

Integrating multi-source big data to infer building functions

Ning Niu, Xiaoping Liu, He Jin, Xinyue Ye, Yu Liu, Xia Li, Yimin Chen & Shaoying Li

To cite this article: Ning Niu, Xiaoping Liu, He Jin, Xinyue Ye, Yu Liu, Xia Li, Yimin Chen & Shaoying Li (2017): Integrating multi-source big data to infer building functions, International Journal of Geographical Information Science, DOI: [10.1080/13658816.2017.1325489](https://doi.org/10.1080/13658816.2017.1325489)

To link to this article: <http://dx.doi.org/10.1080/13658816.2017.1325489>



Published online: 09 May 2017.



Submit your article to this journal [↗](#)



Article views: 138



View related articles [↗](#)



View Crossmark data [↗](#)



ARTICLE



Integrating multi-source big data to infer building functions

Ning Niu^a, Xiaoping Liu ^a, He Jin^b, Xinyue Ye^c, Yu Liu ^d, Xia Li ^a, Yimin Chen^a and Shaoying Li ^e

^aGuangdong Key Laboratory for Urbanization and Geo-simulation, School of Geography and Planning, Sun Yat-sen University, Guangzhou, China; ^bDepartment of Geography, Texas State University, San Marcos, TX, USA; ^cDepartment of Geography and Computational Social Science Lab, Kent State University, Kent, OH, USA; ^dInstitute of Remote Sensing and Geographical Information Systems, Peking University, Beijing, China; ^eSchool of Geographical Sciences, Guangzhou University, Guangzhou, China

ABSTRACT

Information about the functions of urban buildings is helpful not only for developing a better understanding of how cities work, but also for establishing a basis for policy makers to evaluate and improve the effectiveness of urban planning. Despite these advantages, however, and perhaps simply due to a lack of available data, few academic studies to date have succeeded in integrating multi-source 'big data' to examine urban land use at the building level. Responding to this deficiency, this study integrated multi-source big data (WeChat users' real-time location records, taxi GPS trajectories data, Points of Interest (POI) data, and building footprint data from high-resolution Quickbird images), and applied the proposed density-based method to infer the functions of urban buildings in Tianhe District, Guangzhou, China. The results of the study conformed to an overall detection rate of 72.22%. When results were verified against ground-truth investigation data, the accuracy rate remained above 65%. Two important conclusions can be drawn from our analysis: 1. The use of WeChat data delivers better inference results than those obtained using taxi data when used to identify residential buildings, offices, and urban villages. Conversely, shopping centers, hotels, and hospitals, were more easily identified using taxi data. 2. The use of integrated multi-source big data is more effective than single-source big data in revealing the relation between human dynamics and urban complexes at the building scale.

ARTICLE HISTORY

Received 31 October 2016
Accepted 27 April 2017

KEYWORDS

Multi-source big data;
density-based method;
building functions

1. Introduction

Buildings constitute the fundamental components of a city and are hubs of citizen activities. Knowledge of the functions performed by the buildings that make up our cities thus constitutes an important basis for urban planning practice, providing scholars and practitioners with actual use information at a much finer spatial scale than that informing the category 'land use.' Conventionally, geographers and scientists in other fields have focused on studying either the physical characteristics of land (treating it as a surface) or the social characteristics of land (viewing land in terms of use), and have done so at relatively broad spatial scales. Remote sensing techniques (using airborne images or images obtained from satellites) have

traditionally been used to classify the latter category (land use) on the basis of the former (i.e., common physical characteristics) by comparing spectral, spatial, and radiometric characteristics of remote sensing images (Moran *et al.* 1997, Mesev 1998, Yang and Lo 2002, Rogan and Chen 2004, Schmit *et al.* 2006, Huang *et al.* 2013, Liu *et al.* 2014). This dependency means that few methods are based solely upon social functions (Pan *et al.* 2013). Further, despite the widespread use and development of such methods, smaller-scale building functions (i.e., residential, school, and commercial) have been difficult to distinguish to date, as such images tend to span large geographic areas (Pei *et al.* 2014). In order to address this problem, additional information – for instance, contextual information and expert knowledge (De Wit and Clevers 2004, Platt and Rapoza 2008, Hu and Wang 2013) – has had to be used to augment such studies and aid in land-use classification. The images obtained from remote sensing techniques (both satellite and airborne) also provide an excess of physical information about a city that may not be directly relevant to the classification of urban land use (Zhan *et al.* 2014).

In recent years, the extraction of specific, useful information from ‘big data’ samples has emerged as an important new trend. This can be seen in studies estimating urban land use (Qi *et al.* 2011, Soto and Frías-Martínez 2011) and urban human mobility (Kang *et al.* 2012, Yuan *et al.* 2012, Wu *et al.* 2014). The use of big data has in particular been important in studies seeking to identify patterns in the daily mobility of given populations in relation to the locations they frequent. Such patterns – for instance, by confirming that people usually leave home in the morning and return home in the evening on a work day, with the opposite pattern occurring in relation to workplaces – provide researchers with insights into the social functions of a given location. Observing such daily activities via the use of taxi GPS data, social media data and other similar sources makes it possible for scholars to differentiate people’s activities and thereby to identify the different social functions fulfilled by a given piece of land. The objective of this paper is to integrate of multisource big data and high-resolution Quickbird building footprints with the aim of inferring urban building functions. We anticipate that the proposed method will be particularly useful for urban planning, allowing planners to identify actual land use in large cities in China and other rapidly developing countries at the scale of the building.

2. Related work

Big data are frequently arisen in urban research. The principles of big data – preparing, sharing, and analyzing complex information – have facilitated the development of innovative conceptual, analytical and technical skills to handle them with geospatial data (Berman 2013). As a result of rapid development in urban studies, a number of academic studies have, in this way, been able to analyze the relation between land use and human behavior using various kinds of Big Data. For instance, Liu *et al.* (2012) applied the ‘source-sink areas’ theory from ecology in order to identify broad-scale land-use classes using taxi GPS data in Shanghai. In contrast, Yuan *et al.* (2012) were able to infer the functions of various buildings in Beijing using a topic-based inference model. These scholars categorized point-of-interest (POI) data in combination with taxi data, a method that allowed them to identify human mobility patterns and classify land use accordingly. In a further study using taxi data, Pan *et al.* (2013) addressed the social function of land in Hangzhou, using taxi trace data from 4000 taxis. Similarly, Liu *et al.* (2015a) used a two-level hierarchical polycentric city structure model in order to study

spatial interaction in Shanghai city. Using mobile phone call data in Singapore, Pei *et al.* (2014) applied a semi-supervised fuzzy c-means clustering approach in order to identify different types of land use, generating results with an overall detection rate of 58.03%. Integrating land-use, transportation, and environment models and using POI data, the human activity analysis undertaken by Jiang *et al.* (2015) was able to evaluate disaggregated land use at the census block level.

The rapid development of mobile internet, Facebook, Twitter, WeChat, and Weibo, has made social media available to end-users via smartphones, tablet computers, and other devices. The smartphone has, in particular, become an important platform from which interactions between individuals and their geographic space can be observed.

