Journal of Computers Vol. 30 No. 6, 2019, pp. 181-194
doi:10.3966/199115992019123006014

Delineating Urban Functional Areas with Sina Weibo Check-in Data: A Matching Time Series Distance Based LST-SVM Multi-classifier Method

Tianyu Xia[1,2*], Kai Yang[1,2], Wanggen Wan[1,2], Xiang Feng[1,2], Xiaoqing Yu[1,2]

[1] School of Communication and Information Engineering, Shanghai University, Shanghai, China

[2] Institute of Smart City, Shanghai University, Shanghai, China
xiatianyu1203@gmail.com, yangkaijyjy@163.com, wanwg@staff.shu.edu.cn, fengxiang0727@126.com

Received 12 April 2018; Revised 23 August 2018; Accepted 5 October 2018

Abstract. Delineating the distribution of different urban functional areas is a hot topic in urban studies. The accuracy of describing urban functional areas is particularly important. This paper aims to improve the estimated accuracy of delineating urban functional areas according to the assumption that social media activities in these buildings with similar functionality not only have similar spatiotemporal patterns but also are strictly correlated to the temporal information. We propose a novel Matching Time Series (MTS) method to calculate the distance of the time-series data and use this method to modify the least squares twin support vector machine (LST-SVM) Multi-classifier Method for classifying the building objects with similar functionality as a functional area. According to the time series dataset built based on Sina Weibo check-in data, we compare with the dynamic time warping (DTW) distance based *k*-medoids method from different aspects. The results show that the accuracy is improved from 29.96% to 82.68%, which verifies the superiority of our proposed approach in improving the estimated accuracy of delineating urban functional areas. By separating time series dataset into weekdays and weekends, we also obtain relatively high classification accuracy respectively, and it contributes to analyze the distribution of urban functional buildings more clearly.

Keywords: Delineating urban functional areas, Sina Weibo check-in data, matching time series, Least squares twin support vector machine

1 Introduction

With the fast development of the urban, delineating the distribution of different urban functional areas is an important theme in urban studies and planning. For many years, the description of urban functional areas relies mainly on socio-demographic data [1]. However, it is laborious and time-consuming to process these data and the results cannot reflect the dynamic spatiotemporal characteristics of urban functional areas [2].

With the advent of the era of big data, a large number of massages involve people's activity data and movement data, which usually contain geographic information and are conveniently recorded by devices such as smartphones, GPS navigators, smart cards, and location-based APPs [3]. These two types of data have been widely used to delineate urban functional areas. Movement data is adopted to explore the association between the spatio-temporal patterns of human movements and the functional regions [4-7]. And activity data is suitable for finding the connection between the spatio-temporal activity pattern and the urban functional area [7-10]. Yuan et al. [11] segment a city into disjointed regions according to major roads and developed a topic-based inference model to delineate urban functional areas in Beijing based on movement data and point of interests (POIs) data. This method regards a region as a document, a function as a topic, categories of POIs as metadata, and human mobility patterns as words. Yuan et al.

[*] Corresponding Author

[12] captured the latent activity trajectory (LAT) data from GPS trajectory dataset generated by Beijing taxis and proposed a data-driven framework combining POIs data to discover function zones in a city. However, the data used in these two methods needs long-term observation, high time and financial costs. And the accuracy of these two methods for describing urban functional areas will be affected by the POIs data because it is difficult to precisely match human movements with POIs [3].

In recent years, with the widely applying of social media, such as Facebook, Twitter, Tencent QQ, and Sina Weibo, social media data has been widely utilized in travel behaviors [13], urban communities [14], and human mobility [15-16]. And it is also useful for analyzing and delineating urban functional areas. Furthermore, more and more social media registered users like recording and sharing their location and travel demands via check-in [17]. Temporal and spatial referenced check-in data has became as a new data source for studying urban socioeconomic dynamics. For example, Zhi et al. [3] introduced a novel low-rank approximation (LRA)-based model to detect the functional regions with the data from about 15 million social media check-in records during a year-long period in Shanghai. Zhen et al. [18] developed a new approach that uses Sina Weibo check-in data to analyze activity intensity, closeness, and connection to comprehensively delineate the boundaries of the Yangtze River Delta urban agglomeration (YRDUA) in China. Chen et al. [19] assumed that social media activities in buildings of similar functions have similar spatiotemporal patterns, and applied a dynamic time warping (DTW) distance based *k*-medoids method to group buildings with similar social media activities into functional areas. These studies use the clustering method to describe the urban functional area. The DTW distance based *k*-medoids method has emerged as a popular method for time-series data clustering [20] and could obtain relatively high accuracy of describing urban functional areas on some dataset, but it is limited by the defects of the DTW algorithm. Meanwhile, comparing to movement data and activity data, using social media data for describing urban functional area can greatly reduce the time and financial costs. However, looking for a robust method to improve the accuracy of describing urban functional areas is still the hot topic of current research.

In this paper, we study how to describe urban functional areas and improve the accuracy of describing urban functional areas. We propose a novel method for delineating functional urban areas based on the Sina Weibo check-in data. Comparing with existing studies, our method has the following advantages.

First, we build the time-series dataset which reveals the spatiotemporal distributions of Sina Weibo users in hourly intervals based on the Sina Weibo check-in data. Then we propose a novel matching time-series (MTS) method to calculate the distance of the time-series data. Our method assumes that social media activities in buildings of similar functionalities not only have alike spatiotemporal patterns but also are strictly correlated to the temporal information. Comparing with the classic dynamic time warping (DTW) algorithm to calculate the distance of the two time-series data, the MTS method can calculate the distance more accurately and more reasonably to capture the inherent heterogeneity even within the same urban functionality types.

Second, our analysis advances the use of spatiotemporal classifying techniques in urban and geographical studies. Despite the extensive use of clustering analysis in existing studies [19-22], few studies have explored classifying methods for time-series geographical data. Classifying analysis is different from clustering analysis that it is the supervised learning process of the training sample and the datasets include the labeled and unlabeled samples. Therefore, we can use classifying analysis to build prediction model by training labeled sample data, then the model is able to predict the unlabeled data belongs to which category and we can obtain relatively high accuracy of classification results. In the field of pattern recognition, a variety of methods for classifying has been developed. Among these methods, the least squares twin support vector machine (LST-SVM) Multi-classifier has emerged as a popular method for classification [23-24]. The advance of the LST-SVM Multi-classifier method is that it uses the kernel function to separate the data in infinite dimensional space, and during the feature extraction processing the input data vector could be an infinite dimensional vector in theory. Meanwhile, it also can speed up calculation and reduce computational complexity of classic classification method SVM [28].

