Geo-spatial Information Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tgsi20

Multi-scale cross-city community detection of urban agglomeration using signaling big data

Wenbo Yu, Zhenfeng Shao, Xiao Huang, Deren Li, Yewen Fan & Xiaodi Xu

To cite this article: Wenbo Yu, Zhenfeng Shao, Xiao Huang, Deren Li, Yewen Fan & Xiaodi Xu (20 Jun 2023): Multi-scale cross-city community detection of urban agglomeration using signaling big data, Geo-spatial Information Science, DOI: 10.1080/10095020.2023.2197763

To link to this article: https://doi.org/10.1080/10095020.2023.2197763

Published online: 20 Jun 2023.

GEO-SPATIAL INFORMATION SCIENCE
https://doi.org/10.1080/10095020.2023.2197763

Taylor & Francis
Taylor & Francis Group

Multi-scale cross-city community detection of urban agglomeration using signaling big data

Wenbo Yu[a], Zhenfeng Shao[b], Xiao Huang[c], Deren Li[b], Yewen Fan[b] and Xiaodi Xu[a]

[a]School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China; [b]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China; [c]Department of Geosciences, University of Arkansas, Fayetteville, AR, USA

ABSTRACT

Many existing efforts have taken advantage of large-scale spatial-temporal data to partition cities via constructed human interaction networks. However, few studies focus on communities emerging between adjacent cities in big urban agglomerations, which we call "cross-city" communities. In this study, we introduce a novel framework to detect cross-city communities in urban agglomerations under different scales leveraging a large number of fine-grained mobile signaling data aiming to break the original administrative boundaries. Taking the Pearl River Delta (PRD) urban agglomeration in China as study area, we investigate the existence of potential communities at three scales, i.e. city-group level, city level and sub-city level. The partition results are expected to benefit transportation planning, urban zoning and administrative boundary re-delineation. The results from our study highlight the necessity of considering cross-city communities and their scale effects when examining urban spatial interactions.

ARTICLE HISTORY
Received 2 March 2022
Accepted 29 March 2023

KEYWORDS
Cross-city communities; community detection; mobile big data; human interaction network; scale effect

1. Introduction

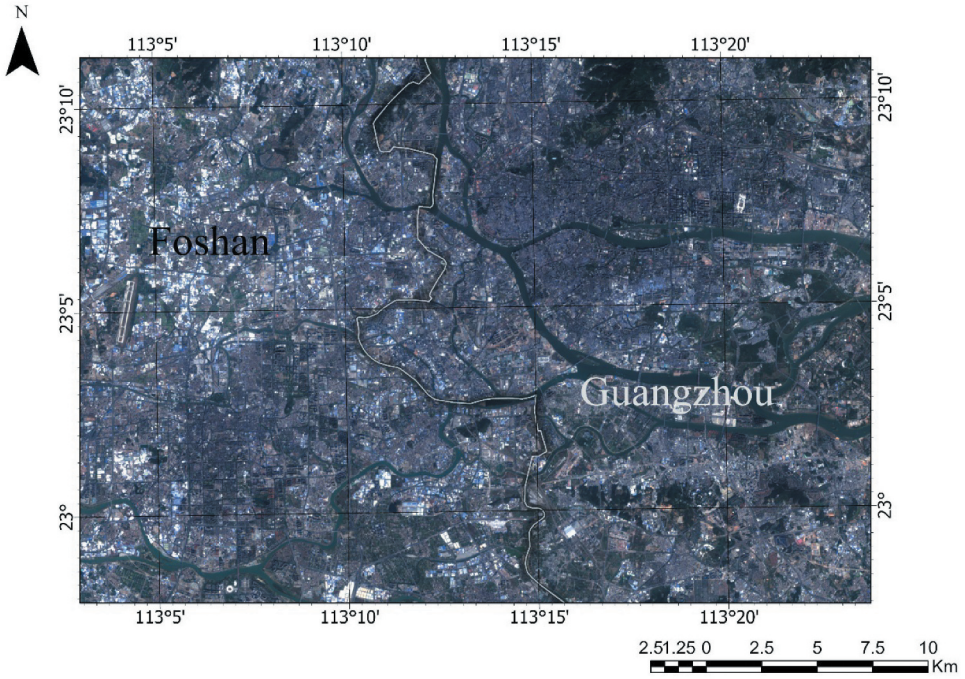
Urban agglomeration, a term that describes the clustering of certain cities and several other closely interrelated cities (Fang and Yu 2017), has become increasingly crucial in the development of most countries (Boix, Veneri, and Almenar 2012). The United Nations' forecast suggests that 75% of the world's population will be living in cities by 2050, and the largest 40 urban agglomerations, despite their small spatial coverage, will have 18% of the total population, 66% of global economic activities, and approximately 85% of technological innovations (United Nations Human Settlements Programme, 2016). Such trends indicate that urban agglomerations are becoming a vital geographic entity for nations to sustain economic growth and development (Rahayu, Haigh, and Amaratunga 2018).

The arrangement of urban space is of great importance to effective management and governance within the boundary of a city (Zhou et al. 2016). The agglomerations, given their complex and dynamic nature (Fang and Yu 2017), have posed great challenges to urban development. The increased interconnectivity of trade, commerce, social connections as well as political activities blurs the boundaries between cities and peripheral regions and makes traditional city boundaries, often imposed by administrative needs, greatly obsolete. Besides, traditional means of zoning city areas also fall short of fully appreciating the newly emerged urban spatial form (Ebner 2008; Gu 2023). In this situation, still taking the administrative boundary of a city (district or county) as the spatial unit to study urban areas lacks advantages. For example, some adjacent areas of cities in urban agglomerations have merged to some extent and formed more stable communities, which we call "cross-city" communities in this work. Taking these cross-city communities into account will benefit public services such as the development of transportation, public security and other infrastructure in urban planning. For business, this is also conducive to the location selection of the company and its service sites. As a result, a new boundary demarcation framework is urgently needed to understand urban zoning within the urban agglomeration efficiently, in a rational manner that is more consistent with people's interactive behavior.

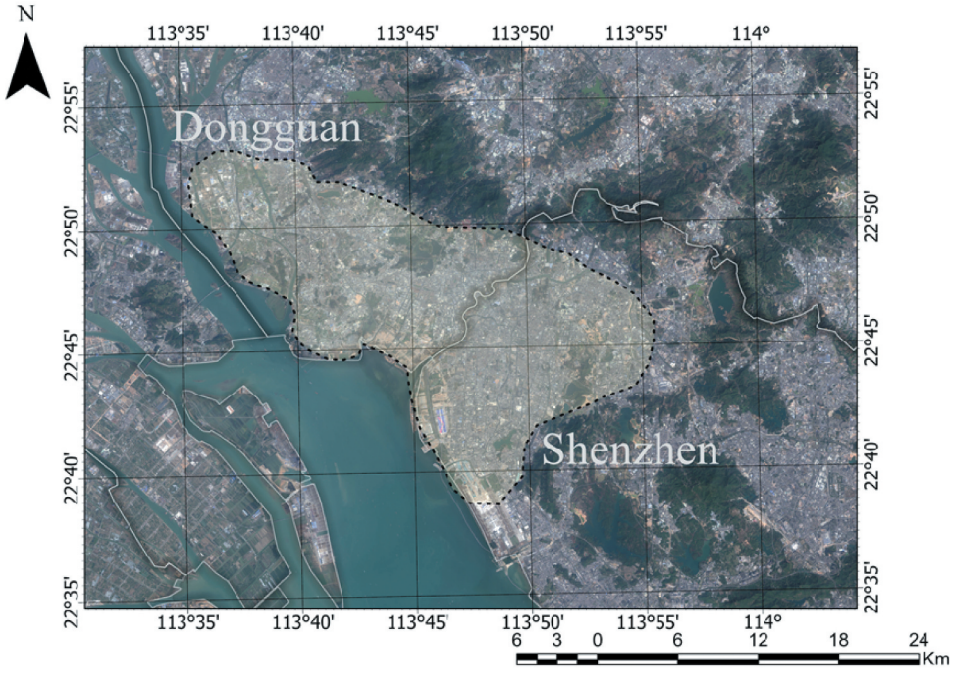
Existing efforts have attempted to detect urban communities using a variety of methods and data (Tang et al. 2015; Grauwin et al. 2015; Frias-Martinez and Frias-Martinez 2014; Brelsford et al. 2019; Thomas 2012). Earlier studies tend to partition urban spaces based on certain natural or man-made features of land surfaces, for example with remote sensing images. With the development of urban perception technology, it has become possible to study the

