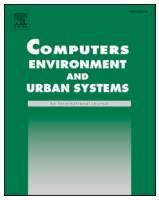
Computers, Environment and Urban Systems 107 (2024) 102045

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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



A topology-based approach to identifying urban centers in America using multi-source geospatial big data

Zheng Ren [a], Stefan Seipel [a,c], Bin Jiang [a,b,*]

[a] Faculty of Engineering and Sustainable Development, University of Gävle, Gävle, Sweden
[b] Urban Governance and Design Thrust, Society Hub, Hong Kong University of Science and Technology (Guangzhou), China
[c] Division of Visual Information and Interaction, Department of Information Technology, Uppsala University, Uppsala, Sweden

ARTICLE INFO

Keywords:
Urban centers
Topological representation
Wholeness
Big data
Nighttime light imagery
Complexity

ABSTRACT

Urban structure can be better comprehended through analyzing its cores. Geospatial big data facilitate the identification of urban centers in terms of high accuracy and accessibility. However, previous studies seldom leverage multi-source geospatial big data to identify urban centers from a topological perspective. This study attempts to identify urban centers through the spatial integration of multi-source geospatial big data, including nighttime light imagery (NTL), building footprints (BFP) and street nodes of OpenStreetMap (OSM). We use a novel topological approach to construct complex networks from intra-urban hotspots based on the theory of centers by Christopher Alexander. We compute the degree of wholeness value for each hotspot as the centric index. The overlapped hotspots with the highest centric indices are regarded as urban centers. The identified urban centers in New York, Los Angeles, and Houston are consistent with their downtown areas, with overall accuracy of 90.23%. In Chicago, a new urban center is identified considering a larger spatial extent. The proposed approach can effectively and objectively prevent counting those hotspots with high intensity values but few neighbors into the result. This study proposes a topological approach for urban center identification and a bottom-up perspective for sustainable urban design.

1. Introduction

Conventionally, urban centers refer to the core areas in a city with high density of infrastructure, population, and services (Christaller, 1933 [1966], Burger & Meijers, 2012). Within urban centers, socio-economic activities are more concentrated than other urban areas (Anas, Arnott, & Small, 1998). For example, central business districts (CBDs) are the most recognized urban centers demarcated by urban planning authorities based on census data, economic statistics, and land use data (Alonso, 1964; McDonald & Prather, 1994; Murphy & Vance, 1954). Urban structures have become increasingly complex with the influx of urban populations and the rapid development of infrastructure (Batty, 2008; Batty & Longley, 1994). Some fast-developing cities tend to show the polycentric structures (McMillen, 2003; Meijers, 2005; Roth, Kang, Batty, et al., 2011). There is also a substantial increase in the volume of big data generated from both individuals and the urban environment. Compared to conventional census data or small data, big data is more suitable for studying urban structures. One the one hand, big data is typically collected at the individual level using bottom-up approaches that utilize the intelligence of crowds (Goodchild, 2007), which results in a finer spatial resolution than conventional data. On the other hand, data analytics for big data are different from that for small data, with big data using fractal geometry (Mandelbrot, 1982) and heavy-tailed distributions (Chen, 2012; Goodchild & Mark, 1987) to illustrate the heterogeneity of urban structure. Small-data analytics are still dominated by the Euclidean geometry and Gaussian statistics (Schmitt et al., 2023) that use mean values and standard deviations to characterize the homogeneity of urban patterns.

The existing studies have utilized various geospatial big data and methods to identify urban centers and their spatial extents. For instance, the nighttime light image data have been widely used to delineate urban centers (Chen et al., 2017; Ma, Lang, Yang, Shi, & Ge, 2020; Yang, Chen, Guo, Zheng, & Wu, 2021). Sun, Fan, Li, and And (2016) and Liu et al. (2021) used social media check-in data and mobile phone data respectively to identify urban functional centers. The above studies tend to use the different geospatial big data separately without data fusion. Zhou, He, & Zhu (2022) identified the polycentric urban structure based on multi-source data fusion. Fusing different sources of data helps to avoid

* Corresponding author.
E-mail addresses: zheng.ren@hig.se (Z. Ren), stefan.seipel@hig.se (S. Seipel), bin.jiang@hig.se, binjiang@hkust-gz.edu.cn (B. Jiang).

https://doi.org/10.1016/j.compenvurbsys.2023.102045
Received 7 February 2023; Received in revised form 4 October 2023; Accepted 8 October 2023
Available online 20 October 2023

misrepresentations of urban centers from each data source. Some novel methods such as the contour trees (Chen et al., 2017; Sun & Fan, 2021) and terrain analysis (Ma et al., 2020) have been adopted to identify urban centers using remote sensing data. However, the uncertainty still exists. Firstly, the thresholds for delineating urban center extents are derived by the trial-and-error experiments and local knowledge (Chen et al., 2017) when using nighttime light image data. Secondly, some lit pixels such as airports, crowded trunk roads, and shopping centers outside the city can be mis-counted into urban centers. Although these areas can be excluded by pre-defined indices such as standard deviation and elongation of feature shapes (Yang et al., 2021), the uncertainty and subjectivity still exist when we define these indices. There are two types of urban center detection methods: geospatial analysis and topological analysis. Since the topology can reveal the deeper urban elements relationships (Xie et al., 2023), the spatial relationships between urban centers should be considered to derive urban centers.

Our urban environment is essentially a complex network with all elements connected to each other (Alexander, 1965; Batty, 2008). This study regards the hotspots as nodes and their spatial relationships as links based on Christopher Alexander's center theory (Alexander, 2002–2005). In the theory of centers, the wholeness or living structure is defined as a recursive structure that consists of coherent centers forming a whole. The wholeness can be understood as a scaling structure (Jiang, 2015; Ma, Omer, Osaragi, Sandberg, & Jiang, 2019) that consists of far more small hotspots than large ones, forming a hierarchy. The coherence means that the centers are not isolated but inter-connected by the links according to their spatial configurations. Those hotspots with high intensity values but few neighbors can be automatically excluded from the final result by adopting topological representations (Jiang, 2018) based on their inherent spatial configurations. We use natural city clustering algorithm to objectively derive the thresholds when creating hotspots (Jiang, 2015; Jiang & Yin, 2014). The thresholds are decided by the data pattern itself. We use a mathematical model and its topological representation (Jiang, 2015, 2018) to quantify the degree of wholeness of centers. For individual center, the one with the highest degree of wholeness can be regarded as the top urban center. For a set of centers, the degree of wholeness is characterized by the ht-index (Jiang & Yin, 2014). In this way, we can derive urban centers by a 'two-stage' identification (before and after topological representation). The hierarchy of the urban centers can be also characterized by the ht-index.

