As depicted in Figure 17(a), the core of this functional zone is the new CBD of Beijing.

Figure 11 and Figure 12 show the arriving/leaving transitions matrix of c_0 and c_2 during weekdays and weekends respectively, where the x-axes are time of day (by hour) and y-axes are the functional zones that people come from and leave for. Both c_0 and c_2 can generally be considered the working areas, since trends reveal for both of them that people come at the morning peak time (8-9am) and leave in the early evening (5-6pm). The results also indicate that c_6, c_7, c_8 are residential areas since most people come to c_0 and c_2 in the morning are originated from these zones (Figure 11 (a) and (c)).

Figure 13 and Figure 14 show transition matrices of c_5 and c_1 on weekdays and weekends. It's clear that on weekdays, most people reach and leave these areas after work (5pm-6pm), while during weekends, people (mostly from the residential areas such as c_6 and c_8) come to and leave for these zones throughout the day, which is a typical pattern of

[5] http://en.wikipedia.org/wiki/Xidan
[6] http://en.wikipedia.org/wiki/Beijing_Financial_Street
[7] http://en.wikipedia.org/wiki/Gongzhufen



(a) functional zone c_1 (b) functional zone c_4

Fig. 17. Functionality intensity of the emerging and developed commercial functional zones

the commercial area. Another signal showing the commercial function of c_5 is that people go to c_5 more often on weekends (as shown in Figure 12 (b) and Figure 14).

With regard to the other identified functional zones, since the frequency densities of POIs are much lower than the above functional zones, we identify their semantic functions with more consideration of functionality intensity and frequent mobility patterns derived for each functional zone in addition to the POI configurations.

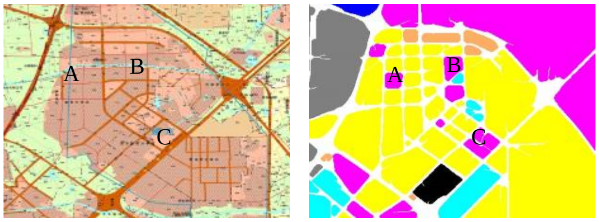
Historical Interests/Parks[c_4]. If we only consider the POI configuration, the characteristic of this cluster does not reveal obviously. However, by considering the functionality intensity estimated by mobility patterns, we find that they are places of historic interests in Beijing. As shown in Figure 17(b), famous historical sites like the Forbidden City and the Temple of Heaven are located in these areas. In addition, some parks like the Purple Bamboo Park[8], Happy Valley[9], and Wangxinghu Park[10] are also successfully clustered into this functional zone.

Nature areas[c_3]. These areas have the fewest POIs in most POI categories. Actually, a lot of forests and mountains cover this cluster, e.g., the Xishan Forest Park, Century Forest Park, and Baiwang Moutain.

Figure 15 and Figure 16 show that people come to c_3 following similar temporal patterns as c_4, but the diversity and quantity are reasonably weaker than c_4, since many POIs in c_4 are very famous scenic spots. In addition, people travel to c_3 and c_4 more often on weekends, which coincides with expected tendencies at that time.

5.2.3 Calibration for Urban Planning

The discovered functional zones provide calibration and reference for urban planning. For example, Figure 18 presents the comparison between governmental land use planning (2002-2010) and the results of our method in 2011. This area forms an emerging residential area as planned by the government, while some small regions become developing commercial areas, such as A, B and C after 2 years' development.



(a) (b)

Fig. 18. (a) governmental land use planning (2002-2010) (b) discovered functional zones in 2011

6 Related Work

6.1 Urban computing with taxicabs

In recent decades, urban computing has emerged as a concept where every sensor, device, person, vehicle, building, and street in urban areas can be used as a component to probe city dynamics and further enable a city-wide computing for serving people and their cities. The increasing availability of GPS-embedded taxicabs provides us with an unprecedented wealth to understand human mobility in a city, thereby enabling a variety of novel urban computing research recently. For example, [20] and [19] studied the strategies for improving taxi drivers' income by analyzing the pick-up and drop-off behavior of taxicabs in different locations. [21] aimed to find the fastest practical driving route to a destination according to a large number of taxi trajectories.

The work presented in this paper is also a step towards urban computing, but unlike the above-mentioned research, , we focus on the discovery of functional zones in a city, which we have never seen before in this research theme.