CONTACT Zhenfeng Shao shaozhenfeng@whu.edu.cn

human-earth relationship from a social and data-driven perspective. Such related typical methods include clustering the urban space units based on certain constructed human mobility pattern features, or segmenting a human-mobility network associating human movements with urban space. The former tends to break the urban space apart and then groups similar space units together in a bottom-up way, while the latter treats the entire city as a whole at first and then try to separate the more tightly connected parts out, which is an up-bottom way. Although many studies have succeeded in detecting urban communities and revealing its spatial structures, few efforts focus on such cross-city communities, and the reasons we assume are as follows. 1. Cross-city communities are something of a novelty. Its formation requires the rapid development of cities in urban agglomerations so that their built-up areas are connected to each other (As shown in Figure 1 where the built-up areas of adjacent cities are merging). Satellite cities that rely



(a)



(b)

Figure 1. (a) Satellite images suggests that the built-up areas of Guangzhou and Foshan, China are merging. (b) the merging area of northeast Shenzhen and southwest Dongguan, China (the yellow area).

on transportation to connect may not be the case. 2. Discovering cross-city communities requires fine-grained geospatial data, which is difficult to obtain in traditional datasets. This is made possible by emerging data sources such as mobile signaling data or social media data brought about by urban perception technologies. 3. The above process involves big data processing, which also requires big data technology developing in recent years.

In this study, we proposed a novel framework to detect fine-grained urban communities under multiple scales. We used massive mobile phone signaling data to extract human trajectories and segmented the constructed human interaction network to detect communities by a modified spectral clustering method. In summary, our contributions are as follows: 1. We demonstrated the existence of cross-city communities in a fine-grained and data-driven way in this case study in PRD urban agglomeration in China. 2. We studied and discussed the scale-effect on the detected communities. 3. We compared the communities detected on working days and holidays and found most of the communities are relatively stable.

2. Related works

Urban partitioning or community detection is not a new research field in urban science. In fact, it has been a popular topic for a long period. However, few researches focus on cross-city community detection in urban agglomerations as no available data sources and inadequate urban agglomeration evolution. Existing efforts have attempted to detect communities using a variety of methods and data. Researches in earlier years tended to divide urban space based on certain properties of the land surface in the city. With the development of urban perception technology, it has become possible to study the human-earth relationship from a data-driven perspective, and a number of urban community detection studies based on these new data sources (mobile phone data, social media data, etc.) have emerged. These methods can be divided into spatial clustering method and network segmentation method.

The spatial clustering method divides the urban space into several subunits, and then clustering the units into different groups according to their features, so as to realize the division of urban space. Tang et al. made use of taxi GPS data and DBSCAN method to reveal the urban human mobility patterns and its spatial distribution in Harbin, China (Tang et al. 2015). Grauwin et al. used mobile phone data to detect users' phone usage pattern, and confirmed the connection between temporal activity profile and land usage in three global cities (Grauwin et al. 2015). Tao et al. constructed a four-dimensional tensor from taxi trajectories to identify human's activity patterns and reclassified the urban regions based on the space clustering formed by the space factor matrix and core tensor (Tao et al. 2019). Spatial clustering methods are usually easy to implement, but the results are greatly affected by the feature composition, which may lead to completely different clustering results, thus weakening the comparability between different studies. Besides, it ignores the relationships between different spatial units, which leads to the failure of making full use of the information from the data.

The network segmentation method constructs a network regarding the spatial units as nodes and the relationships between places as edges and segment the network into sub-networks to partition urban spaces. Related algorithms include spectral clustering, modularity optimization (Girvan and Newman 2002; Blondel et al. 2008), Infomap (Rosvall and Bergstrom 2008) and Label propagation algorithm (LPA) (Raghavan, Albert, and Kumara 2007), etc. For example, Vanessa et al. focused on the characterization of land use by using spectral clustering methods based on geolocated tweet data in three metropolitan cities (Frias-Martinez and Frias-Martinez 2014). Christa et al. used millions of tweet data to develop a spatially embedded network of communication, and then use Louvain (a modularity method) to explore regional and urban delineation in the United States (Brelsford et al. 2019; Jia et al. 2019). Ye et al. employ the Infomap to identify the hierarchical community in city roads based on OpenStreetMap (OSM) roads and points-of-interest (POI) data in Guangzhou, China (Hong and Yao 2019). Huang et al. used LPA and other community detection methods to analyze the relationship between POIs and network communities of human mobility on urban taxi system in Shanghai and Beijing (Huang et al. 2018). The network partitioning methods take into account the relationships between spatial units throughout human activities thus can result in better regionalization results as generally believed.

In this study, we deploy a modified spectral clustering method on NCut to partition the constructed human interaction network, and the reasons are as follows: 1. The spectral clustering method is based on strict mathematical theory and its results are easy to interpret. 2. The method is easy to implement and to be modified and integrated with various constraints. 3. This is an unsupervised data-driven approach that avoids human interference. 4. This method is a stable classical method and has been proved effective in a variety of related studies. In the following sections, we will introduce our data and study area in section 3, our methods in section 4 and experimental results in section 5. Sections 6 and 7 are discussion and conclusions.

3. Data and study area

3.1. Study area

The Pearl River Delta (PRD) urban agglomeration (Figure 2), located in the Guangdong province in China, consists of nine cities that include Guangzhou, Foshan, Zhaoqing, Shenzhen, Dongguan, Huizhou, Zhuhai, Zhongshan, and Jiangmen, with a total area of about 55,368.7 square kilometers and a total population of 64.5 million in 2019, accounting for 30.8% of the area and 44.11% of the population of Guangdong Province. The GDP (Gross Domestic Product) of PRD urban agglomeration is 8.95 trillion Yuan ($1.38 trillion) in 2020, accounting for 80.83% of that in Guangdong Province and 8.8% of that in China. The PRD urban agglomeration is one of the most dynamic economic zones in the Asia-Pacific region, an advanced manufacturing and modern service industry base, and a regional technological innovation and research center in China. The PRD urban agglomeration is the core hinterland of the Guangdong-Hong Kong-Macao Greater Bay Area (Greater Bay Area, GBA), which is one of the four largest bay areas in the world. In recent years, cross-city transportation within the GBA has been significantly developed.

3.2. Data description

In this study, we used mobile phone signaling data that cover the aforementioned nine cities in the PRD urban agglomeration. Our data can be divided into two parts, including a workday set on September 29, 30 and a holiday set on October 3, 4, both in the year of 2020. These data were obtained from China Unicom, one of the China's largest mobile communication companies, and all the data were anonymized by the supplier. The two datasets contain 1.5 billion and 1 billion records, respectively. The fields of the data include "time", "user ID", "latitude", "longitude", and "base station ID" (as shown in Table 1). These data are generated by communication between the user's mobile phone and the base stations. Specifically, when a user makes calls, sends or receives text messages, uses internet services, moves, or stays at the same place for more than half an hour, his/her phone will search for and communicate with the nearest base station, and a signaling record that contains the geolocated information of the base station and the user's ID will be stored in the station. Therefore, if the user keeps moving, a series of corresponding signaling records can be retrieved to construct user's trajectories.

4. Method

4.1. Proposed framework

Our proposed framework is shown in Figure 3. In this study, we first extract human trajectories from a massive phone signaling dataset. Then we build a human interaction network. Here we add some spatial constraints to the interaction network to construct the similarity matrix for spectral clustering algorithm. In this progress the bandwidth of Gaussian kernel for spatial constraints is gradually increased to

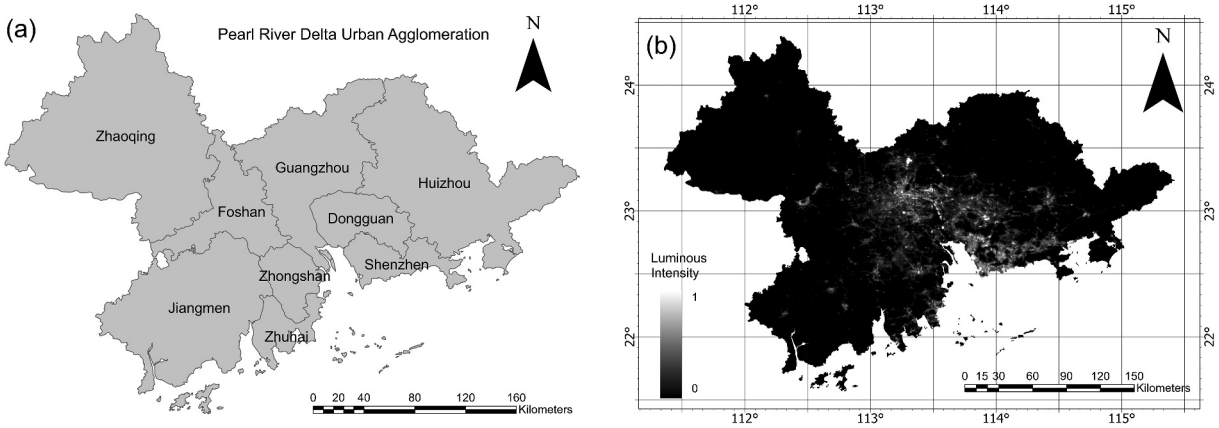


Figure 2. The Pearl River Delta urban agglomeration, China(a). The luminous image of the Pearl River Delta. The gray value is normalized to 0–1 after histogram equalization.

