Physica A 531 (2019) 121719

Contents lists available at ScienceDirect

Physica A

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

Detecting the regional delineation from a network of social media user interactions with spatial constraint: A case study of Shenzhen, China

Tao Jia [a,b,*], Xuesong Yu [a], Wenzhong Shi [b], Xintao Liu [b], Xin Li [c], Yang Xu [b]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, 430072, China
[b] Department of Land Surveying and Geo-informatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China
[c] School of urban design, Wuhan University, Wuhan, 430072, China

HIGHLIGHTS

- A weighted TAZ network is constructed from social media check-in data.
- Regions are derived using a network partitioning method with spatial constraint.
- Agreement between derived regions and administrative districts are examined.
- Connections among the derived regions are analyzed from the aspect of check-in flow.

ARTICLE INFO

Article history:
Received 4 October 2018
Received in revised form 16 May 2019
Available online 13 June 2019

Keywords:
Social media check-in data
Spatial constraint
Regional delineation
Network partitioning

ABSTRACT

Regions are subdivisions of the earth's surface, and many systems of regionalization were proposed. Recently, with the availability of geotagged data, it raises the question of whether regions formed by human interactions agree with government districts. Thus, using network partitioning method with spatial constraint, we derive regional delineations at different spatial scales and examine their agreement with administrative districts. Experiments were conducted using the social media data of Shenzhen, China. Aggregately, the results show that the derived regions become inconsistent with administrative districts by increasing the spatial effect value, which can be largely attributed to the involvement of long human movements. However, the regions tend to keep stable when more long edges are included, which suggests the limitation of long movements effect. Individually, most northern administrative districts display high inconsistency with the derived regions, whereas most southern districts show high consistency. Besides, regions far from the downtown are less connected to the rest of the city, regions near the downtown are more connected, and particularly, regions in Nanshan, Futian, and Luohu are highly connected with each other, which form the backbone of total flows irrespective of spatial effect value. The results were finally validated at specific areas and compared with those using other methods, another dataset, and different spatial units, which suggest the feasibility of our regions for decision making in urban planning and management.

* Corresponding author.
E-mail address: tao.jia@whu.edu.cn (T. Jia).

https://doi.org/10.1016/j.physa.2019.121719
0378-4371/

1. Introduction

Region in geography is usually defined as a part of the earth's surface with some characteristics that make it unique to other areas [1]. These characteristics are not only related to culture, economy, topography, or politics, but also are associated with human interactions in the flow of people, information, or goods [2]. In this respect, regions are also spatial categories, which not only help to manage public resources but also are beneficial for understanding many problems [3]. There were many systems of regionalization proposed in a broad range of fields including geography, public health, psychology, urban planning, and transportation. For instance, geographers investigated the area related temporal variation and derived the regions with different land uses using GPS trajectory data [4]; the hospital service areas and referral regions were derived with a network optimization method using inpatient flow data [5]; the cognitive regions were investigated by the psychologists at relatively large scales [6,7]; urban planners proposed the functional economic areas using migration flows from the UK Census 2001 [8]; transport researchers extracted the functional transport regions using the Czech transport census 2010 [9]. As a typical example, administrative districts are mainly drawn by the government in a top-down way [10], but it is still unclear whether they agree with the regions formed by human interactions in a bottom-up way.

Many regionalization methods were proposed to derive the regions, and basically, they can be classified as three types, including non-spatial clustering, spatial clustering, and network partitioning [11,12]. Non-spatial clustering method derives regions based on attribute similarity, where a spatial process can be carried out after the regions are derived [13] or it can be embedded into the clustering process implicitly [14]. This type of regionalization method is effective for understanding the spatial distribution and dependence of attribute patterns, but the regions can be fragmented in geographic space owing to the difficulty in controlling the spatial effect. Spatial clustering method derives regions by merging adjacent spatial units with similar attribute patterns [15,16], for instance, the automatic zoning procedure method [17,18]. This type of regionalization method can be used to derive a specified number of regions with irregular shapes, but it cannot handle a large size of spatial units due to the high computational cost. Network partitioning method, which is imported from network science to solve the geographical problem, aims to divide the spatial network into communities such that vertices between different communities are less connected than expected [19–24].

With the proliferation of technologies, the network partitioning method has been widely used in various datasets to delineate the regions, such as the social connections from check-in data [25–31], mobile phone data [11,32–34], banknotes tracking data [35]. These studies explored the cohesiveness of derived regions and their agreement with the government districts. On the one hand, Ratti et al. [32] reported a remarkable consistency via a direct application of the spectral-based partitioning method [22] on a large-scale telecommunication network of the UK by ignoring the spatial arrangement of nodes; Thiemann et al. [35] concluded that considerable geographic information encoded in the network topology can reflect the states borders of the US; similar results were also reported by Sobolevsky et al. [33] who used large-scale telecommunication networks in an extensive set of countries and by Kallus et al. [27] who adopted large-scale social networks at the continent or country level. Recently, communities were also derived and analyzed at the urban scale [28,30,36], and particularly, Yin et al. [31] derived delineations at multiple spatial scales from the country level to the city level by restricting movement ranges using Twitter check-in data. On the other hand, Expert et al. [11] argued for a careful treatment of spatial effect on the network topology and thereafter they proposed a modularity measurement by incorporating the spatial effect implicitly, based on which the regions were derived and their agreement were explored.

Therefore, this study aims to derive regions at different spatial scales using a network partitioning method with spatial constraint and further explore their agreement with administrative districts. The major differences from previous studies are in the following aspects. (1) Our work imposes the spatial constraint on the network topology explicitly, which helps to examine the effects of long human movements. In other words, a spatial effect value k is introduced, which is used to constrain the network by removing edges whose topological distance are greater than k; (2) this study proposes a network partitioning method to derive the stable regionalization at different spatial scales, which can depict the variation of network modularity values as the number of regions decreases; (3) this study uses irregularly defined spatial units as the network nodes, which can avoid the problem of cell size encountered in most previous studies and improve the quality of delineation. Experiments were conducted in Shenzhen, China. We constructed the spatial networks from social media check-in records with respect to different k values, and then our method was applied to derive the regional delineations. On the basis of these delineations, we explored their agreement with administrative districts and characterized their connections from the aspect of check-in flows. Lastly, our results were validated at specific areas and compared with those using other network partitioning methods, those derived from the mobile phone data, and those using the spatial units in terms of regular grids or random voronoi.

The remainder of this paper is organized as follows. In Section 2, we introduce the datasets and the procedure to derive the spatial network. In Section 3, we describe the metrics and method for delineating the regions in detail. In Section 4, we present the regional delineations and their characteristics. In Section 5, several issues are discussed. Conclusions are drawn in Section 6.

2. Datasets and preprocessing

2.1. Datasets

Three datasets of Shenzhen, China are used in this study. The first dataset is social media check-in data, which is obtained from the Sina Company via a research contract allowing the non-commercial use in academic community. The second dataset is administrative district data, which is obtained from the bureau of Shenzhen urban planning. The third dataset is traffic analysis zone (TAZ) data, which is obtained from the Shenzhen transportation committee.

The first dataset is composed of a total number of 1,926,262 check-in records, and it spans one entire year from January 2014 to January 2015 (c.f. Fig. 1a). Each check-in record is associated with a point of interest (POI) with geographic location in terms of latitude and longitude. Besides, it includes the check-in time, the user ID, the place where the user registered, the category and address of the corresponding POI. Statistical analysis on the check-in data reveals implicit human check-in behaviors. As shown in Fig. 1b, power law distributions are reported for both the number of check-in per user and per POI, which indicates the heterogeneity of check-in behavior among users and in space. As shown in Fig. 1c, we can observe: (1) there are far more small values of time interval than large ones; (2) the periodic behavior in human mobility patterns can be reflected very well, which is common in people's daily or weekly activities, for instance, goes to work every weekday or visits to grocery store every weekend [25].

The second dataset consists of 10 administrative districts in Shenzhen, as shown in Fig. 1a with the black line. It includes 8 traditional administrative districts of Futian, Luohu, Nanshan, Yantian, Baoan, Longgang, Longhua, and Pingshan, and 2 new functional districts of Guangming and Dapeng. The administrative districts are drawn by the government authority in a top-down way, and it is still unclear whether they agree with the regions formed by social media user interactions in space, which is the major question of our study. In this respect, they are directly used to compare with the regional delineations at different spatial scales.