The present study makes three major contributions to the current literature. First, multi-source geospatial big data are integrated spatially and statistically to detect urban centers. A multi-modal data fusion on decision level is conducted so that the results are derived and reinforced by the common parts of three datasets. Second, we adopt topological representation to view the fragmented hotspots as a complex network, so that the urban centers are identified based on the spatial configuration and the statistical character of each hotspot. The hotspots are derived using a bottom-up approach to avoid the subjectivity to define the spatial unit. Third, the natural cities boundaries are derived in the US and our methods have been conducted in four largest metropolitan areas in the US, showing the methods can be applied to different cities. The reminder of this paper is organized as follows. Section 2 introduces the theoretical foundation of topological representation. Section 3 states methodology about how we process the data. Section 4 presents the results of identification. In Section 5, we further discuss the results, implications and limitations of the proposed approach in urban design and planning. Finally, in Section 6 we conclude our work and point out future direction.

2. Alexander's theory of centers and the topological representation

The idea for this study comes from Alexander's theory of centers (Alexander, 2002–2005). The centers refer to all kinds of coherent entities that exist in space and those centers are nested with each other, forming a structure called wholeness (Jiang, 2015). The notion of wholeness has previously been mentioned in the field of physics, biology, ecology, and architecture (Bohm, 1980; Salingaros, 1997). Alexander (2002–2005) defined wholeness mathematically as a recursive structure that objectively exist in space , matter and human minds. The space can be either large-scale space like cities or small-scale space such as architecture façade and paintings (Jiang & Huang, 2021). Alexander described wholeness a "*quality without a name*" because it is too subtle to be seen. Jiang (2015) developed a graph-based model to quantify the degree of wholeness. A structure with a high degree of wholeness is called a living structure. Alexander summarized and distilled 15 structural properties to create living structures (Fig. 1). A living structure manifests some, if not all, of those 15 properties.

We illustrate how 15 properties work using a strict fractal patten: Sierpinski carpet. The fractal geometry (Mandelbrot, 1982) aims to study the things that are so irregular and fragmented that they cannot be measured using Euclidean geometrics. A strict fractal can be well characterized using a mathematic formula and it has the idealized self-similar property. In Fig. 2a, the Sierpinski carpet is created recursively by taking out 1/9 size of the original square each time. However, the Sierpinski carpet is an example of strict fractals that never exists in nature. Mandelbrot (1982) summarized the notion of statistical fractals to characterize the self-similar shapes in nature, such as coastlines, trees, and clouds using the power law exponents. Further, we can continue to broaden the definition of fractals by introducing the idea of living structure, as we called the third definition of fractal (Jiang & Yin, 2014): A pattern or a space is regarded as fractal or living if there are far more small substructures than large ones, and this pattern recurs at least twice. The fractal and living in this study are used interchangeably.

Under the third definition, the carpet is a living structure. Why it is living can be answered using the 15 properties. For example, the carpet exhibits the levels of scale property, for it has a total of 73 (1 + 8 + 64) centers and three hierarchical levels out of those substructures form the scaling pattern of far more small squares than large ones (Fig. 2a). The strong centers property is also manifested in the carpet, with the strongest center located at the middle and surrounded by the other centers to form a cohesive unit. The thick boundaries are not obvious and neither are the positive space, deep interlock and ambiguity. The alternating repetition, echoes and good shape properties are also held in the carpet for the pattern is repetitive and self-similar. The carpet also exhibits contrast and gradients in terms of square sizes. The roughness is a property that differs from Euclidean geometry and fractal geometry. The real Sierpinski carpet has infinite repetitions, which makes the surface rough, and it cannot be measured using Euclidean measurements. The void is largest square before splitting by 1/3 of the original length. The simplicity and inner calm can be seen from the simple square shape. The not-separateness refers to a center not being separable from its surroundings. The carpet manifests 12 out of 15 properties to be a living structure.

The wholeness is a *de facto* complex network that consists of numerous centers. The grey squares in the Sierpinski carpet are the abstraction of the real hotspots that are used to illustrate the topological representation in a simple way. Accordingly, we glue the dispersed hotspots together using a topological approach according to their spatial configurations and their associated intensity values. The relationships between hotspots are slightly different from the original topological representation (Jiang, 2018), but share the same philosophy. We include the cross-level relationships throughout all scales instead of only two consecutive scales. We deem that nearby centers are connected to each other, regardless of their scales. The hotspots are not isolated, but surrounded and supported by many other hotspots. The 73 squares underlie a complex network with nodes and links (Fig. 2c). Three levels of nodes are derived using head/tail breaks (Jiang, 2013). We build up Thiessen polygons based on nodes to derive their 'territories'. Hence, the largest territory is equal to the original carpet in red polygon, the medium territories are the green polygons, and the smallest territories are the

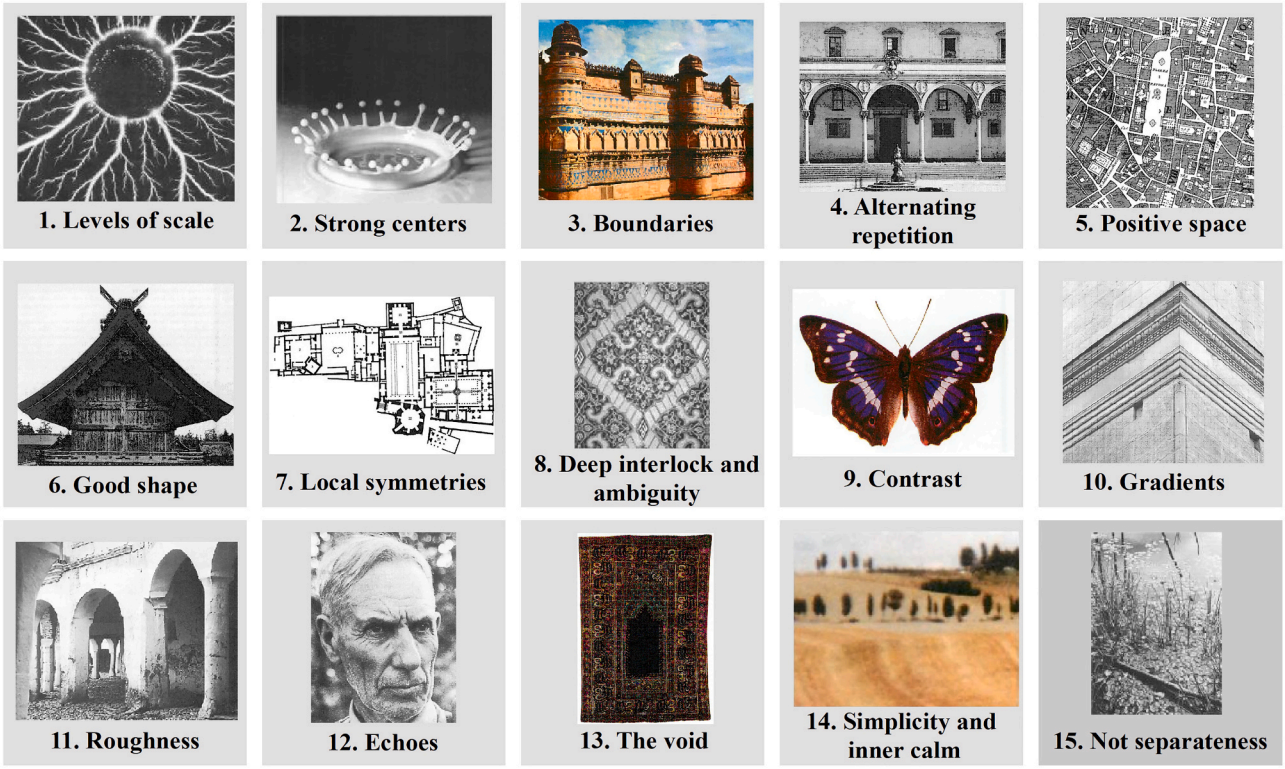


Fig. 1. 15 properties with 15 example patterns (Alexander, 2002–2005). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

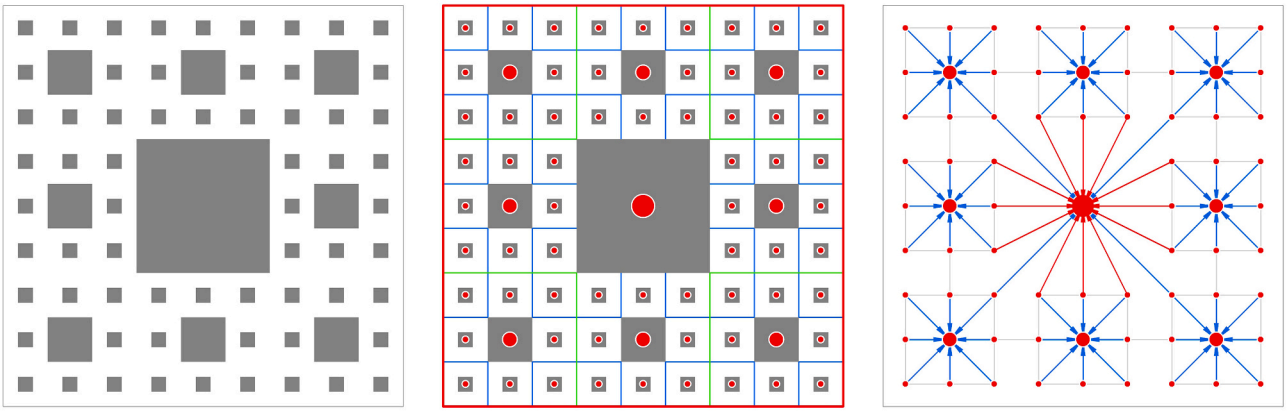


Fig. 2. The fractal Sierpinski carpet, hierarchical levels of the carpet and complex networks based on its spatial configuration. (Note: The Sierpinski carpet, with three scales – 1/3, 1/9, 1/27 – is shown in (a) the centers of each square are represented by the red dots with sizes representing their hierarchical levels. Their 'territories', represented by Thiessen polygons, are shown in (b); (c) shows the complex networks based on (b) and grey lines indicate relationships within the same scales, blue lines indicate relationships across two consecutive scales, and red lines indicate relationship across all scales).(For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

blue polygons shown in Fig. 2b. There are three kinds of links inside the network. Those centers within same hierarchical levels are mutually pointed if their territories neighbor each other (grey links). The squares in two consecutive hierarchical levels are directed from small ones to larger ones if their territories are intersected with each other (blue links). Those squares across all levels are same with consecutive scales (red links). In this way, all the fragmented pieces within a space are glued together, forming a whole, a complex network.

3. Methodology for urban center identification

The study area is the mainland USA, excluding two geographically isolated states: Alaska and Hawaii. We primarily chose the four largest metropolitan areas in terms of population for the case study, including New York, Los Angeles, Chicago, and Houston. We used three big datasets with one raster and two vector formats. We first created hotspots within each city and then created topological representations based on those hotspots. The data and data processing are described in details in the following sections.

3.1. Three geospatial big datasets

The first dataset is the nighttime light imagery (NTL) collected from the Visible and Infrared Imaging Suite (VIIRS) onboard the Suomi National Polar-orbiting Partnership (S-NPP) satellite. S-NPP is a pilot mission for the next generation polar-orbiting operational environmental satellite system of USA. We obtained the monthly Cloud-free Day Light Band (DNB) composite through the academic sector at the Colorado School of Mines (NOAA/NCEI, 2021). The cloud-free DNB NTL data we used has undergone stray-light correction procedure (Elvidge, Baugh, Zhizhin, Hsu, & Ghosh, 2017). The data gathered in January 2021 and the resolution is 15-arc-s (about 500 m at the Equator). We projected the raster using the projection system of 'Contiguous Albers'. Finally, we clipped the projected image using the US boundary.

The second dataset, all American building footprints (BFT), is released from Microsoft Bing Maps, which contains 129,591,582 building footprints. Those building polygons are derived from a two-stage machine learning method (Microsoft, 2022). The first building extraction stage is to recognize the building pixel from aerial images using deep neural networks (DNNs). The second stage is to polygonise the pixels to vector buildings. Ren, Jiang, and Seipel (2019) used the same data source, which contains around 125 million buildings for predicting human activities. Compared to other data sources, Microsoft's data is sourced directly from the recent aerial images using the deep learning method, which provides trustful and abundant data to use.

The third dataset is the streets shapefile from OpenStreetMap (OSM). We downloaded the latest OSM street dataset from the Geofabrik website (accessed on 2021.02.01). We extract the street nodes from OSM data. The street nodes refer to the street junctions together with the dangling street ends. We use an ArcGIS model (Ren, 2016) to extract massive street nodes from the street network. This model first converts the street into segments at the intersections and then converts all segments into start and end points. Finally, it removes the points with two-times of duplicates to derive the final street nodes.

3.2. Data processing

The data processing framework consists of two major parts (Fig. 3). The first part is to define the city boundaries and create hotspots (center candidates) inside each city in a recursive manner. The second part is to

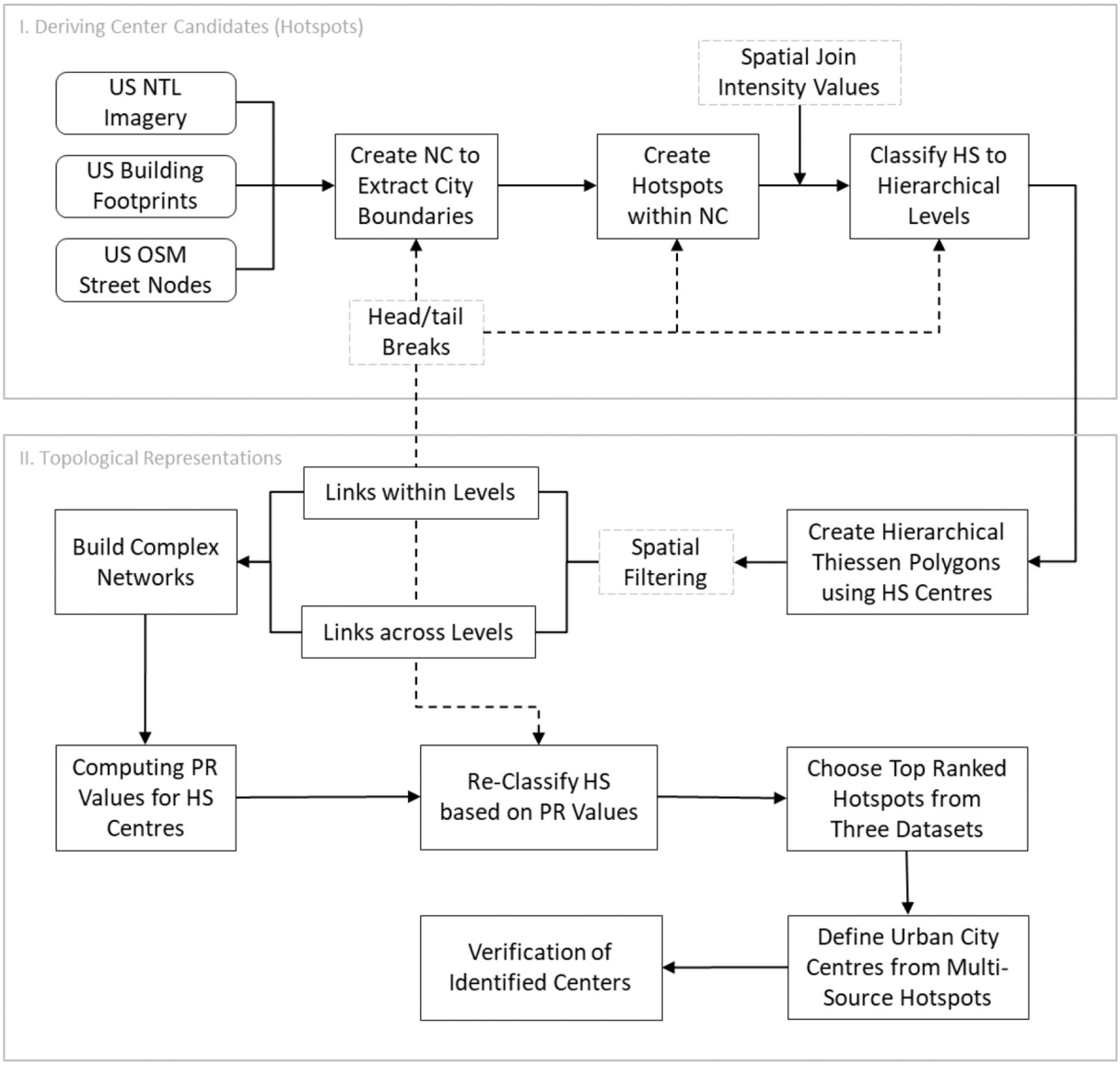


Fig. 3. Data processing framework for identifying urban centers.
(NC is the abbreviation of natural cities and HS is the abbreviation of hotspots.)

construct complex networks based on the spatial configuration of hotspots and compute the centric index, PageRank (PR) score. In this study, we adopted a recursive thinking to derive urban center candidates. Namely, we regard the hotspots at city level as the cities at the country level. For the NTL data, the lit areas with high intensity values are likely to be urban areas. For the point datasets, the urban areas have higher point density.

We use the natural city clustering method for both raster and vector datasets (Jiang & Yin, 2014; Ren et al., 2019). For the NTL data, we chose the intensity value greater than the mean value as the natural cities based on the head/tail division rule. We merge the selected pixels into polygons to create clusters. For the point data, we first built up triangulated irregular network (TIN) and calculated the mean size of the triangles of TIN. Those smaller triangles are dissolved into clusters. Following the same manner, we can also derive hotspots, as shown in Fig. 4. Calculating the mean values of both pixel values and size of triangles is based on the head/tail division rule (Jiang & Liu, 2012). Since our geographic world is fractal in nature (Batty, 2008; Jiang & Liu, 2012), the pixel values and triangle sizes are both heavy-tailed distributed. They can be further classified using head/tail breaks, a data classification scheme for data with a heavy-tailed distribution (Jiang, 2013). If a dataset follows heavy-tailed distribution, we can calculate the mean values of the head parts (greater than the mean) recursively until there is no longer any head part. The partition induces a metric called ht-index (Jiang & Yin, 2014), described by the formula: $ht = m\ (r) + 1$, where m (r) denotes the number of valid means during head/tail breaks.

Previous studies have shown that city sizes and their ranks possess inverse relationships at global, country and even city levels, called Zipf's law (Jiang, Yin, & Liu, 2015; Ma et al., 2021; Zipf, 1949). If we use c to represent the city size and n to represent its rank number, Zipf's law can be denoted by the following equation: $c = n^{-1}$. Zipf's law is more commonly seen as a probability distribution (PDF): $Pr\ (X = x) = x^{-2}$, where the size or population of cities is exact x rather than greater than x. Zipf's law is a special case of power law distribution with the power law exponent of two (Chen, 2012). More generally, a power law distribution can be expressed by: $y = bx^{-\alpha}$, where b is constant and α is the power law exponent. The α is estimated based on maximum likelihood estimation (MLE) method with a lower bound and the Kolmogorov-Smirnov (KS) test (Clauset, Shalizi, & Newman, 2009). The Python package 'power law' (Alstott, Bullmore, & Plenz, 2014) is a fast way to estimate the power law exponents. Alternatively, we can also use the Matlab scripts (Clauset et al., 2009) for the power law detection.

In this study, we generated a series of natural cities and hotspots based on multiple mean values, then calculated their power exponents. We employed an automated method to select one dataset from multiple mean values, starting with the utilization of Zipf's law, which dictates that the power law exponent is approximately 2.0 (with a margin of ±0.1) (Jiang et al., 2015). In cases where multiple datasets exhibit a power law exponent near 2.0, we proceed to compute and contrast the size of the largest hotspot within each dataset. Subsequently, we selected the dataset where the largest hotspot is the smallest.

In this section, we use 40 generated hotspots to illustrate the process in detail (Fig. 5). Fig. 5a shows the original hotspots. We also create centroid points with the intensity values (pixel value or number of points) (Fig. 5b). The next step is to apply head/tail breaks on the intensity values to derive four classes (Fig. 5c). The centroid points are used to create Thiessen polygons. The low-level centroids have smaller Thiessen polygons, shown in blue, and the high-level centroids have larger Thiessen polygons, shown in red in Fig. 5d. Based on the polygon-polygon topology, the complex networks are created with directed links (shown in Fig. 5e). At the same levels, the centroids are mutually pointed if their Thiessen polygons are adjacent to each other. At two different levels, the lower-level centroids point to the higher-level centroids if their Thiessen polygons are intersected. Based on the directed graph, we can compute the degree of wholeness of individual hotspots. Finally, we apply head/tail breaks again to visualize the hotspots based on the degree of wholeness values (shown in Fig. 5f), the largest node is the top ranked urban center.

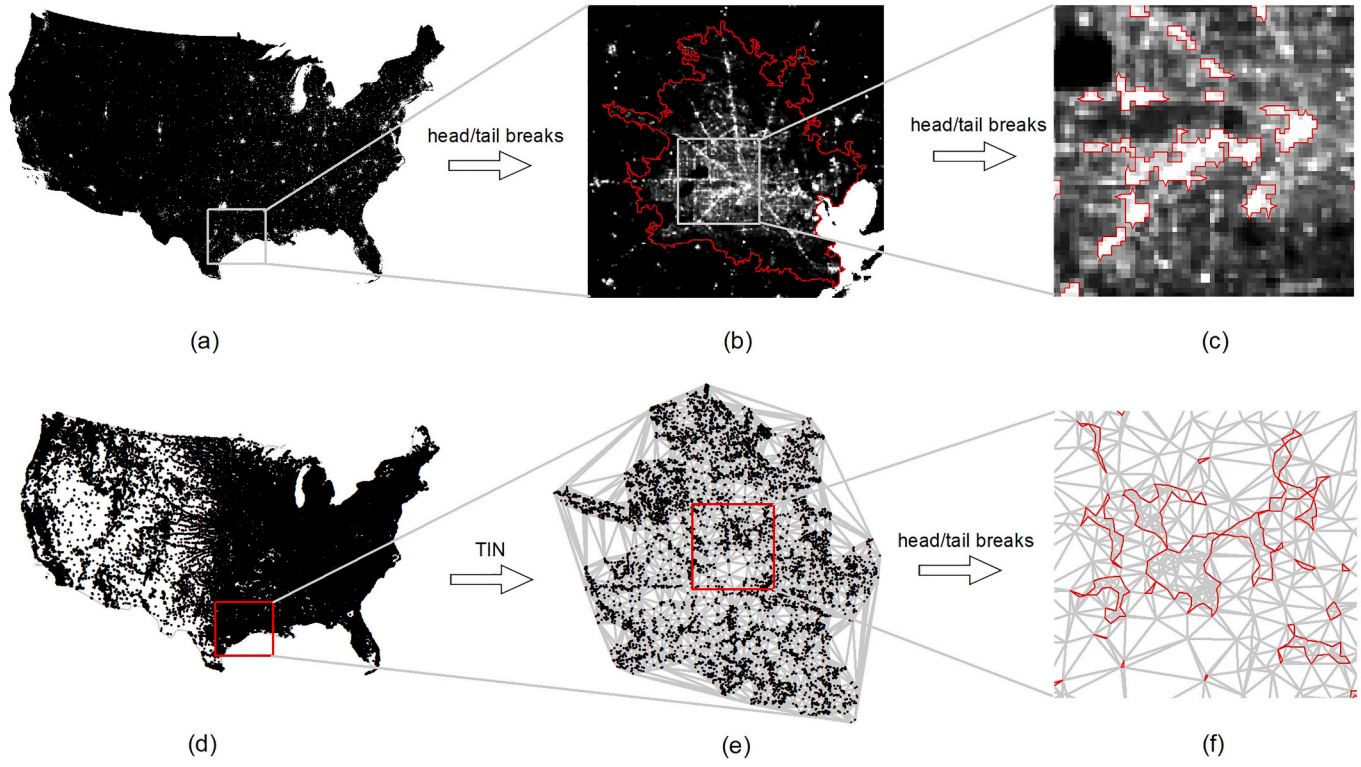


Fig. 4. The hotspot generation process from raster and vector datasets. (Note: (a), (b), (c) are raster process; (d), (e), (f) are points vector process.) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

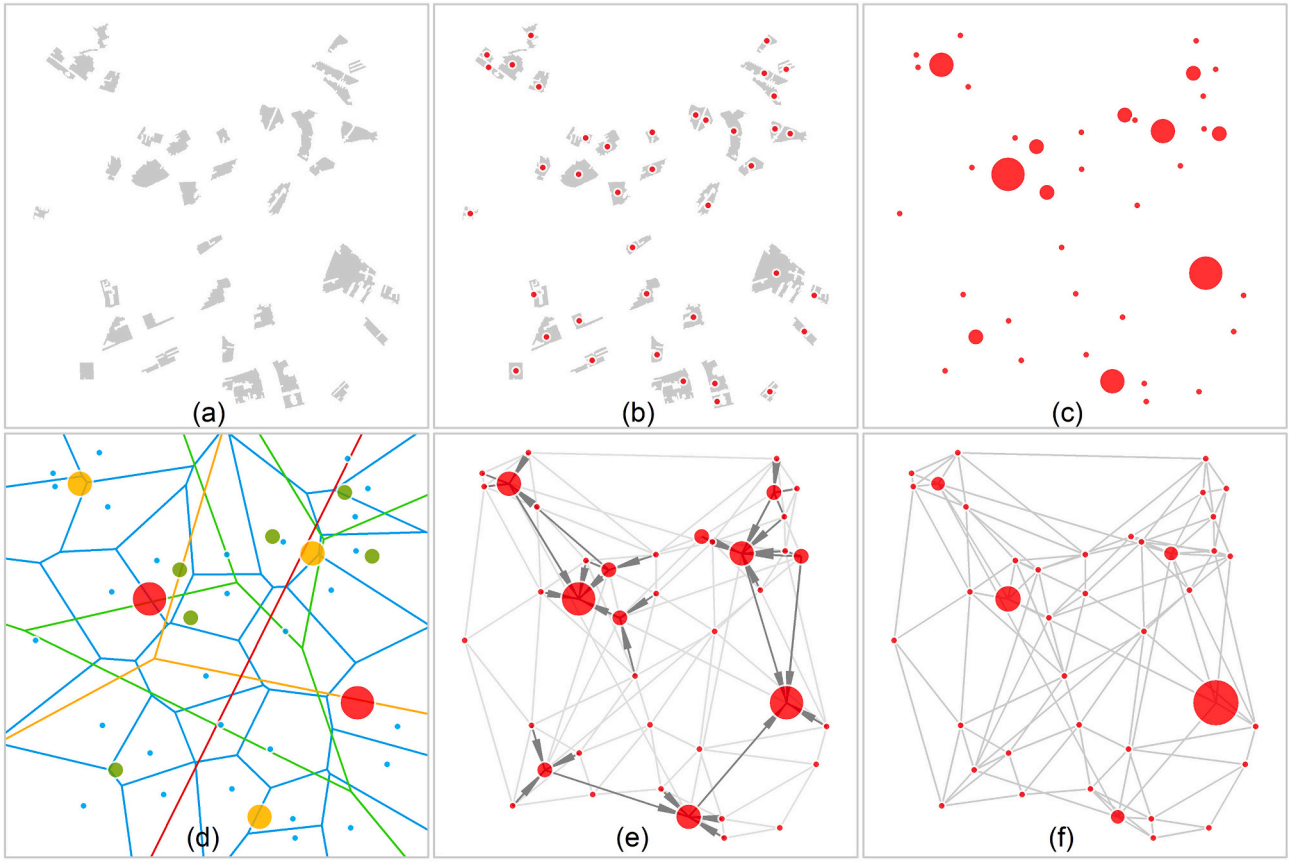


Fig. 5. Finding true urban centers from 40 generated hotspots.(For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The importance of a center is not decided by itself, but relies on its surrounding centers. This is like the importance of a web page, which is not only decided by the number of visits, but by how many other important web pages point to it. The Google's PageRank algorithm is used to calculate the rank of the centers in this study (Page & Brin, 1998). The way of computing PageRank (PR) scores capture the essence of wholeness (Langville and Meyer, 2006; Jiang, 2015). The PR score can be regarded as the degree of wholeness value of individual center. The degree of wholeness for all centers can be quantified by ht-index (Jiang & Yin, 2014). For a high degree of wholeness city, their hotspots possess higher ht-index in terms of PR scores. The last step is to integrate the common parts of three datasets. We expect the true centers have highest PR values in all datasets and they are spatially overlapped. The optimal centric index is expressed as follows:

$$\text{Copt} = \sum_{j \in D} max(r(j)) \tag{1}$$

where C is the centric index, D is intersected hotspots of three datasets, r (j) is the individual PR in each dataset. We first adopt intersection of three datasets in ArcGIS and then use these intersected polygons to label the hotspots that are intersected with those polygons in each dataset. We take the sum of the degree wholeness values of those labelled hotspots in each dataset. Finally, the hotspots that are overlapped and have the highest centric index are regarded as true urban centers. In addition, we adopt intersection over union operation (IoU) (Ma et al., 2021) in ArcGIS to see the intersection extents of three datasets within the center. The IoU metric can be denoted by the following equation:

$$IoU_{ij} = \frac{A_i \cap A_j}{A_i \cup A_j} \tag{2}$$

where $A_{i,\,j}$ are the areas of dataset i, j, the numerator refers the intersection operation and the denominator refers to the union operation. The typical range of IoU is [0,1]. The larger IoU indicates there are more overlapping areas, while a smaller IoU indicates that the two datasets are spatially separated.

4. Identification of US urban centers in four metropolitan areas

We applied head/tail breaks on the NTL data at the country level to define city boundaries in the USA. We derived 12 mean values after head/tail breaks so that the ht-index of US NTL is 13. We examined the power law exponents on the derived natural cities based on top five mean values. We chose the first mean value as the cut-off, with the power law exponent of 1.91, closest to 2.0. We did not calculate the power law exponents for all mean values because, with the higher means, the extents of the natural cities shrink. We prefer a larger spatial extent that contains as much data as possible. In total, we derived 79,725 natural cities in the USA. We selected the boundaries of New York, Los Angeles, Chicago, and Houston for further generation of hotspots (Fig. 6). The New York boundary contains New York City, Philadelphia, and Washington. Chicago's boundary contains Chicago and Milwaukee.

After deriving the city boundaries, we adopt same procedure to generate inner-city hotspots. We generated hotspots based on the top five mean values each time and conducted power law fitting. Finally, we selected hotspots with α values close to 2.0 as the final hotspots. Table 1 shows the power law statistics of hotspots in four cities. We also kept the number of hotspots of four cities similar, for consistency. The number of hotspots of the four cities ranged from 190 to 231. All the hotspots in the four cities conform to Zipf's law with alpha values close to 2.0 (± 0.06).

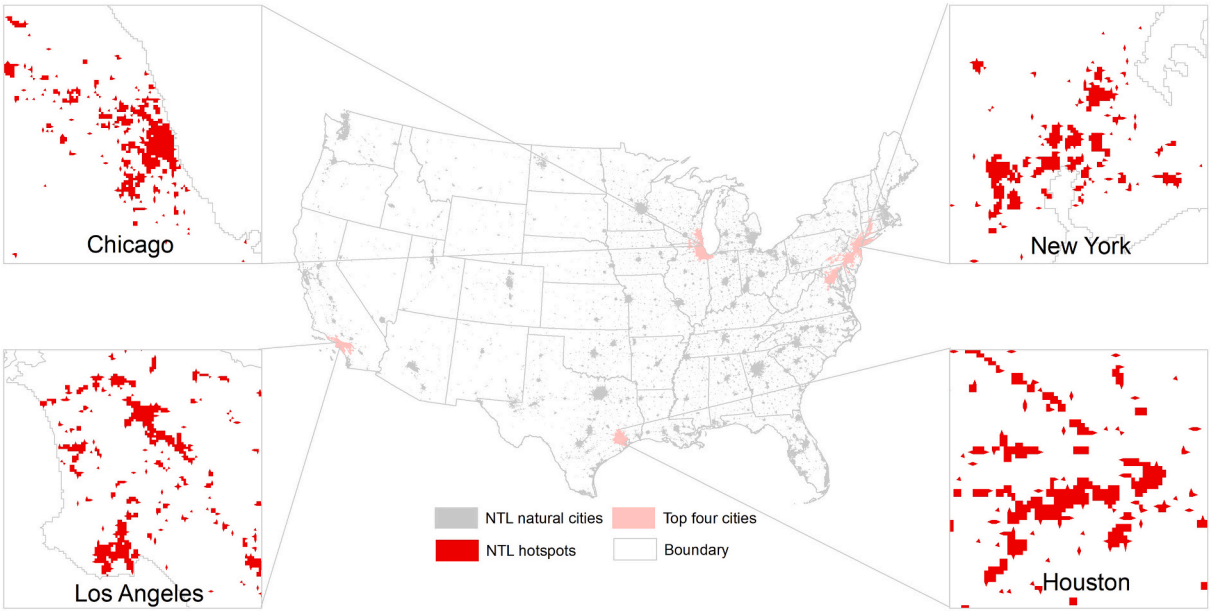


Fig. 6. Natural cities extracted from night-time light image. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Power law statistics for the chosen NTL hotspots in four cities.

City	α	Xmin	P	NC#
New York	1.97	0.83	0.34	230
Los Angeles	2.06	1.51	0.09	231
Chicago	2.04	1.27	0.82	190
Houston	2.04	2.20	0.51	218

Note: The power law detection is based on the method adopted by Clauset et al. (2009).

Based on the total intensity value of hotspots, we adopted head/tail breaks in four cities. The hierarchical levels of four cities vary from four to six; surprisingly, the smallest city (Houston) has the highest ht-index of six (Table 2). The degree of wholeness (PR) value is attributed to each hotspot. We adopted head/tail breaks to visualize the hierarchies of hotspots and derive ht-indices. Table 2 shows that all four cities become more living in terms of ht-index increasing at least one. New York, Los Angeles, and Chicago have the same six hierarchical levels based on the degree of wholeness. The highest ht-index of wholeness is seven in Houston, even though it has smallest population among four cities.

Fig. 7 shows the variations of two phases of the hotspots in four cities. The left panels are the original phase of the hotspots in terms of size and the right panels show the future phase of the hotspots in terms of PR value. In New York, some hotspots located at the lower left shrink and some hotspots in Philadelphia and Manhattan area are enhanced; those enhanced hotspots are likely to be the urban centers in the greater New York area. In Los Angeles, one big hotspot at upper left of the left panel shrinks after the adaptation process and true centers are detected in the right panel. In Chicago, city centers move from right to left, which suggests a better location considering a larger extent, including Milwaukee. The center of Houston moves east slightly but remains close to the original hotspots. We can witness variations from the left to right panels. The right panels for four cities possess higher ht-indices, which suggests a more living structure than the old hotspots.

Table 2
Ht-index for the NTL hotspots in four cities.

City	Ht hotspots	Ht city centers	Ht variation
New York	5	6	1
Los Angeles	4	6	2
Chicago	4	6	2
Houston	6	7	1

In addition, the buildings and street nodes are also used to identify the urban center. The same procedure has been conducted as NTL data. For the US building footprints, the hotspots of New York, Los Angeles, and Chicago have power law exponents close to 2.0 ± 0.3. Houston has a larger exponent of 2.69, which means the hotspots is more heterogenous than others. New York City has the most living hotspots, with an ht-index of 10. Houston is more living than Los Angeles and Chicago, although it has a small spatial extent. For the OSM data, we also found that New York, Chicago, and Los Angeles possess Zipf's law with power law exponents close to 2.0 ± 0.05. As for Houston, the scaling exponent is 2.15, which is slightly larger than others. The hierarchical levels change slightly in New York, with the ht-index only increasing one level. The other three cities encountered ht-indices increasing by two hierarchical levels. The Chicago has the highest ht-index, with 8. In order to avoid repetition, we append the results of building and streets in Appendix A.

In order to compare the differences between the original hotspots and urban centers after conducting topological representations, we also calculated the correlations between original intensity values, such as pixel values, number of buildings and number of street nodes, and PR scores. The high correlation indicates there are less variations of two datasets and *vice versa*. Table 3 shows the three groups of correlations in four cities. There is higher correlation of OSM data in Los Angeles, which might suggest the original street networks already have a high degree of living structure. On average, the building footprints in four cities have higher correlation, indicating slight changes before and after topological representation.

Finally, we focus on the top-ranked hotspots and their spatial extents. We combine the top ranked PR centers in three datasets together to explore the optimal centers. The optimal centric indices are 0.034, 0.059, 0.035, and 0.034 in New York, Los Angeles, Chicago, and Houston, respectively. We select the hotspots that contribute to the optimal centric indices as true urban centers. Fig. 8 shows the identified urban centers. The New York boundary contains four major cities: New York City, Philadelphia, Baltimore and Washington. The identified urban center is in Philadelphia, in the middle between New York City

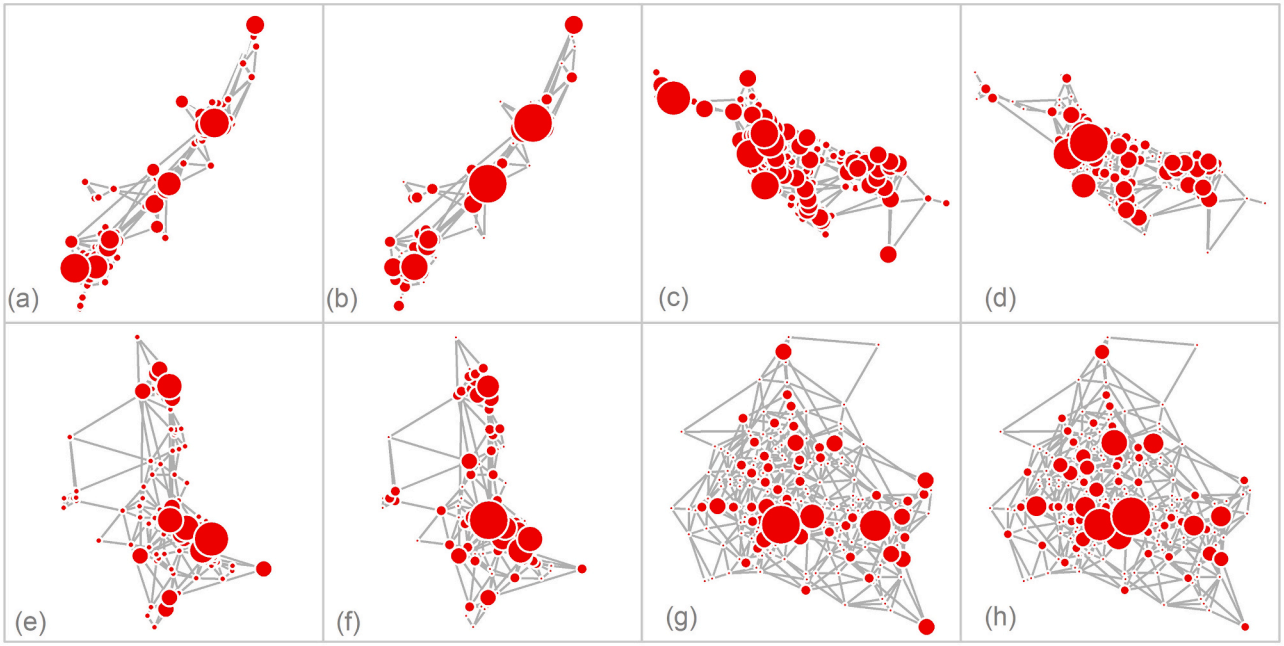


Fig. 7. The hierarchical levels of hotspots before and after topological representation of NTL in the four largest American cities; (a) and (b) are New York, (c) and (d) are Los Angeles, (e) and (f) are Chicago, and (g) and (h) are Houston.(For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Correlations between intensity value and PR.

City	NTL	BFP	OSM
New York	0.34	0.68	0.69
Los Angeles	0.41	0.70	0.84
Chicago	0.36	0.57	0.31
Houston	0.59	0.76	0.34

and Baltimore. The identified center takes up 97.23% of the downtown area. The center in Los Angeles also covers 97.97% of the downtown area. In Chicago, the top-ranked center is located at the upper left of the old downtown area, yet the second center covers 100% of the downtown area, known as the Loop. The result indicates that, morphologically, the city center of Chicago could be at the northwest of the old center considering a larger area. In Houston, the top hotspot is too large to show the detail, so we recursively conduct the topological representation inside the largest OSM hotspots. Three-quarters (75.5%) of the downtown area is covered by the identified urban center. The results show an overall accuracy of 90.23% in three cities.

Morphologically, different data sources take different proportions of the urban area. Whether they are overlapped or dispersed can be analyzed using the IoU method. