Landscape and Urban Planning 160 (2017) 48–60

Contents lists available at ScienceDirect

Landscape and Urban Planning

journal homepage: www.elsevier.com/locate/landurbplan

Research Paper

Delineating urban functional areas with building-level social media data: A dynamic time warping (DTW) distance based *k*-medoids method

Yimin Chen [a], Xiaoping Liu [a,*], Xia Li [a], Xingjian Liu [b], Yao Yao [a], Guohua Hu [a], Xiaocong Xu [a], Fengsong Pei [c]

[a] Guangdong Key Laboratory for Urbanization and Geo-simulation, School of Geography and Planning, Sun Yat-sen University, Guangzhou, China
[b] Department of Urban Planning and Design, University of Hong Kong, Hong Kong
[c] School of Urban and Environmental Sciences, Jiangsu Normal University, Xuzhou, China

HIGHLIGHTS

- A novel method is presented for delineating urban functional areas.
- Heterogeneity is found in building clusters with similar urban functions.
- Concentric structures are found in urban villages that were deemed disordered.

ARTICLE INFO

Article history:
Received 18 April 2016
Received in revised form 1 December 2016
Accepted 3 December 2016
Available online 29 December 2016

Keywords:
Urban functional areas
Social media data
Dynamic time warping
k-Medoids

ABSTRACT

This paper presents a novel method for delineating urban functional areas based on building-level social media data. Our method assumes that social media activities in buildings of similar functions have similar spatiotemporal patterns. We subsequently apply a dynamic time warping (DTW) distance based *k*-medoids method to group buildings with similar social media activities into functional areas. The proposed method is applied in the Yuexiu District, Guangzhou, China. We carry out two clustering experiments with $k = 2$ and $k = 8$. In the experiment with $k = 2$, buildings are separated into two groups based on density values. Buildings with higher density are situated mainly within the traditional city core and urban villages in the northern part of study area. The results for $k = 8$ suggest that most buildings have mixed functions. In addition, heterogeneity can be discerned even in the clusters with similar urban functions. Concentric spatial structures are observed in urban villages that were previously deemed disordered. We also assess the diversity of urban functions at the community level and identify several potential 'central places' based on hot spot analysis. Our analysis provides an alternative way of characterizing intra-city urban spatial structure and could therefore inform future planning and policy evaluation.

1. Introduction

Delineating urban functional areas is one of the long lasting questions in urban studies and planning. While early studies proposed static models, such as concentric and sector models, to describe urban functional areas, recent efforts have viewed the formation of urban functional areas as the results of the evolving self-adaptive urban systems (Chen, Li, Liu, Ai, & Li, 2016; Dear & Flusty, 1998; Liu et al., 2014). Conventional methods of delineating urban functional areas heavily rely on remote sensing images (Heiden et al., 2012; Van de Voorde, Jacquet, & Canters, 2011). Although remote-sensing based methods are capable of capturing physical changes of urban functional areas, they do not provide sufficient socioeconomic information relating to urban functional areas. More recently, spatially referenced social media data emerged as a new data source for studying socioeconomic dynamics in the cities, such as human mobility (Hasan, Schneider, Ukkusuri, & González, 2013; Shi, Chi, Liu, & Liu, 2015), travel behaviors (Yuan, Raubal, & Liu, 2012), urban communities (Gao, Liu, Wang, & Ma, 2013), and urban land use (Pei et al., 2014). Attempts have also been made to delineate urban functional areas using

* Corresponding author at: School of Geography and Planning, Sun Yat-sen University, 135 West Xingang RD., Guangzhou 510275, China.
E-mail address: liuxp3@mail.sysu.edu.cn (X. Liu).

http://dx.doi.org/10.1016/j.landurbplan.2016.12.001
0169-2046/

social media data. For instances, Yuan, Zheng, and Xie (2012) developed a clustering method based on Latent Dirichlet Allocation (LDA) to delineate urban functional areas in Beijing using floating car trajectories and point of interests (POIs) data. Rösler and Liebig (2013) identified urban functional areas in the city of Cologne by analyzing temporal distributions of Foursquare check-in records. Zhi et al. (2016) detected intra-urban functional areas based on the clustering results of check-in records during a year-long period, and discussed the associations between these functional areas and the spatiotemporal activity patterns. Furthermore, Zhong, Huang, Arisona, Schmitt, and Batty (2014) inferred building functions in Singapore using smartcard records and Points-of-Interest (POIs). They determined the trip purposes by clustering the smartcard records according to six activity prototypes derived from household survey data, which allowed them to link these results with the distributions of POIs to deduce the building functions based on a probabilistic model. Still, Long and Shen (2015) used smart card data and POIs to characterize the spatial pattern of urban functional areas in Beijing at the TAZ (traffic analysis zones) level.

This paper presents a novel method for delineating functional areas within cities based on building-level social media data. Comparing with existing studies that delineate urban functional areas with social media data, our method has the following advantages. First, our method employs a new time-series social media dataset with high spatiotemporal resolution. This time-series dataset reveals the spatiotemporal distributions of social network users in hourly intervals. Our method assumes that social media activities in buildings of similar functions have similar spatiotemporal patterns. We subsequently apply a dynamic time warping (DTW) distance based *k*-medoids method to group buildings with similar social media activities into functional areas. Comparing with conventional types of social media data, the new dataset employed in this study offers a unique advantage — the ability to capture the inherent heterogeneity even within the same urban function types. Furthermore, while previous research using POIs data rely on predefined land use and functional types (Jiang, Alves, Rodrigues, Ferreira, & Pereira, 2015; Yuan, Zheng et., 2012; Zhong et al., 2014), the urban functions in our study are 'emerged' from spatiotemporal distributions of social network users.

Second, our analysis advances the use of spatiotemporal clustering techniques in urban and geographical studies. Despite the extensive use of clustering analysis in existing studies (Senthilnath et al., 2012), few studies have explored clustering methods for time-series geographical data (Salmon et al., 2011). Clustering time-series (or sequential) data is more difficult than the clustering of non-sequential data, as the order of elements in the time-series need to be considered when measuring the similarity between data samples (Petitjean, Ketterlin, & Gançarski, 2011). In the field of pattern recognition, however, a variety of methods for time-series data clustering has been developed. Among these methods, *k*-medoids with DTW distance has emerged as a popular method for time-series data clustering (Fu, 2011; Rani & Sikka, 2012). As a modification of the well-known *k*-means method, the *k*-medoids method updates the center of a cluster using the median cluster member itself instead of the mean position of all members. This characteristic makes the *k*-medoids method less sensitive to outliers in the data (Park & Jun, 2009). Furthermore, the DTW distance is a similarity measure transformed from the optimal alignment (i.e., a warping path) between two time series (Rakthanmanon et al., 2013). The DTW distance is found to be more robust to time-series data clustering than other conventional measures such as the Euclidian distance (Fu, 2011; Petitjean et al., 2011).

Third, the DTW distance based *k*-medoids method in our study is computationally more efficient than previous methods used to delineate urban functional areas, such as the family of probabilistic topic models (PTM) (Yuan, Zheng et., 2012; Zhang & Du, 2015). In PTM, the problem of urban functional area delineation is considered analogous to the problem of inferring the topics in an archives of documents, in which a restructuring procedure is usually applied for transforming geographical factors into relevant 'document elements' (e.g., words, corpus and vocabulary). Computation in these models is always extensive due to model training and the transformation of features in semantic spaces. Moreover, the performance of PTM is highly sensitive to the input of reliable prior information (knowledge) and also the fine-tuning of parameters. By contrast, the DTW distance based *k*-medoids method is directly driven by raw data. This method can be implemented without much human intervention (e.g., input prior information or restructuring procedures), and has a much lower computational cost.

The proposed method is illustrated with a case study of Guangzhou, one of the largest cities in China. The remainder of this paper is organized as follows. In the next section, we introduce the study area and data source. We then detail individual steps in our DTW distance based *k*-medoids method. Our method is then implemented to cluster individual buildings based on social media activities and also assess the diversity of urban functions. The article concludes with a discussion of implications for urban planning, methodological limitations as well as directions for future research.

2. Study area and data

Yuexiu, an old urban district in Guangzhou, is selected as our case study area (Fig. 1). In recent decades, Yuexiu has become one of the most densely populated areas in Guangzhou, functioning as the city's political, commercial and cultural center. Most relevant to our study here, Yuexiu is characterized by complex urban morphology and high levels of mixed land use. On the one hand, multiple urban functions, such as residential, commercial, and small-scale manufacturing may be housed in one building. On the other hand, the same urban function may take place in buildings of different types. For example, ordinary residential housing, upscale gated communities, as well as "urban villages" are all main residential types in Guangzhou (See more details about urban villages in Lin, De Meulder, and Wang (2011)).

Our main dataset is the hourly density maps of social network users on Tencent, one of the largest online social media platforms in China with more than 800 million users. We refer this data as Tencent user density (TUD) hereafter. This dataset is produced by mapping locations of active smartphone users who are using Tencent products, such Tencent QQ (an instant messenger software), WeChat (a mobile chatting service), Tencent Maps (a desktop and web mapping service) and other location based services. Due to its large user base, the TUD data could provide a representative depiction of population dynamics. In addition, TUD data also entail a much finer spatial-temporal resolution than other conventional sources of population information, such as the decennial census. Recently, Tecent release TUD through a map service (Easygo; http://ur.tencent.com/articles/100) and provide the public with real-time TUD information. We therefore implement a web crawler to acquire the TUD data for an entire week from June 15 to 21, 2015, with a spatial resolution of 25 m and a temporal resolution of one hour (Fig. 1(c)).

