Understanding China's urban functional patterns at the county scale by using time-series social media data

Article *in* Journal of Spatial Information Science · September 2022

DOI: 10.1080/14498596.2022.2125095

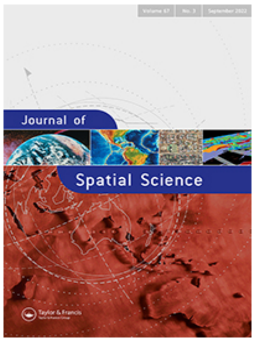
7 authors, including:

Qingfeng Guan
China University of Geosciences
183 PUBLICATIONS **3,854** CITATIONS
SEE PROFILE

Ruifan Wang
The Hong Kong Polytechnic University (PolyU)
5 PUBLICATIONS **94** CITATIONS
SEE PROFILE

Yao Yao
China University of Geosciences
129 PUBLICATIONS **5,241** CITATIONS
SEE PROFILE

Chen Qian
William & Mary
5 PUBLICATIONS **73** CITATIONS
SEE PROFILE



Journal of Spatial Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tjss20

Understanding China’s urban functional patterns at the county scale by using time-series social media data

Qingfeng Guan, Jianfeng Zhou, Ruifan Wang, Yao Yao, Chen Qian, Yaqian Zhai & Shuliang Ren

JOURNAL OF SPATIAL SCIENCE
https://doi.org/10.1080/14498596.2022.2125095

Understanding China's urban functional patterns at the county scale by using time-series social media data

Qingfeng Guan[a], Jianfeng Zhou[a*], Ruifan Wang[a], Yao Yao[a,b,c], Chen Qian[d], Yaqian Zhai[a] and Shuliang Ren[a]

[a]School of Geography and Information Engineering, China University of Geosciences, 430078, Wuhan, Hubei Province, China; [b]Department of Data Technology and Product, Alibaba Group, 311121, Hangzhou, Zhejiang, China; [c]Center for Spatial Information Science, the University of Tokyo, 277-8568, Kashiwa-shi, Chiba, Japan; [d]Department of Engineering Systems and Environment, University of Virginia, 22904, Charlottesville, VA, USA

ABSTRACT

Understanding urban functions help with government planning and resource allocation to promote economic development. Few studies have reflected the influence of population movement and migration on urban functions overtime on a large scale. This study adopted the time-series social media data and points of interest to analyse urban functions' distribution at the county scale. By employing dynamic time warping distance and K-Medoids to cluster cities, the result of China's hourly population distribution over different periods achieves a high accuracy (Pearson's R = 0.821, R2 = 0.668). From the clustering results, time-series population data effectively reflect cities' socioeconomic characteristics and identify the spatial distribution of China's urban functional patterns. We select four representative urban agglomerations to deeply analyse their urban function patterns. Furthermore, urban functions can also influence changes in time-series populations. This study explores the correlations between time-series population mobility and urban functions, which could help analyse urban functions and urban socioeconomic conditions.

KEYWORDS

Urban function; social media data; population mobility; dynamic time warping; K-Medoids

1. Introduction

Urban function, which has been a heavily debated topic in urban geography in recent years, focuses on measuring the roles of different regions in cities from socioeconomic systems (Zhou and Bradshaw 1988). Existing studies build an urban function analysis framework by analysing the proportions of various industries in the economy and employment (Zeng and Shen 2015), and explore the industrial functions, central functions, and gateway functions of the city (Wang 2007, Yu *et al.* 2019). The research on urban functions take administrative divisions as the basic research unit, treat all research units as urban function research goals (Lu *et al.* 2011),

CONTACT Yao Yao ✉ yaoy@cug.edu.cn; yaoy@csis.u-tokyo.ac.jp

*Current affiliation for Jianfeng Zhou is Foshan Surveying Mapping and Geoinformation Research Institute, 528000, Foshan, Guangdong, China

Supplemental data for this article can be accessed online at https://doi.org/10.1080/14498596.2022.2125095

and analyse urban structure and urban evolution (Moody *et al.* 2019). Analysis of urban structures, spatial distribution, and other urban characteristics helps regional planning, resource allocation, and government decision-making processes (Tu *et al.* 2017, Yao *et al.* 2018).

Previous studies have mainly focused on the delineation of urban function based on geospatial data, such as points of interest (POIs) (Yuan *et al.* 2014), smart-card data (Long and Shen 2015), social media check-in record (Chen *et al.* 2017) and satellite image (Gong *et al.* 2020). These intraurban areas' functional division studies revealed the close relationship between urban function and population movement (Zhi *et al.* 2016).

Time-series social media data can effectively reflect a broad range of urban population movements (Wang *et al.* 2019a). The existing studies applied workday and weekend time-series social media data to delineate urban functions (Chen *et al.* 2017). During the Chinese Spring Festival, the massive population movement was exploited to assess Chinese urbanisation and the urban population (Xu *et al.* 2017, Wei *et al.* 2018, Pan and Lai 2019). Without sufficient reliable data and models, existing studies did not use the Chinese Spring Festival population movement to delineate urban function across China (Wang *et al.* 2019b). In this study, holiday (Chinese New Year's Eve) time-series social media data reflect the Chinese Spring Festival population movement.

Time-series data from mobile applications can reflect population movements and urban socioeconomic characteristics on a large scale. The existing large-scale urban function studies rarely considered the effects of population movements and migration on urban functions over time. This study explores and analyzes population mobility differences between cities of different functional types within the study area. We first utilise Tencent user data and census data to map the real-time population distribution over time and calculate the urban population's changes over time in the region. Then, we calculate term frequency-inverse document frequency (TF-IDF) based on POIs to describe the relative abundance of different urban functions, revealing the relationships between urban functions and changes in population mobility.

2. Related work

2.1. Intracity functions delineated

As location-based services (LBS) overgrows, multifaceted geospatial data, such as POIs, mobile communications and check-in data, and public transportation trajectories from global positioning system (GPS), can be obtained from the public services or Internet (Tu *et al.* 2017, Yao *et al.* 2018). These forms of geospatial data have been applied to many areas of urban studies, such as urban land use (Liu *et al.* 2012, Pei *et al.* 2014), urban community exploration (Papadopoulos *et al.* 2012, Yue *et al.* 2019), urban population investigations (Kang *et al.* 2012, Patel *et al.* 2017), and urban structure analysis (Sevtsuk and Ratti 2010). Collecting data through LBS to conduct research on social economy and population activities is an important application of geospatial data (Huang *et al.* 2018).

Scholars have used these datasets to explore urban functions. Yuan *et al.* (2014) proposed using a discovered regions of different functions (DRoF) framework in a city, GPS trajectory datasets and POI data are used to divide Beijing into functional zones, and they revealed the relationship between different functional areas in city and population mobility. Long and Shen (2015) characterised the urban function spatial pattern at the traffic analysis zones (TAZs) in Beijing based on smart-card data and POIs. Zhi *et al*. (2016) detected the functional regions by using a low-rank approximation (LRA)-based model and data from approximately 15 million social media records over a year-long period in Shanghai and explored the relationship between the spatiotemporal activity patterns and these functional areas. Chen *et al*. (2017) employed a time-series social media dataset and POI data to cluster urban functions within the Yuexiu District of Guangzhou, which indicated the changes in the number of social media users at different times are closely related to regional functions.