6.2 Map Segmentation

Grid-based map segmentation is extensively used in geospatial-related analysis. [23] provided a method for predicting drivers' destination by mapping past trips into grid-cells and learning the destination probabilities with respect to each cell. [24] proposed an approach to suggest maximum profit grid for taxi drivers by constructing a *spatio-temporal profitability map* with a grid-based segmentation, where the probabilities are calculated using historical data. Compared with a grid-based segmentation, our solution, which considers high-level roads as the boundary, is more natural for studying the human mobility on a map, as described in Section 1.

[25] proposed a novel road network partition approach based on road hierarchy. Specifically, road networks are first divided into areas by high level roads, then the partition process is recursively performed for each area. The partition process is implemented by finding the strongly connected components after the removal of the intersection nodes connected to high level roads as well as the terminals of high level road segments themselves. Figure 19(b) presents the results of this approach for a portion of the Beijing road network. However, this method does not work in our scenario since 1) our desired region is bound by high level

[8] http://bit.ly/RwzcWs
[9] http://bit.ly/14YDRUG
[10] http://bit.ly/14YDgCi

road segments and may contain several strongly connected components, and 2) we aim to segment the whole area instead of just the road nodes into regions, i.e., we need a mapping from any locations represented by latitudes and longitudes (within the bounding box of the road network) to the region IDs.

Morphology operators are widely used in geographical information systems as well as image processing. [26] employed mathematical morphology for map simplification from remote sensing images by extracting skeletons from the image and converting the structure into vectors. Similar works are presented in [27], [28] and [29]. Different from the above methods which are based on remote sensing images, we aim to segment the urban area represented by vector-based model into regions, instead of simplifying the map or extracting structures from the map. Figure 19(c) plots the result using the proposed morphological-based algorithm, as compared to the grid-based method and hierarchy-based method with respect to the same area of Beijing.

6.3 Discovery of functional zones

Functional zones [30] have been studied in traditional fields of GIS and urban planning for years, as their discovery can benefit policy making, resource allocation, and related research. As early as 1970, [31] provided a case study on functional regions within Central London using surveyed data of taxi flows collected in 1962, which is part of the London Traffic Survey. [14] gave a good survey on related works which are mainly based on clustering algorithms. Some algorithms classify regions in urban area based on remote-sensing data, as thoroughly compared in [32]. Other network-based clustering algorithms (e.g., spectral clustering), however, employed interaction data, such as economic transactions, communication records [33] and people's movement between regions. In [33], the authors exploited telecommunication data to partition the Great Britain into regions. They first rasterized the map into pixels, then built a transition matrix using telecommunication data with respect to each partitioned pixel, which can be regarded as the adjacent matrix of a graph. The technique used for partitioning the map is the spectral method based on modularity, which is usually utilized in community detection. As a result, the detected boundaries coincided well with either the official administrative boundaries of these regions, or results from existing literatures.

As the capital of China, Beijing has experienced profound changes especially during the past two decades. [34] presented a historical review of urban planning in Beijing, with a focus on the period during 1979-1995, where they indicated that the "new urban-planning ideas, complex landuse and transportation patterns" are blended by the evolving form. The work reported in this paper, however, focuses on contemporary Beijing and potentially enables calibration for urban planners in the near future.

Recently, a series of work has aimed to study the geographic distribution of some topic in terms of user-generated social media [35, 36]. For example, [37] studied the distributions of some geographical topics (like the beach, hiking, and sunsets) in the USA using geo-tagged photos acquired from Flickr. [38] explored the space-time structure of topical content from numerous geo-tweets. The social media generated in a geo-region is still used as static features to feature a region. Meanwhile, a few works have reported that human mobility can describe the functions of regions. For instance, [39] observed that the getting on/off amount of taxi passengers in a region can depict the social activity dynamics in the region.

Our work is different from the research mentioned above in the following aspects. First, to the best of our knowledge, our method is the first one that simultaneously considers location semantics (e.g., POIs) of a region and mobility semantics (i.e., human mobility intentions) between regions when identifying functional zones. Second, rather than directly using some clustering algorithm, we propose a topic-model-based solution which represents a region with a distribution of socio-economic activities. Moreover, it reduces data sparsity by clustering regions into functional zones. We demonstrate the advantage of our method over just using the clustering approach in our experiments.