The third dataset consists of 491 TAZs as shown in Fig. 1d. The spatial extent varies from the largest zone (59 km^2) in the suburb to the smallest zone (0.39 km^2) in the downtown, and the mean size equals to 2 km approximately, which is very close to the grid size used in a previous study [31]. Compared with the grid used in previous studies, they can avoid the trouble on choosing the grid size and can improve the quality of delineation. Hence, in this study, they are used as the spatial units to derive the regions.

2.2. Data preprocessing

The raw check-in data are firstly preprocessed to filter and remove the invalid records, which are defined in the case of information loss (which indicates the absence of check-in location, check-in time or user ID) or information invalidity (which indicates the check-in location is out of the TAZ boundary). Secondly, the bots and single check-in behaviors are detected, and the corresponding check-in records are removed. The bots behavior is defined as the check-in behavior on the same POI with many times or on different POIs with the same time interval, while the single check-in behavior denotes the situation that the user has only one check-in record during the study period. Thirdly, check-in trajectory is obtained from the new check-in data by connecting check-in records of each user in a chronological order. After that, check-in network can be built from the check-in trajectory, where node is the check-in location and edge is the trajectory segment. Finally, the check-in network is overlaid with the TAZ data, and the TAZ network can be derived by removing the within-TAZ edges and aggregating the across-TAZ edges. Specifically, in the TAZ network, node is the TAZ, edge is the across-TAZ edge linking two different TAZs, and edge weight is the number of across-TAZ edges.

So far, the whole preprocess to extract the TAZ network has been illustrated as shown in Fig. 2a. Statistically, we have derived a total number of 1,072,399 new check-in records from the raw 1,926,262 records. Then, they are assembled into 175,086 check-in trajectories, which are further built into check-in network with 13,969 nodes and 125,735 edges. Eventually, by intersecting the 491 TAZs with the check-in network, we obtain the TAZ network with 417 nodes and 31,829 edges. As shown in Fig. 2b, the TAZ edge length can be roughly approximated by a heavy-tailed distribution, which implies the existence of significant long human movements. Besides, the TAZ edges are classified into five categories according to their length using equal quantiles (c.f. Fig. 2c), which will be used to examine the edge composition of the derived regions.

3. Metrics and method

3.1. Metrics

3.1.1. Network modularity

Network modularity (Q) measures the quality of a partition of the TAZ network in terms of its deviation from a null model. In this context, we adopted both a commonly used random null model [21] and a uniform null model [38], and thus the corresponding network modularity are defined as:

$$Q^{random}(P) = 1/2w \sum_{C \in P} \sum_{i,j \in C} A_{ij} - k_i * k_j/2w \tag{1}$$

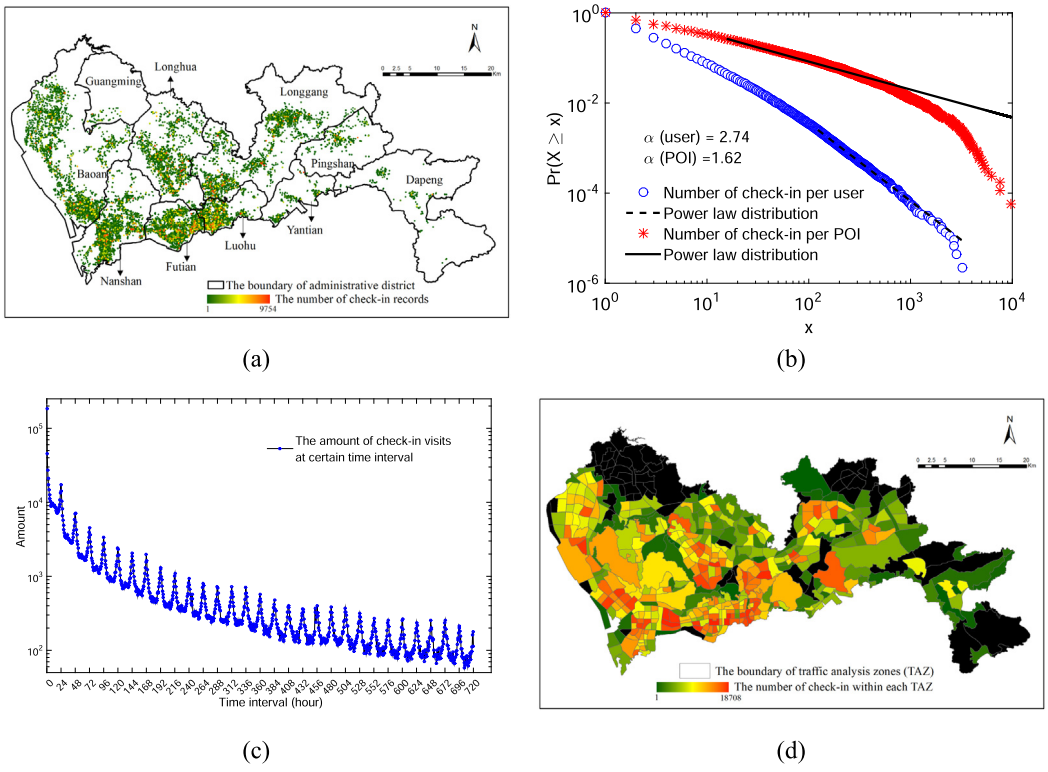


Fig. 1. The three dataset used in this study: (a) Check-in records overlaid with the administrative districts, where green dot indicates a small number and red dot means a large number; (b) the power law distributions of the number of check-in per user and per POI, for which significance tests are conducted with p values of 0.14 and 0.12, respectively [37]; (c) the periodic behavior in human mobility patterns, which resembles very well with the distribution of human returning probability [25]; (d) TAZ data visualized by the number of check-in records, where green color indicates a small number and red color means a large number.

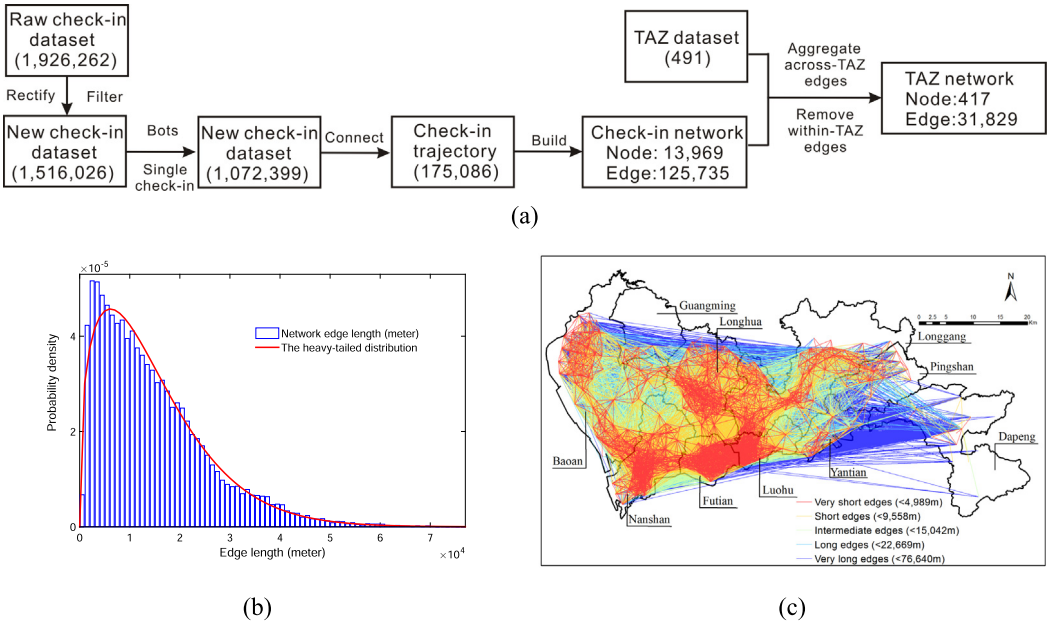


Fig. 2. Demonstration of the data preprocessing: (a) the procedure; (b) the heavy-tailed distribution of edge length in the TAZ network; and (c) the TAZ network with edge length classified by five categories, including very short, short, intermediate, long, and very long.

$$Q^{uniform}(P) = 1/2w \sum_{C \in P} \sum_{i,j \in C} A_{ij} - 2w/N(N-1) \tag{2}$$

Where, in the TAZ network, w is the total edge weight, C is a community of TAZs, P is a set of communities, i and j are two TAZs, A_{ij} is the weight of an edge between i and j, k_i (or k_j) is the sum of the weight of incident edges on TAZ i (or j), N is the number of nodes.