These studies addressed the functional division of intra-urban areas and revealed a close relationship between urban function and population mobility at a fine scale, while time-series social media data were confirmed to reflect urban population movement effectively. During the Chinese Spring Festival, population movement is the most representative event of Chinese migration and demographic patterns (Zhao *et al*. 2017), and it is a large-scale population migration across China (Li *et al*. 2016, Wang *et al*. 2019a). However, none of the previous studies applied Chinese Spring Festival population mobility to delineate urban functions on a large scale.

2.2. Large-scale urban functions delineated

Some studies have delineated urban functions on a large scale. Based on POIs and blocks determined by road networks, Song *et al*. (2018) characterised Chinese urban forms and reclassified cities into four hierarchies. They examined how development stages of the city impact urban forms. Gong *et al*. (2020) mapped and accurately assessed the Essential Urban Land Use Categories (EULUC)-China using 10-metre satellite images, OpenStreetMap, night-time lights and POIs. The above studies effectively identified large-scale urban land use but did not conduct the urban scale's overall distribution of the national urban function patterns.

The emergence of mobile application data provides new opportunities for a deeper understanding of urban socioeconomic conditions (Goodchild 2008, Liu *et al*. 2015). Mobile application data better reflect population movement and migration within the city than data from GPS, traffic tracking, and transit cards (Shaw *et al*. 2016). Wang *et al*. (2019a) used mobile application datasets collected before and during the Chinese Spring Festival. By capturing the population mobility and analysing the output and inflow of labour force between cities, they obtained the urbanisation levels of different regions, indicating that the characteristics of urban population change over time, reflecting the city's economic, political and other aspects of the region. These studies showed that mobile application data could reveal large-scale population movement during the Chinese Spring Festival.

3. Study area and data

China is selected as the study area in this study. To delineate China's county-scale urban functions, data for the cartographic boundaries of 41,263 county-scale administrative districts were obtained from the official government website (http://www.stats.gov.cn). The Sixth National Population Census of China is the latest national census of China, reporting the population number and primary social demographic values (e.g. gender, age) at the county level. We extrapolate the 2016 census data for each county according to China's natural population growth from 2010 to 2016.

3.1. Gaode POI data

POI data are used to analyse urban functions, e.g. retailing, residential, urban village, or wholesaling. Different types of POIs spatial distributions and interactions reflect urban functions and provide auxiliary information for interpreting the results (Chen *et al*. 2017, Gong *et al*. 2020). Gaode POI data were collected via Gaode mapping (https://www.amap.com) application programming interfaces (APIs). The Gaode POI data supply information on the name, address, contact, category/tag of individual places. More than 200 various POI types are included in the raw data, which we reclassified into the following ten categories: transport (TR), leisure tourism (LT), communal facility (CF), enterprise (ET), residential community (RC), train station and airport (TS), government agency (GA), technical facility (TF), financial service (FS) and recreation (RE). These categories serve as secondary data when interpreting county-scale functions.

3.2. The real-time Tencent user density (RTUD) dataset

The real-time Tencent user density (RTUD) dataset from 25 January 2016, to 14 February 2016, with a temporal resolution of 1 hour was obtained from application programming interfaces (APIs) furnished by Tencent EasyGo Software (http://easygo.qq.com), the most widely used social application service provider in China. In China's first-tier cities such as Beijing, Shanghai, and Guangzhou, Tencent users account for more than 93% of the total population. The penetration rate of mobile phones in rural areas of China, especially in remote and backward areas, has reached 91%, and 95.39% of rural mobile phone users are using social software such as WeChat and QQ (Long and Liu 2017, McDonald 2016, Wang *et al*. 2019b, Yao *et al*. 2017). EasyGo's positioning technology based on human trajectory features (Wang *et al*. 2011) realises accurate positioning of Tencent users on all platforms and provides public query services for location congestion.

In this study, the RTUD dataset has a temporal resolution of 1 hour, indicating that it can capture the absolute number of users and changes in the Tencent user distribution per hour. We calculate the respective average RTUD data from each workday, weekend, and Chinese New Year's Eve. Hence, the RTUD dataset is compressed to 72 dimensions. Figure 1 shows RTUD data at 10:00 on weekdays. During this time period, RTUD data in Western China is relatively sparse. However, this study collected RTUD data for different time periods, and the amount of RTUD data in western China increased significantly during the New Year's Eve. Although Tencent's related applications are extremely popular in China (Yao *et al*. 2017), the use of RTUD data still has certain population biases and link biases (Olteanu *et al*. 2019). Tencent's active users reached 9 billion in 2016, making RTUD data collected through EasyGo a reliable data source reflecting the overall population distribution (Zhang *et al*. 2019).

RTUD data records the location of smartphone users when using Tencent apps. It is undeniable that it is more difficult to collect RTUD data in sparsely populated areas, but

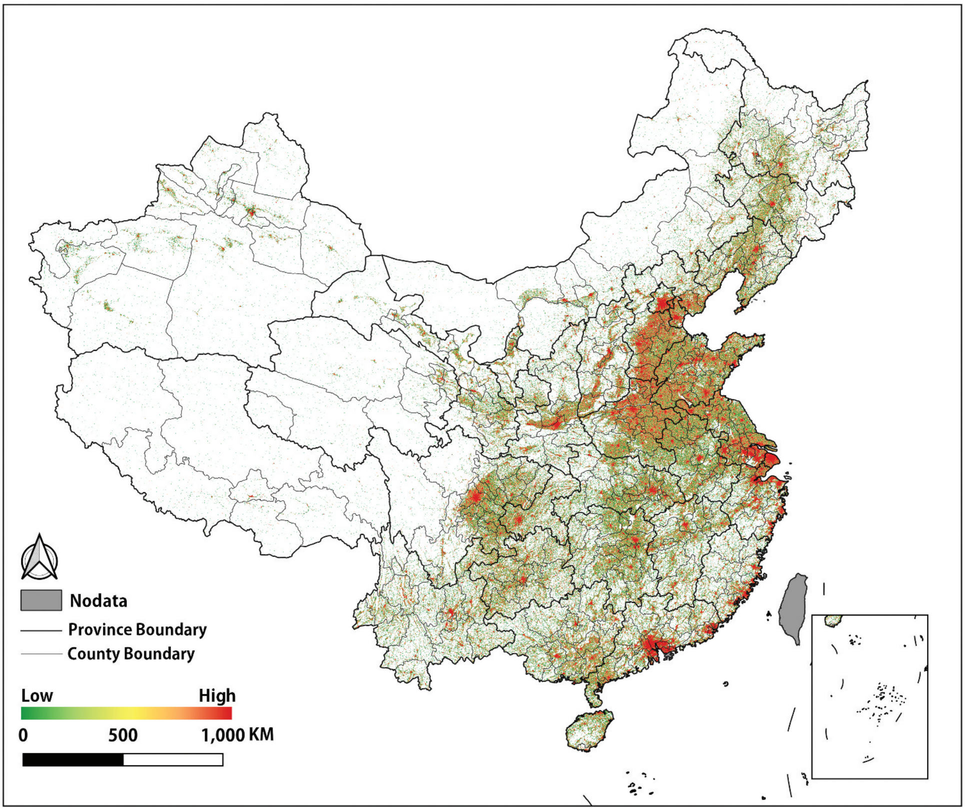


Figure 1. RTUD data of China (unit: person, temporal resolution: 1 hour) on 18 January 2016 (workday), at 10 a.m.

population change is a dynamic process (Wang *et al*. 2019a). The collection of RTUD data for long periods and special holidays can prevent RTUD data from solving the problem of data collection in sparsely populated areas to a certain extent.