Third, we use the proposed MTS method to modify the LST-SVM Multi-classifier method for time-series data classifying, and consider the building objects with similar functionality as a functional area based on the classification result. Through comparing with classic clustering methods, the MTS distance based LST-SVM Multi-classifier method in our study is more efficient and clearly to describe urban functional areas. By separating time series dataset into weekdays and weekends, we also use this method to describe urban functional areas and obtain relatively high classification accuracy respectively. This method uses the idea of classification to describe the urban functional area, which not only has a lower computational cost but also can effectively help us analyze the distribution of urban functional buildings.

Our proposed method is illustrated through a case study of Shanghai and the remainder of this paper is organized as follows. In Section 2, we propose the MTS method and the MTS distance based LST-SVM Multi-classifier method. In Section 3, we introduce the data source and analyze the results of the experiments. Finally, in section 4, we conclude our work and recommend further research.

2 Methodology

This paper aims to improve the estimated accuracy of delineating urban functional areas. According to the assumption that social media activities in these buildings with similar functionality have a similar spatiotemporal pattern and are strictly correlated to the temporal, we propose a novel Matching Time Series (MTS) method to calculate the distance of the time-series data and use this method to modify the LST-SVM Multi-classifier Method for classifying the building objects with similar functionality as a functional area.

2.1 Matching Time Series Method

In the MTS distance based LST-SVM Multi-classifier method, the MTS method is proposed to calculate the distance between the two time-series data. The well-known method DTW distance describes the length of the optimal alignment (i.e., the warping path) between two given time-series data [25], but in order to find the similar time-series data, it will stretch the time-series data. For example, the two time-series data *A* = [2 2 2 2 2 2 2 3 4 3 2 2] and *B* = [2 3 4 3 2 2 2 2 2 2 2 2], the DTW distance of the *A* and *B* is 0. Although the two-time series are very similar in the space, if we consider the temporal distribution in a day and the functional information of a building, that would be no reason. For instance, for the same time-series data *A* and *B*, as shown in Fig. 1, if we assume that the *A* and *B* represent the crowd flows frequency at the different buildings during one day respectively, we will find that the more people go to the building *A* at afternoon, and the more people go to the building *B* in the morning, suggesting that these two buildings have different functionality during one day. If we still use the DTW distance, it will break the temporal rule and consider the two buildings have similar functionality. Obviously, the DTW distance of time-series is not suitable for describing the similarity of urban functionality buildings.

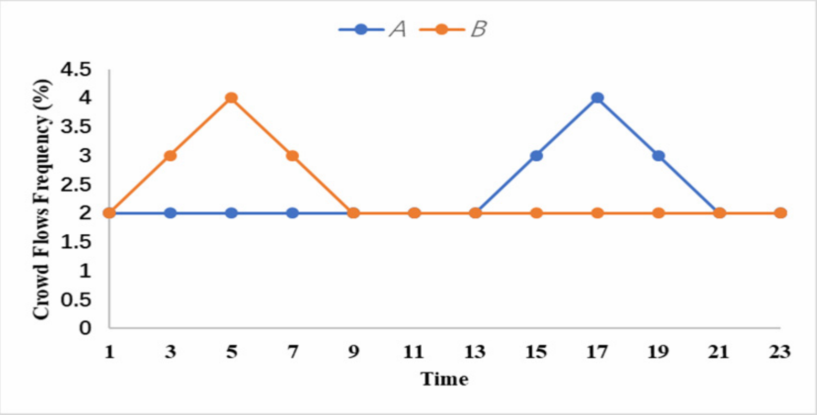


Fig. 1. The two time-series data *A* and *B*

In order to solve the above problem, we propose the MTS method to calculate the distance of two time-series data $x_A = \left(x_{A_1}, x_{A_2}, \cdots, x_{A_N}\right)$ and $x_B = \left(x_{B_1}, x_{B_2}, \cdots, x_{B_N}\right)$. The algorithm as following steps:

(1) Normalizing the time-series data. In order to make sure the processed data conforms to the standard normal distribution, we use the zero-mean normalization method to normalize each data $x_i^* = \left(x_i - \mu\right)/\delta$, where $\mu = \sum_{i=1}^{N} x_i / N$, $\delta^2 = \sum_{i=1}^{N} \left(x_i - \mu\right)^2 / N$, which suggests that the mean is 0 and the standard deviation is 1. Then getting the normalized data $x_A^* = \left(x_{A_1}^*, x_{A_2}^*, \cdots, x_{A_N}^*\right)$ and $x_B^* = \left(x_{B_1}^*, x_{B_2}^*, \cdots, x_{B_N}^*\right)$.

(2) Calculating the sum of squared differences to get the Euclidean distance *S* between normalized time-series *A* and *B*, $S=\sqrt{\sum_{i=1}^{N}\left(x_{A_i}^{*}-x_{B_i}^{*}\right)^{2}}$.

(3) Using the sum of the squared differences of two normalized time-series slope to describe the similarity degree of the fluctuation. Using *k* for the slope between two points $\left(x_1,y_1\right)$ and $\left(x_2,y_2\right)$, and the time-series dataset reveals the spatiotemporal distributions of Sina Weibo users in hourly intervals, thus $x_2-x_1=1$, $k=y_2-y_1$, for the normalized time-series data x_A^{*} , $k_{A_i}=x_{A_{i+1}}^{*}-x_{A_i}^{*}$, $i\in\left(1,N-1\right)$, the similarity degree of the two time-series fluctuation $F=\sqrt{\sum_{i=1}^{N-1}\left(k_{A_i}-k_{B_i}\right)^{2}}$.

(4) According to the positive and negative of the slope to determine the similarity of time series in shape, counting the number *C* of the same symbol of slope in the k_A and k_B .

(5) Acquiring the distance between two time-series data *D*:

$$D=\frac{S+F+p*\left(23-C\right)}{3}. \tag{1}$$

Because of the *C* in the range from 0 to 23 and in order to match the range of *S* and *F*, a coefficient of *p* should be multiplied to make sure they have the same influence. The value of *p* is determined by the maximum and minimum values of *S* and *F*:

$$p=\frac{\max\left(S_{\max},F_{\max}\right)-\min\left(S_{\min},F_{\min}\right)}{23}. \tag{2}$$

The MTS method from the Euclidean distance, fluctuation, and shape to calculate the distance between the two time-series data, and it will be used to determine the similarity of the two time-series data. The smaller the distance value, the greater the similarity or vice versa.

2.2 LST-SVM Multi-classifier Method

The well-known support vector machine (SVM) method has advantages over other existing data classification approaches. In order to speed up the calculation and reduce the computational complexity of SVM, the least squares twin support vector machine (LST-SVM) method was proposed, which is different from SVM seeking one non-parallel hyper-plane for the binary classification problem and will seek two non-parallel hyper-planes $x^T w_1+b_1=0$ and $x^T w_2+b_2=0$ by solving two linear equations [26-28].

The multi-class approaches based on LST-SVM are four main categories, including "One-versus-All LST-SVM", "One-versus-One LST-SVM", "All-versus-One LST-SVM" and "Directed Acyclic Graph (DAG) LST-SVM", in this article we use the One-versus-One LST-SVM Multi- classifier method.