As our analysis focuses on building-level urban functions, a first step involves gathering building outlines and footprints. Based on visual interpretation of 2015 Quickbird image, a total of 17,231 building objects are identified for subsequent analysis (Fig. 2). Each of the building objects is regarded as the basic unit in clustering analysis. The TUD dataset is subsequently aggregated at the building level. Figs. 2 and 3 demonstrate four representative buildings, including two residential buildings and two shopping malls, with corresponding weekly TUD records. While buildings of the same

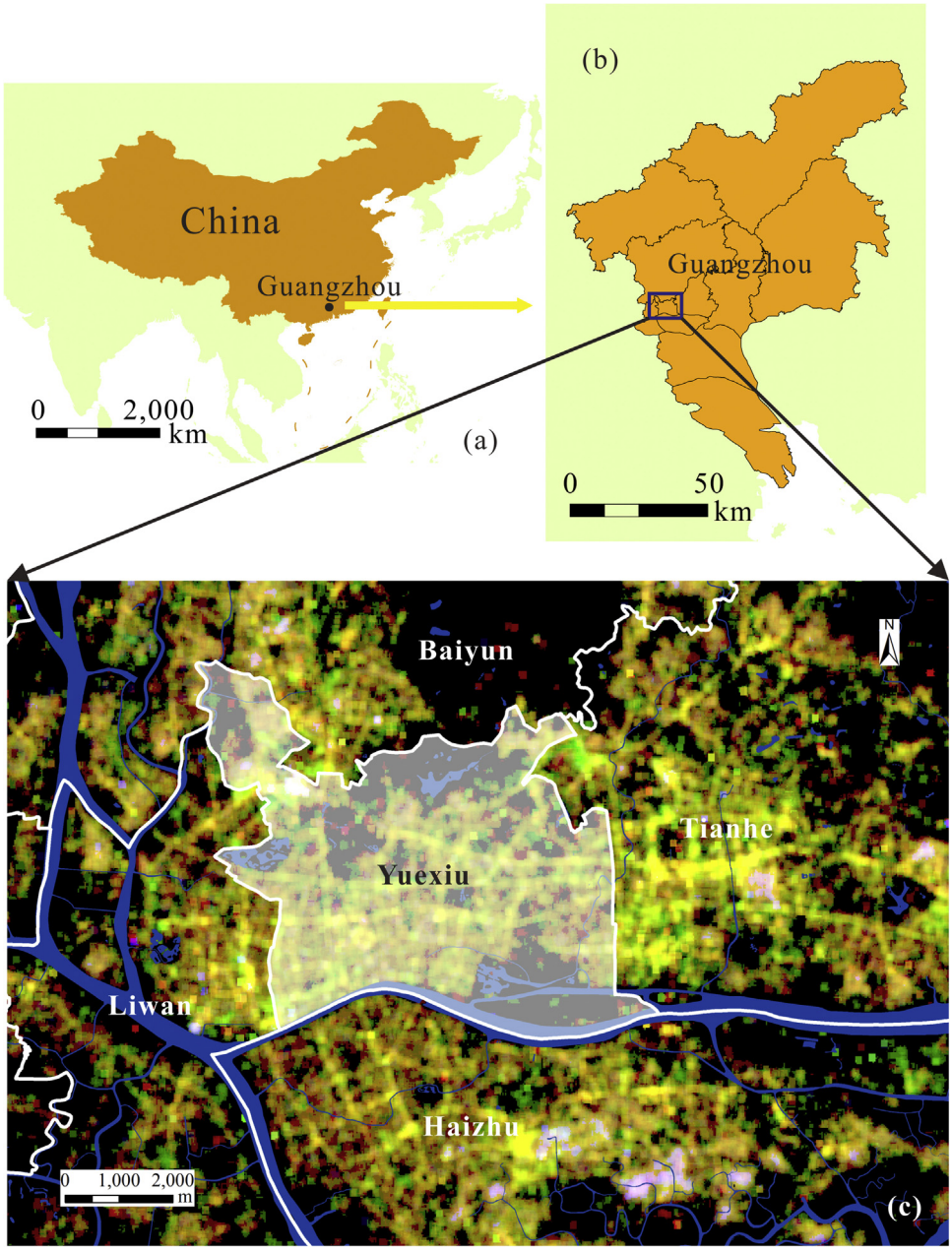


Fig. 1. The case study area: Yuexiu District of Guangzhou, China (23.137°N, 113.258°E). The red-green-blue composite of the Tencent user density in the city core of Guangzhou on June 17, 2015 (red band: 6:00; green band: 12:00; blue band: 20:00).(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

type have similar temporal profiles, these profiles differ between residential buildings and shopping malls (Figs. 2 and 3). Therefore, one critical assumption in our analysis is that social media activities in buildings of similar functions have similar spatiotemporal patterns. Furthermore, the TUD records exhibit significant periodicity during the week. Therefore, we average TUD profiles for weekdays and weekends. This will reduce the computational burden, while does not generate significant information loss. Similar averaging methods are frequently adopted in studies involving time-series data with clear periodicity (Liu et al., 2015).

In addition, our analysis employs a POIs database to supplement information about name, address and category/tag of individual places. The POIs data utilized in this work consists of 17 categories, namely residential community, elementary school, middle/high school, technical school, university (college), research institution, clinic, hospital, enterprise, financial services, government agency, hotel, park, square, retail, recreation and restaurant. These categories of POIs will be used as auxiliary data in the interpretation of building function types.

3. Methodology

3.1. Dynamic time warping (DTW) distance

In our DTW distance based *k*-medoids method, the DTW distance refers to the length of the optimal alignment (i.e., the warping path) between two given time series. A greater DTW distance suggests a more pronounced difference between two time series. For a better understanding of our method, we illustrate the calculation of the DTW distance with an example that involves two building

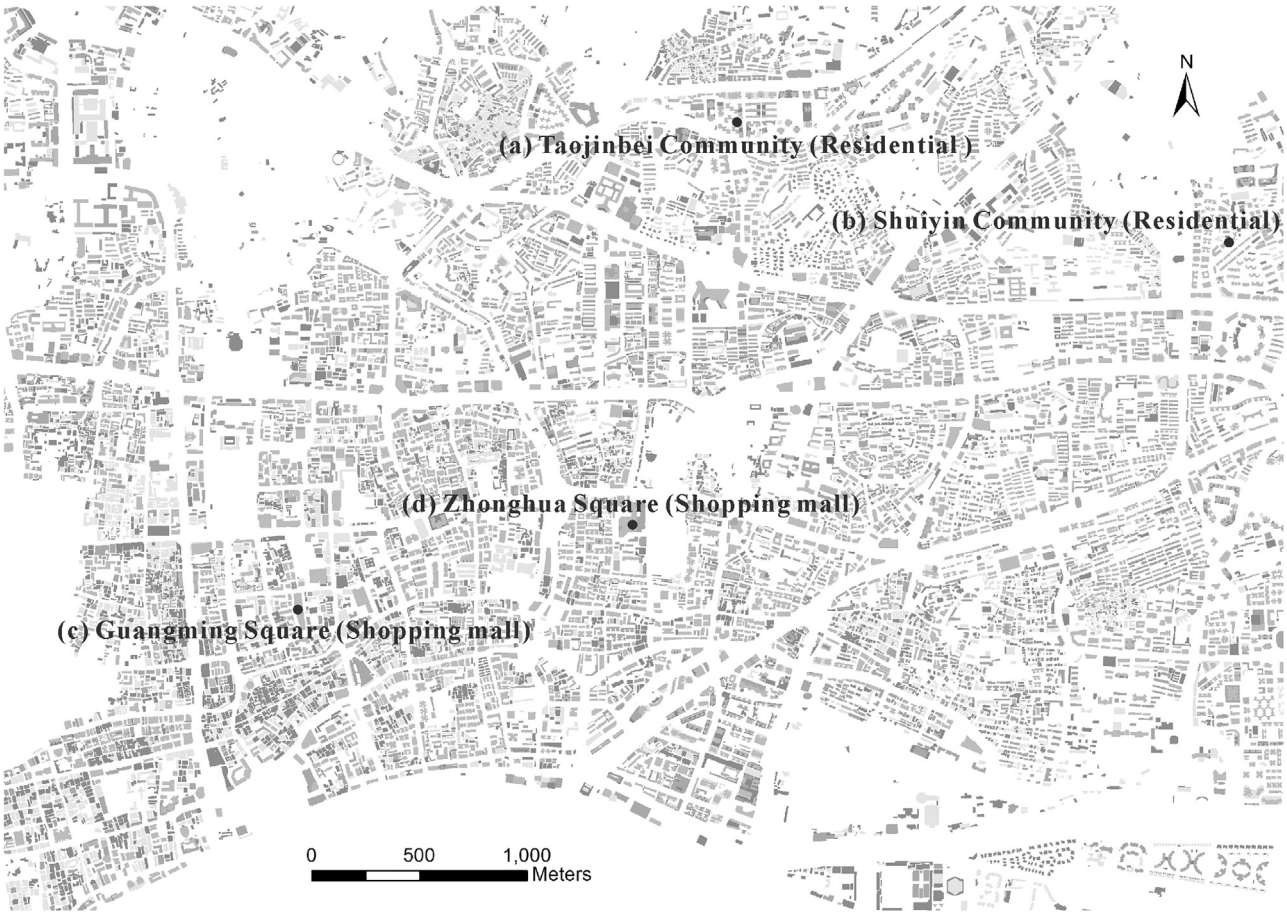


Fig. 2. Extracted building objects in the Yuexiu District with the locations of four selected samples.

objects, namely b_P and b_Q. In this example, the TUD time series for b_P and b_Q can be denoted as $P = p_1, p_2 \ldots p_m$ and $Q = q_1, q_2 \ldots q_n$ ($m = n$ as the length of the TUD records is the same for all building objects). When determining the DTW distance for P and Q, the first step is to establish a distance matrix D of $m \times n$ elements. The value of individual elements (d_{ij}) in this matrix can be calculated as follows:

$$d_{ij} = \sqrt{(p_i - q_j)^2} \tag{1}$$

Where p_i is the TUD value in the i^{th} hour in P and q_j is the TUD at the j^{th} hour in Q; and d_{ij} denotes the TUD difference between these two points. This allows the alignment between P and Q to be generated by finding a warping path (W). A warping path is a series of neighboring elements in the distance matrix that links the lower-left corner of the matrix (i.e., d_{11}) with the upper-right corner (i.e., d_{mn}) and achieves the least cumulative d_{ij} values along the path. A warping path is mathematically defined as $w_1, w_2 \ldots w_k$ where $\max(m, n) \le k \le m + n - 1$ and $w_k = (i^{\text{th}}, j^{\text{th}})$, whereby the length of this path is the so-called DTW distance.

The warping path should be subjected to three conditions, namely boundary condition, continuity condition and monotonicity condition. The boundary condition mandates that the starting point and ending point of the path should be the elements in the lower-left and upper-right corners of the matrix, i.e., d_{11} and d_{mn}, respectively. This condition ensures the alignment between the first and last element pairs of the P and Q time series. The continuity condition, which is also known as the step size condition, stipulates that w_k should be at the neighboring elements (including those diagonally positioned) of $w_{k\text{-}1}$ in the matrix. Finally, the monotonicity condition restricts w_k to be monotonically spaced in time. Given $w_{k\text{-}1} = (i_{k\text{-}1}, j_{k\text{-}1})$, $w_k = (i_k, j_k)$ should be subjected to the following constraints: $i_k \ge i_{k\text{-}1}$ and $j_k \ge j_{k\text{-}1}$. In other words, the path can have forward direction only. While many paths that satisfy these conditions can exist, only the one with the minimum cumulative d_{ij} is of interest, i.e.:

$$min \frac{\sum d_{ij}}{k} \tag{2}$$

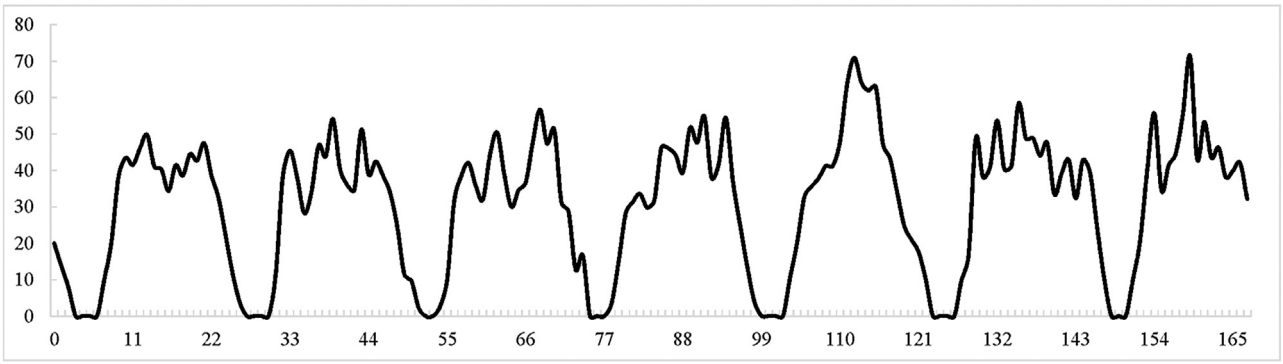
The minimization of Eq. (2) can be accomplished using dynamic programming to evaluate the following recurrence:

$$d_{cum,ij} = d_{ij} + \min\left\{d_{cum,i-1j-1}, d_{cum,i-1j}, d_{cum,ij-1}\right\} \tag{3}$$

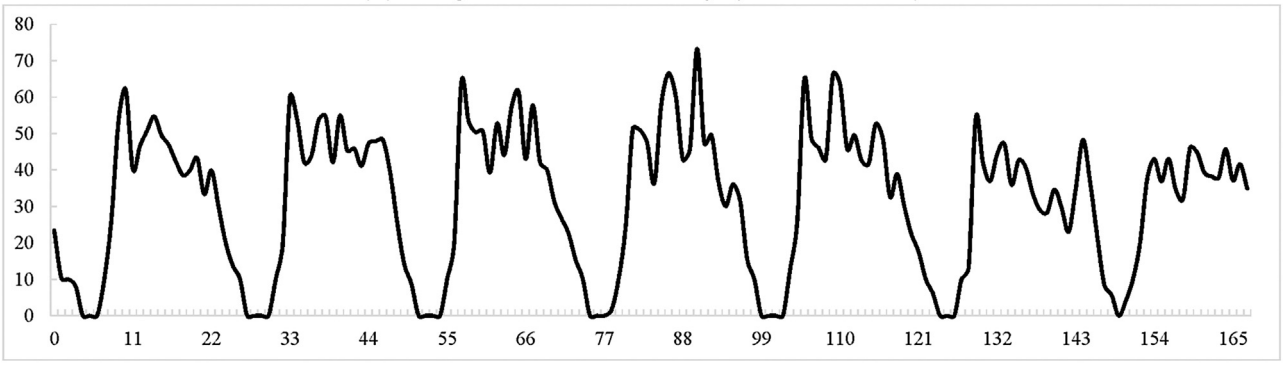
Where $d_{cum,ij}$ is the sum of current d_{ij} and the minimum of the cumulative distances of the previous elements. The resulting $d_{cum,ij}$ denotes the DTW distance between P and Q. More information of DTW distance can be found in Rakthanmanon et al. (2013).

3.2. DTW distance based k-medoids method

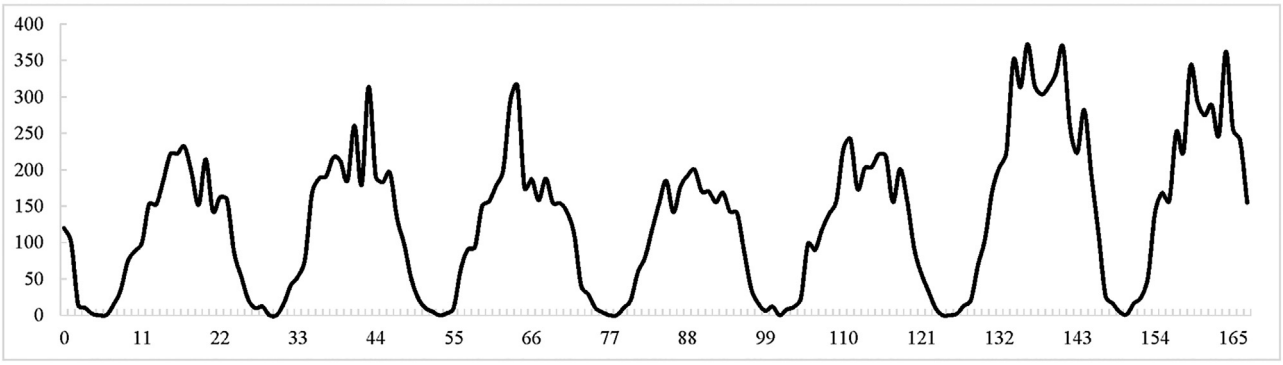
As discussed, the k-medoids method is similar to the well-known k-means for performing clustering analysis. However, these two methods differ in how they update the center location for a certain cluster. In the k-means approach, the center of a cluster is indeed virtual, because it denotes the mean position of the members that are currently within the cluster. However, the k-medoids method treats the center as the median of the cluster, which thus coincides with one of the actual members. Owing to this difference, k-medoids is more robust in responding to the outliers in



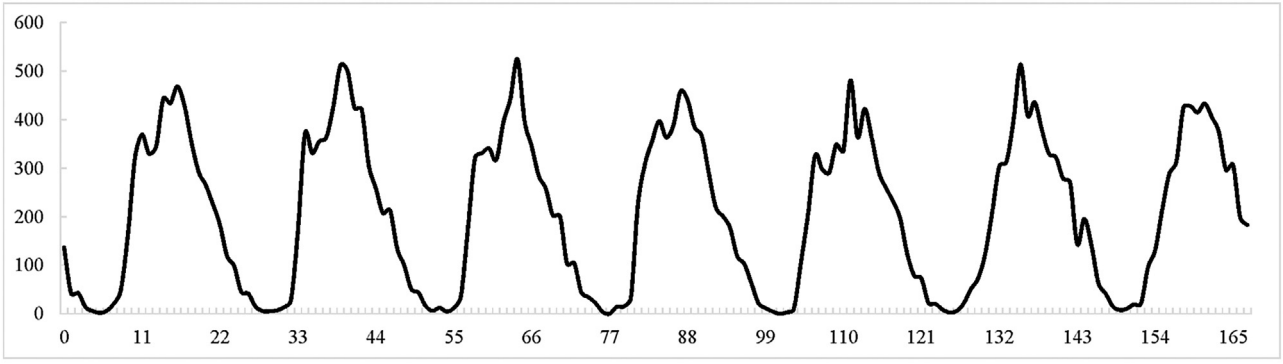
(a) Taojinbei Community (Residential)