4. Methodology

This study's workflow is shown in Figure 2. (1) This study employed census data and RTUD to map the hourly population distribution of China during workdays, weekends, and Chinese New Year's Eve. The correlation between the RTUD and the actual population distribution was estimated. (2) A time series for the population distribution of workdays, weekends and Chinese New Year's Eve was built. Based on dynamic time warping (DTW) distance and the K-Medoids clustering method, the spatiotemporal information from a time series of population distribution was extracted. (3) The POI dataset provides auxiliary information for interpreting the clustering results. Utilising the TF-IDF for POI, this study delineated urban function within each cluster. (4) We systematically explored the association between urban function and temporal population distribution.

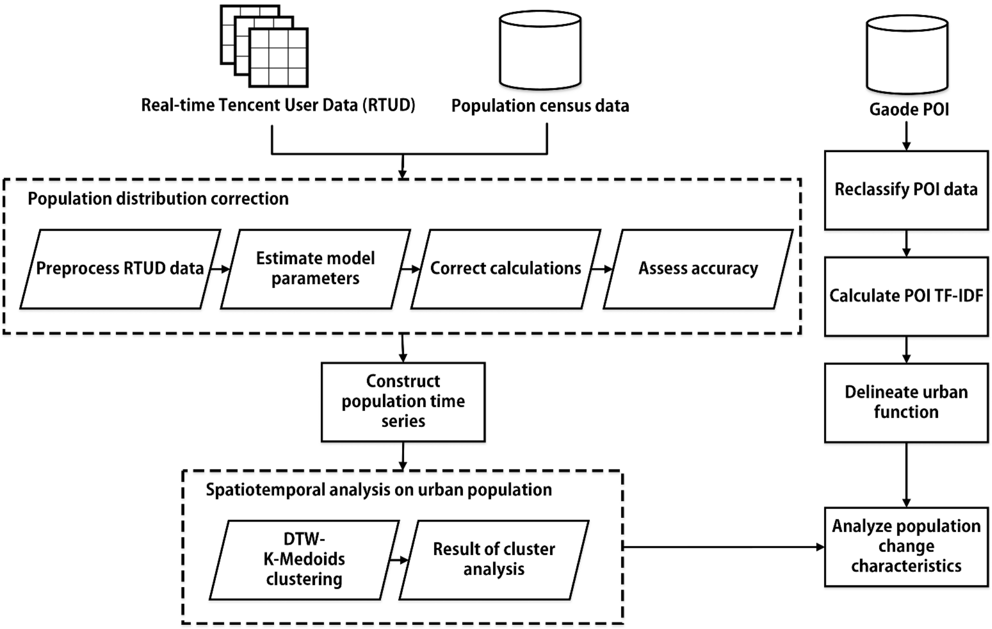

Real-time Tencent User Data (RTUD)
Population census data
Gaode POI
Population distribution correction
Preprocess RTUD data
Estimate model parameters
Correct calculations
Assess accuracy
Reclassify POI data
Calculate POI TF-IDF
Construct population time series
Delineate urban function
Spatiotemporal analysis on urban population
DTW-K-Medoids clustering
Result of cluster analysis
Analyze population change characteristics

Figure 2. The workflow of delineating urban function based on RTUD data.

4.1. Mapping population distributions

The original RTUD dataset has several defects, including noise, coordinate offsets, data redundancy, and loss of spatiotemporal information. After preprocessing, the dataset is compressed into a 72-dimensional hourly time series. The mean density value at the county scale is obtained. Census population data for workdays, weekends, and Chinese New Year's Eve were employed to estimate the correlation between the RTUD and the actual population distribution. The estimation of the population distribution for an administrative unit based on the RTUD data is as follows (Deville *et al.* 2014):

$$P_i = \alpha\left(X_{i,j}\right)^{\beta} \tag{1}$$

where $X_{i,j}$ is administrative unit i in the RTUD dataset at time j, and P_i represents the census population data of administrative unit i. This equation can be transformed into:

$$\log(p_i) = \log\alpha + \beta\log\left(X_{i,j}\right) \tag{2}$$

Population-weighted least square estimate in standard linear regression model can be applied to calculate the two parameters: α and β.

The population distribution was adapted to match the total estimated population $\sum_j p_{i,j}$ with the general population from census $\sum_i P_i$.

$$p_{i,j} = \frac{\sum_i P_i}{\sum_j \hat{p}_{i,j}}\hat{p}_{i,j} \tag{3}$$

4.2. Clustering cities with DTW distance based on K-Medoids

This study characterises population mobility patterns based on mobile application time-series data. The DTW distance measures two given time series similarity on optimal alignment (i.e. the warping path). Compared with Euclidean distance, DTW distance significantly affects downstream analysis based on the specific features of the data and noise in the time series (Rakthanmanon *et al.* 2013). Therefore, DTW distance extracts mobile application time-series data features more effectively.

Given two time series, a time-series data Q($Q = q_1, q_2, q_3, \ldots, q_n$) of length n and a time-series data P($P = p_1, p_2, p_3, \ldots, p_m$) of length m, to match two time-series data using warping path, we construct a matrix with n rows and m columns as follows:

$$\begin{bmatrix} d(q_1, p_1) & d(q_1, p_2) & \cdots & d(q_1, p_m) \\ d(q_2, p_1) & d(q_2, p_2) & \cdots & d(q_2, p_m) \\ \vdots & \vdots & & \vdots \\ d(q_n, p_1) & d(q_n, p_2) & \cdots & d(q_n, p_m) \end{bmatrix}$$

Where the $\left(i^{th}, j^{th}\right)$ element of the matrix corresponds to the squared distance,

$$d\left(q_i, p_j\right) = \sqrt{\left(q_i - p_j\right)^2} \tag{4}$$

Which is the alignment between points q_i and p_j. The core of DTW is to find a warping path which can be calculated through the matrix that minimises the total cumulative distance between Q and P. The warping path (W), characterising a mapping between Q and P, is a set of consecutive matrix elements. $w_k = d\left(q_i, p_j\right)$ is defined as the k^{th} element of W. Therefore:

$$W = w_1, w_2, w_3, \cdots, w_K \max(m, n) \leq K + n - 1 \tag{5}$$

By definition, the optimal path is that which minimises the warping cast (Berndt and Clifford 1994):

$$DTW(Q, P) = Min\left\{\sum\nolimits_{k=1}^{K} w_k\right. \tag{6}$$

The optimal path starts with $w_1 = d(q_1, p_1)$ and ends in $w_K = d(q_n, p_m)$. When $w_k = d\left(q_i, p_j\right)(k > 1)$, w_{k-1} must be $d\left(q_{i-1}, p_j\right)$, $d\left(q_i, p_{j-1}\right)$ or $d\left(q_{i-1}, p_{j-1}\right)$.