3.1.2. Mixing parameter

Mixing parameter [39] quantifies the extent to which the network contains certain assumed community pattern. In this context, it is used for assessing whether the administrative partition can be uncovered in a given network with spatial effect value k. It can be defined as:

$$m = \frac{\sum_i k_i^{Ext}}{\sum_i k_i} \tag{3}$$

Where, k_i^{Ext} is the number of links connecting node i to the nodes in other communities, k_i is the total number of links connecting to node i. Mixing parameter is typically used as a global property, and a network with $m > 0.5$ tends to indicate the disappearance of the assumed community pattern [40]. This is because the community pattern in a strong sense requires each node should have more connections within the community.

3.1.3. Adjusted rand index

Adjusted Rand Index (*ARI*, [41]) measures the similarity between two partitions of the data, which is adjusted for chance partitioning of data elements. In this context, it is used for measuring the agreement between the derived regions and the administrative districts, and high value (greater than 0.5) indicates a good consistency. It can be formulated as:

$$ARI(P, Q) = \frac{2(N_{00} * N_{11} - N_{01} * N_{10})}{(N_{00} + N_{01})(N_{01} + N_{11}) + (N_{00} + N_{10})(N_{10} + N_{11})} \tag{4}$$

Where, P and Q are two partitions of the TAZ data (for instance, P is the partition of TAZ according to the derived regions, and Q is the partition of TAZ according to the administrative districts), N_{00} is the number of pairs that are in different clusters in both P and Q, N_{11} is the number of pairs that are in the same cluster in both P and Q, N_{01} is the number of pairs that are in the same cluster in P but in different clusters in Q, N_{10} is the number of pairs that are in different clusters in P but in the same cluster in Q.

3.2. A network partitioning method with spatial constraint

In this study, we propose a network partitioning method to obtain the regional delineations (communities) at different spatial scales. To examine the influence of spatial constraint on regional delineation, we introduce the spatial effect value k, which is used to spatially constrain the network by discarding edges with topological distance greater than k. Thus, a high value indicates a weak spatial constraint on the network by removing long edges. To delineate the network, we need to maximize the network modularity [19–21,24], and a high value means much more human interactions within regions.

Firstly, a topological distance matrix is computed for the TAZ network, where two TAZs are neighbors to each other when they have a topological distance one. Based on the topological distance matrix, the edges with topological distance greater than k can be removed, which leads to a TAZ network with spatial effect value k. Secondly, a Louvain algorithm [24] is used to extract the communities, which is composed of two steps: (1) it traverses all the neighbor nodes of each node in the TAZ network, calculates the modularity gain by assigning the node to the community of its neighbor node using Eq. (1), and joins the node to the community with the maximum value of modularity gain. This step is repeated until the community of each node is stable; (2) it generates a new network by taking the communities as nodes and updating the edge weight among communities. In this unfolding step, it checks whether the new network is the same as the previous one, and it goes to perform step (1) if they are different, otherwise, the communities are returned as the ones with the maximum modularity value. Thirdly, many sets of communities (e.g., 1000) can be obtained by applying the Louvain algorithm to the same TAZ network repeatedly, and these different sets of communities are used to generate a consistent network, where the edge weight between two nodes is increased by one if they belong to the same community. Thereafter, we again apply the Louvain algorithm using Eq. (2) to the consistent network, which can derive a stable regional delineation. It should be noted that the nodes are gradually merged together as communities in a bottom up way, and hence the network modularity value varies as the number of communities decreases. For convenience, the detailed pseudo-code is provided as follows.

```
Algorithm：A network partitioning method with spatial constraint to generate communities
----------
Input: TAZ network TN, Spatial effect value k
Output: Communities

Function NetworkPartitioning (TN, k)
    Foreach edge in TN.edges
        [startNode, endNode] = GetNodes(edge);
        mDist = topoDist(startNode, endNode);
        If (mDist <= k) Then
            STN.add (edge);
    While (repetitions < 1000)
        MyCommunities = ExtractCommunities (STN, "Girvan-Newman");
        Foreach community in MyCommunities
            Foreach node1, node2 in community
                CN.addEdgeWeight(node1, node2);
    Communities = ExtractCommunities (CN, "Uniform");
    Return Communities;

Function ExtractCommunities (N, ModularityType)
    If (ModularityType == "Girvan-Newman") Then CalculateModularity = GNModularity;
    If (ModularityType == "Uniform") Then CalculateModularity = UniformModularity;
    MyCommunities = SetEachNodeAsACommunity (N);
    SumQGain = Double.Max; ThreshHold = Double.Min;
    While (SumQGain > ThreshHold)
        SumQGain = 0;
        NodeList = GetRandomNodeOrder (N);
        Foreach node in NodeList
            Obtain each TryCommunities by assigning node to the community of its neighbor in MyCommunities;
            Use CalculateModularity to compute the modularity value of each TryCommunities;
            Calculate the modularity gain value of each TryCommunities;
            Find BestCommunities with the largest value of modularity gain BestQGain from all TryCommunities;
            If (BestQGain > 0) Then
                MyCommunities = BestCommunities;
                SumQGain = SumQGain + BestQGain;
    If (SumQGain <= ThreshHold) Then
        Generate Unet by taking communities in N as nodes;
        If (Unet != N) Then
            MyCommunities = ExtractCommunities (Unet, ModularityType);
    Return MyCommunities;
```

4. Results

4.1. Derivation of the regional delineations with different values of spatial effect

As elaborated in our method, different regional delineations can be derived at different spatial scales with different spatial effect values k. It should be noted that our aim is to explore the agreement between the delineation and the administrative districts, and thus it is meaningless to derive the regions from a network if it does not contain the assumed

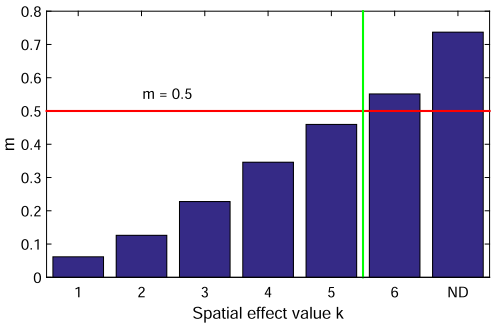


Fig. 3. The values of mixing parameter (m) for networks with different spatial effect values k (Note: ND means the network diameter, and the entire network is used if k = ND).

community pattern of administrative districts. Using mixing parameter, we firstly examine whether the networks with increasing values of k contain the community pattern comparable to the administrative districts. As shown in Fig. 3, the results suggest that the community pattern disappears for the network with k equals to 6. In this respect, we examine the networks with the values of k increasing from 1 to 5.

As shown in Fig. 4, we present the regional delineations with different values of spatial effect. What these delineations have in common is that they contain at least 99% of the total check-in records and occupy at least 68% of the spatial extent of administrative districts owing to the absence of check-in records in a few TAZs. However, a visual comparison with the 10 administrative districts indicates their different spatial organizations.

As shown in Fig. 4a, 33% of edges including all short ones are reserved when the value of k is 1. Meantime, a total number of 19 regions are obtained with the maximum modularity value of 0.89. The number of derived regions is larger than the number of administrative districts, and they show a different spatial organization of TAZs. Through a visual comparison with the 10 administrative districts (c.f. Fig. 4b), we can see that most regions are constrained by the administrative districts and a few regions across over the boundaries of administrative districts. For instance, 4 regions and 3 regions are completely within Baoan and Longgang respectively; 1 region crosses over the boundary of Futian and Luohu; and the region in Pingshan expands into Yantian. However, Guangming, the new functional administrative district, is not completely delineated due to the absence of check-in data.