(b) Shuiyin Community (Residential)



(c) Guangming Square (Shopping mall)



(d) Zhonghua Square (Shopping mall)

Fig. 3. Temporal changes of Tencent user density (*y* axis) by hour (*x* axis) in the four building samples.

the dataset (Park & Jun, 2009). The DTW distance based *k*-medoids method is executed through the following steps:

(1) Determine the number of clusters, i.e., the value of *k*;

(2) For initialization, select *k* samples from all building objects as the initial centers of the *k* clusters;

(3) Assign each building sample to the nearest cluster center based on the DTW distance;

(4) Within each cluster, find the median member, i.e., the member with the minimum average DTW distance to the remaining members, and select this member as the new cluster center;
(5) Repeat steps 3–4, until none of the building samples change their memberships or the number of iterations reaches the preset value.

3.3. Initialization of cluster centers

The performance of either the *k*-medoids or *k*-means is affected by the correct initialization of the initial *k* cluster centers. The conventional approach based on random initialization is problematic as it cannot effectively produce a representative set of initial cluster centers for large datasets (De Amorim & Mirkin, 2012). The current study would suffer similar setbacks, as our dataset comprises of more than 17,000 building objects. Therefore, we replace the random initialization with the modified iterated anomalous pattern (AP) method (Mirkin, 2012). The AP initialization method produces one center at a time. To meet the requirement for obtaining *k* cluster centers, the AP method can be iteratively run k' times (where $k' \geq k$) to provide a sufficient number of initial centers. The AP method also requires the specification of a reference point, which remains static throughout the entire initialization procedure. The reference point is typically chosen by using the grand mean position of the entire dataset. This position can be identified by first dividing the dataset into *n* sub-sample groups to retrieve the *n* mean positions from each group, which allows for the calculation of the mean of these *n* sub-sample mean positions. In the next step, the most distant data point from the reference point is selected as the tentative center *c* of the AP cluster *S*. Thus, all data points are assigned to either the tentative center or the reference point according to the minimum distance rule. The tentative center *c* is subsequently updated using the same strategy as that implemented in the conventional *k*-means algorithm, and the data points are re-assigned as well. Once the AP cluster *S* becomes stable, its center *c* is recorded and can be used as an initial center for the formal *k*-means clustering. Before commencing another iteration of AP initialization, the AP cluster *S* has to be subtracted from the entire dataset. In this manner, the AP method produces k' AP clusters along with their centers. Given the $k' \geq k$ condition, the *k* output centers with the greatest contributions (i.e., those contributing the greatest proportions of the cluster members to the entire dataset) are selected as the initial centers in the formal clustering procedure. More details about the AP initialization method can be found in Mirkin (2012).

We modify the original AP method to match the structure of the DTW distance based *k*-medoids method. These modifications comprise of two main steps:

(1) Select the grand median sample, instead of the grand mean position as the reference point. After finding the median in each sub-sample group (i.e., the sample characterized by the shortest DTW distances to the other members in the same group), the grand median sample is determined by selecting the median of these *n* sub-sample group medians; and
(2) Update the center of the AP cluster using the median sample, instead of the mean of all samples in the cluster.

3.4. Evaluation of clustering

As the ground truth cannot be fully known, the number of clusters *k* is determined by repeatedly executing the clustering algorithm. We evaluate the quality of individual rounds based on the silhouette metric (da Cruz Nassif & Hruschka, 2013). Given a clustering result with *k* clusters, for a building object *i*, let $a(i)$ be the average DTW distance to the building objects in the same cluster, and let $b(i)$ be the smallest average DTW distance to the building objects in the second nearest neighboring cluster. Under these conditions, the silhouette metric ($s(i)$) for building object *i* is given by:

$$s(i) = \frac{b(i) - a(i)}{\min\{a(i), b(i)\}} \tag{4}$$

A greater positive value of $s(i)$ suggests a better assignment for the building object *i*. On the other hand, a negative $s(i)$ suggests that the building object *i* should be assigned to the neighboring cluster instead of the current one. The average $s(i)$ of all building objects is the measure of the overall quality of clustering results.

We employ a concurrent POIs dataset to provide auxiliary information to interpret the clustering results. As mentioned before, it is useful to obtain the land use structure on a fine scale by analyzing the POIs dataset. We adopt the enrichment factor (Verburg, de Nijs, Ritsema van Eck, Visser, & de Jong, 2004) to characterize the relative abundance of different land use types using POIs:

$$F_{i,l} = \frac{n_{l,i}/n_i}{N_l/N} \tag{5}$$

Where $F_{i,\,l}$ denotes the enrichment of POIs in the l^{th} category for building *i*; $n_{l,\,i}$ is the number of POIs in the l^{th} category in the vicinity of the location of building *i* (e.g., 10 m), N_k represents the total number of POIs in the l_{th} category; n_i is the number of all POIs in the vicinity of the location of building *i*; and *N* is the total number of POIs in the entire study area. A higher value of $F_{i,\,l}$ indicates a larger number of POIs in the l^{th} category at the location of building *i*. Moreover, the enrichment factor is a normalized measure that removes the effects of the imbalanced POIs size in different categories. In particular, a $F_{i,\,l}$ value of 1 suggests that the enrichment of POIs in the l^{th} category is equal to the regional average, whereas $F_{i,\,l} > 1$ (or < 1), indicates that the enrichment of POIs in the l^{th} category is greater (or lower) than the regional average. The building-level clustering results can be further aggregated into, for example, the community level so that the macro urban structure can be unveiled. We use Shannon index to characterize the function diversity at the community level:

$$H_i = -\sum_k p_k \ln(p_k) \tag{6}$$

Where H_i is the function diversity for the i^{th} community and p_k is the proportion of buildings with function *k*.

4. Implementation and results

4.1. Implementation of the DTW distance based k-medoids method

A few trials are performed to determine an appropriate value for the number of clusters (*k*) in our case study. This is achieved by running the DTW distance based *k*-medoids method with different values of *k* (ranging from 2 to 20) and evaluating the results based on the silhouette metric. For comparison purpose, similar experiments are performed using the Euclidean distance based *k*-medoids method. Fig. 4 shows that the silhouette metric values obtained by the DTW distance based *k*-medoids method is consistently higher, suggesting a better performance than the Euclidean distance based *k*-medoids method. For clustering results with the DTW distance, the highest silhouette metric value (exceeding 0.5) is obtained when $k = 2$. As the value of *k* increases, the silhouette metric value decreases significantly, reaching 0.3 at $k = 5$. However, its value gradually increases again to the second highest peak level (exceeding 0.4) when *k* is within [8,10]. Therefore, in subsequent analysis, two clustering schemes are implemented using $k = 2$

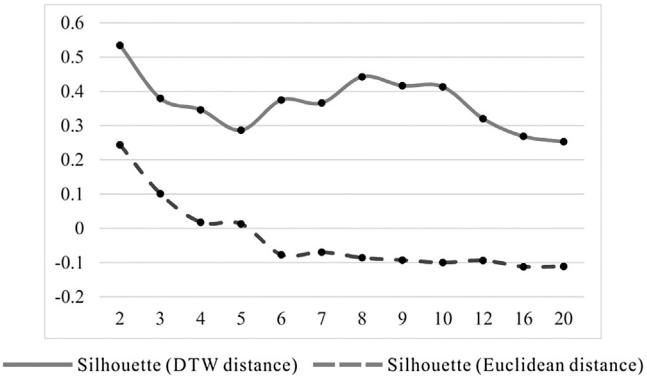


Fig. 4. The silhouette metric values of different experimental results (*y* axis) by varying the number of clusters *k* (*x* axis).

and $k = 8$, as these *k* values correspond to higher silhouette metric values.

4.2. Clustering results

4.2.1. Overview

The clustering results for $k = 2$ are shown in Fig. 5. Fig. 5(a) and (b) demonstrates different temporal patterns for weekdays and weekends. For Cluster 1, the TUD levels peak at approximately 15:00 irrespective of the day of the week; whereas for Cluster 2 dual peaks are noted at approximately 10:00 and 15:00. Fig. 5 also indicates that Cluster 1 comprises of buildings with higher population densities than those in Cluster 2, with a higher 24-h average density values for weekdays (79.09/hm^2 vs 39.66/hm^2) and weekends (88.57/hm^2 vs 40.41/hm^2). For Cluster 1, buildings are mainly located in the old city core as well as in the urban villages (Fig. 5c). Moreover, Cluster 1 includes the three most important business areas in the Yuexiu—Beijing Road (a traditional shopping center in Guangzhou), Zhonghua Square (a local shopping mall) and Guangzhou Railway Station (a retail and wholesale market for garments and clothing) – all of which are areas with the highest population densities in the Yuexiu District. This coincides with the higher TUD levels measured on the weekends in Cluster 1, as these are the peak time for shopping and social activities. In addition, buildings with higher density values in Cluster 1 are found in several urban villages, such as Yaotai, Wangshengtang, Dengfeng, and Xikeng (Fig. 5c). This suggests that the time-series TUD data can capture specific activity modes in urban villages (Lin et al., 2011).