The ‘crowdsourcing’ is a representative form of these social media data. Social media amateurs in mapping produce the geospatial data rather than depending on professionals. These amateurs collaborate voluntarily without financial compensation. Thus, open datasets with low monetary cost become more and more available than ever; real-time mapping and change detection have also become prevalent in academia and industry (Goodchild and Michael. 2007, Heipke 2010). Recognizing these, a number of scholars have attempted to use social media check-in data to classify land use in cities. For example, Zhan *et al.* (2014) undertook a comprehensive investigation of the validity of utilizing large-scale social media check-in data in order to infer land-use types using data mining techniques in New York City. Similarly, Frias-Martinez and Frias-Martinez (2014) classified urban land use by clustering Twitter data collected in Manhattan, New York City (USA), London (UK), and Madrid (Spain). In a recent study, Zhong *et al.* (2014) integrated two data sources (surveys and smartcard data) to infer the building functions for two areas in Singapore, demonstrating that practical land-use information can be obtained through the use of probabilistic modelling. Their study addressed small land-use areas with dimensions 1500 m by 2000 m for 2737 buildings, and 5000 m by 3000 m for 3909 buildings.

Pursuant to the above literature review, few studies have to date examined urban structure at the building scale by employing methods that either fused different sources of big data (i.e., taxi GPS data, phone records data, smartcard data, and social media data), or big data with remote sensing data. To do so – that is, to employ big data and remote sensing data from a range of sources – would allow researchers to acquire information about a range of different aspects of the geographical environments studied. The integration of multiple types of data thus constitutes a particularly attractive topic for researchers at present (Liu *et al.* 2015b). In addition to meeting this first challenge, Zhong *et al.* (2014) describe two further problems which have not yet been solved within existing scholarship on this topic. Firstly, they noted that different types of mixed buildings in the cities may influence the reference detection and accuracy rates, and methods must be developed that take such effects into account; and secondly, they also pointed out that final reference results depend on people’s travel purpose in the survey, which may in fact lack objectivity.

Being mindful of these problems, in this study we attempted to infer building functions using multisource big data. We see our research as offering an intuitive approach for understanding the building functions present within cities, providing necessary information for urban planners to evaluate and improve the effectiveness of

planning schemes. The major contributions of this paper are four-fold. First, we were able to fuse taxi data, WeChat data, and building footprint areas taken from high-resolution Quickbird images. Second, we proposed a density-based method in order to infer building functions. Third, we applied this methodology through a case study of the Tianhe District of Guangzhou, in China. Finally, we also developed a novel method that was used to verify our experimental results, using ground-truth investigation data. The final results showed an overall detection rate of 72.22% and an accuracy rate that was above 65%.

3. Study area and data

Our case study area, Tianhe District, is located at the geographical center of the Chinese city of Guangzhou (Figure 1). The population of Tianhe District in 2015 was 1,430,000 and its area was 96.33 km² (Bureau of Statistics of Guangzhou 2015, <http://www.gzstats.gov.cn/tjgb/qstjgb/>). With the rapid development of Guangzhou in recent decades, Tianhe District has become the city's new residential, commercial, and educational center, accommodating the highest population density in Guangzhou. The types of urban buildings in Tianhe District include residential buildings, office complexes, shopping centers, schools, hospitals, and hotels, amongst others.

This study used four data sets relating to Tianhe District: a building footprint dataset, a taxi dataset, a WeChat dataset, and a POI database and street view taken from Baidu Map. We used high-resolution Quickbird images from 2015 to support the visual

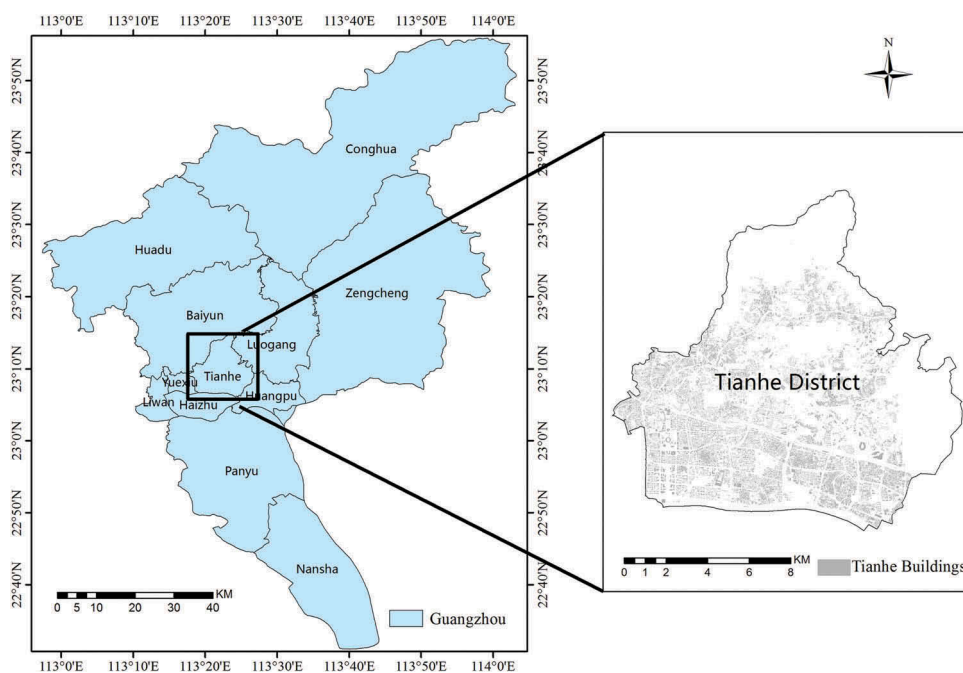


Figure 1. The case study area: Tianhe District, Guangzhou, China.

interpretation required to obtain relevant building footprints. Through this process, we successfully extracted the footprints of 68,997 buildings.

China's urban public transportation system, which is made up of public buses, metro rail services, taxis, and ferries, facilitates the mobility of urban commuters in cities across the country. In 2014, the public transit system served 35% of travelers in broader Guangzhou and 45% of travelers in the urban core, with taxis accounting for 19.52% of intra-city trips (Guangzhou Transportation Committee 2014, <http://www.gzjt.gov.cn/gzjt/web/Publish/PublicMain.aspx>). Compared to other transit modes, taxis constitute a relatively time-efficient option – as a result, they are widely used in Guangzhou, and a large numbers of taxi data records are generated in the city each day. These data records, which are tagged with the time period and geographic locations of a pick-up or drop-off, constitute valuable information for modeling human activity, serving as an alternative measure to describe building functions in major cities. In this study, we selected 25,000 taxis working in Tianhe District on the days of Monday, 5 May 2014 and Saturday, 16 May 2014 – a sample made up of 6 million records. This dataset is provided by a company (Guangdong Ritu Wanfang Science & Technology Co., Ltd). We constructed a dataset to concentrate on the reciprocities between buildings and to simplify taxi trajectories into vectors composed of origins, destinations, and times.

The third dataset used in this study was made up of real-time WeChat user density information. In order to obtain these real-time data, we developed a web crawler based on an API (application program interface) from 'Easygo' (<http://ur.tencent.com>) to record the real-time WeChat user data covering our study area. It is an open platform of WeChat to provide the public with the real-time crowdedness around the location of interest. By executing our program, we gathered the hourly WeChat user density data from June 15 to 21 June 2015, with a spatial resolution of 25 m. Given that people generally have different activities on work days and non-work days, we selected WeChat data on a work day (15 June 2015) and a non-work day (20 June 2015) in the Tianhe District for our research.