7 Conclusion

This paper has proposed a framework for discovering functional zones (e.g., educational areas, entertainment areas, and regions of historic interests) in a city using human trajectories, which imply socio-economic activities performed by citizens at different times and in various places. We have evaluated this framework with large-scale datasets including POIs, road networks, taxi trajectories and public transit data. According to extensive experimental results, our method using both location and mobility semantics outperforms the baselines solely using location or mobility semantics in terms of effectively finding functional zones. Meanwhile, we have found that public transit data can be used as a complement to the taxi trips in representing urban mobility, so as to achieve a better performance for discovering functional zones. In addition, by matching the discovered functional zones against Beijing land use planning (2002-2010), we have shown exemplary calibrated results. The proposed framework provides a powerful tool for computational urban science, and offers emerging implications for human mobility analytics and location-based services.

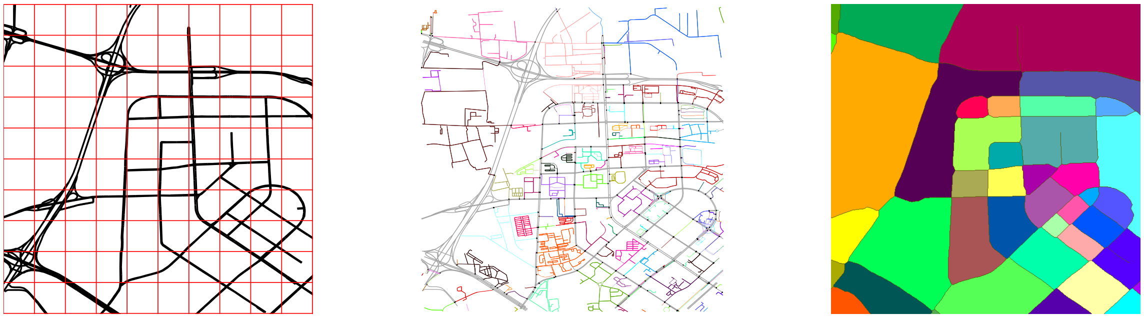
References

[1] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and pois," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '12. ACM, 2012, pp. 186–194.

[2] S. M. Willhelm, *Urban zoning and land-use theory*. Free Press of Glencoe, 1962.

[3] http://www.watchdata.com/transportation/10150.html.

[4] R. Estkowski, "No Steiner point subdivision simplification is NP-Complete," in *Proc. 10th Canadian Conf. Computational Geometry*. Citeseer, 1998.



(a) Grid-based (b) Hierarchy-based (c) Morphology-based

Fig. 19. Different map segmentation methods

[5] J. Snyder, *Map projections–A working manual*. USGPO, 1987, no. 1395.

[6] http://msdn.microsoft.com/en-us/library/aa940991.aspx.

[7] Z. Guo and R. Hall, "Parallel thinning with two-subiteration algorithms," *Communications of the ACM*, vol. 32, no. 3, pp. 359–373, 1989.

[8] A. Rosenfeld and J. Pfaltz, "Sequential operations in digital picture processing," *Journal of the ACM (JACM)*, vol. 13, no. 4, pp. 471–494, 1966.

[9] L. Lam, S. Lee, and C. Suen, "Thinning methodologies-a comprehensive survey," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 14, no. 9, pp. 869–885, 1992.

[10] L. Shapiro and G. Stockman, *Computer Vision*. Prentice Hall, 2001.

[11] D. Billsus and M. J. Pazzani, "Learning collaborative information filters." in *ICML*, vol. 98, 1998, pp. 46–54.

[12] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Advances in artificial intelligence*, vol. 2009, p. 4, 2009.

[13] D. Blei, "Introduction to probabilistic topic models," *Communications of the ACM*, 2011.

[14] C. Karlsson, "Clusters, functional regions and cluster policies," *JIBS and CESIS Electronic Working Paper Series (84)*, 2007.

[15] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[16] D. Mimno and A. McCallum, "Topic models conditioned on arbitrary features with dirichlet-multinomial regression," in *Uncertainty in Artificial Intelligence*, 2008, pp. 411–418.

[17] P. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[18] M. Wand and M. Jones, *Kernel smoothing*. Chapman & Hall/CRC, 1995, vol. 60.

[19] J. Yuan, Y. Zheng, L. Zhang, X. Xie, and G. Sun, "Where to find my next passenger," in *Proc. Ubicomp '11*, 2011, pp. 109–118.

[20] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, and M. Pazzani, "An energy-efficient mobile recommender system," in *Proc. KDD '10*, 2010, pp. 899–908.

[21] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *Proc. KDD '11*, 2011, pp. 316–324.