Table 1. Example signaling data.

Time	User ID	Latitude	Longitude	Cell ID
29 September 2020 00:00:00	827316775	22.78542	113.67184	−463400143
29 September 2020 00:00:00	−1461270008	23.13149	113.38373	−523253145
29 September 2020 00:00:00	−1830412423	23.07440	113.26700	−521618484
29 September 2020 23:59:59	2077573988	22.21457	113.44351	−405072247

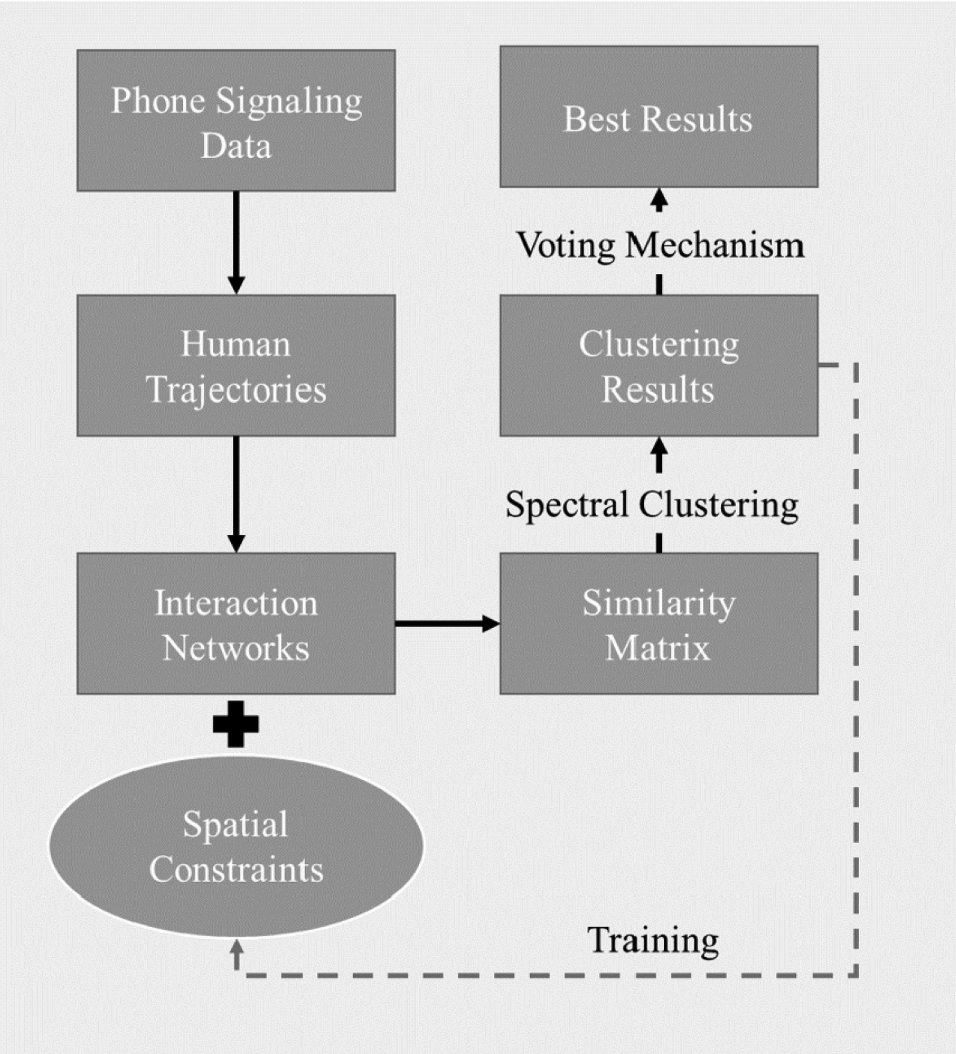


Figure 3. Proposed framework.

train the clustering model. Finally, a voting mechanism is applied to elect the optimal clustering results from all the outputs.

4.2. Obtaining human trajectories

It is challenging to extract the users' trajectories directly from raw signaling data obtained from the base stations. Several preprocessing steps need to be performed. In general, there are three commonly-known issues: (1) Signal occlusion. If a user is moving in an area with no phone signal, the base station cannot receive any message from the phone, resulting in the leapfrog of the user's location. This problem can be mitigated by dividing the "moving chain" with only "continuous trace" left. (2) Signal drift. Sometimes a user's mobile phone may fail to find the nearest base station but find farther ones instead. This problem is manifested in the user's trajectory as a sharp singularity suddenly appearing at a certain point in a continuous trace. In other words, the user's moving speed suddenly becomes abnormally large. (3) Signal that switches repeatedly. If a user is at a cross-area covered by signals from multiple base stations, that user's phone may quickly and repeatedly communicate with all these stations, leading to a signal oscillation pattern in the trajectory.

The problem of signal occlusion and switching signals are solved by the method introduced by Cai and Zhong (2018), in which the steps of handling the signal occlusion are: (a) sorting the user's records according to the timestamp; (b) using the segmentation index to obtain the smallest continuous travel trajectory; (c) using a threshold method to clean each trajectory by removing drifting points; (d) removing the points that switch repeatedly. The calculation of the segmentation index in step (b) follows:

$$S = \mathrm{F}(p,t) \times \left| \ln\left(\frac{v}{\bar{v}} + \frac{1}{10000}\right) \right| + \mathrm{R}(t) \times \arctan\left(\frac{\mathrm{t}}{\bar{t}}\right) \tag{1}$$

where p is the number of all the signaling records of the current user, t is the interval of the current record,

v is the instantaneous speed of the current record, $\bar{v}$ is the average speed of all the records, $\bar{t}$ is the overall interval of all the records, $\mathrm{F}(p,t) = \begin{cases} a \times \frac{t-c}{|t-c|}, p > b \\ 0, p \le b \end{cases}$, $\mathrm{R}(t) = \begin{cases} 0, t \le c \\ d, t > c \end{cases}$, a, b, c, d are constants determined by the elbow law. Besides, the signal drift issue is handled by a speed threshold.

4.3. Constructing human interaction networks

We use the trajectories obtained from the signaling datasets to construct a human interaction network. As users always move from the cell area of one base station to another, we take all the base stations as nodes and the path between two base stations in a trajectory as edges to build a network $G \equiv \{V, E_w\}$, where V is a set of spatial nodes corresponding to the underlying urban regions, E_w is a set of edges, with each representing the connection between a pair of nodes, and the weights are assigned by the accumulated volume of the movements. For two networks on holidays and working days, each one is normalized according to the formula (2) to obtain the human interaction network G_H and G_W, which is

$$w_{ij} = e^{-\frac{\mathrm{avg}^2(v_{ij})}{2v_{ij}^2}} \tag{2}$$

where w_{ij} represents the weight between node_i and node_j, v_{ij} represents the volume of connections between node_i and node_j, and $\mathrm{avg}(*)$ means average.