Moreover, we also utilized a POI database (comprising approximately 40,000 records) and street view data from Baidu Map in order to help infer the building functions of the training samples. The POI database provides detailed, actual information for a given point location, including the **name, address, and category of that location**. To acquire these data for our study, we designed another web crawler that uses a Baidu Map API to record 17 categories of POI data covering the study area. We calculated the proportion of each type, and then selected six highest ratios of POI types: schools (elementary/middle/high/technical schools/universities), office complexes (enterprises, financial services, and government agencies), shopping buildings (supermarkets, malls), hospitals (including clinics), hotels, and residential communities.

4. Methodology

4.1. The architecture

This study utilized a density-based method (Figure 2). The approach integrated three steps: Firstly, indicators of individual-level behavior (taxi passenger pick-up and drop-off locations) were used to establish the relations between urban buildings, which were then clustered using a modified Density-Based Spatial Clustering of Applications with

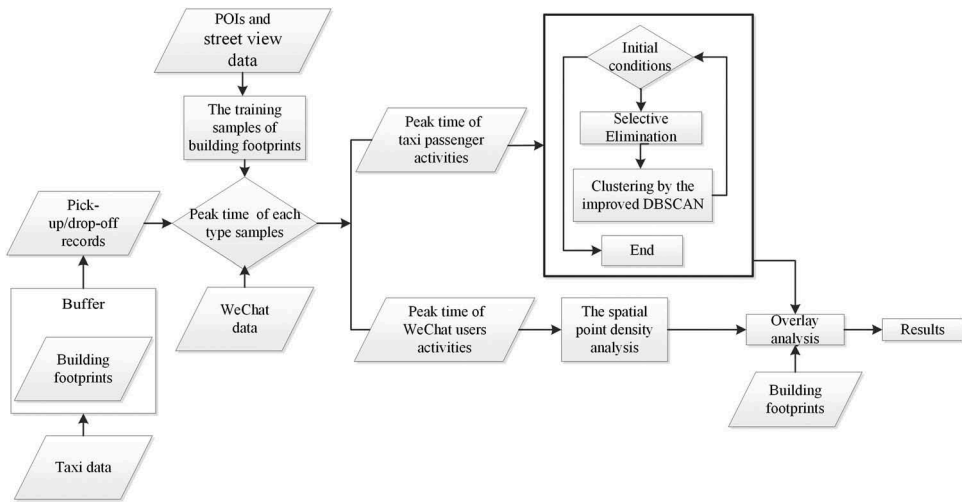


Figure 2. The density-based method used in this study.

Noise algorithm (modified DBSCAN). Secondly, social media data from WeChat was analyzed using spatial point density analysis. Thirdly, building functions were inferred in accordance with a defined set of evaluation rules. The detailed process followed can be described in terms of six steps:

- (1) We established the spatial links between individual-level behavior (understood in terms of both taxi passenger pick-up/drop-off locations and WeChat user activities) and urban buildings.
- (2) The peak activity time for each type of building was then established through analysis of the training samples.
- (3) Taxi passenger pick-up/drop-off records were then clustered using the modified DBSCAN algorithm.
- (4) A spatial point density analysis was then performed in relation to real-time WeChat user location records.
- (5) The clustering results from the third step and the density analysis results from the fourth step were then integrated.
- (6) An evaluation rule for inferring building function was defined.

The first step of our methodology was oriented towards solving the spatial relation problem between urban buildings and human dynamics. In order to address this, it was important that we obtain the ‘peak time’ of each type of building in terms of the activities undertaken in relation to them. This was done using the training samples. We assumed that buildings with similar functions would have similar peak times in terms of taxi passenger pick-ups or drop-offs and WeChat user activity. We then clustered taxi pick-ups or drop-offs, performing a spatial point density analysis on the WeChat data sample. These subsequent steps in our proposed methodology were established on the basis of the initial assumption. The taxi and WeChat data was composed of locations and times, with each taxi record standing for a given taxi passenger’s pick-up or drop-off.

Use of the modified DBSCAN algorithm was considered appropriate in clustering points in this uneven dataset. Given the real-time WeChat user density data, the spatial point density analysis worked well in processing this data. Each type of geospatial big data used in this study can be considered as a remotely sensed representation of a specific human activity (Liu *et al.* 2015b), for example taxi data describes passenger dynamics. We fused the results of the analyses undertaken in the second and third steps with the footprints of buildings as they were ascertained via the fifth step of our research design. We then subsequently defined an evaluation rule for inferring building functions. Lastly, we calculated the overall inference detection and verified this against the correct inference rate. Each step is described in detail in [Section 4.2](#).

4.2. A density-based method

According to POI categories from Baidu Map and building types of Zhong *et al.* (2014), we were able to define seven types of buildings in Tianhe District, namely: residential buildings, urban villages (Lin *et al.* 2011), office complexes, shopping centers, hotels, hospitals, and schools. It was vital that we establish the proportion of buildings of each type in order to determine the inference sequence to be applied in relation to the different types of buildings. We thus calculated the proportions of residential buildings, office complexes, shopping centers, hotels, hospitals, and schools as they are described by the urban land-use classification map of Guangzhou, provided by Guangzhou Urban Planning Bureau. Besides, human activities are similar in residential buildings and urban villages, it was difficult to differentiate between these types of buildings. Wang *et al.* (2009) and Lin *et al.* (2011) indicated that in most Chinese urban villages, buildings (not including unauthorized buildings) usually have no more than four floors and an area of not more than 480 m². Moreover, many urban villages tend to be located close to Central Business Districts (CBDs) with building densities of 70% or greater. According to these differences of residential buildings and urban villages, we distinguished urban villages from residential buildings.

1. Ascertain the spatial ties between individual-level behavior and urban buildings

In general, the pick-up and drop-off sites of taxi passengers occur at a distance from their actual destinations. The walking distance from pick-up or drop-off sites to the building is thus a vital factor affecting taxi passengers' choices. In order to confirm the assumption that most taxi passengers walk some distance from drop-off sites to buildings and from buildings to pick-up sites, we undertook a ground-truth investigation in our case area. The subsequent buffer analysis that we performed allowed us to construct an understanding of the spatial relation between taxi passenger pick-up or drop-off sites and the urban buildings they frequented. The acquired WeChat data were aggregated to construct a dataset composed of location and time data. Each record documents the density of WeChat users at the building level with a spatial resolution of 25 m.