Different cities can have similar functions. K-Medoids are used to cluster cities with the same function. The K-Medoids is a k-means-like algorithm for performing clustering analysis (Park *et al.* 2006). The approach of updating the central location of a particular cluster varies greatly between these two methods. K-means approach treats the centre as the mean position of members. However, K-Medoids chooses the cluster median as a centre that is the minimum of the sum of other entities' distances in the same cluster. We hope that the cluster centre is one of the actual data rather than a calculated average centre, so K-Medoids is selected as the clustering algorithm. This method for selecting the cluster centre is more robust to noise and outliers in the dataset (Park and Jun 2009).

The implementation process of K-Medoids methods is the four steps: First, we confirm the clusters' quantity. Second, we initialise cluster centres and select initial centres of the k clusters. Third, each sample is allocated to the closest cluster centre. Fourth, we update the centre of each cluster to the nearest cluster centre. Finally, we repeat steps 3 and 4 until none of the clusters changes its memberships or iterations reach the preset values.

The initialisation of clusters affects the performance of K-Medoids. The conventional way of selecting initial cluster centres at random is problematic, for it unable to effectively generate the typical initial cluster centres of large datasets. (Amorim and Mirkin 2012). The modified iterated anomalous pattern (AP) method is a more effective initialisation method than random initialisation (Mirkin 2012). This study employs the AP method to initialise the cluster centre.

To determine the appropriate number of cluster k, the silhouette index was used to evaluate the clustering result. The silhouette index involves the similarity of an entity to the same cluster and separation from other groups (Chiang and Mirkin 2010). The silhouette index $S(i)$ for an entity is defined as (Rousseeuw 1987):

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{7}$$

where $a(i)$ stands for the mean distance between i and all of the other entities in the same cluster, whereas $b(i)$ stands for the minimum of the mean distances between i and all of the entities in any other cluster. The silhouette index values lie within the range from −1 to 1. The average silhouette index of the clustering result is close to 1, indicating that the work is clustered.

4.3. Computing the TF-IDF of each POI category

TF-IDF is typically used to determine the words in a corpus of documents that might be more appropriate to use in a query (Wu *et al*. 2008). TF-IDF can also effectively characterise different function types in a county using POIs (Yuan *et al*. 2012). We adopt TF-IDF to describe the relative abundance of an urban function using POIs:

$$TF_{ij} = \frac{n_{i,j}}{N_i} \tag{8}$$

$$IDF_i = \log \frac{D}{\{j : POI_i \in \mathrm{d}_j\}} \tag{9}$$

$$TF - IDF = TF_{ij} \cdot IDF_i \tag{10}$$

where TF_{ij} denotes the term frequency of POIs in the *ith* category for county *j*; IDF_j denotes the inverse document frequency of POIs in the *ith* category. $n_{i,j}$ is the number of POIs in the *ith* category for the county *j*, N_i is the number of all POIs in county *j*; *D* is the total number of counties; and $\{\mathrm{j} : \mathrm{POI}_i \in \mathrm{d}_j\}$ is the number of counties that contain the *ith* category of POIs. A higher value of $TF - IDF$ indicates a larger number of the *ith* category POI at the location of county *j*.

5. Results

5.1. Mapping population distribution via RTUD

Previous studies have shown that weekday and weekend population changes are near related to urban functions (Chen *et al*. 2017). Population movement around Chinese New Year's Eve reflects social and economic dynamics in China (Wang *et al*. 2019b). This study uses RTUD data and census data to map hour-by-hour population distributions for workdays, weekends, and Chinese New Year's Eve. The results illustrate that RTUD data reflect population distributions with high accuracy. Table 1 shows the accuracy evaluation of the mapping of census data using RTUD. The weekend results perform best, with a mean Pearson's R of 0.821 and a mean R^2 of 0.668. The overall results have relatively high accuracy, with a mean Pearson's R greater than 0.7 and a mean R^2 greater than 0.5.

The low accuracy of the population mapping results on New Year's Eve is due to the concentration of migrant workers returning home during the Spring Festival and the large-scale interregional population movement around New Year's Eve. It causes discrepancies between the RTUD data collected and the census data on New Year's Eve.

Figure 3 shows the results of the population distribution mapping at 10:00 a.m. on 18 January 2016. The population distribution trend in China is generally higher in the east and lower in the west, with populations clustering around major cities (Zhang *et al*. 2015). Population clustering is evident in China's most economically developed cities (e.g.

Table 1. Accuracy assessment of population mapping on weekdays, weekends and Chinese New Year's Eve

Day	Pearson R	R^2
Weekdays	0.814	0.661
Weekends	0.821	0.668
Holiday (Chinese New Year's Eve)	0.772	0.578

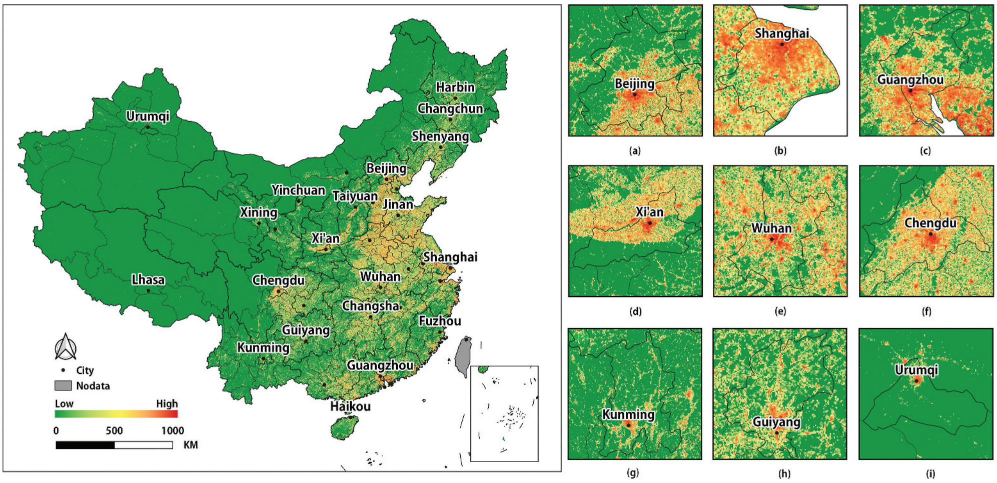


Figure 3. Spatial distribution of China's population mapping results (10:00 a.m. on a weekday): (a) Beijing, (b) Shanghai, (c) Guangzhou, (d) Xi'an, (e) Wuhan, (f) Chengdu, (g) Kunming, (h) Guiyang, and (i) Urumqi.