For the *K*-class classification problem, the One-versus-One LST-SVM is one of the popular Multi-classifier methods. It generates *K*(*K*-1) binary classifiers and each class contains (*K*-1) non-parallel hyper-plans according to the result of that *i*th class is trained with *j*th class, then the *i*th class as a positive and *j*th class as a negative class and vice versa [24, 29].

In the linear case, the LST-SVM Multi- classifier method as following steps:

(1) According to the hyper-planes $f_{ij}=\left(w_{ij}x\right)+b_{ij}=0$ and $f_{ji}=\left(w_{ji}x\right)+b_{ji}=0$, where $w_{ij},w_{ji}\in R^n$ are normal vectors to the hyperplane in n-dimensional real space and $b_{ij},b_{ji}\in R$ are bias terms, the objective functions for linear cases is obtained, if the *i*th class is trained with the *j*th class or vice versa:

$$\min(w_{ij},b_{ij},\xi_{ij})\ \frac{1}{2}\|X_i w_{ij}+e_{i1}b_{ij}\|^2+\frac{c_i}{2}\xi_{ij}^T\xi_{ij},$$

$$s.t.\quad (X_j w_{ij}+e_{j1}b_{ij})+\xi_{ij}=e_{j1}, \tag{3}$$

$$\min(w_{ji}, b_{ji}, \xi_{ji})\ \frac{1}{2} \| X_j w_{ji} + e_{j1} b_{ji} \|^2 + \frac{c_j}{2} \xi_{ji}^T \xi_{ji},$$

$$s.t.\quad (X_i w_{ji} + e_{i1} b_{ji}) + \xi_{ji} = e_{i1}, \tag{4}$$

where the matrices X_i and X_j are comprised of the data points of the *i*th and *j*th class respectively. $e_{i1} \in R^{l_i}$ and $e_{j1} \in R^{l_j}$ are two vectors of ones. $c_i > 0$ and $c_j > 0$ represent penalty parameters, where $\xi_{ij} \in R^{l_j}$ and $\xi_{ji} \in R^{l_i}$ are slack variables [24].

(2) The Lagrangian function corresponding to $f_{ij} = \left(w_{ij} x\right) + b_{ij} = 0$ is determined as:

$$L(w_{ij}, b_{ij}, \xi_{ij}, \alpha_i) = \frac{1}{2} \| X_i w_{ij} + e_{i1} b_{ij} \|^2 + \frac{c_i}{2} \xi_{ij}^T \xi_{ij} - \alpha_i^T ((X_j w_{ij} + e_{j1} b_{ij}) + \xi_{ij} - e_{j1}), \tag{5}$$

where $\alpha_i > 0$ is a Lagrangian multiplier.

(3) Differentiating the parameters w_{ij}, b_{ij}, ξ_{ij}, α_i in equation (5) according to Karush–Kuhn–Tucker (KKT) conditions:

$$\frac{\partial L}{\partial w_{ij}} = X_i^T (X_i w_{ij} + e_{i1} b_{ij}) - X_j^T \alpha_i = 0, \tag{6}$$

$$\frac{\partial L}{\partial \mathrm{b}_{ij}} = e_{i1}^T (X_i w_{ij} + e_{i1} b_{ij}) - e_{j1}^T \alpha_i = 0, \tag{7}$$

$$\frac{\partial L}{\partial \xi_{ij}} = c_i \xi_{ij} - \alpha_i = 0, \tag{8}$$

$$\frac{\partial L}{\partial \alpha_i} = (X_j w_{ij} + e_{j1} b_{ij}) + \xi_{ij} - e_{j1} = 0. \tag{9}$$

(4) The Lagrangian parameter is obtained according to equation (6) to (9):
Equation (6) and (7) lead to:

$$\begin{bmatrix} X_i^T \\ e_{i1}^T \end{bmatrix} [X_i \ \ e_{i1}] \begin{bmatrix} w_{ij} \\ b_{ij} \end{bmatrix} - \alpha_i \begin{bmatrix} X_j^T \\ e_{j1}^T \end{bmatrix} = 0. \tag{10}$$

Equation (8) and (9) lead to:

$$\alpha_i = c_i (e_{j1} - \begin{bmatrix} X_j \, e_{j1} \end{bmatrix} \begin{bmatrix} w_{ij} \\ b_{ij} \end{bmatrix}). \tag{11}$$

Equation (10) and (11) lead to:

$$\begin{bmatrix} w_{ij} \\ b_{ij} \end{bmatrix} = (\begin{bmatrix} X_j^T \\ e_{j1}^T \end{bmatrix} [X_j \ \ e_{j1}] + \frac{1}{c_i} \begin{bmatrix} X_i^T \\ e_{i1}^T \end{bmatrix} [X_i \ \ e_{i1}])^{-1} \begin{bmatrix} X_j^T \\ e_{j1}^T \end{bmatrix} e_{j1}. \tag{12}$$

According to step (2) and (3), we can also obtain w_{ji} and b_{ji}.

$$\begin{bmatrix} w_{ji} \\ b_{ji} \end{bmatrix} = (\begin{bmatrix} X_i^T \\ e_{i1}^T \end{bmatrix} [X_i \ \ e_{i1}] + \frac{1}{c_j} \begin{bmatrix} X_j^T \\ e_{j1}^T \end{bmatrix} [X_j \ \ e_{j1}])^{-1} \begin{bmatrix} X_i^T \\ e_{i1}^T \end{bmatrix} e_{i1}. \tag{13}$$

(5) The distance of a test data point is calculated from f_{ij} and f_{ji} hyper-plane:

$$d_{ij} = \frac{| w_{ij}x + b_{ij} |}{\| w_{ij} \|} \text{ and } d_{ji} = \frac{| w_{ji}x + b_{ji} |}{\| w_{ji} \|}. \tag{14}$$

The training phase repeat step (1) to step (4) to get *K(K-1)* hyper-planes, *(K-1)* planes for each class, the test phase assigns the class to new data point according to the basis of "max-wins voting" strategy and step (5). For example, if $d_{ij} < d_{ji}$, i.e., the distance of the point from *i*th class is less than the distance from the *j*th class, then the vote is given to the *i*th class, conversely, the number of votes for the *j*th class is increased by one [24].

2.3 MTS Distance Based LST-SVM Multi-classifier Method

For the time-series data, we use the MTS method to remove the noise data and get time feature data as the input data, the processing as described below:

The raw time-series data $X = X_1 + \cdots + X_N$, where X_i is the matrix of the data with the same label, *N* is the number of class.

(1) Calculating the mean μ_{x_i} of each X_i, for each data with the same label using the MTS method to calculate the distance D_{x_i} with μ_{x_i}, then getting the distance matrix $D = D_{x_1} + \cdots + D_{x_N}$.