2. Obtaining the peak time for each type building via the use of training samples

In accordance with the prevalence of each type of building in the case study area (measured as the number of buildings of that type as a proportion of the total number

of buildings), we chose M samples from K building function types (where m_k stands for the number of samples for building function type k). Information about these sample function types was acquired from Points of Interest (POIs) and street view in Baidu Map. Human activities usually exhibit different peak times in relation to different types of buildings (Liu *et al.* 2012). The purpose of selecting these training sample buildings was to obtain the peak time of each type of building addressed in the study. Thus, the average number of taxi passenger pick-ups and drop-offs and the average number of real-time Wechat users for each type of building in the sample group (A_k ($\{A_{k,p}, A_{k,d}, A_{k,w}, k = 1, \dots, K\}$)) was calculated using the formulas (1–3):

$$A_{k,p} = \frac{1}{m_k} \sum_{m_k}^1 (N_{k,p}, t1) \quad (1)$$

$$A_{k,d} = \frac{1}{m_k} \sum_{m_k}^1 (N_{k,d}, t2) \quad (2)$$

$$A_{k,w} = \frac{1}{m_k} \sum_{m_k}^1 (N_{k,w}, t3) \quad (3)$$

Where, $A_{k,p}$ is the average number of pick-ups for building function type k inside the sample building buffers in $t1$ time; $A_{k,d}$ is the number of drop-offs on average for building function type k inside the sample building buffers in $t2$ time. $A_{k,w}$ is the average number of real-time WeChat users for building function type k within the sample buildings in $t3$ time. $N_{k,p}$ stands for the sum number of pick-ups for building function type k in the sample building buffers; $N_{k,d}$ is the total number of drop-offs for building function type k in the sample building buffers. $N_{k,w}$ is the sum number of the real-time WeChat users for building function type k in the sample buildings; $t1$ is the time of pick-ups, $t2$ is the time of drop-offs, $t3$ is the time of WeChat users' activities; m_k is the number of building function type k in the sample buildings, $k \in [1, K]$, and $t1, t2, t3 \in [0, 24]$.

3. Clustering taxi data using the modified DBSCAN algorithm

The distribution of taxi data is not regular, and there are many sources of noise in taxi data (e.g., pick-up/drop-off location deviation). Many algorithms, such as K-means and K-Medoids, have difficulty dealing with these problems. However, the DBSCAN algorithm, first proposed by Ester *et al.* (1996), is capable of solving these issues. DBSCAN algorithm has a simple structure and high computational efficiency; it can process high-density data to generate clusters with any arbitrary shapes as well as eliminate spatial data noise (Ertöz *et al.* 2003). In recent years, it has also been widely used in clustering big data (Pan *et al.* 2013, Shen and Cheng 2015, Tang *et al.* 2015). Hence, we used the DBSCAN algorithm to cluster taxi data in this paper. Important definitions in DBSCAN include:

Definition 1: (Eps-neighborhood of a point) The Eps-neighborhood of a point p , denoted by $N_{Eps}(p)$, is defined as $N_{Eps}(p) = \{q \in D | L(p, q) \leq Eps\}$. Here, $L(p, q)$ is the Euclidean distance from p to q .

Definition 2: (Directly Density-Reachable) Point p is directly density-reachable from point q with respect to Eps and $MinPts$ only if $p \in N_{Eps}(q)$ and $|N_{Eps}(q)| \geq MinPts$.

Definition 3: (Density-Reachable) Point p is density-reachable from point q with respect to Eps and $MinPts$ if there is a chain of points $p_1, \dots, p_n, p_1 = q, p_n = p$.

Definition 4: (Density-Connected) Point p is density-connected to point q with respect to Eps and $MinPts$ if there is a point o such that both p and q are density-reachable from o with respect to Eps and $MinPts$.

Definition 5: (Cluster) Let D be a point database. Cluster C with respect to Eps and $MinPts$ is a nonempty subset of D satisfying the following conditions: 1) $\forall p, q$: if $p \in C$ and q is density-reachable from p with respect to Eps and $MinPts$, then $q \in C$. 2) $\forall p, q \in C$: p is density-connected to q with respect to Eps and $MinPts$.

There are two important parameters in DBSCAN: Eps (the cluster radius parameter) and $MinPts$ (the neighborhood density threshold). Previous studies have shown that DBSCAN is sensitive to the Eps and $Minpts$ parameters, and as such that it is difficult to cluster uneven datasets, an operation that requires larger memory support and IO (computer input and output) consumption (Gan *et al.* 2007). These issues were present in our research, which used approximately 6 million taxi track records, with taxi passenger pick-up and drop-off sites displaying an uneven spatial distribution – as such, large memory support and IO consumption was required. To address these issues, we modified the DBSCAN clustering algorithm.

Previous studies have demonstrated that it is not suitable to use the Euclidean distance alone for time series data clustering (Fu 2011). Further, given that taxi passenger pick-ups or drop-offs happened at the border of streets, and were linked by streets, it was considered reasonable to use street network distance as a measure of distance in the study. According to Daszykowski, *et al.* (2001) advice, we set the value of $Minpts$ is 1/25 of the total number of current data points. We then undertook the following steps in order to solve the sensitivity of parameters (Eps) problem and the uneven datasets problem. Four steps were involved in the initialization of the algorithm: 1) the cycle time was set; 2) a point p was randomly selected; 3) the near distance point group k of point p was identified; 4) all points in the group were sorted in descending order in terms of the distances that each point maintained from point p ; 5) the nearest distance for the points from step 3) was assigned to Eps . In the case that the randomly selected initial point was identified as a noise point or a low-density critical point in step 2), it also then had the potential to generate clusters from its surrounding k points. The clusters that result from such an operation are inaccurate and affect the accuracy of final clusters. When a noise point is used to search its directly density-reachable clusters, the short distance points in the cluster include this noise point. When the noise point's neighborhoods are low-density regions, the value of Eps becomes far less than the average street network distance of all points in the cluster. In order to fix this problem, we defined the

following evaluation function to detect noise, which was tested before proceeding to step 2):

$$D = \frac{1}{m} \sum_m^1 d(p_i, p_n), n \in [1, m] \quad (4)$$

$$D \gg \text{Eps} \quad (5)$$

Where p_i is the initial point, p_n is the point group around p_i , d is the street network distance between points, and D is the average street network distance of all points in the cluster.

Secondly, as explained earlier, when the DBSCAN clustering algorithm processes big data, it requires larger memory support and necessitates higher IO consumption. In addition, the inference of one building type can adversely affect the inferences of other building types. We therefore utilized the ‘Selective Elimination’ method to address these problems, which meant that, for example, when we inferred one building type (e.g., office complexes), the other types (e.g., residential buildings, commercial buildings, and so on) in the nearby records of taxi pick-ups and drop-offs were removed from the calculation.

Lastly, we selected the pick-up and drop-off records from Tianhe taxi dataset at peak time periods, clustering them using this modified DBSCAN algorithm.

4. Analyzing the wechat data and defining the evaluation rules for inferring building functions

We also performed a spatial point density analysis using ArcGIS 10.2, which allowed us to extract WeChat user locations in the peak time period. Here, we assumed that buildings with similar functions have similar peak times in terms of taxi passenger pick-ups or drop-offs and WeChat user activities. According to this assumption, we defined an evaluation rule to infer building functions in the formula (6).

$$\begin{aligned} \text{Result} = & \text{overlap}\{C(\text{taxi}, pt1), D(\text{WeChat}, pt2), \text{building}\} \\ & + \text{overlap}\{C(\text{taxi}, pt1), \text{building}\} + \text{overlap}\{D(\text{WeChat}, pt2), \text{building}\} \end{aligned} \quad (6)$$

Where $pt1$ is the peak time of taxi passenger pick-ups or drop-offs in a given type of building, $pt2$ is the peak time of WeChat user activities in a given type of building, C is the clustering results of the taxi data, and D is the results from the spatial point density analysis of WeChat data.