Figs. 6 and 7 depict the results for $k = 8$. It is found that Cluster 3 is almost identical to Cluster 2 shown in Fig. 5. Therefore, the experiment with $k = 8$ can be considered a further decomposition of Clusters identified in Fig. 5. As noted, we use the POIs database to ease the interpretation of clustering results. We calculate the POIs enrichment factors (Eq. (5)) for all buildings and calculate the average for individual clusters (Table 1). The actual types of individual building clusters (Fig. 6) are determined through the interpretation of enrichment factors and site visits. We take pictures of a number of representative buildings in each cluster, which are uploaded onto Panoramio (http://www.panoramio.com/user/7144992/tags/Field%20survey; We recommend using the map view instead of the satellite view of the Panoramio region to avoid location deviation issues). The locations of these site visits are shown in Fig. 7.

To validate our clustering results, we select 130 buildings through random sampling. Subsequently we use Internet search to collect information about these buildings, and synthesize information from various sources (e.g., news, reviews, pictures, advertisements, and so on). This would allow us to infer the actual building usage and type, which will be used as benchmark to validate our clustering results. We compare the clustering results and actual building usage for the selected 130 buildings and the accuracy rate of the clustering results is relatively high (between 72% and 85%; Fig. 8).

4.2.2. Urban functional areas and their characteristics

As shown in Fig. 7, Cluster 3 ('Residential/Workplace') has the largest spatial extent compared with other clusters. For this cluster, the TUD levels are high during office hours and decline sharply during lunchtime (Fig. 6). This is consistent with patterns observed in Fig. 5b. Moreover, Cluster 3 is characterized by high enrichment factor values for the 'residential community' type, as well as some other POIs types, such as education, enterprise, government agency and hospital (Table 1). This corresponds to a high level of mixed land use. Cluster 1 ('Urban village/Residential') is mainly composed of buildings in urban villages in the northern part of Yuexiu District (Fig. 7). The weekday TUD for Cluster 1 is low during 10:00–18:00 but increases during 6:00–9:00 and 18:00–22:00 (Fig. 6). This resonates with the typical work/leisure activity patterns during the weekdays. The weekend TUD of Cluster 1 increases even further during 10:00-18:00, suggesting that

Table 1
Enrichment factors of POIs in different categories grouped by clusters.

Cluster	RC	ES	MS	TS	UN	RI	ET	FS	GA
1	1.05	0.17	0.81	0.73	0.61	0.09	0.00	0.64	0.64
2	0.55	1.50	0.55	0.67	0.00	0.29	0.00	0.65	0.38
3	1.96	2.56	3.08	1.36	1.74	1.85	1.88	1.01	1.64
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.92	0.61
5	0.56	0.00	4.60	0.00	0.00	1.64	0.00	0.49	0.61
6	0.85	1.36	0.91	0.97	0.69	0.52	0.00	0.89	0.65
7	0.89	1.22	0.44	0.77	0.92	0.60	0.00	0.90	0.73
8	0.00	12.10	0.00	0.96	0.00	0.00	0.00	0.69	1.83

Cluster	CN	HP	HT	PK	SQ	RT	RC	RS
1	2.98	1.66	0.90	4.22	5.80	1.07	1.43	0.80
2	0.31	0.00	0.54	0.00	0.00	1.57	0.43	1.15
3	1.30	2.19	0.78	2.57	2.41	0.68	0.98	0.75
4	2.19	0.00	0.37	0.00	0.00	1.10	1.29	1.50
5	3.14	3.29	1.17	0.00	0.00	1.20	1.02	0.79
6	1.10	0.70	1.26	0.59	0.00	0.98	1.00	1.20
7	1.36	1.04	0.86	0.26	0.09	1.16	1.05	0.96
8	0.00	0.00	0.00	0.00	7.06	1.57	0.07	0.65

Note: RC = residential community; ES = elementary school; MS = middle/high school; TS = technical school; UN = university (college); RI = research institution; ET = enterprise; FS = financial service; GA = government agency; CN = clinic; HP = hospital; PK = park; SQ = square; RT = retail; RC = recreation; RS = restaurant.

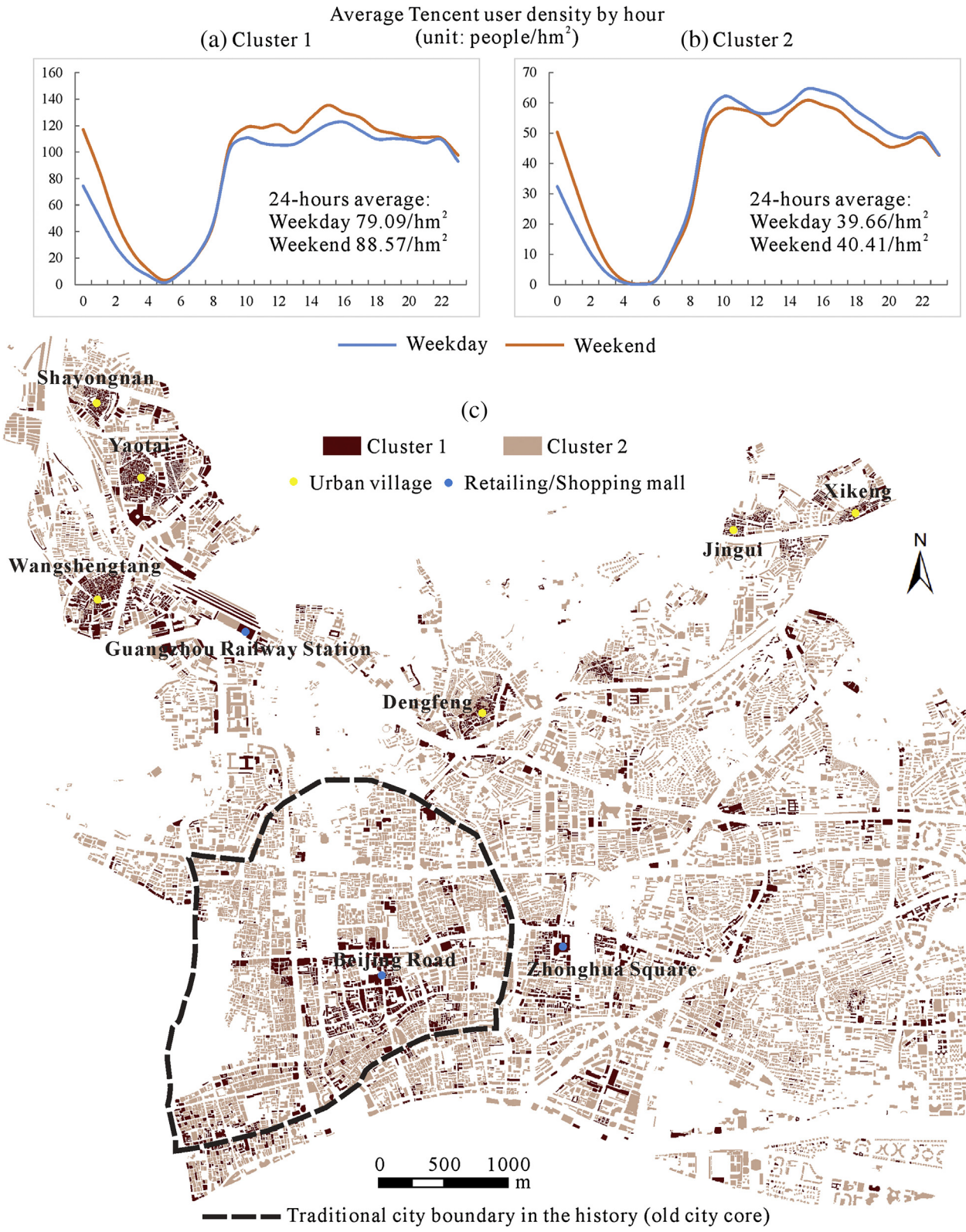


Fig. 5. Clustering results for $k = 2$, whereby (a) and (b) depict the average Tencent user densities (y axis) by hour (x axis) for the two clusters on the weekday and weekend, respectively, and (c) presents the spatial distribution of the two building clusters.

most local residents remain in the area instead of going out. The buildings in Cluster 1 are surrounded by those in Cluster 6 ('Urban village/Workplace/Socializing'), which in turn exhibit typical workplace characteristics and have much higher TUD values during the office hours (Fig. 6). Cluster 6 also has higher enrichment factors in the hotel and restaurant POIs types (Table 1). Cluster 7 ('Residential/Workplace/Socializing') has almost the same TUD patterns as those noted for Cluster 6, with the exception of the slightly lower density levels (Fig. 6). The land use compositions are similar for these two clusters (Table 1). Some buildings in Cluster 7 form the