Beijing, Shanghai, Guangzhou) with subcenters in the surrounding areas. Xi’an, Wuhan, and Chengdu also show the phenomenon of population clustering with subcenters in the surrounding areas, but the scale of population clustering is smaller and less extensive. The populations of Kunming, Guiyang, and Urumqi are mainly converged around the urban centres, and the surrounding areas have not experienced significant population clustering. In general, the population distribution mapping using RTUD data achieves high accuracy. The mapping results accord with China’s population distribution features and effectively reflect the socioeconomic characteristics of Chinese cities.

5.2. *Cluster results based on time series of population distribution*

This study conducted a clustering test with the cluster number between 2 and 10 and used the silhouette index to select the appropriate clustering results. Figure S1 shows the silhouette index corresponding to the different number of clusters k. The first inflection point in the silhouette index variation occurs at k = 3 when the profile factor is more significant than 0.7. With the cluster number increase, the silhouette index gradually decreases, and a second inflection point occurs at k = 6, where the silhouette index is greater than 0.5. The silhouette index decreases to 0.3 when the cluster number increases to 10. In this study, the number of clusters corresponding to the two inflection points with a silhouette index greater than 0.5 was chosen as the number of clusters (Chen *et al.* 2018). Figures 4 and S2 show the clustering results for numbers of clusters k = 3 and k = 6, respectively.

Cluster 3 in Figure 4 is concentrated around the core cities, while there is a spatial difference between Cluster 1 and Cluster 2. As shown, Cluster 1 is distributed in western and northern China areas generally considered to be relatively backward. Economic development areas in Cluster 2 are located primarily near provincial capitals and in Cluster 3 are usually the centre of economic and social activities.

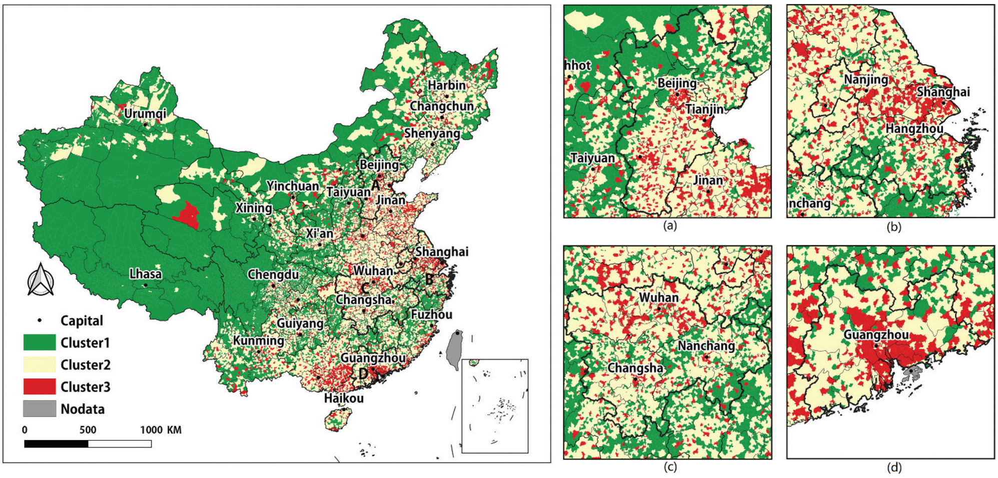


Figure 4. Clustering results for k = 3: (a) Beijing-Tianjin-Hebei, (b) the Yangtze River Delta, (c) the Middle Reaches of the Yangtze River, and (d) the Pearl River Delta.

Cluster 3 occurs within the Beijing-Tianjin-Hebei region (Figure 4a). Cluster 2 mainly lies in the Beijing-Tianjin-Hebei region's southeastern area, while Cluster 1 occupies the northwestern of the Beijing-Tianjin-Hebei region. Cluster 3 also occur within the Yangtze River Delta (Figure 4b) around Shanghai, Nanjing and Hangzhou. Cluster 2 occupies the Yangtze River Delta's northern area, with a small Cluster 1 region in the Yangtze River Delta's southern area. Within the middle reaches of the Yangtze River (Figure 4c), the proportion of Cluster 2 is significantly larger than in other regions, and there are also several regions of Cluster 1. The number of Cluster 3 regions around Wuhan is considerably higher than around Changsha and Nanchang. In the Pearl River Delta (Figure 4d), there is a stratification of the various types of regions, with a small number of Cluster 2 areas in the core city of Guangzhou, Cluster 3 regions in the middle, and Cluster 2 areas and a small number of Cluster 1 areas in the outer reaches.

Figure 5 shows the change curves of the time-series population for each category of clustering results where k = 3. Table 2 shows the calculation results based on the TF-IDF of POIs in different types grouped by the cluster. On workdays, weekends, and Chinese New Year's Eve, time-series population curves for the same category of regions show essentially the same characteristics, with significant differences in the mean population size for each category of areas where Cluster 3 (81,182.07) > Cluster 2 (36,174.06) > Cluster 1 (13,764.08). Population size is closely related to the region's level of socioeconomic

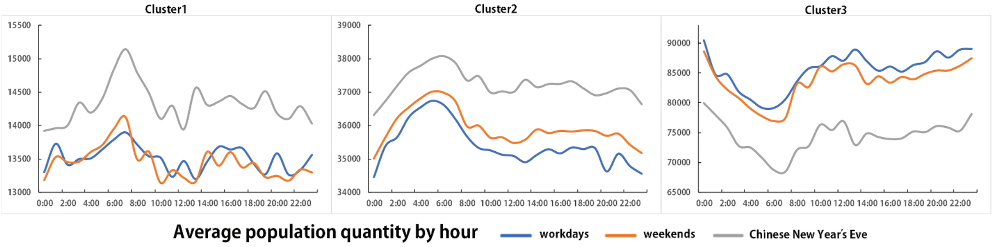


Figure 5. Average population by the hour for the three clusters on workdays, weekends, and Chinese New Year's Eve.

Table 2. TF-IDF of POI in different categories grouped by cluster (k=3). The values that are the top three are shown in boldface.

	Cluster 1	Cluster 2	Cluster 3
TR	0.974	1.384	1.495
LT	1.578	**1.880**	**1.940**
CF	0.968	1.134	1.119
ET	1.147	1.366	1.340
TS	0.046	0.167	0.661
RC	0.735	1.522	**1.981**
GA	**2.083**	**1.868**	1.646
TF	**1.663**	1.673	1.527
FS	**1.618**	**2.091**	**2.102**
RE	0.493	0.673	0.692

TR=Transport, LT=leisure tourism, CF=communal facility, ET=enterprise, RC=residential community, TS=train station and airport, GA=government agency, TF=technical facility, FS=financial service, RE=recreation.

development (Chen *et al*. 2018). According to changes in population size and the concentration of POIs in each category, we classify these three types of areas into regions to be developed (Cluster 1), fast-developing regions (Cluster 2), and well-developed regions (Cluster 3).

Table 2 shows the composition of urban functions in different regions. The results can be obtained by analysing the main functions of different regions in combination with multi-period time-series population curves. Cluster 1 represents regions to be developed. The weekday population increases significantly at 6:00 ~ 7:00 and 14:00 ~ 15:00, while the population sizes are larger in 7:00 ~ 11:00 and 15:00 ~ 18:00. This fluctuation reflects the typical pattern of government work/leisure activities during the workday. In terms of population on Chinese New Year's Eve, the average population on Chinese New Year's Eve Increased by 815.68 and 896.82 compared to weekdays and weekends, respectively. This increase is due to the massive influx of people around Chinese New Year's Eve.