(2) Calculating the mean $\mu_{D_{xi}}$ and the standard deviation $\delta_{D_{xi}}$ of each $D_{x_i} = \left(D_{i_m}, \cdots, D_{i_n}\right)$, where *m, n* are the index of the X_i. Then removing the data which $D_{ij} > \mu_{D_{xi}} + 3\delta_{D_{xi}}$ of each D_{x_i} by the outlier detection based on normal distribution, where $m \le j \le n$.

Through the above steps, we get the input data with time feature. Then in order to get the classifier model, we randomly choose the 80 percent of labeled data to train by LST-SVM Multi-classifier method and use remaining data to test the model precision.

3 Experiment Results and Analysis

3.1 Data Source

Shanghai, one of the biggest urban in China, is selected as our case study area, but the whole urban is too big, so we choose the longitude from 121.3 to 121.7 and latitude from 31.0 to 31.3 of the city for the study as shown in Fig. 2. In recent decades, Shanghai has become one of the central urban in China, functioning as the Chinese economy, transportation, technology, industry, finance, trade, convention, and shipping center. Shanghai is characterized by complex urban morphology and high levels of mixed land use.

Our main data is the check-in data of social network users on Sina Weibo, one of the largest online social media platforms in China with more than 300 million users. This data is produced by mapping locations of active smartphone users who are using Sina Weibo. Due to its large user base, the check-in data not only contains location information but also records users' travel demands, and it could provide a representative depiction of population dynamics. However, the Sina Weibo check-in data in one place for a day is very sparse, we implement a web crawler to acquire the check-in data for five years from 2012 to 2017.

Our analysis employs a POIs database to supplement information about name, address, and category of individual places. The POIs data utilized in this work consists of 8 representative categories, namely traffic (TR), residential community (RC), education (ED), restaurant (RU), shopping (SH), entertainment (EN), scenic spot (SS), and company (CO). These categories of POIs will be used as auxiliary data to label the building functionality types.

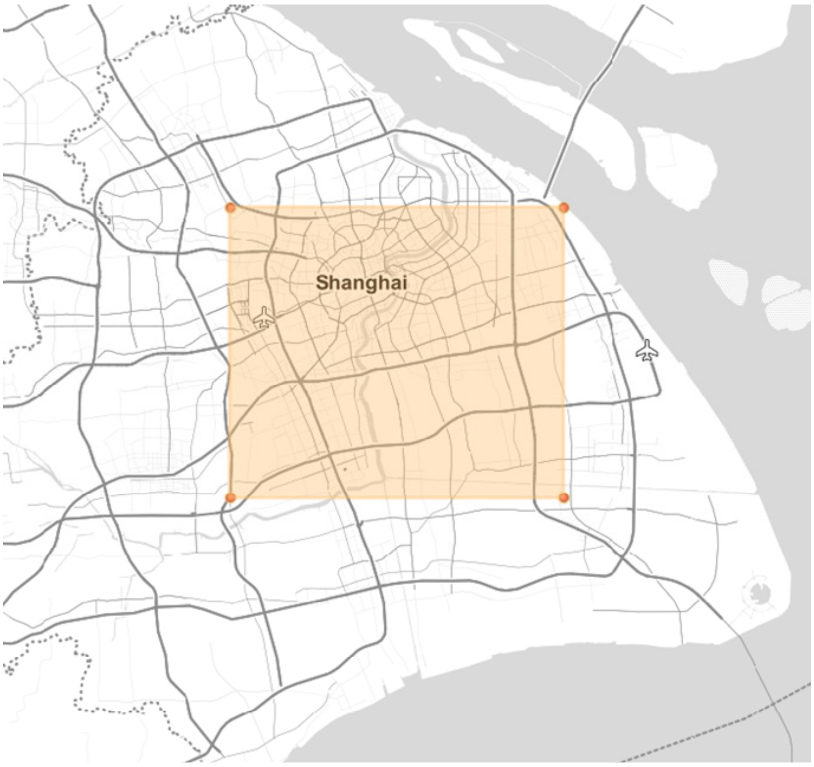


Fig. 2. The selected study area of Shanghai

The critical assumption in our analysis is that the social media activities in buildings with similar functionality have similar spatiotemporal profiles and are strictly correlated to the temporal. We get 16,853 building objects with POI label data, and their 2,570,731 check-ins data from the study area, Fig. 3 shows the spatial distribution for different types of functionality buildings in the study area. We can see that the buildings with different functionality are cross-distributed in Shanghai. For example, Fig. 4(a) and Fig. 4(b) shows the spatial distribution for RC and RU respectively, we can find that most RU is located adjacent to the RC from Fig. 4(c). Due to this distribution characteristic, we delineate urban functional areas from the perspective of functionality buildings distribution.

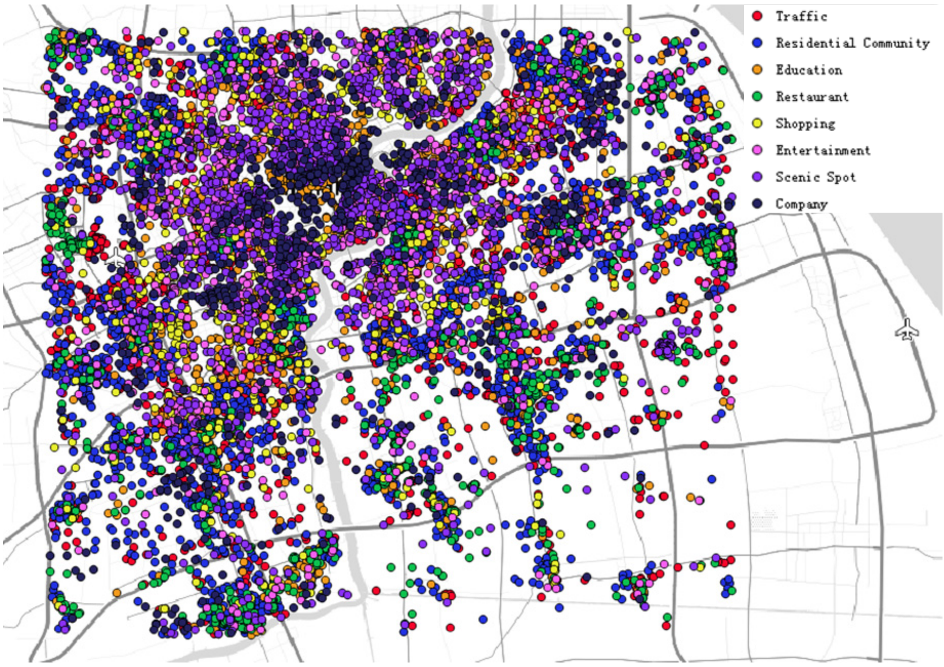
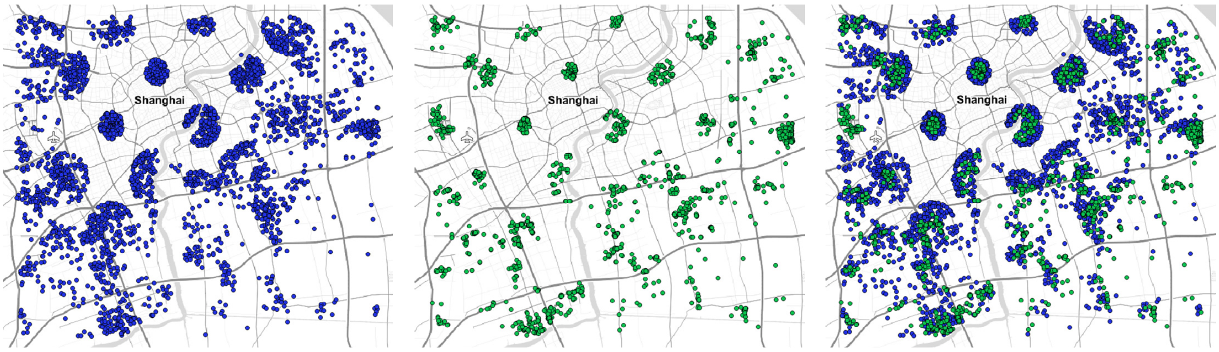