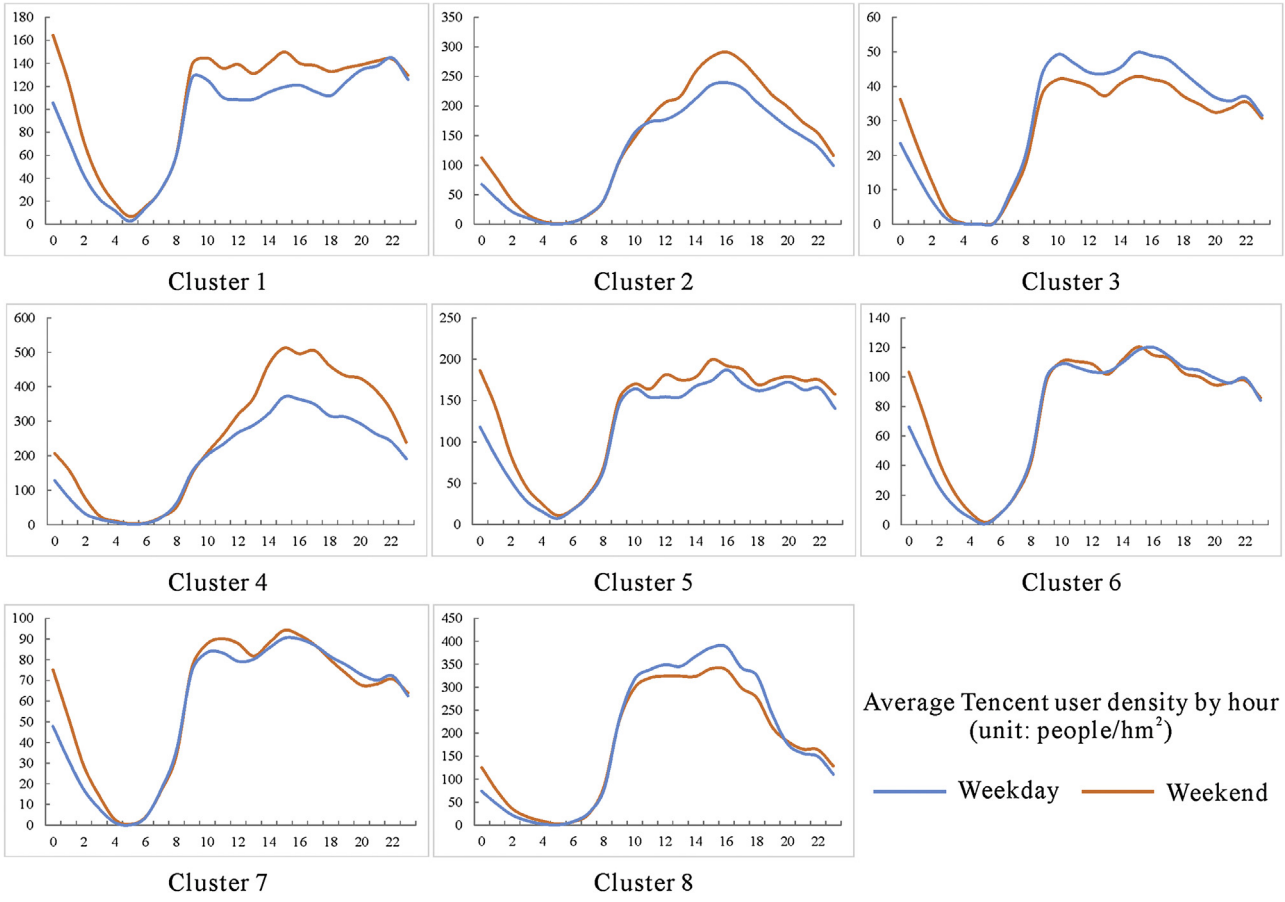


Fig. 6. Clustering results for $k = 8$: the average hourly Tencent user densities for the eight clusters on the weekday and weekend, respectively.

outermost interface that connects the urban villages with the regular residential communities (such as those in Cluster 3), whereas most buildings in Cluster 7 are located in the old city center and near the traditional business area (Fig. 7).

Clusters 2, 4, 5 and 8 comprise mostly of buildings with commercial functions. Despite having different TUD profiles, these clusters have much higher density levels than the other clusters. Buildings in Clusters 2 ('Retail') and 4 ('Retail/Shopping mall (I)') form the Beijing Road business area (Fig. 7). More specifically, Cluster 2 includes buildings with ground level stores along the pedestrian street of Beijing Road, while Cluster 4 comprises of major shopping malls and commercial complex. This distinction is evident in the high enrichment factors values for recreation and restaurant POIs types in Cluster 4 (Table 1). Both Clusters 2 and 4 have TUD peaks at approximately 16:00 (Fig. 6). Cluster 8 includes a relatively small number of 'Retail/Shopping mall (II)' buildings because they are specialty markets where merchants sell small wares (e.g., toys). In addition, it is interesting that most of the 'Wholesale' buildings in Cluster 5 are located within an urban village (i.e., Wangshengtang Village; Fig. 5c). This may due to the proximity of Wangshengtang Village to the Guangzhou Railway Station, a well-known clothing retail and wholesale market. Our site visits and local experiences suggest that many residential buildings in Wangshengtang Village have been transformed into warehouses and even retail stores.

4.2.3. *Community-level urban function diversity*

We further examine the urban function diversity for individual communities. Using both the Shannon index and hotspot analysis (Fig. 9a), we identify areas that are characterized by greater diversity of urban functions. As shown by Fig. 9b, five 'hot spots' are identified: Wangshengtang Village-Guangzhou Railway Station (A), Jingui-Xikeng (B), Beijing Road (C), Zhonghua Square (D) and Yide (E). As already shown, all these hotspots correspond to areas with high TUD values (Fig. 6) and are important business areas (Fig. 7). Therefore, these identified hotspots can be regarded as local centers (or 'central places') in the Yuexiu District. In this regard, our analysis can to some extent address the issue of defining 'vague places' such as 'downtown' or 'city centers' (Lüscher & Weibel, 2013), as we not only are able to determine the locations but also delineate the extent of these places (Fig. 9).

5. Discussion and conclusion

Our analysis has the following practical and empirical implications. First, the spatiotemporal dynamics in TUD data reflect the inherent heterogeneity of urban functional areas. As suggested by Zhong et al. (2014), urban functions are mixed and are not always fully compliant with planned land uses. This is evident in our experiments with TUD data (Figs. 6 and 7). Nevertheless, the general modes of urban functions relating to commercial, working and residential places can be captured in our results based on the spatiotemporal patterns in TUD values. For example, as illustrated by Fig. 6, the TUD profiles related to commercial functions (Clusters 2, 4 and 8) exhibit a relatively simple structure, i.e., continuous increase till approximately 16:00 and a sharp decline afterwards. For the workplace types (Clusters 3, 6 and 7), the corresponding TUD profiles contain both morning and evening peaks. The TUD profiles of Clusters 6 and 7 suggest that more people start to work in the afternoon than in the morning (Fig. 6). Compared with commercial and workplace Clusters, Cluster 1 (Residential) has a more oscillatory