Next, we combined the clustering results of the taxi data, the results from the spatial point density analysis of WeChat data, and Tianhe’s building footprints, undertaking an overlay analysis in ArcGIS 10.2. Then, we overlap the clustering results of the taxi data and Tianhe’s building footprints, and integrated the results from the spatial point density analysis of WeChat data with Tianhe’s building footprints.

Moreover, the inference of buildings functions is from one type to another, we infer building functions for one type; when we infer the building functions for another type, the same building is identified as an another type. If this is the case, we define that the building has mixed functions.

5. Results

Following the methods in [Section 4.2](#), we obtained the proportions of each building type as: residential (54.34%), school (10.13%), office (9.39%), shopping (3.58%), hotel (2.46%), and hospital (15.21%). These proportions were then used to determine the building inference sequence as: 1. residential building (including urban village), 2. school, 3. office, 4. shopping building, 5. hotel, and 6. hospital. Following [Section 4.2](#) (1), we built the spatial relations between taxi passengers' activities and urban buildings. In undertaking such an analysis, it was important to determine the optimal buffer distance for different types of buildings. Therefore, we randomly selected seven buildings in the study area, and let these buildings represent the building types, calculating the number of pick-ups and drop-offs within the buffer zones of these buildings (between 1 m and 40 m) over a period of 3 h (between 9:00 am and 12:00 noon), which is usually the peak time of human activities. The results are shown in [Table 1](#).

[Table 1](#) illustrates the frequency distribution of the number of taxi passengers' pick-ups and drop-offs and walking distance between those points and selected buildings in the case study area. The results demonstrate that most taxi passengers prefer to walk from pick-ups or drop-offs sites to buildings (their destinations) within 30 m, and that this distance increases as the frequency of taxi passenger pick-ups and drop-offs decreases.

In accordance with the procedure set out in [Section 4.2\(2\)](#), we selected 2000 buildings to serve as training samples. The number of training samples for each building type (residential buildings, office complexes, shopping buildings, hotels, hospitals, and schools) were 1087, 203, 187, 72, 49, and 304, respectively. We then calculated the average number of pick-ups and drop-offs in accordance with formulas 1 – 4, extracting the average number of real-time WeChat users from different types of training samples on the Monday and the Saturday in question. The statistical results of the training samples are shown in [Figures 3](#) and [4](#), wherein (a) refers to residential buildings, (b) to office complex, (c) to shopping buildings, (d) to hotels, (e) to hospitals, and (f) to schools. On the X axis, 1–24 represents the time period of Monday, and 25–48 stands for the time period of Saturday; the Y axis of [Figure 3](#) shows the time of taxi passengers' pick-up or drop-off, and the Y axis of [Figure 4](#) displays the number of WeChat users.

From the results reported at [Figures 3](#) and [4](#), we identified the peak time of human activity (taxi WeChat activity) in residential, urban villages, office, shopping centers, hospital, and school buildings, which are also described in [Table 2](#).

Table 1. The number of pick-ups and drop-offs within building buffer zones (0–40 m) over 3 h (9:00 am – 12:00 noon).

Building name	0 – 10 m	10 – 20 m	20 – 30 m	30 – 40 m
Grandview Mall (shopping center)	60; 52; 42	63; 154; 56	158; 312; 256	0; 2; 5
Xian village (urban village)	21; 23; 28	159; 86; 52	180; 134; 110	1; 3; 8
Rui xin yuan (residential building)	10; 8; 9	58; 69; 87	120; 96; 55	0; 1; 5
Tianhe High School (school)	22; 21; 42	101; 90; 60	133; 125; 66	12; 12; 23
Third Affiliated Hospital of Sun Yat-sen University (hospital)	21; 25; 32	127; 112; 75	134; 148; 116	14; 18; 17
Times Plaza (office complexes)	12; 15; 16	80; 64; 107	120; 110; 132	16; 24; 32
Hanting Hotel (hotel)	20; 50; 60	71; 62; 52	74; 62; 41	11; 14; 12

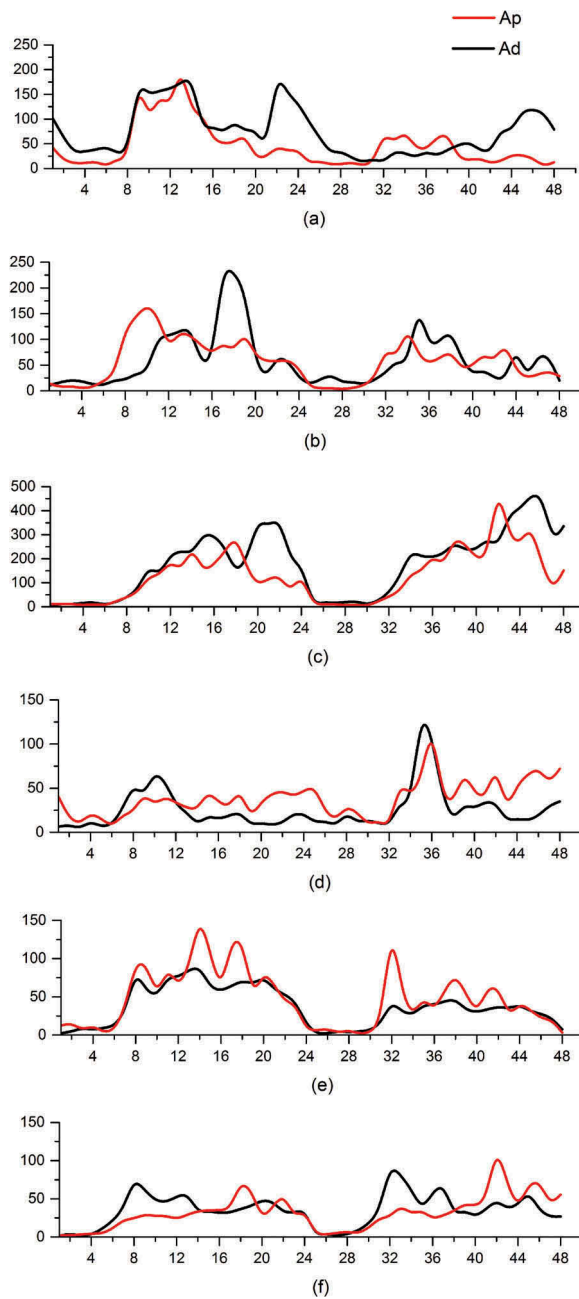


Figure 3. Ap and Ad curves, representing temporal variations in pick-ups and drop-offs, in relation to the 6 building type sample points.