Fig. 3. Spatial distribution for different types of functionality buildings



(a) RC (b) RU (c) RC and RU

Fig. 4. A spatial distribution example of Residential Community (RC) and Restaurant (RU)

For each category, the number of building objects is shown in Table 1. We sorted out the above check-in data with the label and built the time-series dataset which reveals the spatiotemporal distributions of Sina Weibo users in hourly intervals for each building objects. The diurnal average temporal distribution characteristics of different categories are depicted in Fig. 5. These characteristic curves show People Daily travel habits on different functionality buildings. For example, we can observe that the TR and the RU each have two peaks that emerge during different periods throughout the day. The first peak for TR appears in the period from 7 am to 9 am, showing that most people are more likely to appear in traffic-related buildings during this time period; at lunchtime, the RU reaches its first peak. Their second peaks appear in a similar time frame which is from 5 pm to 7 pm, and combining the curve trend of RC, SC, CO, and EN, suggesting that most of the users return home, have dinner or participate in entertainment after work. We refer to this data as building time series (BTS) hereafter.

Table 1. The number of building objects

Category	TR	RC	ED	RU	SH	EN	SS	CO	Total
Number	1267	3955	1992	3134	1696	1847	1835	1127	16853

Note. TR=Traffic; RC=Residential Community; ED=Education; RU=Restaurant; SH=Shopping; EN=Entertainment; SS=Scenic Spot; CO=Company.

3.2 Result and Analysis

In order to verify the accuracy and usefulness of our proposed method, we choose dynamic time warping (DTW) distance based *k*-medoids method (DTW *K*-Medoids) as a comparison method, which has emerged as a popular method for time-series data clustering and some scholars have used this method to describe urban functional areas and get well results [19]. We have done a series of experiments based on the *k*-medoids method and LST-SVM Multi-classifier method to better illustrate the advantages of our proposed method. According to the eight types of functional buildings, each group of experiments uses a different algorithm for grouping buildings with similar social media activities into urban functional areas. We compare the experimental results with the existing building functional label and take the average of 10 times test accuracies as each method final predictive accuracy. Tables 2 and Table 3 show the prediction accuracy of each type of functional building under different methods respectively.

Firstly, we applied the DTW distance based *k*-medoids method (DTW *K*-Medoids) to group buildings in BTS, $k = 8$ [19]. However, the overall clustering accuracy is not good after using this method. Although the accuracy of the RC category is 91.88%, the ED, RU, and SH are very low. There are two reasons for this experimental result. On the one hand, the type of buildings depends on the spatiotemporal trend of the time series, and as discussed in Section 2.1, the inherent defect of DTW algorithm in calculating time series similarity plays a great influence. On the other hand, we built the time-series dataset based on users' check-in data for each building objects, and some buildings have similar time series trend, but the number of check-in data in the same time period may be different or even be big. It will also affect the time-series similarity. In order to better handle the data, we use the zero-mean normalization method to get normalized building time series (NBTS). Then the DTW distance based *k*-