As shown in Fig. 4c, 49% of edges including a few intermediate ones are reserved when the value of k is 2. Meantime, 14 geographically cohesive regions are derived with the maximum modularity value of 0.85. However, they display a slight dissimilarity to the administrative districts. For instance, 2 regions are roughly derived from Longhua, 3 regions are clearly delineated in Nanshan, and 4 regions are delineated in Longgang. Again, a denser TAZ network containing 59% of edges including all intermediate ones is achieved when the value of k is 3, and 11 regions can be derived with a maximum modularity value of 0.77 (c.f. Fig. 4e). Importantly, a visual comparison with the 10 administrative districts suggests that regions at this scale agree well with the administrative districts, although areas in Dapeng, Yantian and Pingshan are merged as one region. For instance, Nanshan, Futian and Luohu can be clearly delineated, and the number of regions in Longgang has decreased from 4 to 2 (c.f. Fig. 4f).

Lastly, with the increment of k values, long edges are gradually included and the TAZ network becomes much denser. As shown in Fig. 4g and i, 8 regions can be derived for both $k = 4$ and $k = 5$. At these two scales, most regions tend to deviate from the administrative districts except for the region in Longhua. For instance, areas in Dapeng, Yantian, Pingshan and Longgang are merged as the largest region, and area in Luohu expands into Longgang to form a large region. Importantly, we can observe that regions do not change too much from $k = 4$ to $k = 5$, which suggests that regional delineation reaches the stability at $k = 5$. In fact, this conclusion can be verified using the TAZ network including all edges. Therefore, the last two regional delineations seem to display higher inconsistency to the administrative districts, which might be largely owing to the inclusion of long edges, but they will remain stable with the inclusion of more long edges. In other words, the long movement effect is limited.

In addition, an in-depth investigation on these regional delineations is conducted by extracting the core regions. A core region is defined as the aggregation of TAZs where any two TAZs belong to the same cluster irrespective of delineations. As shown in Fig. 5, 24 core regions are extracted by intersecting the 5 delineations, and they are separated from each other by non-core regions that lie at the boundaries and have somewhat ambiguous associations. The core regions include most of the popular central business districts (CBD) such as Dongmen CBD in Luohu, Huaqiangbei CBD in Futian, and Shenzhen north center CBD in Longhua. These core regions contain around 97% of the total check-in records and occupy about 60% of the spatial extent. Specifically, three core regions with the largest values of check-in density are Futian center CBD, Dongmen CBD, and Huaqiangbei CBD, which contain about 27% of the total check-in records but occupy only 3% of the entire area. These findings verify that human activities are much more likely to be concentrated on core regions or particularly the CBD regions and indicate the usefulness of our regional delineations.

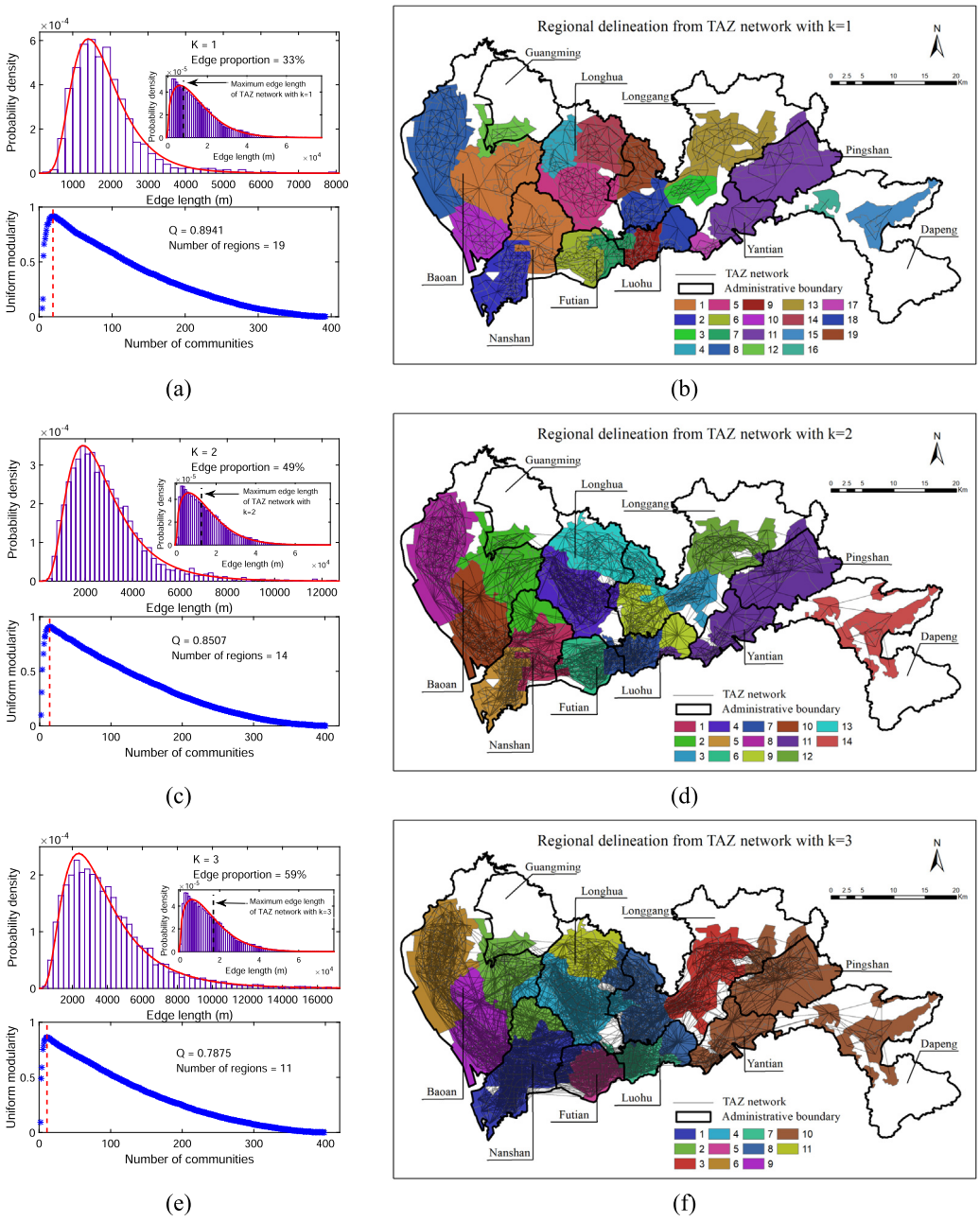


Fig. 4. Regional delineations with different values of spatial effect: (a), (c), (e), (g), and (i) show the procedure of deriving regional delineations with different k values, where the upper figures display the probability distributions of edge length for the TAZ network with k and the entire TAZ network (insect, where the dotted line indicates the maximum edge length of TAZ network with k), and the lower figures display partitioning procedure with k; (b), (d), (f), (h), and (j) draw the optimal regional delineations corresponding to the maximum modularity with different k value.

4.2. Agreement of the regional delineations with the administrative districts

The regional delineations have shown different spatial organizations of TAZs, and hence it is of importance to examine their agreement with the administrative districts in detail. To do so, we use Adjusted Rand Index (*ARI*) to measure the agreement between the derived regions and the administrative districts. The *ARI* measures the agreement by penalizing both false positive and false negative decisions during clustering. As shown in Fig. 6a, we present the agreement between the derived regions with different values of spatial effect and the administrative districts. It can be clearly observed that the regional delineation with $k = 3$ (c.f. Fig. 4f) gives the highest agreement value of *ARI* (62%), while the regional delineation