4.4. Network partitioning and community structure detection

We further determine the clusters of strongly connected base stations as spatial nodes, which can be regarded as a community detection problem (or optimal subgraph partitioning). Here, we define the similarity between two base stations to be proportional to their spatial distance and human interactive strength, where the spatial distance is conducted with the Manhattan distance with the Gaussian kernel for filtering, and the interactive strength is set from weights of edges in the human interaction network. In this way, we take all the base stations as nodes and their similarities as weights to construct an undirected weighted graph and build the similarity matrix of this graph following formula (3). Here, we take the two components of similarity by the same weight and adjust the width of the Gaussian kernel for Manhattan distance to train for the optimal result.

$$s_{ij} = 0.5 \times \mathrm{Gaussian}\big(\mathrm{Manhattan}\big(\mathrm{node}_i, \mathrm{node}_j\big)\big) + 0.5 \times w_{ij} \tag{3}$$

In this study, we adopt the spectral clustering method to segment the graph. With the similarity matrix, we can construct the Laplacian matrix of this graph:

$$L = D - S \tag{4}$$

where L is the Laplacian matrix, D is the degree matrix, and S is the similarity matrix. Then, we generate the smallest k eigenvalues of the Laplacian matrix and the corresponding eigenvectors and finally use the k-means algorithm to cluster the eigenvectors. The optimal subgraph cutting method used is the ***NCut***, an algorithm that normalizes the Laplacian matrix to unify the dimensions of each subset and improve the accuracy and speed of the algorithm (Von Luxburg 2007). The basic form of the method is shown in formula (5), where the set $\{A_1, A_2, \ldots, A_k\}$ is the node subset contained in the k subgraphs obtained by ***NCut*** on the graph $G \equiv \{V, E_w\}$, which satisfies $A_i \cap A_j = \emptyset$ and $A_1 \cup A_2 \cup \cdots \cup A_k = V$, $\overline{A_i}$ is the complementary set of A_i on V, $W(A,B) = \sum_{i \in A, j \in B} w_{ij}$, $vol(A_i)$ is the sum of the weights of the edges between all nodes in A_i.

$$\boldsymbol{NCut}(A_1, A_2, \ldots, A_k) = \frac{1}{2}\sum_{i=1}^{k} \frac{W\big(A_i, \overline{A_i}\big)}{vol(A_i)} \tag{5}$$

In order to obtain less fragmented results without changing the interactive connection characteristic, we use the Calinski-Harabaz score (CH score, Caliński and Harabasz 1974) to measure the quality of partitioned results. The method for calculating the CH score is:

$$\mathrm{s(k)} = \frac{\mathrm{Tr}(B_k)}{\mathrm{Tr}(W_k)} \frac{m-k}{k-1} \tag{6}$$

where m is the number of training samples, k is the number of clusters, B_k is the covariance matrix between clusters, W_k is the covariance matrix within the cluster, and $\mathrm{Tr}(*)$ is the trace of the matrix. The CH score method is suitable for measuring unsupervised clustering results, and the algorithm is fast and intuitive.

5. Experimental results

5.1. Collective mobility patterns in PRD cities

We first derived statistics on the characteristics of travel volume during working days and holidays in the study area. According to our algorithm, we extracted 2 million and 1.26 million user tracks from 1.5 billion mobile signaling data on working days and 1 billion data on holidays, respectively, using Apache Spark. We further calculated the average travel volume of users per second on these two periods, and the travel volume dynamics over time are shown in Figure 4, where (a) presents the pattern on working days while (b) presents the pattern on holidays. These discrete points are highly continuous in time, which

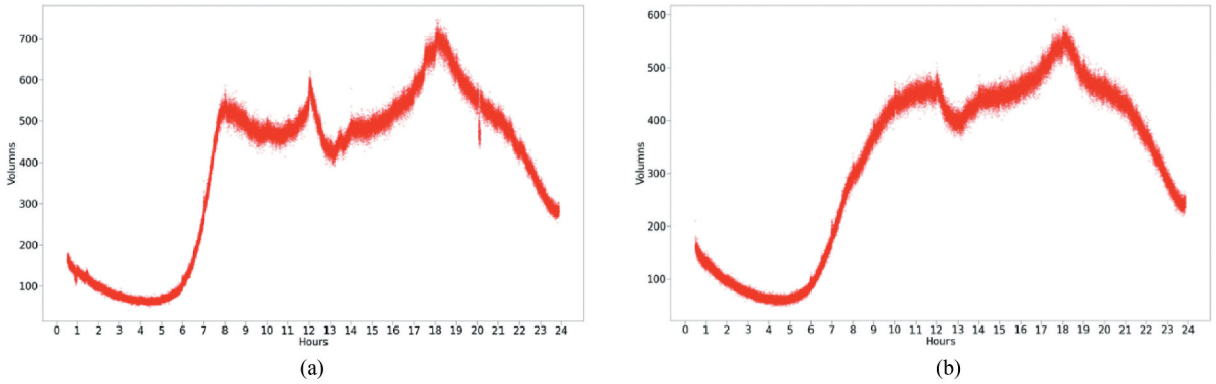


Figure 4. The variation pattern of crowd travel volume with time. (a) working days (b) holidays.

proves that human social activities have strong regularity. In addition, differences in travel volume on working days and holidays can be clearly observed. On working days, people present more concentrated travel patterns, evidenced by its sharpness compared to holidays. Note that during holidays, although people's travel is no longer subject to strict social norms such as going to or leaving work, strong consistency in the timing of travel during the entire holiday period can be observed, indicating that in the absence of strict social schedules, people's travel behavior still presents a regularity to some extent.

We then use the travel trajectories during holidays and working days to build human interaction networks. When people move from the coverage area of one base station to another, their positions transferring between stations are recorded, thus forming a data structure with the base station at the endpoints as the nodes and the paths as edges. All these structures are combined to form a large graph structure. We obtained a graph on the working days with total nodes of 37,462 and edges of 1,923,888 and a graph on holidays with total nodes of 36,701 and edges of 1,671,865. Figure 5(a) shows the interaction network of holidays in the PRD urban agglomeration where the color and transparency of the edges are adjusted according to the weights normalized by formula (2). We then screened the network by weights greater than 0.99 as the network skeleton, shown in Figure 5(b), where different colors represent different cities. The size of the node changes according to the PageRank algorithm. The higher the score, the larger the node size.

5.2. A fine-grained clustering of the PRD urban agglomeration

We segment the networks of holidays and working days and obtain clustering results at various scales (as shown in Figure 6). In order to obtain less fragmented results without changing the interactive connection characteristic, we use the CH score to measure the quality of partitioned results. Besides, we adjust the parameter σ of the Gaussian kernel in filtering the Manhattan distance to achieve the best clustering results.

It can be seen from Figure 6 that when the number of clusters (n_clusters) continues to increase, the CH scores show an overall downward trend, indicating

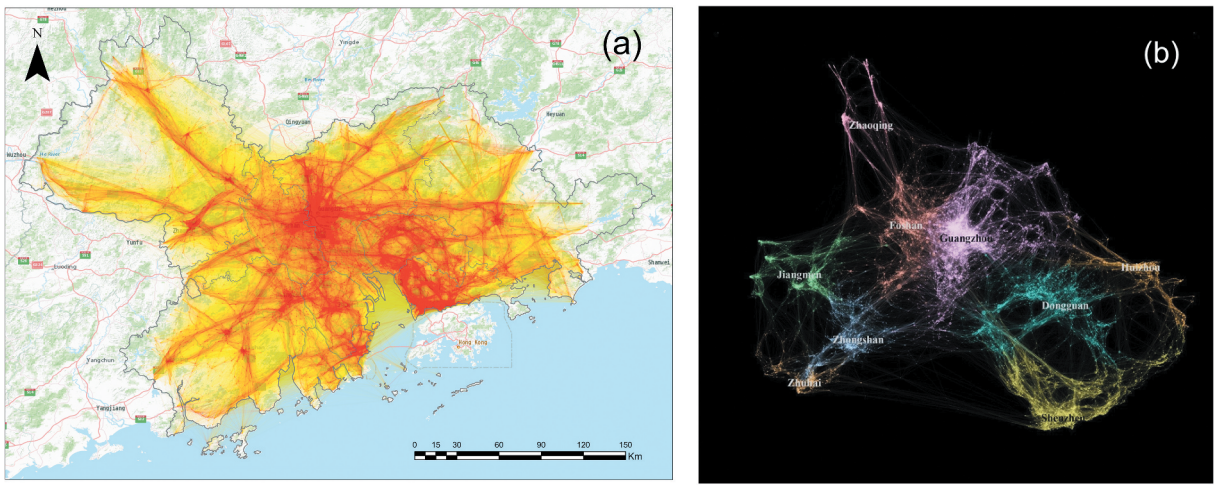


Figure 5. (a) Human interaction networks in PRD urban agglomeration; (b) network skeleton of holidays in PRD urban agglomeration (both holidays as an example).