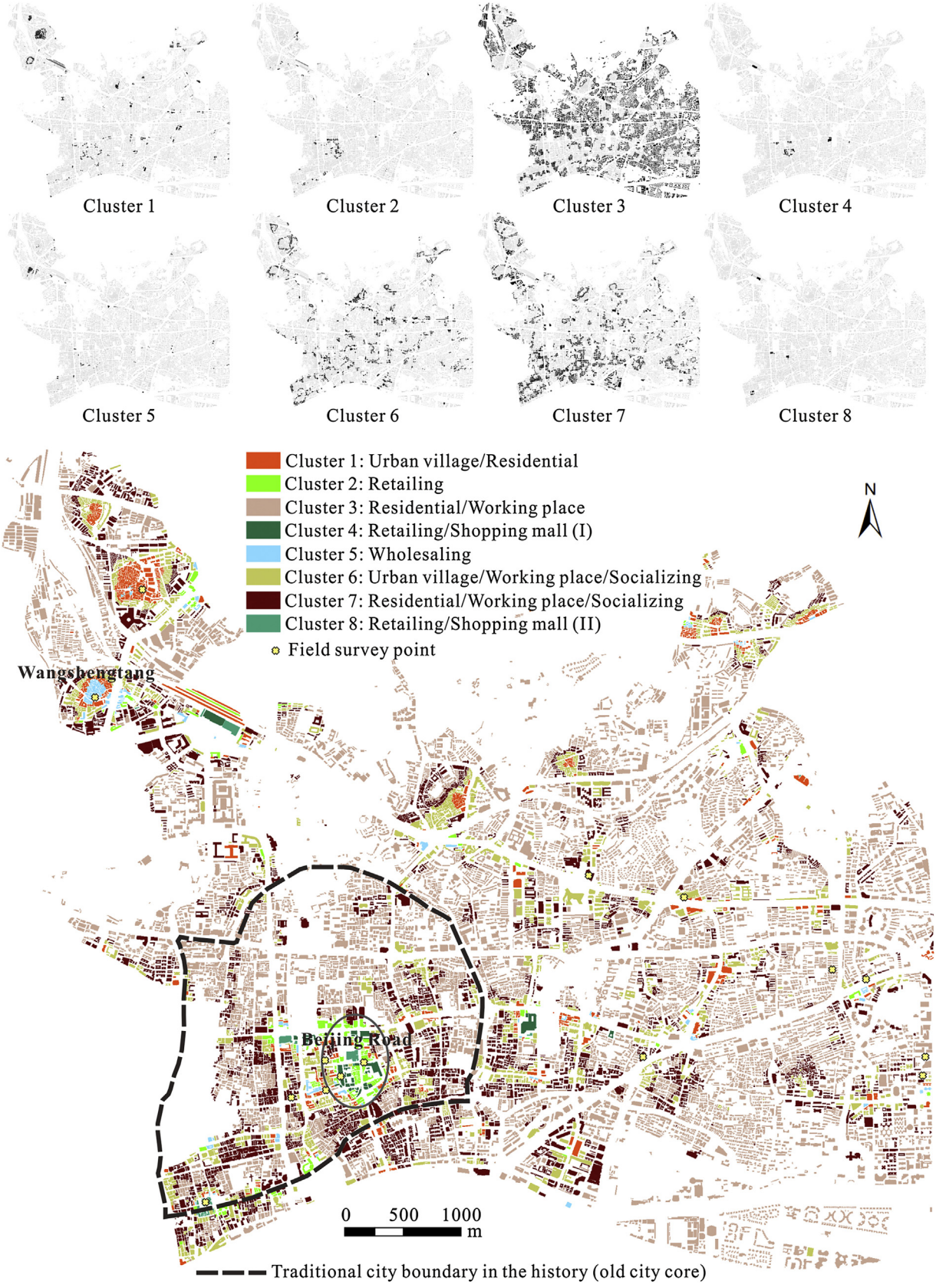


Fig. 7. Clustering results for $k = 8$: the spatial distributions of the building clusters.

TUD profile. Moreover, heterogeneity is found even in the clusters with similar urban functions. For example, Clusters 2, 4, 5 and 8 in Fig. 7, all of which are buildings with commercial functions, have distinct TUD temporal profile patterns (Fig. 6). Technically, this is analogous to the characteristic of 'same land cover types with different spectral signatures' in the land cover classification of remote

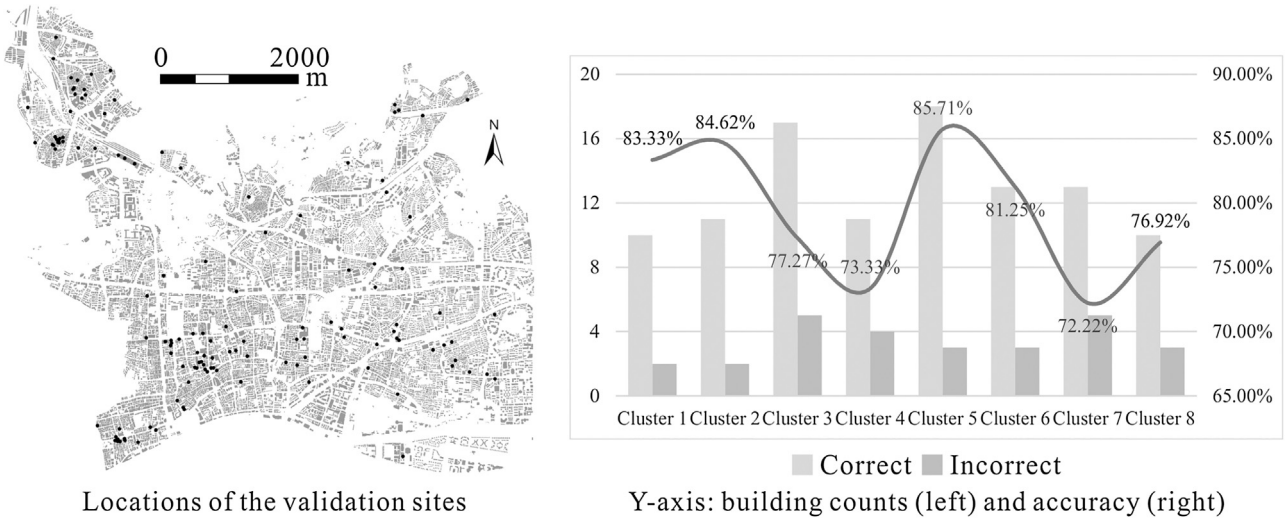


Fig. 8. The locations of the validation sites and the accuracy of the clustering results.

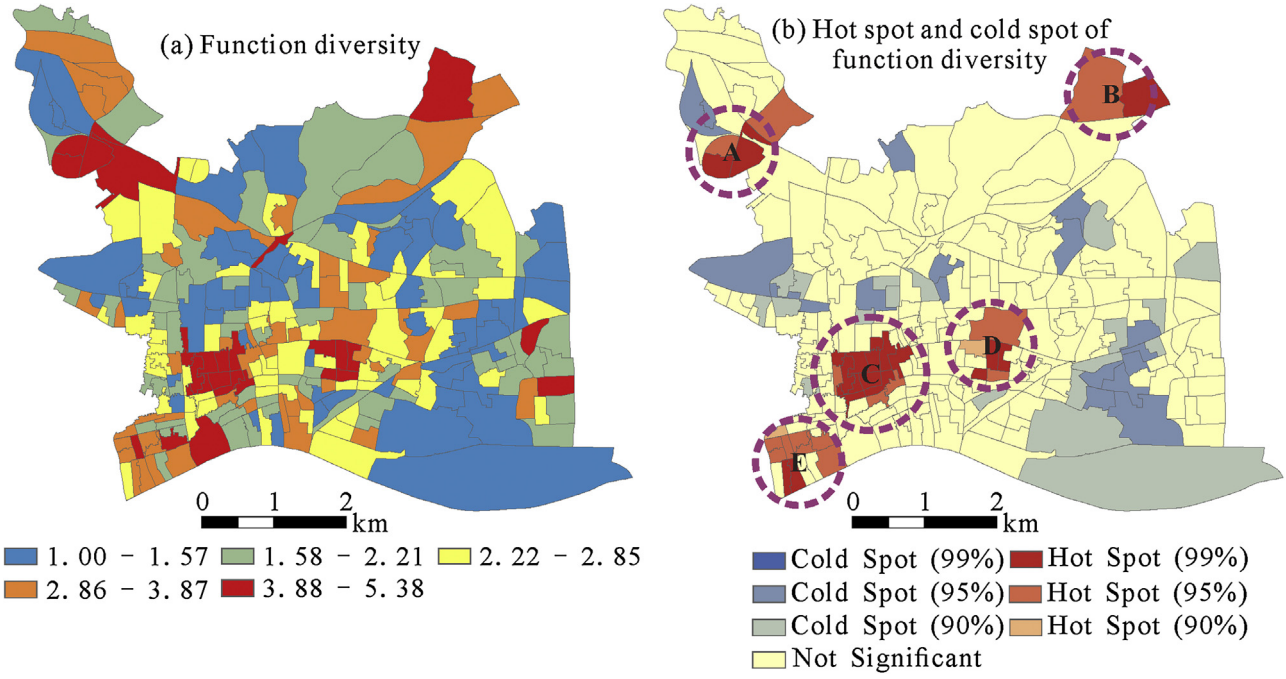


Fig. 9. The community-level urban function diversity in the Yuexiu District.

sensing images (Small, 2006). These results reflect how commercial activities in different forms (e.g., shopping malls and wholesale) can entail distinct population dynamics.

Second, timely and accurate depictions of urban functional areas are critical for the implementation and evaluation of urban policies. Previous studies suggest that actual urban developments can greatly deviate from original urban plans (Tian & Shen, 2011). Therefore, our method is of particular relevance as it would provide additional insights for planners and policy makers to compare planned and observed urban functional areas (Laurian et al., 2010). In addition, the TUD dataset offers an important foundation to elucidate new urban structure types. For instance, the clear concentric spatial structures can be observed in the urban villages in the north of Yuexiu District (Fig. 7). However, to the best of our knowledge, few studies have explored the internal spatial structure of urban villages and existing analyses mostly label urban villages with randomness and spontaneity (Lin et al., 2011). This misconception may be partially due to the lack of reliable data about urban villages at the fine-scale. With social media data, however, it is possible to take further steps in investigating the formation of the spatial structures in urban villages, as well as their latent impacts on residents' activity modes. Findings of such studies can help address social issues such as inequality (Wu, 2009) and urban regeneration (Hin & Xin, 2011) in Chinese cities.

Third, we have explored the usefulness of a new social media dataset (TUD) in delineating building-level urban functional areas. We believe that the TUD data also has promising potential to assist many other research on urban issues. As a fine-scale proxy for the distributions of urban population, the TUD data can support the analysis of living environments and health, such as population exposures to air contaminants (Li, Zhou, Kalo, & Piltner, 2016). The TUD data is also useful for analyzing the accessibility and equity of urban facility and service provisions (Páez, Scott, & Morency, 2012). Moreover, the TUD data can be incorporated with addi-

tional sources of social media data, such as floating car data (Liu, Kang, Gao, Xiao, & Tian, 2012) and mobile phone signals (Jiang et al., 2013). The integration of multi-source social media data can provide more comprehensive information of human mobility and built-environments (Liu et al., 2015). For example, a combination of the TUD data and the mobile phone data may generate a much refined depiction of population dynamics within cities, providing valuable information for urban transportation planning, disaster management, and emergency evacuation (Jiang et al., 2013).