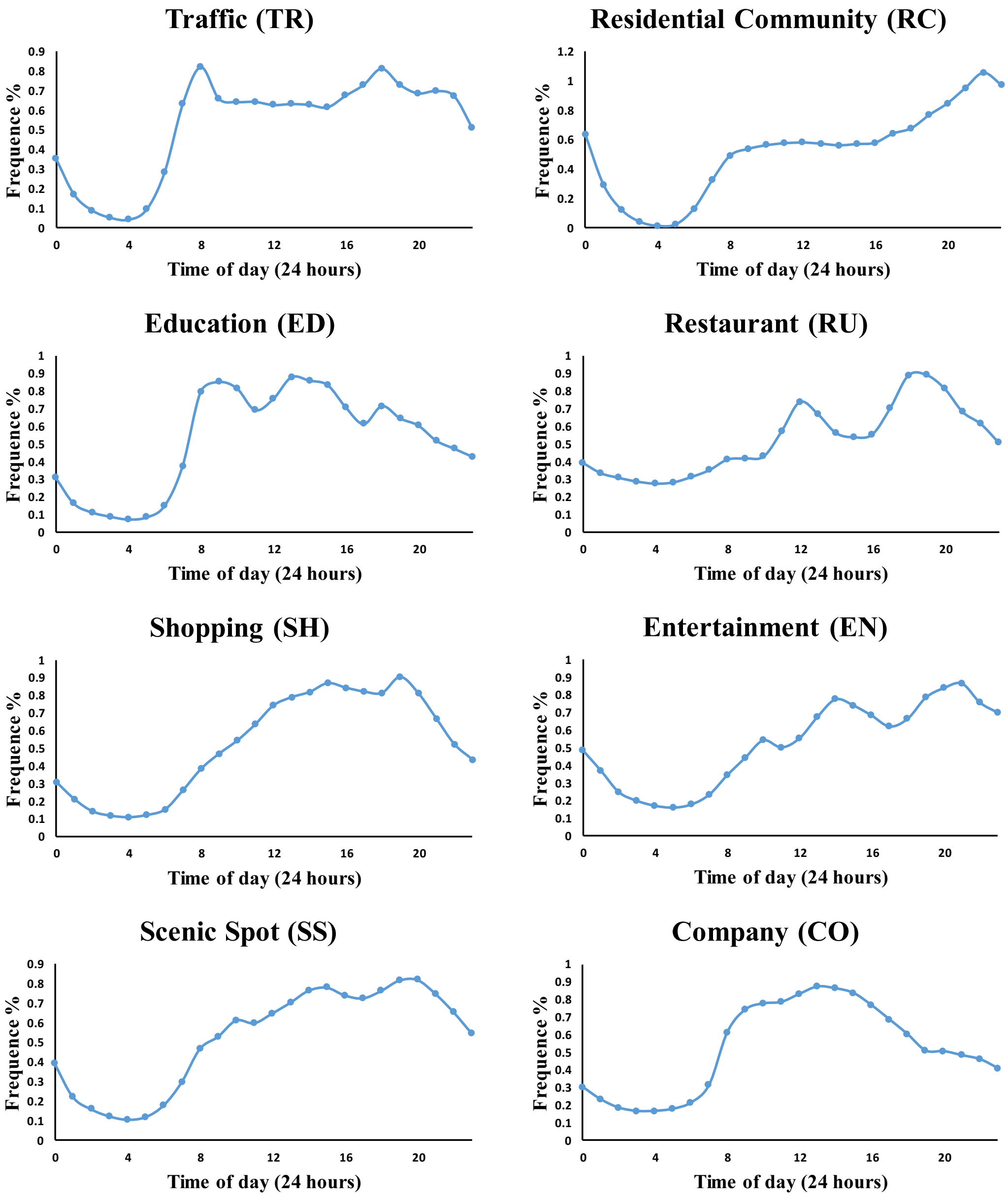


Fig. 5. Diurnal average temporal distributions of different functionality building categories

Table 2. Accuracy of each algorithm based on *k*-medoids method

Algorithm	TR	RC	ED	RU	SH	EN	SS	CO	All
DTW *K*-Medoids (%)	14.63	91.88	0.21	0.37	0.12	60.47	3.60	34.87	29.96
NDTW *K*-Medoids (%)	16.10	32.77	22.84	42.65	26.58	12.39	16.51	25.82	25.78
MTS *K*-Medoids (%)	42.13	42.18	58.11	51.73	58.57	29.85	27.06	77.99	45.87

Table 3. Accuracy of each algorithm based on LST-SVM Multi-classifier method

Algorithm	TR	RC	ED	RU	SH	EN	SS	CO	All
NDTW LST-SVM (%)	30.88	75.22	49.78	75.42	32.53	29.96	18.08	20.04	48.66
MTS LST-SVM (%)	76.96	96.87	84.02	97.17	79.08	73.86	48.01	70.53	82.68

Note. TR = Traffic; RC = Residential Community; ED = Education; RU = Restaurant; SH = Shopping; EN = Entertainment; SS = Scenic Spot; CO = Company.

medoids method is still utilized to cluster the buildings in NBTS dataset (NDTW *K*-Medoids), *k* = 8. As seen in Table 2, after data normalization, although each category is identified, the accuracy of each building functional categories is still very low. It suggests that the DTW distance is not appropriate to calculate the distance of time-series in the BTS dataset.

Then we use the matching time series (MTS) method proposed in this paper to replace the DTW method to calculate the distance between two time-series. Then the MTS distance based *k*-medoids method (MTS *K*-Medoids) is applied to cluster urban functionality buildings in BTS dataset, *k* = 8. Comparing to the prior method in Table 2, we can find that each building functional category is identified and the accuracy of each category have a great improvement. By comparing the MTS distance based *k*-medoids method with the traditional DTW distance based *k*-medoids method, we can find that using the MTS method to determine the similarity of time series is valid.

Next, we use the MTS distance based LST-SVM Multi-classifier method (MTS LST-SVM) classify buildings with similar functionality as a functional area. Using the MTS method to remove the noise data and get input data with time features from the BTS dataset $X = X_1 + \cdots + X_N$, where X_i is the matrix of the data with the same functionality label, *N=8*. Then we randomly choose the 80 percent of data to train the model by LST-SVM Multi-classifier method and use remaining data to test the model precision. Table 3 shows the predicted accuracy of the functional building categories obtained by this method. Meanwhile, in order to illustrate using the MTS distance to remove the noise time series is effective, we use the time series distance calculated by the traditional DTW method based on the NBTS dataset to replace it. Then we use the same way to remove the time series noise and randomly choose the 80 percent of data to train the model by LST-SVM Multi-classifier method and use remaining data to test the model precision. The result obtained by the DTW distance based LST-SVM Multi-classifier method (DTW LST-SVM) is also shown in Table 3. Observing the Table 3, we can also find the MTS method is superior to the traditional DTW method in calculating the similarity of time series.

Comparing Table 2 and Table 3, the overall predicted accuracy of MTS LST-SVM method is 82.26%, which is better than other methods, and the accuracy of each category is relatively high. In addition to CO, other categories of building classification accuracy have been greatly improved. Especially, the accuracy of RC and RU is up to 95%. Fig. 6 shows the test results for the RC class (the blue dots indicate a successful classification of RC buildings, and other color points suggest that some buildings have been misjudged). The above results show that the MTS distance based LST-SVM Multi-classifier method proposed in this paper is more effective than the traditional clustering method in improving the accuracy rate of determining the function of urban buildings. Combining latitude and longitude data and the building function judged by our method, it can help us better study and delineate areas in a city with mixed distribution of functional buildings.

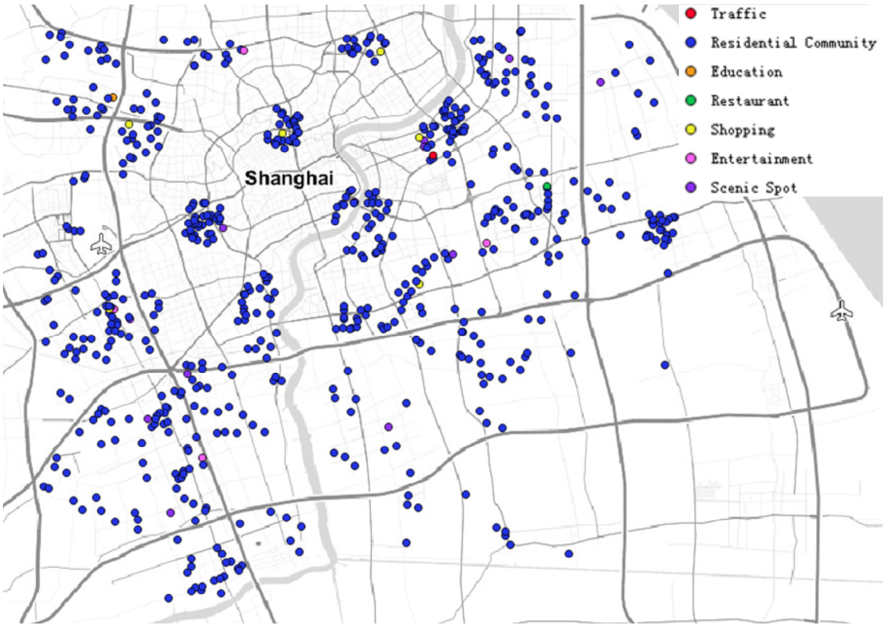


Fig. 6. The test result for the RC by the MTS based LST-SVM Multi-classifier method