Cluster 2 represents fast-growing regions. The peak in population size occurs at 6:00, and there is little variation in population size fluctuations after the population size decreases between 8:00 and 10:00. The population between 19:00 and 20:00 on workdays decreased by 694.58, consistent with the characteristics of residents' interregional movement to work. The average population increased by 511.12 over the weekend, reflecting the influx of people due to tourism. There is a high concentration of POIs for leisure tourism (1.880) in this regional category, showing that leisure tourism is the second most important function after financial services (2.091). There is also a concentration of POIs for companies (1.366), in line with the characteristics of urban functions in fast-developing regions.

Cluster 3 represents well-developed regions. Corresponding to the high concentration of POIs with a residential function (1.981), population size decreases in 0:00 ~ 8:00 and increases in 18:00 ~ 24:00, in line with the change in residential functional regions. The population size increases by 10,111.63 from 7:00 to 10:00, and the peak occurs in 10:00 ~ 13:00, reflecting the attractiveness of the region's concentration of recreational functions to the population. Population size on weekends decreased by 3,789.38 compared to that on weekdays, related to residents venturing out of the area for entertainment on weekends. Population size decreased by 13,832.85 and 10,042.97 on Chinese New Year's Eve compared to weekdays and weekends, respectively. This decrease is due to the large number of migrant workers returning home, causing a decline in population size on Chinese New Year's Eve.

From the clustering results where k = 3, we can see that the time-series population data effectively reflects various urban functions. Furthermore, the combination of the time-series population changes characteristics of the city, and the level of POI clustering can be used to analyse urban development.

5.3. Urban functional areas and their population characteristics

The clustering results where k = 6 (Figure S2) are further decomposed based on clustering results where k = 3 (Figure 4). From the perspective of spatial inheritance relations, 72.93% of the Cluster 1 and Cluster 2 regions in Figure S2 belong to Cluster 1 in Figure 4; 84.15% of the Cluster 4 and Cluster 5 regions in Figure S2 belong to Cluster 2 in Figure 4; and 98.76% of the Cluster 3 and Cluster 6 regions in Figure S2 belong to the Cluster 3 in

Figure 4. For clustering results where k = 3 and k = 6, the overall spatial distribution is essentially the same, but there are spatial differences in the distribution of urban functions in different regions.

The time-series population change curves (Figure S3) and POI TF-IDF calculations (Table 3) can be obtained from clustering results where k = 6. Regions with similar time-series population curves were found to have a similar composition of urban functions. The population curve pattern of Cluster 2 is consistent with that of Cluster 4, with peaks in population sizes at 7:00 and 15:00 ~ 18:00. Both patterns have smaller population sizes between 10:00 and 12:00. The population on weekdays and weekends is higher than on Chinese New Year's Eve. The composition of each category of POI in Clusters 2 and 4 is similar, with government and public services having the highest clustering levels in the region, followed by finance, science, education, and leisure tourism. Clusters 3 and 6 have consistent patterns of time-series population change curves, with peaks of population size at 7:00 and between 11:00 and 13:00. The population on weekdays and weekends is higher than on Chinese New Year's Eve. Clusters 3 and 6 also have similar compositions in terms of the level of concentration of each POI category, with larger concentrations in financial services and residential functions and smaller concentrations in leisure tourism and transportation.

The larger the average population size, the greater the concentration of functions. Cluster 4 has an average population of 25,351.98, which is greater than Cluster 2 at 13,855.33. The average TF-IDF of the various city functions in Cluster 4 is 1.345 higher than in Cluster 2 at 1.202. The average population of Cluster 6 is 103,460.35, which is greater than that of Cluster 3 at 56,986.16. The average TF-IDF of the various city functions of Cluster 6 is 1.478, which is higher than that of Cluster 3 at 1.412. In summary, Clusters 6 and 4, where the average population is more extensive, have larger concentrations of various urban functions than Clusters 3 and 2.

Among regions with similar urban functions, the more remarkable that the average population size is, the slighter that the contrast is in the degree of urban functional concentration. The TF-IDF value for the government and public social services function in Cluster 4 is the highest at 2.048, which is 0.02 higher than financial services and 0.217 higher than leisure tourism. The TF-IDF value for the government and public social services function in Cluster 2 is the highest at 2.280, which is 0.418 higher than science and 0.5 higher than financial services. The TF-IDF value for financial assistance in Cluster 3

Table 3. TF-IDF of POI in different categories grouped by cluster (k=6). The values that are the top three are shown in boldface.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
TR	0.620	1.006	1.460	1.282	1.422	1.500
LT	**1.275**	1.602	**1.929**	**1.831**	**1.893**	**1.937**
CF	0.589	1.053	1.115	1.148	1.129	1.118
ET	0.807	1.195	1.345	1.363	1.373	1.330
TS	0.027	0.042	0.329	0.097	0.191	0.981
RC	0.379	0.723	**1.866**	1.219	1.611	**2.003**
GA	**1.601**	**2.280**	1.700	**2.048**	**1.814**	1.621
TF	**1.151**	**1.826**	1.572	1.768	1.642	1.505
FS	0.827	**1.780**	**2.110**	**2.046**	**2.106**	**2.095**
RE	0.299	0.515	0.690	0.644	0.682	0.692

TR=Transport, LT=leisure tourism, CF=communal facility, ET=enterprise, RC=residential community, TS=train station and airport, GA=government agency, TF=technical facility, FS= financial service, RE= recreation is the highest at 2.110, which is 0.181 higher than the residential function and 0.244 higher than leisure tourism. The TF-IDF value for financial services in Cluster 6 is the highest at 2.095, which is 0.092 higher than the residential function and 0.158 higher than leisure tourism. The difference in the level of urban functional concentration in Cluster 4 and Cluster 6 is smaller than Cluster 2 and Cluster 3.

6. Discussion

This study combines RTUD data and census data to map the hourly population distribution on weekdays, weekends and Chinese New Year's Eve, and the results obtained perform well (Pearson's R = 0.821, $R^2 = 0.668$). In general, this study's population distribution results are aligned with the characteristics of China's population distribution, which is higher in the east than in the west (Zhang *et al*. 2015). The phenomenon of population concentration is evident in regions with strong economic development, such as Beijing-Tianjin-Hebei, the Yangtze River Delta and the Pearl River Delta. Based on the population distribution in different cities, there is a correlation between cities' population size and their socioeconomic conditions (Kiessling and Landberg 1997). Using Tencent's user data, we can accurately obtain real-time population distributions, which reference studies related to urban populations.

This study extracted the hourly population distributions for weekdays, weekends, and Chinese New Year's Eve by DTW distance. Based on the K-Medoids clustering method, we analysed the clustering at the county scale in China. We then combined the data with the Gaode POIs to calculate the proportion of urban functions in the cities.