The procedures used in this paper both in order to cluster taxi pick-ups or drop-offs and to perform spatial point density analysis in relation to the WeChat data were established on the basis of [Section 4.2](#). We also undertook a spatial overlay analysis of taxi data, WeChat data, and the building footprints of Tianhe District in terms of formula (6), the results of which are presented at [Figure 5\(a–e\)](#), for (a1-2) residential buildings

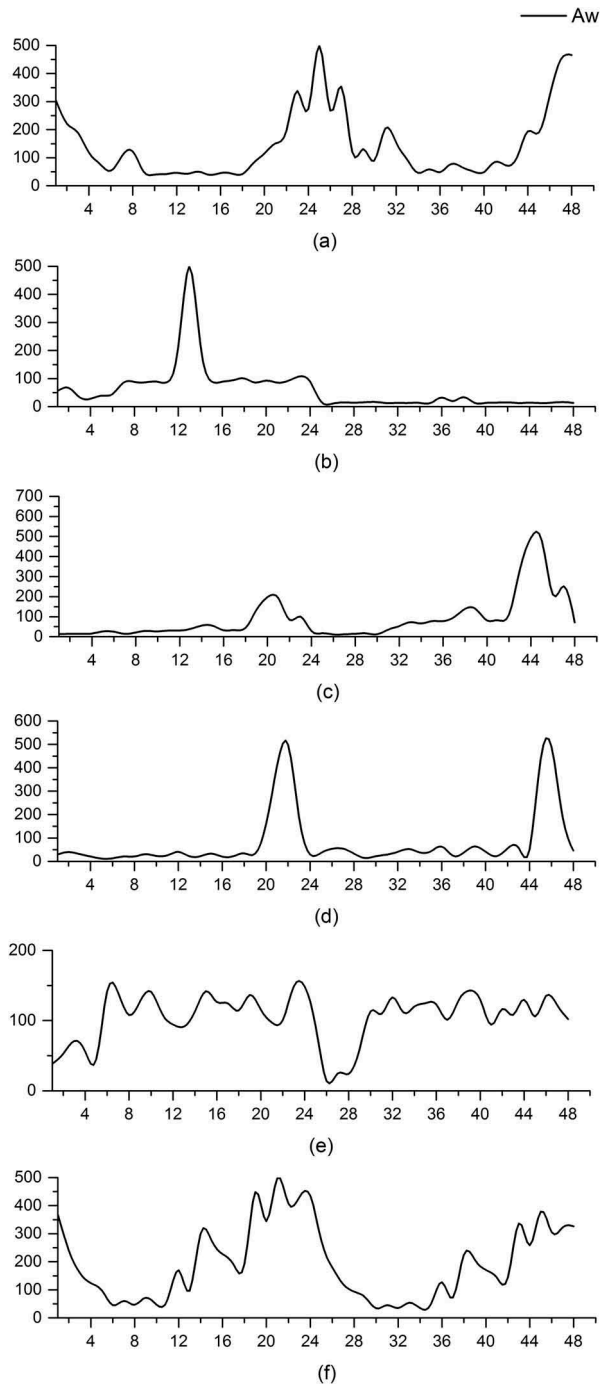


Figure 4. Aw curves, representing temporal variations in the number of WeChat users for the 6 building type sample points.

Table 2. Taxi passenger and WeChat user activity peak times for 6 building type sample points.

Building type	The peak time of taxi passengers' activities	The peak time of WeChat users' activities
Residential buildings	Pick-ups 7:00–9:00 a.m. on Monday, drop-offs 10:00 p.m.–3:00 a.m. on Monday and Saturday	11:00 p.m.–3:00 a.m. on Monday and Saturday
Office	Pick-ups 5:00–7:00 p.m. on Monday	12:00 p.m.–2:00 p.m. on Monday
Shopping centers	Pick-ups 9:00 p.m.–12:00 a.m. on Saturday	7:00 p.m.–12:00 a.m. on Saturday
Hotels	Pick-ups 10:00–11:00 a.m. on Monday and Saturday	6:00–8:00 a.m. on Monday
Hospital	Drop-offs at 7:00–9:00 a.m. on Monday	-
School	Pick-ups and drop-offs 5:00–6:00 p.m. on Saturday	10:00 p.m.–12:00 a.m. on Saturday

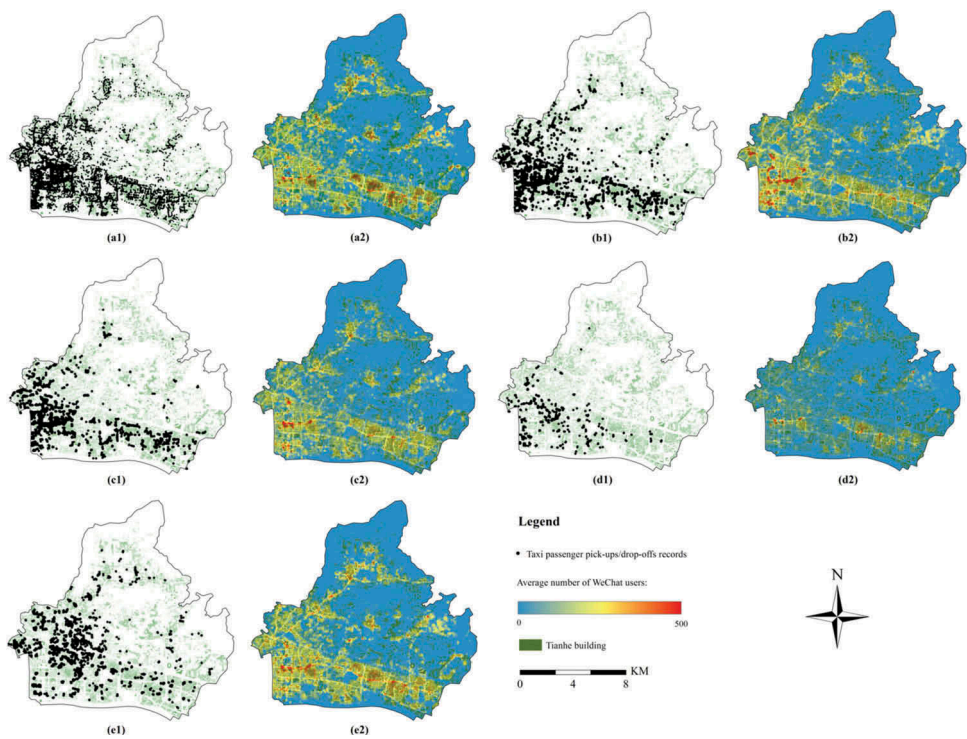


Figure 5. Results of spatial overlay analysis (taxi data and WeChat data) at peak times for (a1-2) residential buildings (including urban villages), (b1-2) office complexes, (c1-2) shopping buildings, (d1-2) hotels, and (e1-2) schools.

(including urban villages), (b1-2) office complexes, (c1-2) shopping buildings, (d1-2) hotels, and (e1-2) schools.

The final inference results are shown in Figure 6. These were arrived at in accordance with the evaluation function in Section 4.2(4). As shown in Figure 6, 49,823 buildings were able to defined and labeled (using color coding) in terms of their possible buildings functions. The overall detection rate (the number of identified buildings as a proportion of the total number of buildings in the study area) for these results was 72.22%. Our results suggest that most residential buildings were located in the area that is south of Guangyuan Expy and north of Linjiang Ave. A




Figure 6. Inferred building functions in the Tianhe District (Guangzhou, China).

high number of residential buildings were found to be located along Linjiang Ave. Urban villages were scattered throughout the urban core and suburban areas, surrounding the new CBD and high-rise residential areas. Office complexes and shopping centers were found to be located in the new CBD (e.g., the Zhujiang New Town and Zhongxin Square) or mixed with residential buildings. Hotels accounted for a low proportion of all buildings in Tianhe District (Guangzhou City) and were shown to be mingled with office complexes and shopping buildings. Hospital buildings have a unique function, and are as such were shown to be mostly built around residential buildings. Most schools were found to be located north of Guangyuan Expy. Some mixed functional buildings were also identified along the main road (i.e., Guangyuan Expy, Guangzhou Street).