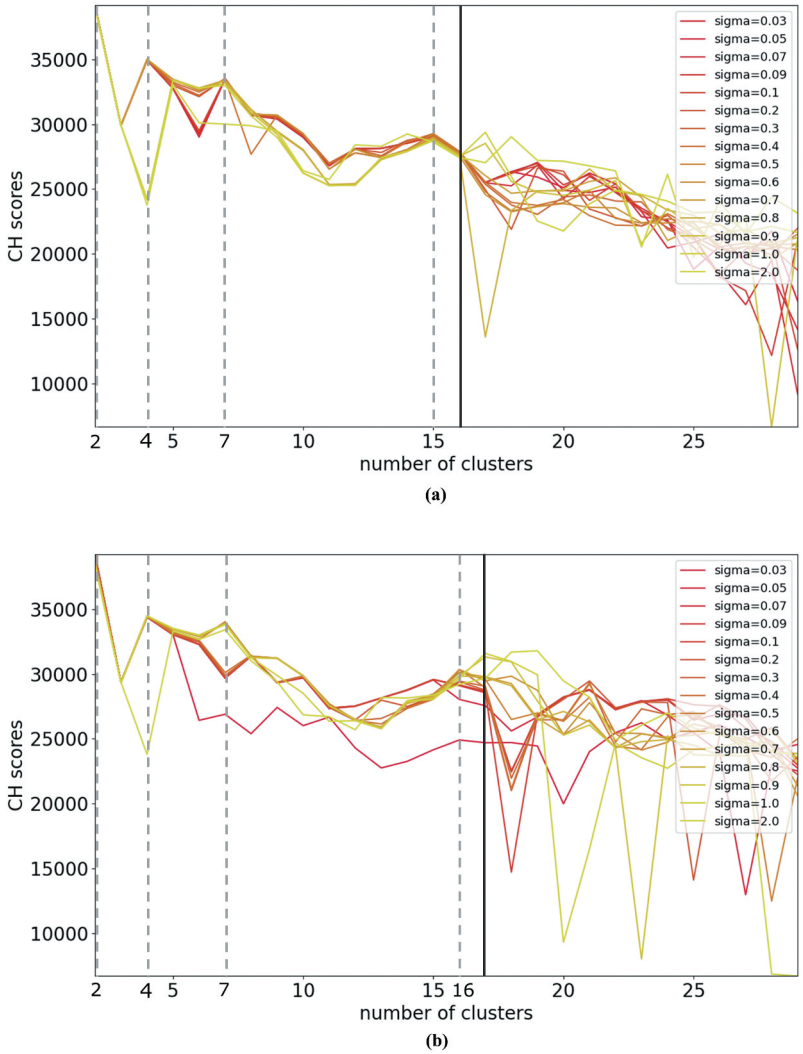


Figure 6. Training process of n_clusters and σ. (a) working days; (b) holidays.

that the clustering results gradually become fragmented. Note that as the number of clusters increases, the CH score presents many peaks, which may suggest the optimal clustering result achieved by the algorithm in different geographic scales. In this study, we trained the two parameters, i.e.σ, a parameter of Gaussian kernel function for the filtering of Manhattan distance, and n_clusters that represents the number of clusters. The strategy of parameter adjustment follows a voting mechanism on "most votes" and "highest score". "most votes" means that we choose the n_clusters where a set of more concentrated peaks appear. "highest score" means to select the parameter pairs of σ and n_clusters that makes the highest CH score on a "most vote peak". During our training the sigma is set from 0.01 to 2.0 and n_clusters from 2 to 29. With to the voting mechanism, we derived our parameter choices shown in Table 2.

For holidays and working days, we found their optimal parameter pairs of n_clusters and σ, respectively. It can be seen from Figure 6 that on working days when n_clusters is 2, 4, 7, 15 and σ is 0.2, 0.7, 0.2, 0.4, the CH scores reach the highest, while during holidays, when n_clusters is 2, 4, 7, 16 and σ is 0.07, 0.8, 0.5, 0.4, CH scores reach the highest. In Figure 6 on the left of the black line, the "results" show strong regularity while on the right the polylines become chaotic due to the vanish of commonly-stable

Table 2. The corresponding CH scores to the selected parameter pairs of n_clusters and σ on the voting mechanism.

Working Days			Holidays		
n_clusters	σ	CH score	n_clusters	σ	CH score
2	0.2	39183.1	2	0.07	39221.8
4	0.7	35017.1	4	0.8	34471.2
7	0.2	33532.7	7	0.5	34027.7
15	0.4	29279.0	16	0.4	30342.8

clustering results. This is how the voting mechanism works: searching for a more stable result. From Figure 6 we can also indicate that the training mechanism gets regular results more easily on working days than on holidays, which may due to people travel more regularly on working days thus leads to more consistent and compact spatial ranges while in the case of holidays, people move more casually.

5.3. Detecting communities and redrawing urban boundaries

As n_clusters increases, the CH score shows an overall downward trend with multiple peaks that may suggest the optimal combination of algorithm parameters under different scales. Here to prove this we divided the clustering results, or communities in the context of social studies of urban agglomerations obtained during holidays and the working days into three categories, i.e. city-group level, city level, and sub-city level according to n_clusters. The city-group level means that most communities contain more than one city, and the internal connections of these groups are greater than the connections between groups. As for city-level clusters, multiple single communities whose boundaries are consistent with the administrative boundaries began to appear in a notable manner. These communities may be more independent than other communities in terms of political, economic or geographic conditions. For sub-city clusters, small communities with a scope smaller than existing cities (e.g. administrative districts and functional regions in a city) begin to appear. Given this standard, we classified the clustering results on holidays and working days shown in Table 3:

In this study, we used the Voronoi diagram to divide the study area into continuous mosaic polygonal areas. These polygons are formed based on discrete base station location points according to the shortest distance principle of mobile communication (Das et al. 2006; Voronoi 1908). The distribution of the base stations is highly correlated with urban population, which benefits urban community detection and boundary drawing.

5.3.1. City-group level

The clustering results at the city-group level provide essential knowledge on how cities are connected by human mobility. These cross-city communities are usually relatively stable. Figure 7 shows the communities of the PRD urban agglomeration at the city-group level during working days and holidays with both n_clusters of 2 (a1, b1) and 4 (a2, b2). From a large scale we can notice some impressive large city groups such as Guangzhou-Foshan (b1, b2–2, GF), Zhongshan-Zhuhai-Jiangmen (b1, b2–4, ZZJ), Dongguan-Shenzhen-Huizhou (all-3, DSH). These large city groups won't change whether it's working days or holidays, which indicates that these communities formed from human movements are strongly stable. From a larger perspective, we can also measure how close these city groups are to each other, that is, city group GF and ZZJ can form a bigger one, together with Zhaoqing city (a1, a2–1), while city group DSH is relatively more independent.

As for city boundaries, Guangzhou and Foshan have formed a very stable city-group community (b1, b2–2). It can be seen from the satellite imagery in Figure 1(a) that the built-up area in the core part of these two cities have got fused. This makes sense because with the development of these cities, the construction of adjacent cities in space and function are constantly close to each other, leading to more frequent shuttling of people between these cities, thus blurring the original boundaries and forming a new cross-city life structure.

Table 3. Categories results under different scales.

	City-group Level	City Level	Sub-city Level
Working Days	2,4	7	15
Holidays	2,4	7	16

5.3.2. City level

When n_clusters is 7, the communities' boundaries match better with the administrative boundaries as shown in Figure 8. It is notable that the inner centripetal forces of Zhaoqing(⑥), Jiangmen (④), and Shenzhen(③) are relatively strong, evidenced by their consistent community boundaries with the administrative boundaries. We also notice that Guangzhou and Foshan still form a very stable community(②), indicating that the interaction and communication of these two cities have been very strong. Besides, they have shown a strong gravitational effect on the northern part of Zhongshan and the northwestern part of Huizhou. In addition, the main urban areas of Zhuhai and Zhongshan, as well as Dongguan and Changning County of Huizhou, also present strong ties in terms of human spatial interactions.