The clustering results show that the urban functions of the Beijing-Tianjin-Hebei region differ significantly between the southeastern and northwestern areas (Ren and Fang 2017). There is a large concentration of the well-developed regions around Beijing and Tianjin, while to the south of Beijing and Tianjin lie mainly fast-developing areas where financial services are the primary function and to the north of Beijing lie mostly undeveloped areas where government and public services are the primary functions (Sun *et al*. 2013). The well-developed regions are primarily located around Nanjing, Shanghai and Hangzhou, while the southern area is mainly undeveloped, and the northern area is home to the fast-developing areas. Different functional cities in the Pearl River Delta show stratification (Li *et al*. 2012). With a small number of fast-developing regions in Guangzhou as the core, a layer of well-developed areas is distributed worldwide. The fast-growing regions dominate the outermost layer. The results show that Tencent user data can effectively reflect cities' socioeconomic characteristics and identify the spatial distribution of urban functions in China.

Further analysis of time-series population curves (Figure 5) and TF-IDF calculations for each urban function (Table 2) indicate that the use of time-series population distribution data can reflect the urban function. For example, a region's time-series population curve reflects typical government work/rest characteristics, corresponding to the central government and public service functions in the area. The concentration of leisure tourism functions in fast-developing regions causes an increase in population size on weekends. Moreover, the concentration of entertainment functions in the well-developed areas causes peaks in the population on weekends to occur earlier. These results indicate that urban functions also influence the characteristics of time-series population changes. The above results show that Tencent's user data are closely related to urban functions, providing support for studies of large-scale urban functions.

This study also identified that regions with similar characteristics of time-series population change are similar in urban functions through Tencent's time-series population data. In regions with similar urban functions, the higher the level of clustering of regions with high average populations, the smaller the difference is in the clustering of urban functions. Studies have shown that population concentrations have a significant, positive impact on Chinese cities' economic growth (Chen *et al*. 2018). This study shows that, under certain conditions, the balance of urban functions is positively correlated with population concentration. By promoting the balanced development of urban functions, increased population concentration can positively affect the city's economy. Therefore, in regions with poorly developed urban functions, poor socioeconomic conditions, and small population size, construction could be considered to complement supplementary urban functions, providing favourable conditions for developing urban economies.

This study remain has some shortcomings. This study only analyzes the time-series demographic characteristics of working days and rest days in the collected RUTD data. There are certain differences in the time-series demographic characteristics of different periods and seasons, and there are also certain differences in the urban function. In the follow-up research, we can consider collecting a longer period of RTUD to analyse its reflection on urban functions. The single RTUD data still has certain shortcomings in covering the population and area. In the follow-up research, we will consider adding auxiliary data such as mobile phone signalling data to calculate the time-series population data. In the results of urban function division, China's township administrative regions are not distinguished. The analysis of the urban functions of townships with different levels of urbanisation can be one of the directions for further research. When determining the impact of urban functions on time-series population changes, changes in a single urban function have not been analysed. For instance, when examining the effect of leisure tourism functions on weekend population increases, we cannot estimate the impact of a single leisure tourism function on increased population size during the weekends. Future studies could focus on employing more fine-grained data on POI and could delve into the relationships between urban functions and time-series population changes. Tencent's user data are closely associated with urban functions and socioeconomic conditions within cities, and subsequent studies of the fine-scale distribution of poverty in China could be considered.

7. Conclusion

Studies of urban functions have primarily emphasised the division of functions within cities. Due to limitations in data and methodologies, existing studies of urban functions on a large scale have rarely considered the effects of population movement and migration during the Chinese Spring Festival. This study delineates urban functions at the county scale in China by combining time-series Tencent user data, census data, and Gaode POI data. We map the hourly population distribution on weekdays, weekends, and Chinese New Year's Eve by combining RTUD and census data, and we obtain highly accurate results (Pearson's R = 0.821 R^2 = 0.668). Through clustering the time-series population data, we classify the county scale urban functions in China. The urban functions of the Beijing-Tianjin-Hebei cluster are found to be significantly different in the southeastern and northwestern areas of the region, as are the southern and northern areas of the Yangtze River Delta cluster. The Pearl River Delta cluster reflects a stratification phenomenon.

Further analysis of urban function and time-series population data shows the interaction between time-series population changes and urban function and that regions with similar characteristics of time-series population changes have similar functional urban structures. This study shows that Tencent's user data effectively reflects urban functions and urban socioeconomic factors, supporting studies related to urban functions and urban socioeconomic conditions. Our results could improve understanding of the urban function and contribute valuable material to better urban planning.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Key Research and Development Program of China [2019YFB2102903], the National Natural Science Foundation of China [41801306, 41671408]; the "CUG Scholar" Scientific Research Funds at China University of Geosciences (Wuhan) [2022034] and a grant from State Key Laboratory of Resources and Environmental Information System.

ORCID

Qingfeng Guan http://orcid.org/0000-0002-7392-3709
Yao Yao http://orcid.org/0000-0002-2830-0377
Chen Qian http://orcid.org/0000-0002-4028-4752
Shuliang Ren http://orcid.org/0000-0003-3776-3266

References

Amorim, R.C.D. and Mirkin, B., 2012. Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering. *Pattern Recognition*, 45 (3), 1061–1075. doi:10.1016/j.patcog.2011.08.012

Berndt, D.J. and Clifford, J., 1994. Using dynamic time warping to find patterns in time series. *KDD Workshop*, 10 (16), 359–370.

Chen, Y., et al., 2017. 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, 48–60. doi:10.1016/j.landurbplan.2016.12.001

Chen, L., Xun, L., and Yao, Y., 2018. Effects of population agglomeration on urban economic growth in China. *Acta Geographica Sinica*, 73 (6), 1107.

Chiang, M.T. and Mirkin, B., 2010. Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads. *Journal of Classification*, 27 (1), 3–40. doi:10.1007/s00357-010-9049-5

Deville, P., et al. 2014. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111 (45), 15888–15893. doi:10.1073/pnas.1408439111

Gong, P., et al. 2020. Mapping essential urban land use categories in *China* (EULUC-*China*): preliminary results for 2018. *Science Bulletin*, 65 (3), 182–187. doi:10.1016/j.scib.2019.12.007

Goodchild, M.F., 2008. Commentary: whither VGI? *GeoJournal*, 72 (3–4), 239–244. doi:10.1007/s10708-008-9190-4

Huang, H., et al. 2018. Location based services: ongoing evolution and research agenda. *Journal of Location Based Services*, 12 (2), 63–93. doi:10.1080/17489725.2018.1508763

Kang, C., et al. 2012. Towards estimating urban population distributions from mobile call data. *Journal of Urban Technology*, 19 (4), 3–21. doi:10.1080/10630732.2012.715479

Kiessling, K.L. and Landberg, H., 1997. *Population, economic development, and the environment*. New York: Oxford University Press.

Li, J., et al. 2016. Spatial-temporal analysis on Spring Festival travel rush in China based on multisource big data. *Sustainability*, 8 (11), 1184. doi:10.3390/su8111184

Li, T., Cao, X., and Huang, X., 2012. The relationship between spatial structure of accessibility and population change in Pearl River Delta. *Geographical Research*, 31 (9), 1661.