Fourth, the DTW distance based *k*-medoids method is applicable to other urban studies involving time-series or sequential data, such as the classification of multi-temporal images (Salmon et al., 2011), pattern recognition in mobile phone signals (Jiang et al., 2013), and crime trend analysis (Rani & Rajasree, 2014). Despite the increasing availability of these datasets, their effective use remains a challenge. This study suggests that the DTW distance based method performs better than the Euclidean distance based method for delineating urban functional areas. Nevertheless, the performance of the presented method can be further improved to enhance its ability to handle datasets with high dimensions.

In this paper, we present a DTW distance based *k*-medoids method for delineating urban functional areas based on building-level social media data. More specifically, we collect online Tencent user density (TUD) data for Yuexiu District in Guangzhou for a one-week period. Assuming that buildings with similar urban functions will have similar temporal TUD profiles, we use a DTW distance based *k*-medoids method to perform time-series clustering TUD data and aggregate individual buildings with similar social media activities. Our method is illustrated with two clustering experiments (Fig. 4). In the experiment with $k = 2$, buildings are separated into two groups based on the density values (Fig. 5). Buildings with higher density are situated mainly within the traditional city core and urban villages in the northern part of the Yuexiu District. A more detailed depiction of urban functional areas is obtained by running our analysis with a higher *k* value ($k = 8$). The results for $k = 8$ suggest that most buildings have mixed functions. In addition, heterogeneity can be discerned even in the clusters with similar urban functions (Clusters 2, 4, 5 and 8 in Fig. 7). Concentric structures are observed in urban villages that have previously been characterized by randomness and spontaneity. By calculating the function diversity at the community level and performing a hotspot analysis, our study also identifies several potential 'central places'. These findings will help improving our understanding of the city and provide useful information for urban planning. In future studies, we plan to expand time-series clustering to broader geographical contexts. Moreover, we attempt to incorporate more detailed information about individual activities, thus opening up possibilities to explore the relationship between human mobility and the built environments.

Acknowledgements

We thank Prof. Wei-Ning Xiang and the anonymous reviewers for their valuable comments and suggestions. This research was supported by the Key National Natural Science Foundation of China (Grant No. 41531176), the National Natural Science Foundation of China (Grant No. 41601420, 41371376, 41501177), the Guangdong Natural Science Foundation (Grant No. 2015A030310288), and the Fundamental Research Funds for the Central Universities (16lgpy03).

References

Chen, Y., Li, X., Liu, X., Ai, B., & Li, S. (2016). Capturing the varying effects of driving forces over time for the simulation of urban growth by using survival analysis and cellular automata. *Landscape and Urban Planning*, *152*, 59–71.

da Cruz Nassif, L. F., & Hruschka, E. R. (2013). Document clustering for forensic analysis: An approach for improving computer inspection. *IEEE Transactions on Information Forensics and Security*, *8*(1), 46–54.

De Amorim, R. C., & Mirkin, B. (2012). Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering. *Pattern Recognition*, *45*(3), 1061–1075.

Dear, M., & Flusty, S. (1998). Postmodern urbanism. *Annals of the Association of American Geographers*, *88*(1), 50–72.

Fu, T.-C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, *24*(1), 164–181.

Gao, S., Liu, Y., Wang, Y., & Ma, X. (2013). Discovering spatial interaction communities from mobile phone data. *Transactions in GIS*, *17*(3), 463–481.

Hasan, S., Schneider, C. M., Ukkusuri, S. V., & González, M. C. (2013). Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics*, *151*(1–2), 304–318.

Heiden, U., Heldens, W., Roessner, S., Segl, K., Esch, T., & Mueller, A. (2012). Urban structure type characterization using hyperspectral remote sensing and height information. *Landscape and Urban Planning*, *105*(4), 361–375.

Hin, L. L., & Xin, L. (2011). Redevelopment of urban villages in Shenzhen, China-an analysis of power relations and urban coalitions. *Habitat International*, *35*(3), 426–434.

Jiang, S., Fiore, G. A., Yang, Y., Ferreira Jr, J., Frazzoli, E., & González, M. C. (2013). A review of urban computing for mobile phone traces: Current methods, challenges and opportunities. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing* (pp. 1–9).

Jiang, S., Alves, A., Rodrigues, F., Ferreira, J., & Pereira, F. C. (2015). Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems*, *53*, 36–46.

Lüscher, P., & Weibel, R. (2013). Exploiting empirical knowledge for automatic delineation of city centres from large-scale topographic databases. *Computers, Environment and Urban Systems*, *37*, 18–34.

Laurian, L., Crawford, J., Day, M., Kouwenhoven, P., Mason, G., Ericksen, N., et al. (2010). Evaluating the outcomes of plans: Theory, practice, and methodology. *Environment and Planning B: Planning and Design*, *37*(4), 740–757.

Li, L., Zhou, X., Kalo, M., & Piltner, R. (2016). Spatiotemporal interpolation methods for the application of estimating population exposure to fine particulate matter in the contiguous us and a real-time web application. *International Journal of Environmental Research and Public Health*, *13*(8), 749–768.

Lin, Y., De Meulder, B., & Wang, S. (2011). Understanding the 'Village in the City'in Guangzhou economic integration and development issue and their implications for the urban migrant. *Urban Studies*, *48*(16), 3583–3598.

Liu, Y., Kang, C., Gao, S., Xiao, Y., & Tian, Y. (2012). Understanding intra-urban trip patterns from taxi trajectory data. *Journal of Geographical Systems*, *14*(4), 463–483.

Liu, X. P., Ma, L., Li, X., Ai, B., Li, S. Y., & He, Z. J. (2014). Simulating urban growth by integrating landscape expansion index (LEI) and cellular automata. *International Journal of Geographical Information Science*, *28*(1), 148–163.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., et al. (2015). Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers*, *105*(3), 512–530.

Long, Y., & Shen, Z. (2015). Discovering functional zones using bus smart card data and points of interest in Beijing. In *Geospatial analysis to support urban planning in Beijing*. pp. 193–217. Springer.

Mirkin, B. (2012). *Clustering: A data recovery approach*. CRC Press.

Páez, A., Scott, D. M., & Morency, C. (2012). Measuring accessibility: Positive and normative implementations of various accessibility indicators. *Journal of Transport Geography*, *25*, 141–153.

Park, H.-S., & Jun, C.-H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, *36*(2), 3336–3341.

Pei, T., Sobolevsky, S., Ratti, C., Shaw, S.-L., Li, T., & Zhou, C. (2014). A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*, *28*(9), 1988–2007.

Petitjean, F., Ketterlin, A., & Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, *44*(3), 678–693.

Rösler, R., & Liebig, T. (2013). Using data from location based social networks for urban activity clustering. In *Geographic information science at the heart of Europe*. pp. 55–72. Springer.

Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., et al. (2013). 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), 1–31.

Rani, A., & Rajasree, S. (2014). Crime trend analysis and prediction using mahanolobis distance and dynamic time warping technique. *International Journal of Computer Science & Information Technologies*, *5*(3), 4131–4135.

Rani, S., & Sikka, G. (2012). Recent techniques of clustering of time series data: A survey. *International Journal of Computer Applications*, *52*(15), 1–9.

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

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

Shi, L., Chi, G., Liu, X., & Liu, Y. (2015). Human mobility patterns in different communities: A mobile phone data-based social network approach. *Annals of GIS*, *21*(1), 15–26.
Small, C. (2006). Comparative analysis of urban reflectance and surface temperature. *Remote Sensing of Environment*, *104*(2), 168–189.
Tian, L., & Shen, T. (2011). Evaluation of plan implementation in the transitional China: A case of Guangzhou city master plan. *Cities*, *28*(1), 11–27.
Van de Voorde, T., Jacquet, W., & Canters, F. (2011). Mapping form and function in urban areas: An approach based on urban metrics and continuous impervious surface data. *Landscape and Urban Planning*, *102*(3), 143–155.
Verburg, P., de Nijs, T., Ritsema van Eck, J., Visser, H., & de Jong, K. (2004). A method to analyse neighbourhood characteristics of land use patterns. *Computers, Environment and Urban Systems*, *28*(6), 667–690.
Wu, F. (2009). Land development, inequality and urban villages in China. *International Journal of Urban and Regional Research*, *33*(4), 885–889.
Yuan, J., Zheng, Y., & Xie, X. (2012). Discovering regions of different functions in a city using human mobility and POIs. In *The 18th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 186–194).
Yuan, Y., Raubal, M., & Liu, Y. (2012). Correlating mobile phone usage and travel behavior—A case study of Harbin, China. *Computers, Environment and Urban Systems*, *36*(2), 118–130.
Zhang, X., & Du, S. (2015). A linear dirichlet mixture model for decomposing scenes: Application to analyzing urban functional zonings. *Remote Sensing of Environment*, *169*, 37–49.
Zhi, Y., Li, H., Wang, D., Deng, M., Wang, S., Gao, J., et al. (2016). Latent spatio-temporal activity structures: a new approach to inferring intra-urban functional regions via social media check-in data. *Geo-spatial Information Science*, 1–12.
Zhong, C., Huang, X., Arisona, S. M., Schmitt, G., & Batty, M. (2014). Inferring building functions from a probabilistic model using public transportation data. *Computers, Environment and Urban Systems*, *48*, 124–137.