When we compare the results of working days and holidays, we notice that Longmen County in the north of Huizhou(a-⑤) has a closer connection with the Guangzhou-Foshan city group(a-②) on the working days, while it is slightly more connected with Dongguan(b-⑦) on holidays, which indicates that Longmen people tend to work in Guangzhou while may travel more to Dongguan on holidays. Zhongshan can be divided into three parts: the northern towns, the central main urban area, and the southern towns. The central and southern parts are more

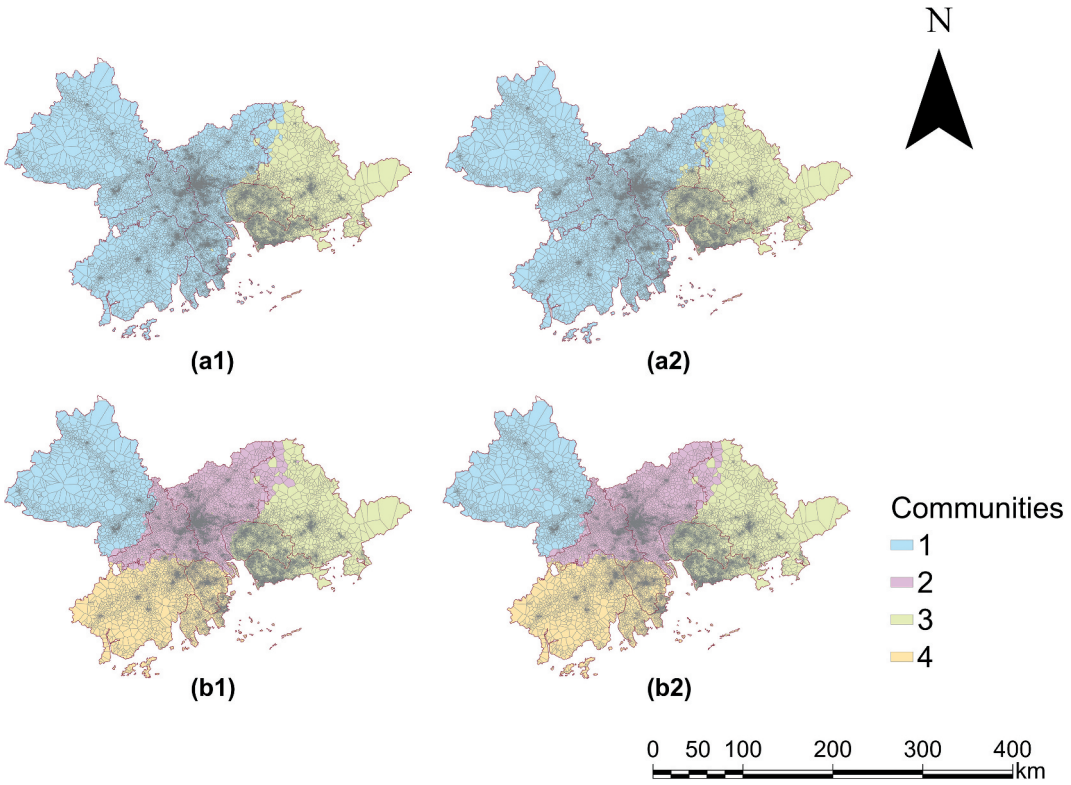


Figure 7. The communities of the PRD urban agglomeration in the scale of city-group level. (a1) two communities on the working days (a2) two communities on holidays, (b1) four communities on the working days (b2) four communities on holidays.

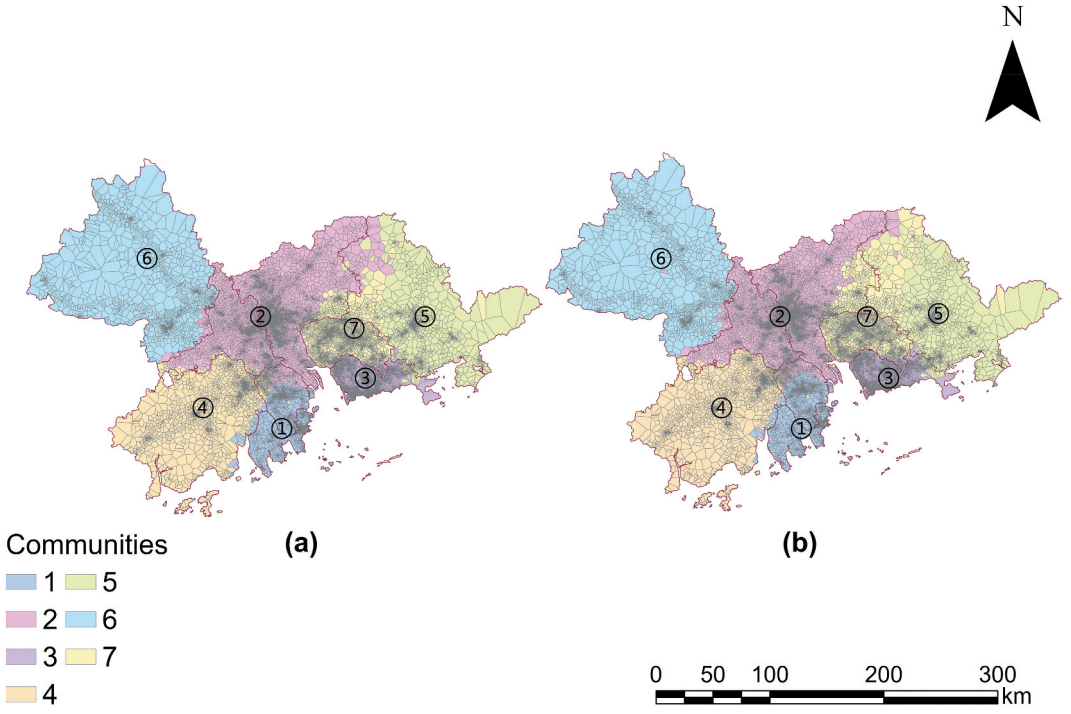


Figure 8. The communities of the PRD urban agglomeration in the scale of city level. (a) seven communities on the working days (b) seven communities on holidays.

closely connected with Zhuhai(①), while the northern towns are closely connected with the Foshan-Guangzhou city group(②). All these evidence indicates that the city has both its endogenesis, which promotes the stability of its internal spatial structure, and a gravity to the surrounding areas.

5.3.3. Sub-city level

In this level we can notice some interesting cross-city communities. We advocate a data-driven approach that breaks the boundaries of existing urban units and unearths naturally formed urban communities in space. On this scale, more refined and small

communities can be discovered. A conventional scenario is that one large city is divided into smaller parts based on the spatial distribution of its population, which means the boundaries of communities and counties (districts) match well, such as Zhaoqing with community 4, 6, 15, Jiangmen with community 3, 5, etc. This is quite normal and in line with expectations. However, what we concern are those communities that break the original boundaries, such as community 1, 2, 7, 10, and 14.

As shown in Figure 9, these cross-city communities are often located on the fringes of cities and cover a part of area of adjacent cities. For example, community 1 is Zhuhai city with the southern part of Zhongshan; community 2 is Dongguan city with the western part of Huizhou; community 7 is the southern part of Huizhou with the eastern part of Shenzhen; community 14 is the western part of Shenzhen with the southwestern part of Dongguan, etc. Taking community 14 as an example, it can be seen from satellite image Figure 1(b) where the community is located crossing the two cities. These cross-city communities actually means that the human interactions between these neighboring cities are actually stronger than that within a single city. Other than re-delineate the administrative boundaries, we can take some inspirations from these new communities in terms of transportation planning, urban services, etc. For example, increasing the traffic routes in these communities, considering these community factors when selecting locations for hospitals, schools, fire services, etc. Merchants can also refer to these factors, such as express service stations, takeaway distribution points, etc. Because these communities are actually dense-traveling areas, we may have to consider them as essential factors for us to better understand and plan our cities.

Due to the limitations on the time span of our data, we are not able to clarify or predict whether these communities will change or vanish in a longer future. However, all the available data we used, which include a holiday set and a working day set, indicate that these communities are quite stable. Although people move very differently on working days and holidays, the spatial units are connected by them due to the working, affinal and cultural relationships in a period of time, which takes regularity. Nevertheless, we still look forward to seeing a longer period of data-driven community detection work in a rapidly developing urban agglomeration and the evolution of communities could be revealed then.