However, the spatiotemporal patterns are different on weekdays and weekends for building objects with related functionality. Therefore, we separate BTS dataset for weekdays and weekends. This will reduce the computational burden, while does not generate significant information loss. And it helps us better understand the characteristics of urban functional areas in time [19]. Table 4 shows the number of functionality buildings for each category on weekdays and weekends. We can see that the total amount of functional buildings marked on social media on weekdays is more than the weekend. Particularly, the number of RS and CO change is the most obvious, showing that people eating out and going to the company on the weekend is relatively small. As seen in Fig. 7, the diurnal average temporal distributions of different functionality building categories are different on weekdays and weekends. For instance, the two obvious peaks of TR is disappear and the trend of the day has no major fluctuations on weekends, suggesting that the phenomenon of city morning and evening rush hour is not obvious on weekends; by observing the distribution changes of SH, EN, and SS, we can also find differences that their peaks on weekends are earlier than weekdays, showing that most of the people on weekdays are used to going entertainment and shopping after work, but people like doing them in the afternoon on weekends.

Table 4. The number of building objects on the weekdays and weekends

Time	TR	RC	ED	RU	SH	EN	SS	CO	Total
Weekdays	1267	3955	1991	2796	1689	1805	1778	1127	16408
Weekends	1250	3920	1944	2317	1662	1712	1675	997	15477

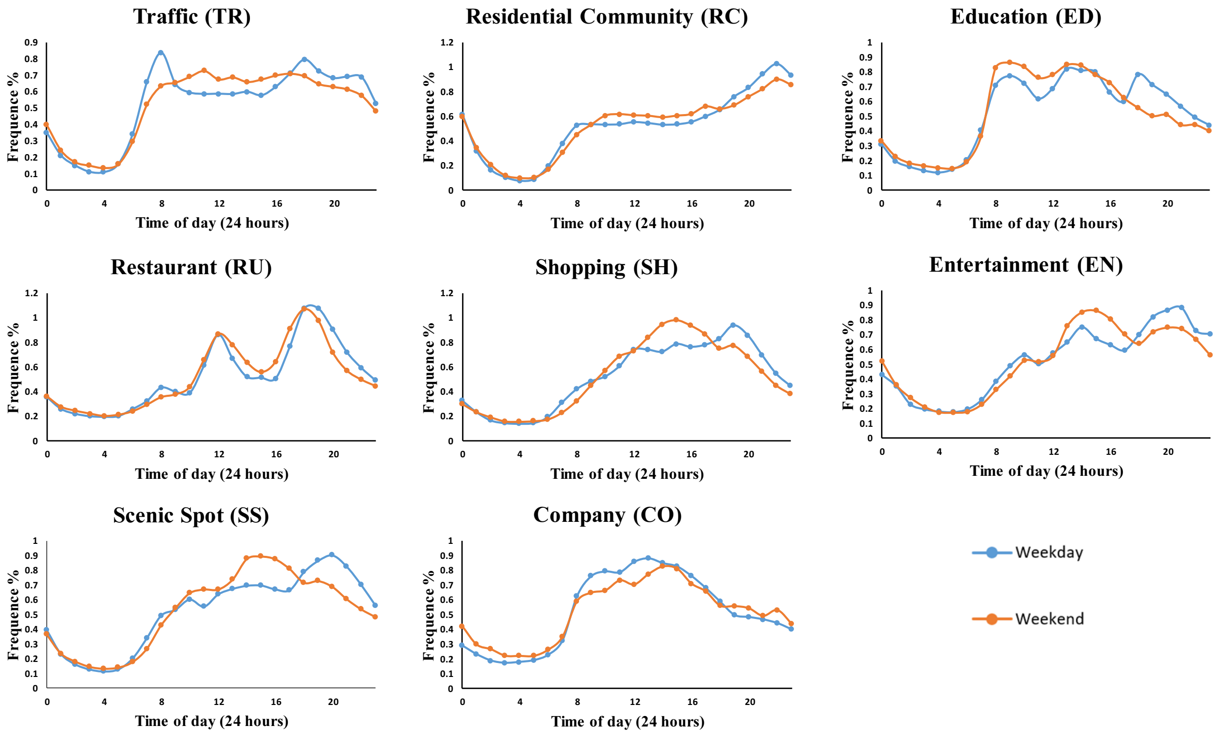


Fig. 7. Diurnal average temporal distributions of different functionality building categories on weekdays and weekends

Based on the difference of time-series characteristics described above on weekdays and weekend, we also use the MTS based LST-SVM Multi-classifier method to classify buildings with similar functionality as a functional area according to the datasets of these two time periods respectively. Table 5 lists the classification accuracy rate using this method on weekdays and weekends. The overall classification accuracy of weekdays and weekends is 84.03% and 84.84% respectively, and they are all higher than the overall accuracy of no time division. The accuracy of SS is lower than other functionality objects on weekdays, and the reason is that the trend and frequency of SS are very similar to the EN and SH, thus it is easy to misjudge when doing multiple classifications. For the same reason, CO, SS, and SH are very close on weekends, and their classification accuracies would be affected by each other.

Table 5. Classifying accuracy of LST-SVM on the weekdays and weekends

Time	TR	RC	ED	RU	SH	EN	SS	CO	All
Weekdays (%)	83.94	96.73	85.81	91.70	75.11	73.45	60.26	86.31	84.03
Weekends (%)	70.83	96.02	93.72	91.53	82.41	85.64	63.24	53.75	84.84

Note. TR = Traffic; RC = Residential Community; ED = Education; RU = Restaurant; SH = Shopping; EN = Entertainment; SS = Scenic Spot; CO = Company.

By dividing the BTS dataset into weekdays and weekends, not only we are able to describe the urban functional areas more clearly, but also it helps us better to analyze the change of urban population flow and strengthen urban planning through combining the functionality buildings distribution.

4 Conclusion

In this paper, we use the BTS dataset obtained from Sina Weibo check-in data and propose an MTS distance based LST-SVM Multi-classifier method for delineating urban functional areas. We assume that social media activities in buildings with similar functionality have similar spatiotemporal profiles and are strictly correlate to the temporal, and subsequently verify the effectiveness of this method by a series of experiments. Firstly, in our dataset, comparing with the method that using a dynamic time warping (DTW) distance based k-medoids to group functionality buildings, the MTS method can calculate the distance of two time-series more accurately and more reasonably to capture the inherent heterogeneity even within the same urban functionality types. We implement the MTS based k-medoids method and the DTW distance based k-medoids method for clustering functional buildings in BTS dataset respectively. The results show that the overall accuracy is improved from 29.96% to 45.87%. Thus, our proposed MTS method, in terms of calculating the distance and determining the similarity of two time-series, is effective and generalized. Secondly, in order to improve the estimated accuracy of delineating urban functional areas, we use classification ideas to delineate urban functional areas. Compared our method with the previous clustering method the accuracy of the MTS based LST-SVM Multi-classifier method is achieved to 82.68%, which verifies the superiority of our approach over the MTS distance based k-medoids method. Last, we separate BTS dataset into weekdays and weekends, and the classification results are 84.03% and 84.84% respectively. It demonstrates that utilizing different spatiotemporal patterns of these two time periods helps us to analyze the distribution of urban functional buildings more clearly, especially the different functional buildings in the city are cross-distributed.