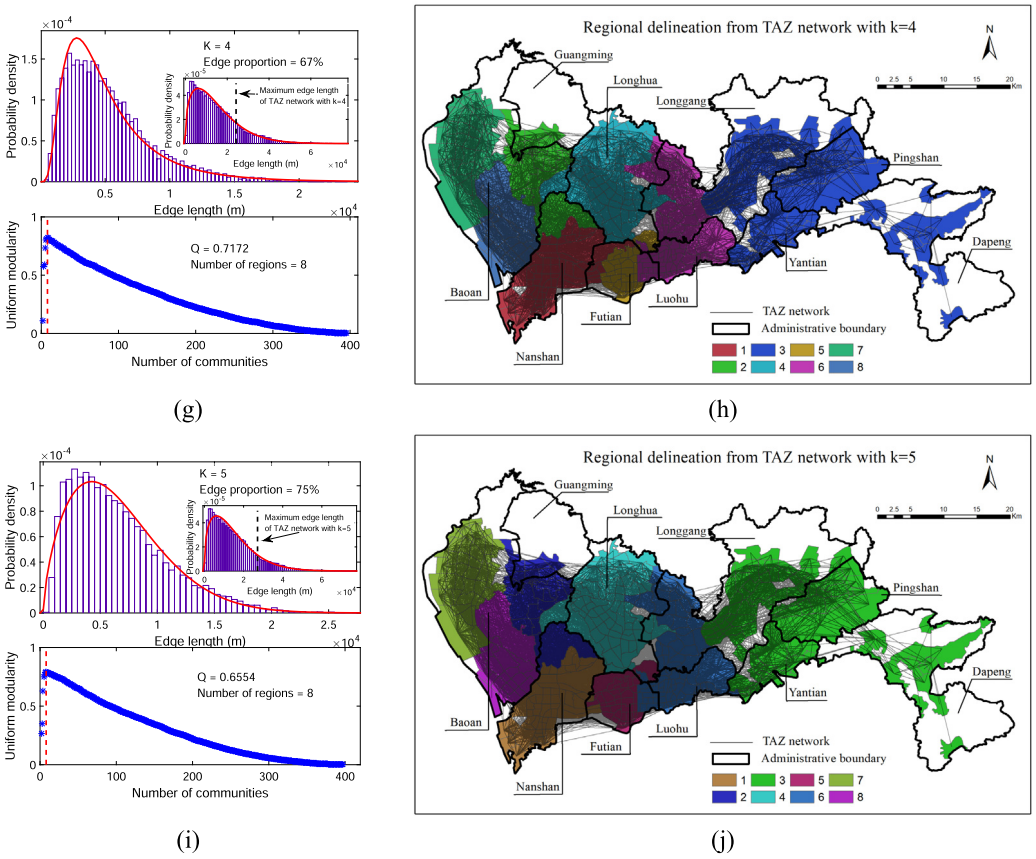


Fig. 4. (*continued*).

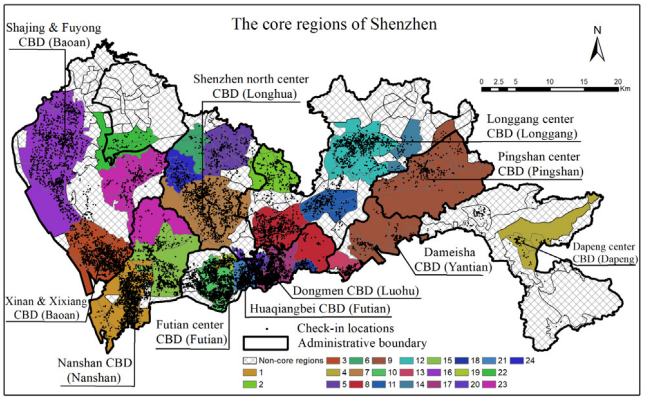


Fig. 5. The core regions extracted by intersecting the five regional delineations.

with $k = 1$ (c.f. Fig. 4j) suggests the lowest agreement value of *ARI* (48%). This finding coincides with our visual comparison as elaborated above, and it supports the arguments in a previous study [35], which assumed that the existence of long movements could weaken the geographical constraint on human mobility and would lead to inconsistent delineations with the administrative districts. However, the inconsistency keeps unchanged when more long edges were included. Besides, the finding does not mean that the delineation with $k = 3$ should be considered as the best one, because each delineation reflects well the check-in movements at the current scale.

Spatially, we intersect the five regional delineations with the administrative districts to extract the core regions, which contains a high percentage of check-in records fluctuating around 99% as k changing from 1 to 5. In this respect, we can examine the consistency within a single administrative district by simply counting the number of core regions, and a

large number of core regions means a low degree of consistency. As shown from Fig. 6b to e, two observations can be reported for the northern districts (including Baoan, Longhua, Longgang and Pingshan) and the southern districts (including Nanshan, Futian, Luohu, Yantian and Dapeng).

Firstly, most northern administrative districts show a high inconsistency with the derived regions, which can be explained largely by the loosely distributed urban infrastructure in a relative large area. For instance, the average number of core regions in Longhua, Baoan, and Longgang reaches around 2, 3, and 4, respectively. Particularly, the number of core regions displays a decreasing trend owing to the concentration of TAZs into large regions. The low degree of consistency in these districts does not necessary mean that these core regions are ineffective. Instead, they can serve as the surrogates for a better understanding on how these administrative districts should be delineated according to social media user interactions. For instance, the three core regions in Baoan keep stable from $k = 1$ to $k = 5$, and they coincide well the Long-term master plan in Shenzhen [34].

Secondly, most southern administrative districts show a high consistency with the derived regions, which can be probably due to the compact distributed urban infrastructure in a relative small area. For instance, Nanshan, Luohu, Yantian, and Dapeng are roughly dominated by only one large core region. Particularly, the number of core regions in these districts keeps relatively stable. However, Futian shows a slight difference from the other districts, and one core region occurs in its eastern area along the boundary, which can be attributed to the strong connection between Huaqiangbei CBD and Luohu district. Nonetheless, this observation suggests that the current district boundary can reflect well the interactions of social media user in real life, although marginal inconsistency can be observed.

4.3. Characteristics of the regional delineations from the aspect of check-in flows

The derived regions have shown significant differences from the administrative districts, but it is still unknown how they connect to the rest of the city and to each other. To measure their connectivity with the rest of the city, we compute the intra-travel ratio, which is defined as the proportion of travels within the region. In this context, a travel is regarded as a movement from one TAZ to a different one. Hence, the larger the value of intra-travel ratio, the much more isolated the region is. To measure their connectivity with each other, we present a network where regions are nodes, travel flows are edges and the width of an edge denotes the number of travel flows between two regions. Through the network analysis, we examine how the regions are connected with each other.

Firstly, intra-travel ratio is calculated and visualized for each derived region with different spatial effect values k. As shown from Fig. 7a to e, we can see clearly that the northern regions in Baoan, Longgang, Pingshan and the southern regions in Yantian, Dapeng are relatively less connected to the rest of the city, although long human movements are gradually included from $k = 1$ to $k = 5$. A potential explanation is that these regions are geographically far away from the downtown, which might prevent the connections. On the contrary, the southern regions in Futian, Luohu and Nanshan are relatively more connected to the rest of the city. These regions locate in the downtown, and importantly they have the check ports departing to Hong Kong. Thus, these factors might enhance the connections.

Secondly, networks are constructed among the regions with different spatial effect values k and the inter-travel flows are visualized into 5 categories using the natural break method. Generally, the inter-travel flow has increased dramatically among the regions with the inclusion of long human movements from $k = 1$ to $k = 5$. Specifically, as shown from Fig. 7a to e, very high inter-travel flows are always observed among the regions in the southern districts including Nanshan, Futian, and Luohu. These inter-travel flows account for 65.3%, 56.0%, 45.6%, 50.9%, and 45.0% of the total inter-travel flows at the spatial effect values of 1, 2, 3, 4, and 5, respectively. They hint the important roles that these regions might play in Shenzhen. Actually, the Long-term master plan addresses that they belong to the central cluster out of 11 functional clusters [34]. The concentration of inter-travel flows on the three districts also coincides with the heterogeneous check-in behavior in space.

5. Discussions and limitations

Human interactions in space is constrained by the boundaries of administrative districts, while the boundaries at the same time can be gradually changed by the human interactions. Hence, this study explores the agreement between the regional delineation from social media user interactions and the administrative districts, and the regional delineations are derived by a network partitioning method with spatial constraint.

5.1. Characteristics of this study

Our study has two characteristics. Firstly, it enriches the literature on regional delineation by using a network partitioning method with spatial constraint. Specifically, it embeds the influence of spatial constraint on human movement into the construction of spatial network. Thus, spatial networks with different values of spatial effect can be derived and partitioned into regional delineations at different spatial scales. Secondly, the agreement between the derived regions and the administrative districts is examined both quantitatively using *ARI* value and spatially using core regions. Importantly, the regional delineations with different spatial effect values are geographically cohesive and provide brand new data sources as supplementary for urban design and resource management.