6. Discussion

6.1. The scale effect of community detection

In this section, we focus on the scale effect of communities in urban spaces. For general clustering methods, there is usually a "best" number of clusters so that the result is optimal under a certain evaluation system. However, different from general situations, there exists a significant scale effect in geographic spatial clustering based on human-spatial interactions. Specifically, under a larger geographical scale (e.g. regional scale), human activities can form a huge network, which contains multiple closely connected sub-networks. These sub-networks can be administrative units at the provincial or municipal levels or relatively

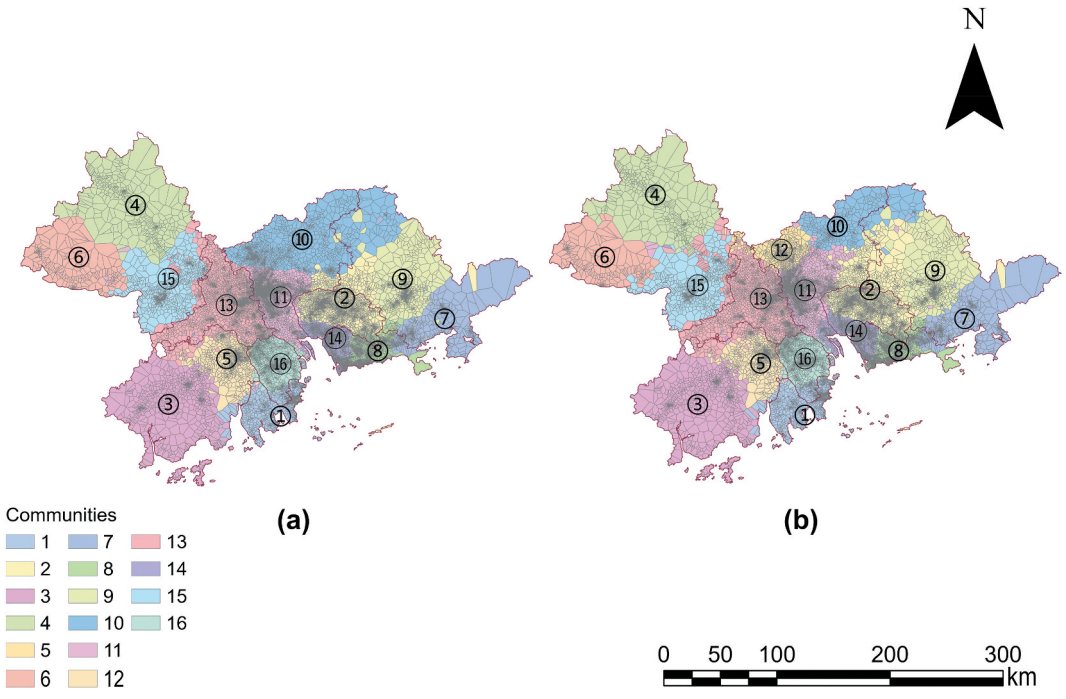


Figure 9. The communities of the PRD urban agglomeration in the scale of sub-city level. (a) working days (b) holidays.

adjacent districts, counties, or villages. These sub-networks of different scales own self-similarity, hierarchy, and recursion with the higher-level "overall network" and can be retrieved under different algorithm strengths, which is very consistent with the concept of "fractal city" (Salingaros 2003). In this section, we take Shenzhen, a sub-network in the study area as an example to discuss the scale effect of the method in this study.

Under the sub-city scale of the PRD urban agglomeration, Shenzhen was divided into three parts on the working days. Its northeast part is closely connected with Dongguan and its east part with Huizhou. However, when we used the data of Shenzhen to build the network, the number of stable communities reaches seven when σ is 0.01, which is relatively close to the number of district-level administrative units in Shenzhen (Shenzhen has nine district-level administrative units), as shown in Figure 10. The boundaries of some communities are well consistent with their administrative boundaries, while there are also some notable exceptions. For example, Nanshan District, Futian District, and the southern part of Bao'an District have formed a larger community. Similarly, Luohu District, Yantian District, and the southern part of Longgang District formed another. Note that Pingshan District, Guangming District and the northeastern part of Bao'an District are divided into three different community, which is consistent with the clustering results of Shenzhen under sub-city level.

Although we cannot obtain more stable communities of smaller scales of Shenzhen on the whole PRD urban agglomeration, it is possible when the network is limited within the range of single city, and the results show some extent of coherence with the larger-scale ones. It can be inferred that when the scope of the network is further reduced, we can still find communities of smaller scales, such as streets and even a group of houses. This reveals how the spatial network constructed by human activities shows its multi-scale characteristics that small communities combine to form larger communities: from several houses, streets to districts and counties, from cities to urban agglomerations and even countries. As a result, when the scope of the network shrinks, we can still manage to find stable communities of smaller scales despite their obscure independence in huge-scale networks. The results from our study highlight the necessity of unraveling the scale effects of communities when mining urban spatial interactions.

6.2. Limitations

The limitations of our work are twofold. One is about the study area. We attempt to use the trajectory of the population to indicate that some relatively stable cross-city communities have been formed in the urban agglomeration, and these communities have obvious scale effects. Although this phenomenon has been verified in a large urban agglomeration in China (the PRD urban agglomeration in this work), the status of this phenomenon from a global perspective cannot be clarified yet, as many other urban agglomerations (such as Atlantic coastal agglomerations in the northeast of the United States and Pacific coastal agglomerations in Japan, etc.) are different from our study area in terms of location, development stage and formation factors. The study of urban agglomerations on a global scale will be a promising research field.

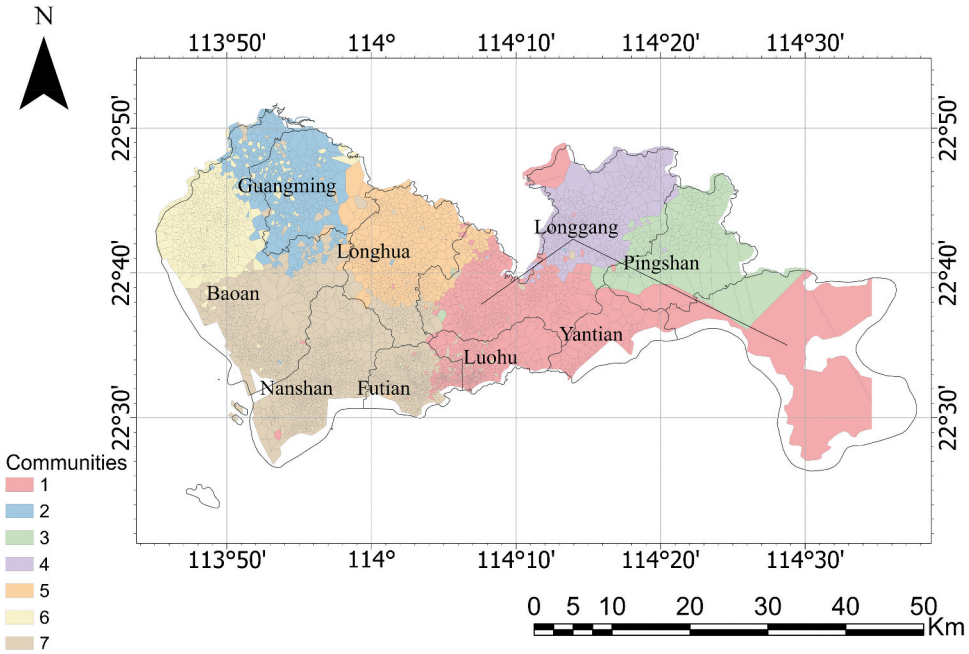


Figure 10. County-level community detection on the network of Shenzhen.

The second is that these multi-scale cross-city communities clarified in our work only come from a cross-section of the development process of the urban agglomeration. In other words, these communities change as urban agglomerations evolve. The surface logic of the formation of cross-city community lies in the cross-city flows of people, behind which is the integration of politics, economy, culture, important events and other fields among cities. Although these elements remain stable within a historical period, which generalizes relatively stable communities, once these elements change, the communities will change. For example, the COVID-19 pandemic since 2019 has reduced the movement of people between cities, which could shrink the cross-city communities that have already emerged. It is also worth studying how these communities evolve over time.

7. Conclusions

In this study, we develop a data-driven method to discover communities of different scales in urban agglomerations. These continuously mosaic polygonal areas can also be applied to delineate the boundaries of cities and districts beyond our study area. We use mobile signaling big data to detect the communities of the PRD urban agglomeration at three scales that include city-group level, city level, and sub-city level, and two periods of time, i.e. working days and holidays. The experimental results show that in the PRD urban agglomeration, three large city groups have formed: i.e. Guangzhou-Foshan, Shenzhen-Dongguan, and Zhuhai-Zhongshan-Jiangmen. On a finer scale, we notice some relatively stable cross-city communities, which is expected to benefit transportation planning, urban services, etc. These communities are formed due to more frequent traveling than that within a single city. Besides, these communities remain stable in both working days and holidays regardless of very different travel behavior of people. This is because the working, affinal and cultural relationships take regularity in a period of time. We also discussed the scale effect of community detection in sub-city scenarios. Our work highlights the necessity of considering fine-grained and multi-scale research frameworks when examining human interactions in urban communities.

Disclosure statement

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

Funding

This work was supported in part by the Guangxi science and technology program (GuiKe 2021AB30019); Sichuan Science and Technology Program (2022YFN0031, 2023YFN0022, and 2023YFS0381), Hubei key R & D plan (2022BAA048); Zhuhai industry university research cooperation project of China (ZH22017001210098PWC), Shanxi Science and Technology Program (202201150401020).