[22] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xing, "Discovering spatio-temporal causal interactions in traffic data streams," in *Proc. KDD '11*, 2011, pp. 1010–1018.

[23] J. Krumm and E. Horvitz, "Predestination: Where do you want to go today?" *Computer*, vol. 40, no. 4, pp. 105–107, 2007.

[24] J. Powell, Y. Huang, F. Bastani, and M. Ji, "Towards reducing taxicab cruising time using spatio-temporal profitability maps," in *Proceedings of the 12th International Symposium on Advances in Spatial and Temporal Databases*, ser. SSTD '11, 2011.

[25] H. Gonzalez, J. Han, X. Li, M. Myslinska, and J. P. Sondag, "Adaptive fastest path computation on a road network: a traffic mining approach," in *Proceedings of the 33rd international conference on Very large data bases*, ser. VLDB '07, 2007, pp. 794–805.

[26] M. Saradjian and J. Amini, "Image map simplification using mathematical morphology," *International Archives of Photogrammetry and Remote Sensing*, vol. 33, pp. 36–43, 2000.

[27] E. López-Ornelas, "High resolution images: segmenting, extracting information and gis integration," *World Academy of Science, Engineering and Technology*, vol. 54, pp. 172–177, 2009.

[28] C. Martel, G. Flouzat, A. Souriau, and F. Safa, "A morphological method of geometric analysis of images: Application to the gravity anomalies in the indian ocean," *Journal of Geophysical Research*, vol. 94, no. B2, pp. 1715–1726, 1989.

[29] M. ANSOULT, P. SOILLE, and J. LOODTS, "Mathematical morphology- a tool for automated gis data acquisition from scanned thematic maps," *Photogrammetric engineering and remote sensing*, vol. 56, no. 9, pp. 1263–1271, 1990.

[30] J. Antikainen, "The concept of functional urban area," *Findings of the Espon project*, vol. 1, no. 1, 2005.

[31] J. B. Goddard, "Functional regions within the city centre: a study by factor analysis of taxi flows in central london," *Transactions of the Institute of British Geographers*, pp. 161–182, 1970.

[32] 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 COM.Geo '11*, 2011, pp. 11:1–11:10.

[33] C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, and S. H. Strogatz, "Redrawing the map of great britain from a network of human interactions," *PLoS One*, vol. 5, no. 12, p. e14248, 2010.

[34] P. Gaubatz, "Changing beijing," *Geographical Review*, pp. 79–96, 1995.

[35] J. Cranshaw, R. Schwartz, J. I. Hong, and N. M. Sadeh, "The livehoods project: Utilizing social media to understand the dynamics of a city." in *ICWSM*, 2012.

[36] F. Kling and A. Pozdnoukhov, "When a city tells a story: urban topic analysis," in *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*. ACM, 2012, pp. 482–485.

[37] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang, "Geographical topic discovery and comparison," in *Proc. WWW '11*, 2011, pp. 247–256.

[38] A. Pozdnoukhov and C. Kaiser, "Space-time dynamics of topics in streaming text," in *Proc. LBSN '11*, 2011, pp. 8:1–8:8.

[39] 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 *IEEE PERCOM Workshops*, 2011, pp. 384–388.

Nicholas Jing Yuan is an associate researcher in Microsoft Research Asia. Currently, his research interests include spatial-temporal data mining, behavioral mining and computational social science. He is a member of ACM and IEEE.

Yu Zheng is a lead researcher from Microsoft Research Asia. His research interests include location-based services, spatio-temporal data mining, ubiquitous computing, and mobile social applications. He is a senior member of IEEE and ACM.

Xing Xie is a senior researcher in Microsoft Research Asia, and a guest Ph.D. advisor for the University of Science and Technology of China.His research interest include spatial data mining, location based services, social networks and ubiquitous computing. He is a senior member of ACM and the IEEE.

Yingzi Wang is currently an graduate student in University of Science and Technology of China. Her recent research interests include spatial-temporal data mining and human mobility analytics.

Kai Zheng is currently a Research Fellow at the University of Queensland. His research interests include efficient spatio-temporal query processing, uncertain data management, and spatial trajectory computing.

Hui Xiong is currently an Associate Professor and the Vice Department Chair of the Management Science and Information Systems department at Rutgers University. His general area of research is data and knowledge engineering, with a focus on developing effective and efficient data analysis techniques for emerging data intensive applications. He is a senior member of the ACM and IEEE.