We selected test samples in relation to a range of different spatial scales in our study area, subsequently comparing our inferred results against the ground-truth investigation data. The results of this verification process are shown in Table 3.

Table 3. Verification of the inferred results at different spatial scales.

Site	Size	N1	N2	N3	Percentage of detection, correctness
	500 × 500 m	205	44	27	78.53%, 65.37%
	1000 × 1000 m	356	20	30	94.38%, 85.95%

N1, N2, and N3 are the number of buildings counted, number of unidentified buildings, and number of inferred incorrect buildings, respectively

6. Discussion

6.1. The size of the training sample influences the inferred detection rate

To examine how the number of samples influences the inferred detection rate, we calculated detection rates for a range of different sample sizes (Figure 7). We found that the best detection rate (65.68%) was acquired in relation to large samples. The detection rate of building functions stabilized when the sample size was more than 1500 – as such, the minimum sample size in such a study ought to be no less than 1500. This analysis indicates that our estimation, which selected 2000 training sample sets, was of a desirable scale.

6.2. Multi-source data integration advantage

From Figure 6 and Tables 3 and 4, we also found the detection and correct rates to be positively correlated with the frequency of human activities. In other words, when the frequency of taxi pick-ups and drop-offs and the density of WeChat users are high in

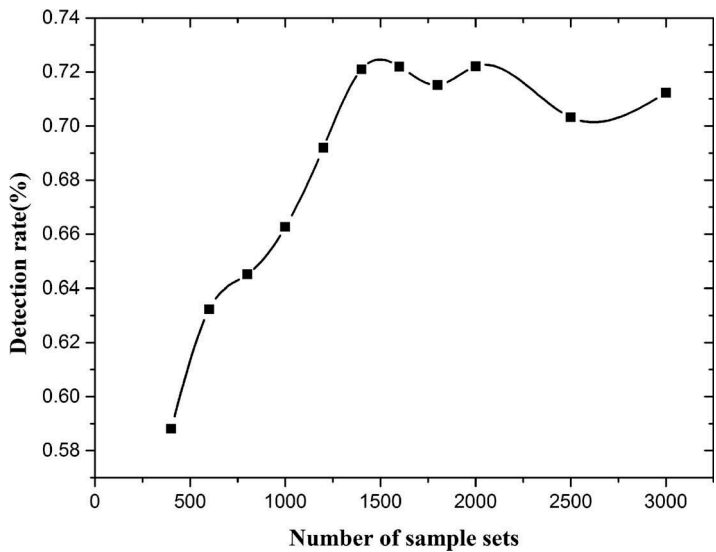


Figure 7. Inferred detection rates for different sample sizes.

Table 4. Detection rates of each type building using different data sources.

Building type	DR1	DR2	DR3
Residential	82.12%	70.32%	65.52%
Urban villages	83.16%	71.14%	68.25%
Office	71.43%	56.73%	53.85%
Shopping centers	72.89%	42.21%	61.11%
Hotels	53.92%	32.11%	57.32%
Hospitals	40.25%	-	40.25%
Schools	42.38%	31.45%	31.36%

DR1, DR2, and DR3 stand for Detection rate (multi-source data), Detection rate (WeChat) and Detection rate (taxi)

buildings, the detection and correct rates are also high. This can be seen in relation to the detection rates for residential buildings, urban villages, office complexes, and shopping buildings, which were all above 70%. Similarly, the detection rates for hotels, hospitals, and schools were 53.92%, 40.25%, and 42.38%, respectively – lower detection rates that reflect a lower frequency and density of activities. The inferred results in the urban core (for example, Pearl River new town, Zhongxin square) were also more accurate than those in the suburban areas.

In summary, the analysis indicates that each type of big data has its own (single attribute) limitations. As such, WeChat data or Taxi data only reflect part of people's travel behavior, leading to the detection rate of some buildings being low and that of others being particularly high when a single data source was used. Fusing multi-source big data provides a greater breadth of information about human activities, and combining sources is of great advantage to research like ours that aims to infer urban building functions. These findings may provide some help for other urban studies (such as urban vibrancy) at a fine scale. Moreover, our research will help to support the development of future urban planning to address locally specific issues.

There are some limitations in our work. Firstly, an increasing number of citizens in Guangzhou use private cars, public buses, special cars that are related to companies like Uber and DiDi, and the metro system. Given this spread, taxi data and WeChat data only account for a portion of human activities. The inference accuracy rates are mainly dependent on the density of pick-up and drop-off records and WeChat users. The inference accuracy rate with respect to hospitals and schools was found to be relatively low; the possible reason may be that people accessing these buildings do not often take taxis and rarely use WeChat. Secondly, the time of multi-source data is not always synchronized. In this study, taxi GPS trace records were collected in 2014; other data (WeChat user location records, Tianhe building footprints, and the POIs from Baidu Map) were collected in 2015. This may have resulted in some errors in the inference process.

7. Conclusions

In this paper, we proposed a density-based method to infer building functions by integrating multi-source 'big data'. After implementing our proposed method in relation to Tianhe District, the final detection rate reached 72.22%, the accuracy rate was above 65%. We also speculatively put forward two factors that may have influenced the inferred detection and accuracy rates. The first factor was the size of the training samples, which we concluded should be at least 1500. The second factor was that the inferred detection and accuracy rates obtained from two sources of big data were markedly better than that using single-source big data.

In summary, the integration and use of multi-source big data in studies of urban land use can assist researchers in acquiring more detailed information about actual land use in a city than can be done using a single source of big data. Our evaluation and comparison confirmed that WeChat data has an advantage in identifying building types through inference-based methods – this was clearly the case in relation to residential buildings, office, and urban villages in this study. Conversely, shopping centers, hotels, and hospitals proved easier to identify using taxi data. The evaluation performed here also indicates that building functions are better inferred when using multi-source big data for urban

environments characterized by a high density of human activities and building functions. Our study provides precise delineating information on actual land use in Tianhe District (Guangzhou, China). This research offers an example of our understanding of cities might be advanced using intuitive approaches. Such approaches are able to provide information that is useful for evaluating the effectiveness of the planning schemes, ultimately assisting policy makers in improving the effectiveness of urban planning.

In the future, we will apply this methodology to more cities in developing and developed countries. Moreover, We will also integrate other big data (i.e., smart card data from bus and metro) and utilize multi-disciplinary knowledge to infer the mixed functions of urban buildings.

Acknowledgments

The authors would like to thank Assoc Prof. Shawn Laffan, and two reviewers who gave us so many useful comments and suggestions for the revision. This study was supported by the National Natural Science Foundation of China (Grant No. 41671398), the Key National Natural Science Foundation of China (Grant No.41531176) and the National Natural Science Foundation of China (Grant No. 41601420).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study was supported by the National Natural Science Foundation of China [Grant No. 41671398], the Key National Natural Science Foundation of China [Grant No. 41531176] and the National Natural Science Foundation of China [Grant No. 41601420].