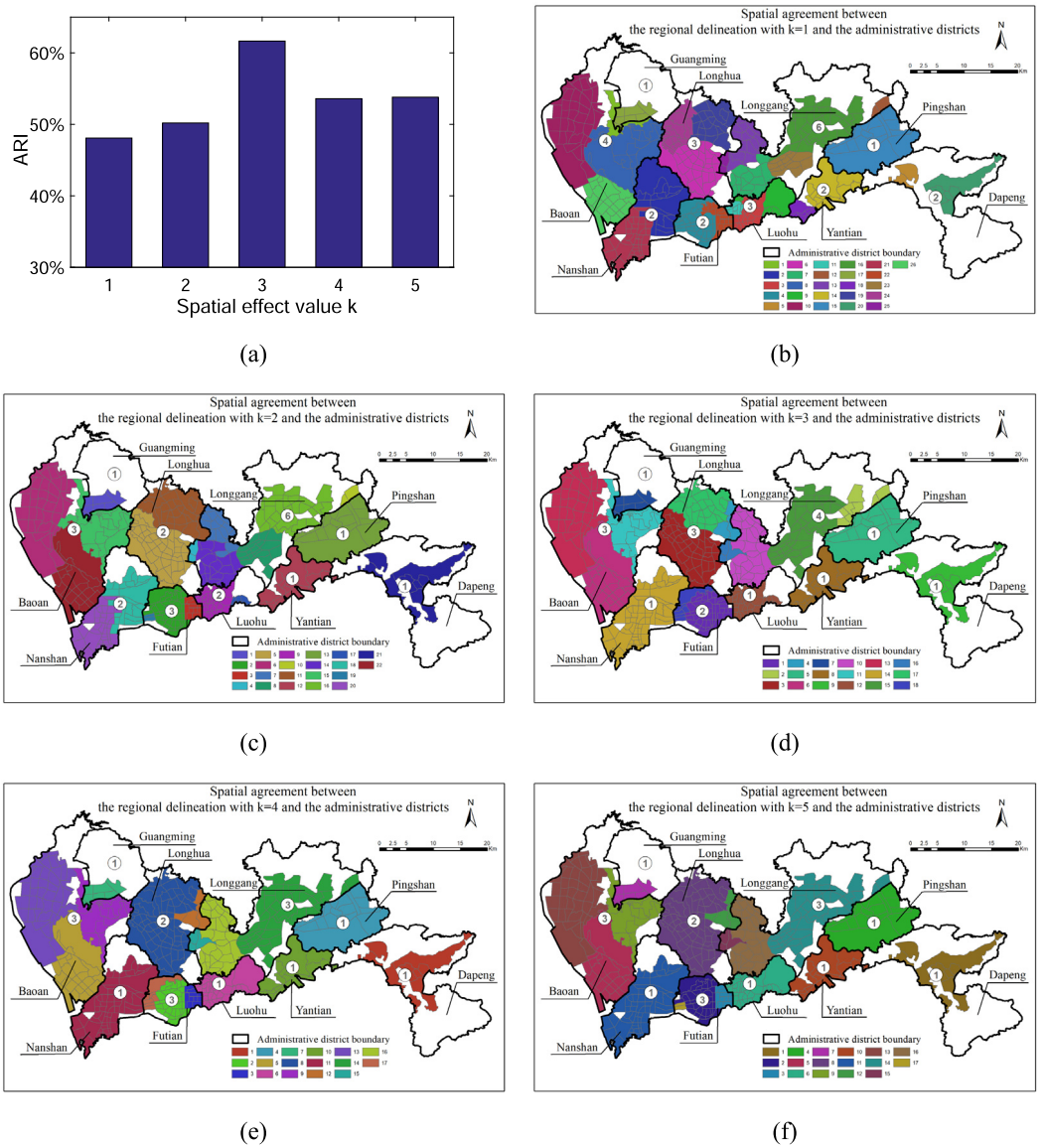


Fig. 6. Agreement between the regional delineations and the administrative districts under different values of spatial effect: (a) the change of *ARI* value with k; (b) $k = 1$; (c) $k = 2$; (d) $k = 3$; (e) $k = 4$; (f) $k = 5$.

5.2. Discussions of this study

Firstly, this study adopts the Louvain-like method, which can improve the robustness of regional delineation compared with the conventional Louvain algorithm [24]. Besides, there were many other types of network partitioning methods proposed in the literature. Typical examples include the Girvan–Newman method [19], the spectral-based method [22], and the Info-Map method [23]. The Girvan–Newman method can derive exact regional delineation, but it is computational inefficient for large-scale network such as our TAZ network with all edges; the other two methods use heuristics strategy to improve the computational efficiency, but the derived regional delineations might be stochastic and unstable [42]. Compared with the above methods, our method is not only computational efficient for large-scale network, but also can derive a robust and stable regional delineation. Importantly, as shown in Fig. 8, the regional delineations with different spatial effect values using our method seem to have larger values of network modularity than those of the other three methods, which suggests the superiority of our method.

Secondly, this study used TAZ data as the spatial units, which is different from the grid data used in other studies [31,43]. The analytic results from different spatial units can be inconsistent due to the modifiable areal unit problem. Thus, to examine how it affects our results, our method was applied to the spatial units of both the grid (a regular

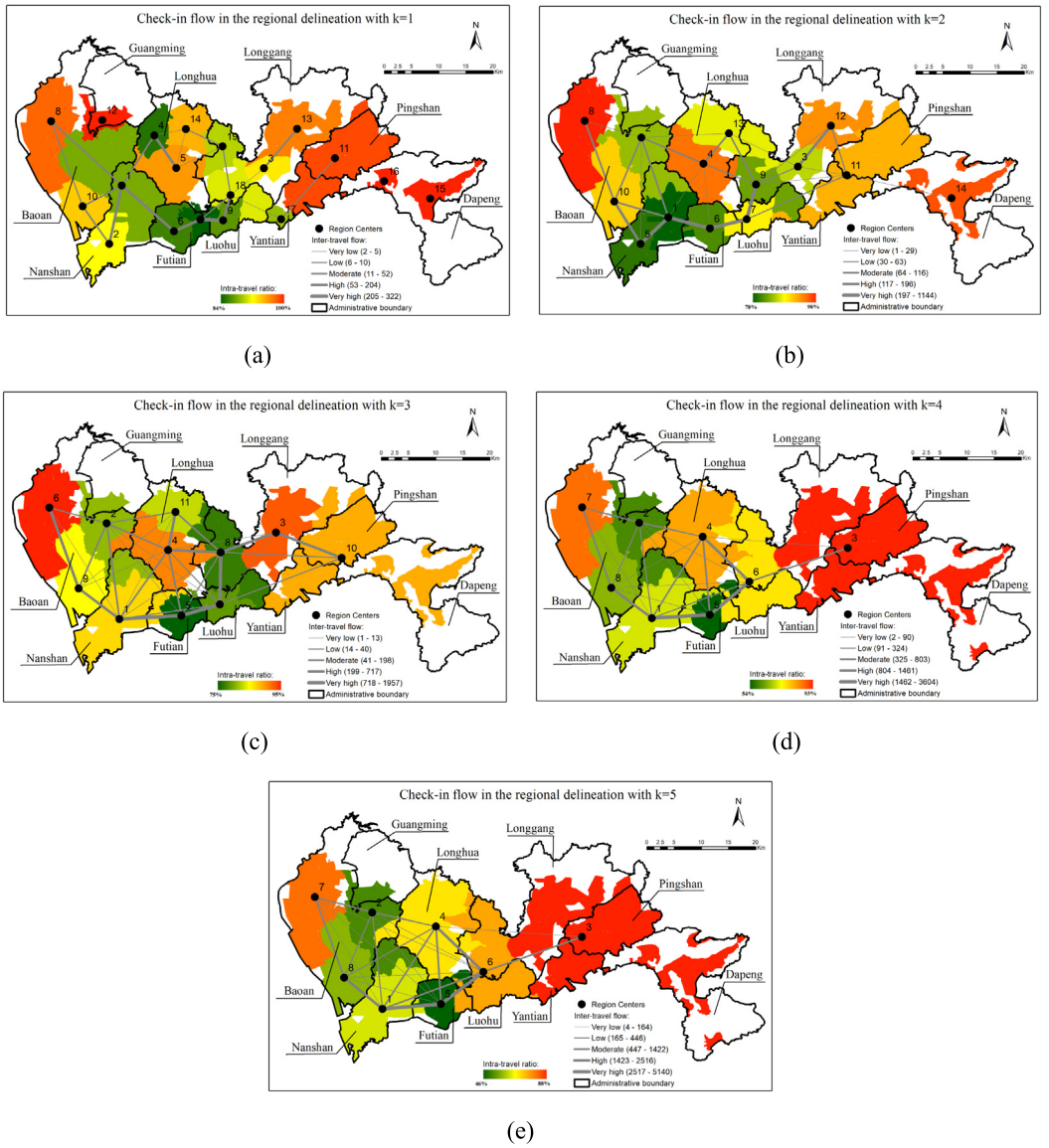


Fig. 7. Check-in flow in the regional delineations with (a) $k = 1$; (b) $k = 2$; (c) $k = 3$; (d) $k = 4$; (e) $k = 5$.