Liu, Y., et al. 2012. Urban land uses and traffic 'source-sink areas': evidence from GPS-enabled taxi data in Shanghai. *Landscape and Urban Planning*, 106 (1), 73–87. doi:10.1016/j.landurbplan.2012.02.012

Liu, 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. doi:10.1080/00045608.2015.1018773

Long, Y., Liu, L., and Ma, X., 2017. How green are the streets? An analysis for central areas of Chinese cities using Tencent Street View. *PLoS One*, 12 (2), e171110. doi:10.1371/journal.pone.0171110

Long, Y. and 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*. Beijing: Springer, 193–217.

Lu, M., Li, G., and Sun, T., 2011. Study on the functional pattern of the Beijing metropolitan region and its changes: based the analysis of data from the economic unit census. *Geographical Research*, 30 (11), 1970–1982.

McDonald, T., 2016. *Social media in rural China*. UCL Press.

Mirkin, B., 2012. *Clustering: a data recovery approach*. CRC Press.

Moody, J., et al., 2019. Transportation policy profiles of Chinese city clusters: a mixed methods approach. *Transportation Research Interdisciplinary Perspectives*, 2, 100053. doi:10.1016/j.trip.2019.100053

Olteanu, A., et al., 2019. Social data: biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2, 13. doi:10.3389/fdata.2019.00013

Pan, J. and Lai, J., 2019. Spatial pattern of population mobility among cities in China: case study of the National Day plus Mid-Autumn Festival based on Tencent migration data. *Cities*, 94, 55–69. doi:10.1016/j.cities.2019.05.022

Papadopoulos, S., et al. 2012. Community detection in social media. *Data Mining and Knowledge Discovery*, 24 (3), 515–554. doi:10.1007/s10618-011-0224-z

Park, H. and Jun, C., 2009. A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, 36 (2), 3336–3341. doi:10.1016/j.eswa.2008.01.039

Park, H., Lee, J., and Jun, C. 2006. A K-means-like algorithm for K-medoids clustering and its performance. *Proceedings of ICCIE*, 102–117.

Patel, N.N., et al. 2017. Improving large area population mapping using geotweet densities. *Transactions in GIS*, 21 (2), 317–331. doi:10.1111/tgis.12214

Pei, T., et al. 2014. A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*, 28 (9), 1988–2007. doi:10.1080/13658816.2014.913794

Rakthanmanon, T., 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. doi:10.1145/2500489

Ren, Y. and Fang, C., 2017. Spatial pattern and evaluation of eco-efficiency in counties of the Beijing-Tianjin-Hebei Urban Agglomeration. *Progress in Geography*, 36 (1), 87.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. doi:10.1016/0377-0427(87)90125-7

Sevtsuk, A. and Ratti, C., 2010. Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. *Journal of Urban Technology*, 17 (1), 41–60. doi:10.1080/10630731003597322

Shaw, S., Tsou, M., and Ye, X., 2016. Editorial: human dynamics in the mobile and big data era. *International Journal of Geographical Information Science*, 30 (9), 1687–1693. doi:10.1080/13658816.2016.1164317

Song, Y., et al. 2018. Are all cities with similar urban form or not? Redefining cities with ubiquitous points of interest and evaluating them with indicators at city and block levels in China. *International Journal of Geographical Information Science*, 32 (12), 2447–2476. doi:10.1080/13658816.2018.1511793

Sun, D., Zhang, J., and Mingdou, Z., 2013. Coupling relationship between urbanization efficiency and economic development level in the Yangtze River Delta. *Progress in Geography*, 32 (7), 1060.

Tu, W., et al. 2017. Coupling mobile phone and social media data: a new approach to understanding urban functions and diurnal patterns. *International Journal of Geographical Information Science*, 31 (12), 2331–2358. doi:10.1080/13658816.2017.1356464

Wang, M., 2007. The effects of urban function on population growth in Shandong Province. *Acta Geographica Sinica*, 2 (62), 127–136.

Wang, Y., et al. 2011. Towards street-level client-independent IP geolocation. *Nsdi*, 11, 27-27.

Wang, Y., et al., 2019a. Delineating urbanization “source-sink”regions in China: evidence from mobile app data. *Cities*, 86, 167–177. doi:10.1016/j.cities.2018.09.016

Wang, Y., et al., 2019b. Migration patterns in China extracted from mobile positioning data. *Habitat International*, 86, 71–80. doi:10.1016/j.habitatint.2019.03.002

Wei, Y., et al., 2018. The rich-club phenomenon of China’s population flow network during the country’s Spring Festival. *Applied Geography*, 96, 77–85. doi:10.1016/j.apgeog.2018.05.009

Wu, H.C., et al. 2008. Interpreting TF-IDF term weights as making relevance decisions. *ACM Transactions on Information Systems*, 26 (3), 13. doi:10.1145/1361684.1361686

Xu, J., et al., 2017. Difference of urban development in China from the perspective of passenger transport around Spring Festival. *Applied Geography*, 87, 85–96. doi:10.1016/j.apgeog.2017.07.014

Yao, Y., et al., 2017. Mapping fine-scale population distributions at the building level by integrating multisource geospatial big data. *International Journal of Geographical Information Science*, 31 (6), 1220–1244.

Yao, Z., et al., 2018. Representing urban functions through zone embedding with human mobility patterns. *IJCAI*, Stockholm, Sweden, 3919–3925.

Yuan, N.J., et al. 2014. Discovering urban functional zones using latent activity trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 27 (3), 712–725. doi:10.1109/TKDE.2014.2345405

Yuan, J., Zheng, Y., and Xie, X. 2012. Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Beijing, 186–194.

Yue, H., et al. 2019. Detecting clusters over intercity transportation networks using K-shortest paths and hierarchical clustering: a case study of mainland China. *International Journal of Geographical Information Science*, 33 (5), 1082–1105. doi:10.1080/13658816.2019.1566551

Yu, J., Li, J., and Zhang, W., 2019. Identification and classification of resource-based cities in China. *Journal of Geographical Sciences*, 29 (8), 1300–1314. doi:10.1007/s11442-019-1660-8

Zeng, C. and Shen, Y., 2015. A study of the functional features of China’s urban service industries. *Geographical Research*, 34 (9), 1685–1696.

Zhang, Y., et al., 2019. Community scale livability evaluation integrating remote sensing, surface observation and geospatial big data. *International Journal of Applied Earth Observation and Geoinformation*, 80, 173–186. doi:10.1016/j.jag.2019.04.018

Zhang, Y., Song, Y., and Chuanyong, Z., 2015. The new urbanization and the possibility of breaking through the “HU Line”. *Journal of East China Normal University (Humanities and Social Sciences)*, 2, 101–112.

Zhao, Z., et al., 2017. Spatiotemporal and structural characteristics of interprovincial population flow during the 2015 Spring Festival travel rush. *Progress in Geography*, 36 (8), 952–964.

Zhi, Y., 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*, 19 (2), 94–105. doi:10.1080/10095020.2016.1176723

Zhou, Y. and Bradshaw, R., 1988. The classification of industrial function of Chinese cities (including attached counties): theory, method and results. *Acta Geographica Sinica*, 55 (4), 287.