Notes on contributors

Wenbo Yu received his Master's degree from Wuhan University in 2022 on the topic of spatiotemporal big data mining. His current research focuses on GeoAI, big data and social sensing.

Zhenfeng Shao received the PhD degree in photogrammetry and remote sensing from Wuhan University in 2004. Since 2009, he has been a Full Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University. He has authored or coauthored over 70 peer-reviewed articles in international journals. His research interests include high-resolution image processing, pattern recognition, and urban remote sensing applications.

Xiao Huang received his PhD degree in Geography from the University of South Carolina in 2020. He is currently an Assistant Professor in the Department of Geosciences at the University of Arkansas, with his expertise in GeoAI, deep learning, big data, remote sensing, and social sensing.

Deren Li is a professor and chair of the Academic Committee of the State Key Laboratory for Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University. He was selected as a member of Chinese Academy of Sciences in 1991 and a member of Chinese Academy of Engineering in 1994. He got his PhD degree from University of Stuttgart, Germany. He was awarded the title of honorary doctor from ETH, Switzerland in 2008. His research interests include photogrammetry and remote sensing, global navigation satellite system, geographic information system, and their innovation integrations and applications.

Yewen Fan is an associate professor at LIEMSARS, Wuhan University. His research interests are GIS applications.

Xiaodi Xu is now pursuing her PhD degree in photogrammetry and remote sensing at Wuhan University. Her research interests include urban vegetation remote sensing and remote sensing image scene classification.

ORCID

Wenbo Yu http://orcid.org/0009-0004-0302-5390
Zhenfeng Shao http://orcid.org/0000-0003-4587-6826

Data availability statement

Data not available due to ethical and legal restrictions. Disclosing users' mobile phone signaling data may raise concerns about personal privacy leaks. The data procurement agreement also prohibits the disclosure of these data legally.

References

Blondel, V. D., J. L. Guillaume, R. Lambiotte, and E. Lefebvre. 2008. "Fast Unfolding of Communities in

Large Networks." *Journal of Statistical Mechanics: Theory & Experiment* 2008 (10): 10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.

Boix, R., P. Veneri, and V. Almenar. 2012. "Polycentric Metropolitan Areas in Europe: Towards a Unified Proposal of Delimitation[m]." In *Defining the Spatial Scale in Modern Regional Analysis*, 45–70. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-31994-5_3.

Brelsford, C., G. Thakur, R. Arthur, and H. Williams. 2019. "Using Digital Trace Data to Identify Regions and Cities[c]." Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Advances on Resilient and Intelligent Cities, November 5, 2019. https://doi.org/10.1016/j.proeng.2018.01.174.

Cai, M., and S. Zhong. 2018. "A Method for Cleaning Mobile Phone Signaling Data." *CN109040989A[P]*.

Caliński, T., and J. Harabasz. 1974. "A Dendrite Method for Cluster Analysis." *Communications in Statistics-Theory and Methods* 3 (1): 1–27. https://doi.org/10.1080/03610927408827101.

Das, G. K., S. Das, S. C. Nandy, and B. P. Sinha. 2006. "Efficient Algorithm for Placing a Given Number of Base Stations to Cover a Convex Region." *Journal of Parallel and Distributed Computing* 66 (11): 1353–1358. https://doi.org/10.1016/j.jpdc.2006.05.004.

Ebner, M. H. 2008. "Jon C. Teaford. The Metropolitan Revolution: The Rise of Post-Urban America. (The Columbia History of Urban Life)." *The American Historical Review* 113 (2): 535–536. https://doi.org/10.1086/ahr.113.2.535 .

Fang, C., and D. Yu. 2017. "Urban Agglomeration: An Evolving Concept of an Emerging Phenomenon." *Landscape and Urban Planning* 162: 126–136. https://doi.org/10.1016/j.landurbplan.2017.02.014 .

Frias-Martinez, V., and E. Frias-Martinez. 2014. "Spectral Clustering for Sensing Urban Land Use Using Twitter Activity." *Engineering Applications of Artificial Intelligence* 35: 237–245. https://doi.org/10.1016/j.engappai.2014.06.019.

Girvan, M., and M. E. J. Newman. 2002. "Community Structure in Social and Biological Networks." *Proceedings of the National Academy of Sciences* 99 (12): 7821–7826.

Grauwin, S., S. Sobolevsky, S. Moritz, I. Gódor, and C. Ratti. 2015. "Towards a Comparative Science of Cities: Using Mobile Traffic Records in New York, London, and Hong Kong[m]." In *Computational Approaches for Urban Environments*, 363–387. Cham: Springer. doi:10.1007/978-3-319-11469-9_15.

Gu, Y., R. Shi, Y. Zhuang, Q. Li, and Y. Yue. 2023. "How to Determine City Hierarchies and Spatial Structure of a Megaregion?" *Geo-Spatial Information Science*1–13. https://doi.org/10.1080/10095020.2022.2161425.

Hong, Y., and Y. Yao. 2019. "Hierarchical Community Detection and Functional Area Identification with OSM Roads and Complex Graph Theory." *International Journal of Geographical Information Science* 33 (8): 1569–1587.

Huang, L., Y. Yang, H. Gao, X. Zhao, and Z. Du. 2018. "Comparing Community Detection Algorithms in Transport Networks via Points of Interest." *IEEE Access* 6: 29729–29738. https://doi.org/10.1109/ACCESS.2018.2841321.

Jia, T., X. Yu, W. Shi, X. Liu, X. Li, and Y. Xu. 2019. "Detecting the Regional Delineation from a Network of Social Media User Interactions with Spatial Constraint: A Case Study of Shenzhen, China." *Physica A Statistical Mechanics & Its Applications* 531: 121719. https://doi.org/10.1016/j.physa.2019.121719.

Raghavan, U. N., R. Albert, and S. Kumara. 2007. "Near Linear Time Algorithm to Detect Community Structures in Large-Scale Networks." *Physical Review E* 76 (3): 036106. https://doi.org/10.1103/PhysRevE.76.036106.

Rahayu, H., R. Haigh, and D. Amaratunga. 2018. "Strategic Challenges in Development Planning for Denpasar City and the Coastal Urban Agglomeration of Sarbagita." *Procedia Engineering* 212: 1347–1354. https://doi.org/10.1016/j.proeng.2018.01.174

Rosvall, M., and C. T. Bergstrom. 2008. "Maps of Random Walks on Complex Networks Reveal Community Structure." *Proceedings of the National Academy of Sciences* 105 (4): 1118–1123. https://doi.org/10.1073/pnas.0706851105.

Salingaros, N. A. 2003. "Connecting the Fractal City." *Fractals (An Interdisciplinary Journal on the Complex Geometry of Nature)* 1–27.

Tang, J., F. Liu, Y. Wang, and H. Wang. 2015. "Uncovering Urban Human Mobility from Large Scale Taxi GPS Data." *Physica A Statistical Mechanics & Its Applications* 438: 140–153. https://doi.org/10.1016/j.physa.2015.06.032.

Tao, H., K. Wang, L. Zhuo, and X. Li. 2019. "Re-Examining Urban Region and Inferring Regional Function Based on Spatial–Temporal Interaction." *International Journal of Digital Earth* 12 (3): 293–310. https://doi.org/10.1080/17538947.2018.1425490.

Thomas, I., C. Cotteels, J. Jones, and D. Peeters. 2012. "Revisiting the Extension of the Brussels Urban Agglomeration: New Methods, New data... New Results?" *Belgeo Revue belge de géographie*, 1–2. https://doi.org/10.4000/belgeo.6074.

United Nations Human Settlements Programme. 2016. "World Cities Report 2016 Urbanisation and Development: Emerging Futures, United Nations Human Settlements Programme (UN-Habitat)." Kenya.

Von Luxburg, U. 2007. "A Tutorial on Spectral Clustering." *Statistics and Computing* 17 (4): 395–416. https://doi.org/10.1007/s11222-007-9033-z.

Voronoi, G. 1908. "Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Deuxième mémoire. Recherches sur les parallélloèdres primitifs[J]." *Journal für die reine und angewandte Mathematik (Crelles Journal)* 1908 (134): 198–287. https://doi.org/10.1515/crll.1908.134.198.

Zhou, M., Y. Yue, Q. Li, and D. Wang. 2016. "Portraying Temporal Dynamics of Urban Spatial Divisions with Mobile Phone Positioning Data: A Complex Network Approach." *ISPRS International Journal of Geo-Information* 5 (12): 240. https://doi.org/10.3390/ijgi5120240.