ORCID

Xiaoping Liu  <http://orcid.org/0000-0003-4242-5392>

Yu Liu  <http://orcid.org/0000-0002-0016-2902>

Xia Li  <http://orcid.org/0000-0003-3050-8529>

Shaoying Li  <http://orcid.org/0000-0003-3050-8529>

References

- Berman, J.J., 2013. *Principles of big data: preparing, sharing, and analyzing complex information*. Newnes: Morgan Kaufmann Publishers Inc.
- Daszykowski, M., Walczak, B., and Massart, D.L., 2001. Looking for natural patterns in data: part 1. Density-based approach. *Chemometrics and Intelligent Laboratory Systems*, 56, 83–92. doi:10.1016/S0169-7439(01)00111-3
- De Wit, A. and Clevers, J., 2004. Efficiency and accuracy of per-field classification for operational crop mapping. *International Journal of Remote Sensing*, 25, 4091–4112. doi:10.1080/01431160310001619580

- Ertöz, L., Steinbach, M., and Kumar, V. 2003. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In: *Third Siam International Conference on Data Mining*, May, San Francisco, CA, USA, 47–58.
- Ester, M., et al., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*, 96, 226–231.
- Frias-Martinez, V. and Frias-Martinez, E., 2014. Spectral clustering for sensing urban land use using Twitter activity. *Engineering Applications of Artificial Intelligence*, 35, 237–245. doi:10.1016/j.engappai.2014.06.019
- Fu, T., 2011. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24, 164–181. doi:10.1016/j.engappai.2010.09.007
- Gan, G., Ma, C., and Wu, J., 2007. *Data clustering: theory, algorithms, and applications*. ASA-SIAM Series on Statistics and Applied Probability; SIAM.
- Goodchild, M.F. and Michael, F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69, 211–221. doi:10.1007/s10708-007-9111-y
- Heipke, C., 2010. Crowdsourcing geospatial data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65, 550–557. doi:10.1016/j.isprsjprs.2010.06.005
- Hu, S. and Wang, L., 2013. Automated urban land-use classification with remote sensing. *International Journal of Remote Sensing*, 34, 790–803. doi:10.1080/01431161.2012.714510
- Huang, K., et al., 2013. An improved artificial immune system for seeking the Pareto front of land-use allocation problem in large areas. *International Journal of Geographical Information Science*, 111, 922–946. doi:10.1080/13658816.2012.730147
- Jiang, S., et al., 2015. Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems*, 53, 36–46. doi:10.1016/j.compenvurbsys.2014.12.001
- Kang, C., et al., 2012. Intra-urban human mobility patterns: an urban morphology perspective. *Physica A: Statistical Mechanics and Its Applications*, 391, 1702–1717. doi:10.1016/j.physa.2011.11.005
- Lin, Y., de Meulder, B., and 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, 3583–3598. doi:10.1177/0042098010396239
- Liu, X., et al., 2014. Simulating urban growth by integrating landscape expansion index (LEI) and cellular automata. *International Journal of Geographical Information Science*, 28, 148–163. doi:10.1080/13658816.2013.831097
- Liu, X., et al., 2015a. Revealing travel patterns and city structure with taxi trip data. *Journal of Transport Geography*, 43, 78–90. doi:10.1016/j.jtrangeo.2015.01.016
- Liu, Y., et al., 2015b. Social sensing: a new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers*, 105, 512–530. doi:10.1080/00045608.2015.1018773
- Liu, Y., et al., 2012. Urban land uses and traffic ‘source-sink areas’: evidence from gps-enabled taxi data in shanghai. *Landscape & Urban Planning*, 106 (1), 73–87.
- Mesev, V., 1998. The use of census data in urban image classification. *Photogrammetric Engineering and Remote Sensing*, 64, 431–436.
- Moran, M.S., Inoue, Y., and Barnes, E., 1997. Opportunities and limitations for image-based remote sensing in precision crop management. *Remote Sensing of Environment*, 61, 319–346. doi:10.1016/S0034-4257(97)00045-X
- Pan, G., et al., 2013. Land-use classification using taxi GPS traces. *IEEE Transactions on Intelligent Transportation Systems*, 14, 113–123. doi:10.1109/TITS.2012.2209201
- 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, 1988–2007. doi:10.1080/13658816.2014.913794
- Platt, R.V. and Rapoza, L., 2008. An evaluation of an object-oriented paradigm for land use/land cover classification. *The Professional Geographer*, 60, 87–100. doi:10.1080/00330120701724152
- Qi, G., et al. 2011. Measuring social functions of city regions from large-scale taxi behaviors. In: *Pervasive Computing and Communications Workshops (PERCOM Workshops)*, 384–388..

- Rogan, J. and Chen, D., 2004. Remote sensing technology for mapping and monitoring land-cover and land-use change. *Progress in Planning*, 61, 301–325. doi:[10.1016/S0305-9006\(03\)00066-7](https://doi.org/10.1016/S0305-9006(03)00066-7)
- Schmit, C., Rounsevell, M.D.A., and La Jeunesse, I., 2006. The limitations of spatial land use data in environmental analysis. *Environmental Science & Policy*, 9, 174–188. doi:[10.1016/j.envsci.2005.11.006](https://doi.org/10.1016/j.envsci.2005.11.006)
- Shen, J. and Cheng, T., 2015. Clustering analysis of officer's behaviours in london police foot patrol activities. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-4/W2, 143–146. doi:[10.5194/isprsannals-II-4-W2-143-2015](https://doi.org/10.5194/isprsannals-II-4-W2-143-2015)
- Soto, V. and Frías-Martínez, E. 2011. Automated land use identification using cell-phone records. In: *Proceedings of the 3rd ACM international workshop on MobiArch, HotPlanet '11*, 28 June, Bethesda, MD, 17–22.
- Tang, J., et al., 2015. Uncovering urban human mobility from large scale taxi GPS data. *Physica A: Statistical Mechanics and Its Applications*, 438, 140–153. doi:[10.1016/j.physa.2015.06.032](https://doi.org/10.1016/j.physa.2015.06.032)
- Wang, Y.P., Wang, Y., and Wu, J., 2009. Urbanization and informal development in China: urban villages in Shenzhen. *International Journal of Urban and Regional Research*, 33, 957–973. doi:[10.1111/ijur.2009.33.issue-4](https://doi.org/10.1111/ijur.2009.33.issue-4)
- Wu, L., et al., 2014. Intra-urban human mobility and activity transition: evidence from social media check-in data. *PLoS One*, 5, e97010. doi:[10.1371/journal.pone.0097010](https://doi.org/10.1371/journal.pone.0097010)
- Yang, X. and Lo, C.P., 2002. Using a time series of satellite imagery to detect land use and land cover changes in the Atlanta, Georgia metropolitan area. *International Journal of Remote Sensing*, 23, 1775–1798. doi:[10.1080/01431160110075802](https://doi.org/10.1080/01431160110075802)
- 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*. ACM, 186–194.
- Zhan, X., Ukkusuri, S.V., and Zhu, F., 2014. Inferring urban land use using large-scale social media check-in data. *Networks and Spatial Economics*, 14, 647–667. doi:[10.1007/s11067-014-9264-4](https://doi.org/10.1007/s11067-014-9264-4)
- Zhong, C., et al., 2014. Inferring building functions from a probabilistic model using public transportation data. *Computers, Environment and Urban Systems*, 48, 124–137. doi:[10.1016/j.compenvurbsys.2014.07.004](https://doi.org/10.1016/j.compenvurbsys.2014.07.004)