In addition, with the fast development of the urban, more and more new buildings are springing up in the city, it is time-consuming and tedious to understand the functionalities of new buildings using traditional human intervention methods, our proposed method can be used to quickly determine its functionality according to the time series trend of the target building and delineate urban functional areas dynamically and clearly. Our future research work will further expand the categories of building functionality and incorporate more social network check-in data in different cities to promote this approach. We plan to extract more information from social media and combine crowd flow to study more urban problems.

Acknowledgments

This work is supported by key projects of Shanghai science and technology commission (No. 17511106802).

References

[1] C. Karlsson, M. Olsson, The identification of functional regions: theory, methods, and applications, The Annals of Regional Science 40(1)(2006) 1-18.

[2] S.-S. Wu, X. Qiu, E.L. Usery, L. Wang, Using geometrical, textural, and contextual information of land parcels for classification of detailed urban land use, Annals of the Association of American Geographers 99(1)(2009) 76-98.

[3] Y. Zhi, H. Li, D. Wang, M. Deng, S. Wang, J. Gao, Z. Duan, Y. Liu, Latent spatio-temporal activity structures: a new approach to inferring intra-urban functional regions via social media check-in data, Geo-spatial Information Science 19(2)(2016) 94-105.

[4] J. Reades, F. Calabrese, C. Ratti, Eigenplaces: analysing cities using the space–time structure of the mobile phone network, Environment and Planning B: Planning and Design 36(5)(2009) 824-836.

[5] G. Qi, X. Li, S. Li, G. Pan, Z. Wang, D. Zhang, Measuring social functions of city regions from large-scale taxi behaviors, in: Proc. 2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011.

[6] Y. Liu, F. Wang, Y. Xiao, S. Gao, Urban land uses and traffic "source-sink areas": evidence from GPS-enabled taxi data in Shanghai, Landscape and Urban Planning 106(1)(2012) 73-87.

[7] J.L. Toole, M. Ulm, M.C. González, D. Bauer, Inferring land use from mobile phone activity, in: Proc. 2012 the ACM SIGKDD International Workshop on Urban Computing, 2012.

[8] R.L. Steiner, Residential density and travel patterns: review of the literature, University of Florida, Gainesville, 1994 (No. 1466).

[9] K. Kockelman, Travel behavior as function of accessibility, land use mixing, and land use balance: evidence from San Francisco Bay Area, Journal of the Transportation Research Board 1607(1997) 116-125.

[10] S. Gao, K. Janowicz, H. Couclelis, Extracting urban functional regions from points of interest and human activities on location-based social networks, Transactions in GIS 21(3)(2017) 446-467.

[11] J. Yuan, Y. Zheng, X. Xie, Discovering regions of different functions in a city using human mobility and POIs, in: Proc. The 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012.

[12] N.-J. Yuan, Y. Zheng, X. Xie, Y.-Z. Wang, K. Zheng, H. Xiong, Discovering urban functional zones using latent activity trajectories, IEEE Transactions on Knowledge and Data Engineering 27(3)(2015) 712-725.

[13] Y. Yuan, M. Raubal, Y. Liu, Correlating mobile phone usage and travel behavior: a case study of Harbin, China, Computers, Environment and Urban Systems 36(2)(2012) 118-130.

[14] S. Gao, Y. Liu, Y. Wang, X. Ma, Discovering spatial interaction communities from mobile phone data, Transactions in GIS 17(3)(2013) 463-481.

[15] S. Hasan, C.M. Schneider, S.V. Ukkusuri, M.C. González, Spatiotemporal patterns of urban human mobility, Journal of Statistical Physics 151(1-2)(2013) 304-318.

[16] L. Shi, G. Chi, X. Liu, Y. Liu, Human mobility patterns in different communities: a mobile phone data-based social network approach, Annals of GIS 21(1)(2015) 15-26.

[17] A. Noulas, S. Scellato, C. Mascolo, M. Pontil, An empirical study of geographic user activity patterns in foursquare, ICwSM 11(70-573)(2011) 2.

[18] F. Zhen, Y. Cao, X. Qin, B. Wang, Delineation of an urban agglomeration boundary based on Sina Weibo microblog "check-in" data: A case study of the Yangtze River Delta, Cities 60(2017) 180-191.

[19] Y. Chen, X. Liu, X. Li, X. Liu, Y. Yao, G. Hu, X. Xu, F. Pei, Delineating urban functional areas with building-level social media data: A dynamic time warping (DTW) distance based k-medoids method, Landscape and Urban Planning 160(2017) 48-60.

[20] S. Rani, G. Sikka, Recent techniques of clustering of time series data: a survey, International Journal of Computer Applications 52(15)(2012) 1-9.

[21] B.P. Salmon, J.C. Olivier, K.J. Wessels, W. Kleynhans, F.V.D. Bergh, K.C. Steenkamp, Unsupervised land cover change detection: Meaningful sequential time series analysis, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 4(2)(2011) 327-335.

[22] J. Senthilnath, S.N. Omkar, V. Mani, N. Tejovanth, P.G. Diwakar, A.B. Shenoy, Hierarchical clustering algorithm for land cover mapping using satellite images, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 5(3)(2012) 762-768.

[23] Y.-H. Shao, W.-J. Chen, W.-B. Huang, Z.-M. Yang, N.-Y. Deng, The best separating decision tree twin support vector machine for multi-class classification, Procedia Computer Science 17(2013) 1032-1038.

[24] D. Tomar, S. Agarwal, A comparison on multi-class classification methods based on least squares twin support vector machine, Knowledge-Based Systems 81(2015) 131-147.

[25] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, E. Keogh, Addressing big data time series: Mining trillions of time series subsequences under dynamic time warping, ACM Transactions on Knowledge Discovery from Data (TKDD) 7(3)(2013) 10.

[26] N. Deng, Y. Tian, C. Zhang, Support Vector Machines: Optimization Based Theory, Algorithms, and Extensions, Chapman & Hall/CRC, New York, 2012.

[27] Y. Tian, Y. Shi, X. Liu, Recent advances on support vector machines research, Technological and Economic Development of Economy 18(1)(2012) 5-33.

[28] M.A. Kumar, M. Gopal, Least squares twin support vector machines for pattern classification, Expert Systems with Applications 36(4)(2009) 7535-7543.

[29] Q. Yu, L. Wang, Least squares twin SVM decision tree for multi-class classification, in: Proc. 2016 IEEE International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2016.