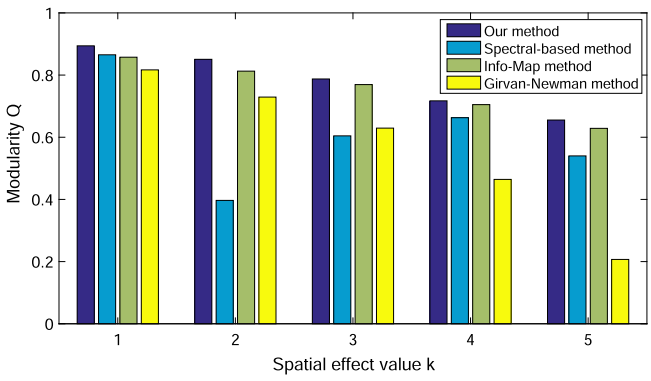


Fig. 8. Comparative results of the network modularity values of the regional delineations using different network partitioning methods.

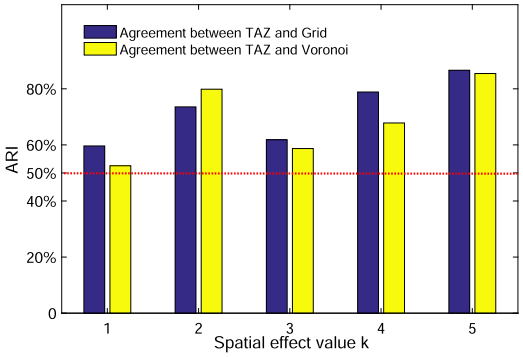


Fig. 9. Agreement between our regional delineations and those derived from the grid network and the voronoi network.

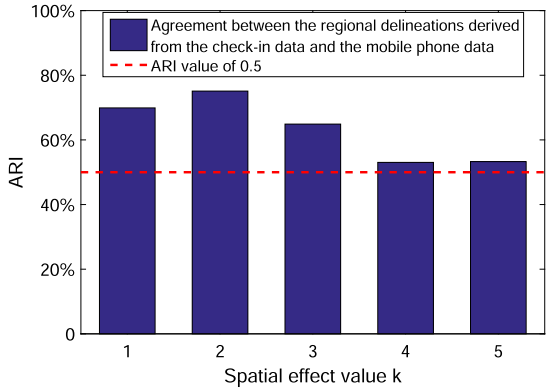


Fig. 10. Agreement between our regional delineations and those derived from the mobile phone data.

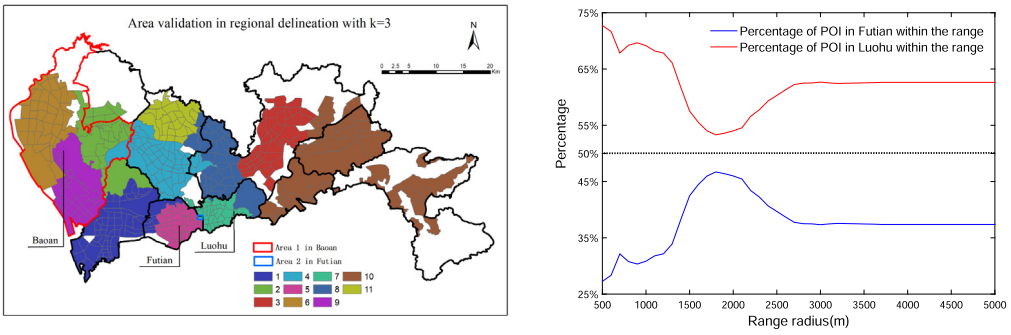
decomposition) and the voronoi (a random decomposition). The size of the grid (2 km) is approximately the same as the average size of the TAZs, whereas the number of the voronoi (491) is the same as the number of the TAZs. Compared with our results, two facts can be observed. On the one hand, the community pattern disappears with $k = 6$ for both the grid network and the voronoi network, which is similar to our TAZ network. On the other hand, the regions derived from the two spatial units display a relative high degree of consistency with our regions from $k = 1$ to $k = 5$, although slight fluctuations are observed (c.f. Fig. 9). The comparative analysis suggests that spatial units can indeed affect the regional delineation via changing the network structure, but the influence is limited in our study.

Thirdly, this study adopted one day mobile phone data [44] to construct the TAZ network and further derived the regional delineations using our method, and the aim is to examine the controversial issue of whether check-in data can be used to represent the spatial interactions at the city level. Specifically, we verify the reliability of our regional delineations by comparing them with those derived from the mobile phone data. As shown in Fig. 10, we can see clearly that the regions derived from the two different types of data display a relative high degree of consistence irrespective of the spatial effect value k, and particularly, they resemble with each other very well with *ARI* value as high as 0.76 when k is 2. Besides, the results from the mobile phone data, such as the agreement with the administrative district and the travel flow patterns, are consistent with our findings. The comparative analysis suggests that social media check-in can be used to derive regional delineation at the city level, although slight differences can be observed using the mobile phone data.

5.3. Limitations of this study

Firstly, the check-in data impose a limitation, although comparative results with the mobile phone data suggest the reliability of our regional delineations. We recognize that there still remain uncertainties in the data. For instance, not all human activities can be associated with the check-in behavior, although they display the periodic behavior of human mobility as shown in Fig. 1c; missing activities might occur between two consecutive check-in locations, although our results suggest that there are far more small values of time interval than large ones.

Secondly, the resolution problem is a very common issue for any community detection method with modularity optimization [45]. In this respect, the resolution problem also exists in our study, but its influence can be limited. This



(a) (b)

Fig. 11. Two areas in regional delineation with $k = 3$: (a) the locations of the two areas; (b) percentage of POIs in Futian and Luohu as a function of distance away from area 2.

is because our method can derive regional delineations with different spatial effect values k. For small values of k, long edges are cut off, large regions can be hardly derived, and hence the resolution problem is limited; for large values of k, long edges are included, large regions can be derived, and hence the resolution problem cannot be avoided.

Thirdly, the regional delineations with different spatial effect values are different from the administrative districts, but there is a limitation on how to explain these differences and justify these regions owing to the absence of detailed human commuting data, which might have practical implications for urban planning and management. Using only check-in data, we attempt to give limited explanation on specific areas. Taking the regional delineation with $k = 3$ as an example, the regions at two specific areas can be explained and justified. As shown the red line in Fig. 11a, Baoan district is clearly delineated into three regions, which agrees very well with the Long-term Master plan in Shenzhen [34]; as shown the blue line in Fig. 11a, a residential area in Futian district is assigned to the region in Luohu, because points of interest such as shops and commercial buildings in Luohu are much more accessible for people in area 2 than those in Futian (c.f. Fig. 11b). Therefore, to explain and justify more areas, further studies are needed.

6. Conclusions

This study explores an open question of whether regional delineations from human interactions agree with the administrative districts. A network partitioning method with spatial constraint is employed to derive the regional delineations from a TAZ network, where the spatial constraint is explicitly imposed on the network by discarding edges whose topological distance are greater than k. In this respect, the regional delineations at different spatial scales were derived and their agreement with administrative districts were examined quantitatively and spatially.

Experiments were conducted in Shenzhen, China. With the increment of k value from 1 to 5, large regions are more likely to occur, and the derived regions and the administrative districts become much more inconsistent. This can be attributed to the involvement of long human movements, which could weaken the geographical constraint on human mobility and reduce the consistency. However, the regions tend to keep stable by including more long edges, which indicates the limitation of long movement effect. Individually, the northern districts are more likely to display a higher degree of inconsistency than the southern districts. Besides, our results suggest that the regions far from the downtown are less connected, the regions near the downtown are more connected, and the regions in Nanshan, Futian, and Luohu are highly connected with each other to form the backbone of the total check-in flows.

Acknowledgments

We would like to thank the constructive comments from the anonymous referees, whose efforts greatly improve the quality of this paper. We would also like to thank the financial support from the National Natural Science Foundation of China with Award number 41401453 and Hong Kong Scholar program with project number G-YZ91.

References

[1] M. Roger, Regional Geography: Theory and Practice, Routledge, New York, USA, 2017.
[2] P. Hall, Looking backward, looking forward: The city region of the mid-21st century, Reg. Stud. 43 (6) (2009) 803–817.
[3] P.R. Good, B. Derudder, F.J. Witlox, The Regionalization of Africa: Delineating Africa's subregions using airline data, J. Geogr. 110 (5) (2011) 179–190.
[4] Y. Liu, F. Wang, Y. Xiao, S. Gao, Urban land uses and traffic 'source–sink areas': Evidence from GPS-enabled taxi data in shanghai, Landsc. Urban Plan. 106 (1) (2012) 73–87.
[5] Y. Hu, F. Wang, I.M. Xierali, Automated delineation of hospital service areas and hospital referral regions by modularity optimization, Health Serv. Res. 53 (1) (2018) 236–255.

[6] A. Friedman, D.R. Montello, Global-scale location and distance estimates: common representations and strategies in absolute and relative judgments, J. Exp. Psychol. Learn. Mem. Cogn. 32 (2) (2006) 333–346.
[7] A. Friedman, The role of categories and spatial cuing in global-scale location estimates, J. Exp. Psychol. Learn. Mem. Cogn. 35 (1) (2009) 94–112.
[8] C. Jones, Spatial economy and the geography of functional economic areas, Environ. Plann. B 44 (3) (2017) 486–503.
[9] S. Kraft, M. Marada, Delimitation of functional transport regions: understanding the transport flows patterns at the micro-regional level, Geografiska Ann. Ser. B 99 (1) (2017) 79–93.
[10] Da Yang, Remaking the Chinese Leviathan: Market Transition and the Politics of Governance in China, Stanford University Press, CA, US, 2004.
[11] P. Expert, T.S. Evans, V.D. Blondel, R. Lambiotte, Uncovering space-independent communities in spatial networks, Proc. Natl. Acad. Sci. 108 (19) (2011) 7663–7668.
[12] H. Zhu, J. Liu, H. Liu, X. Wang, Y. Ma, Recreational business district boundary identifying and spatial structure influence in historic area development: A case study of Qianmen area, China, Habitat Int. 63 (2017) 11–20.
[13] R.P. Haining, S.M. Wise, M. Blake, Constructing regions for small area analysis: material deprivation and colorectal cancer, J. Publ. Health Med. 16 (1994) 429–438.
[14] M.A. Oliver, R. Webster, A geostatistical basis for spatial weighting in multivariate classification, Math. Geol. 21 (1989) 15–35.
[15] J. Han, M. Kamber, A.K.H. Tung, Spatial clustering methods in data mining: a survey, in: H.J. Miller, J. Han (Eds.), Geographic Data Mining and Knowledge Discovery, Taylor & Francis, London, 2001, pp. 33–50.
[16] T. Jia, B. Jiang, Building and analyzing the US airport network based on en-route location information, Physica A 391 (15) (2012) 4031–4042.
[17] S. Openshaw, L. Rao, Algorithms for reengineering 1991 census geography, Environ. Plan. A 27 (1995) 425–446.
[18] H.J. Miller, J. Han, Geographic Data Mining and Knowledge Discovery, CRC Press, London, UK, 2009.
[19] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. 99 (2002) 7821–7826.
[20] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2004) 026113.
[21] M.E.J. Newman, Modularity and community structure in networks, Proc. Natl. Acad. Sci. 103 (2006a) 8577–8582.
[22] M.E.J. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E (2006b) 74, http://dx.doi.org/10.1103/PhysRevE.74.036104.
[23] M. Rosvall, C.T. Bergstrom, Maps of information flow reveal community structure in complex networks, Proc. Natl. Acad. Sci. 105 (1118) (2008).
[24] V.D. Blondel, J.L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of community hierarchies in large networks, J. Stat. Mech. Theory Exp. (2008) P10008.
[25] Z. Cheng, J. Caverlee, K. Lee, D.Z. Sui, Exploring millions of footprints in location sharing services, in: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011, pp. 1–8.
[26] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, C. Ratti, Geo-located Twitter as proxy for global mobility patterns, Cartogr. Geogr. Inf. Sci. 41 (3) (2014) 260–271.
[27] Z. Kallus, N. Barankai, J. Szüle, G. Vattay, Spatial fingerprints of community structure in human interaction network for an extensive set of large-scale regions, PLoS One 10 (5) (2015).
[28] Y. Sun, Investigating "Locality" of intra-urban spatial interactions in New York city using Foursquare data, ISPRS Int. J. Geo-Inf. 5 (4) (2016).
[29] A. Belyi, et al., Global multi-layer network of human mobility, Int. J. Geogr. Inf. Sci. 31 (7) (2017) 1381–1402.
[30] F. Zhen, Y. Cao, X. Qin, B. Wang, Delineation of an urban agglomeration boundary based on sina weibo microblog 'check-in' data: A case study of the Yangtze River Delta, Cities 60 (2017) 180–191.
[31] J. Yin, A. Soliman, D. Yin, S. Wang, Depicting urban boundaries from a mobility network of spatial interactions: a case study of Great Britain with geo-located Twitter data, Int. J. Geogr. Inf. Sci. 31 (7) (2017) 1293–1313.
[32] C. Ratti, et al., Redrawing the map of Great Britain from a network of human interactions, PLoS One 5 (12) (2010) e14248, http://dx.doi.org/10.1371/journal.pone.0014248.
[33] S. Sobolevsky, M. Szell, R. Campari, T. Couronne, Z. Smoreda, C. Ratti, Delineating geographical regions with networks of human interactions in an extensive set of countries, PLoS One 8 (12) (2013).
[34] M. Zhou, Y. Yue, Q. Li, D. Wang, Portraying temporal dynamics of urban spatial divisions with mobile phone positioning data: A complex network approach, ISPRS Int. J. Geo-Inf. 5 (240) (2016).
[35] C. Thiemann, F. Theis, D. Grady, R. Brune, D. Brockmann, The structure of borders in a small world, PLoS One 5 (11) (2010).
[36] J. Cranshaw, R. Schwartz, J. Hong, N. Sadeh, The livehoods project: Utilizing social media to understand the dynamics of a city, in: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media, June, Dublin, Ireland, 2012.
[37] A. Clauset, C.R. Shalizi, M.E.J. Newman, Power-law distributions in empirical data, SIAM Rev. 51 (2009) 661–703.
[38] D.S. Bassett, M.A. Porter, N.F. Wymbs, S.T. Grafton, J.M. Carlson, P.J. Mucha, Robust detection of dynamic community structure in networks, Chaos 23 (2013) 013142.
[39] A. Lancichinetti, S. Fortunato, Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities, Phys. Rev. E 80 (2009) 016118.
[40] Z. Yang, R. Algesheimer, C.J. Tessone, A comparative analysis of community detection algorithms on artificial networks, Sci. Rep. 6 (30750) (2016).
[41] N.X. Vinh, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance, J. Mach. Learn. Res. 11 (2010) 2837–2854.
[42] B.H. Good, Y.A. de Montjoye, A. Clauset, Performance of modularity maximization in practical contexts, Phys. Rev. E 81 (2010) 046106.
[43] X. Liu, L. Gong, Y. Gong, Y. Liu, Revealing travel patterns and city structure with taxi trip data, J. Transp. Geogr. 43 (2015) 78–90.
[44] Y. Xu, S.L. Shaw, Z. Fang, L. Yin, Estimating potential demand of bicycle trips from mobile phone data-an anchor-point based approach, ISPRS Int. J. Geo-Inf. 5 (131) (2016).
[45] S. Fortunato, M. Barthelemy, Resolution limit in community detection, Proc. Natl. Acad. Sci. 104 (1) (2007) 36–41.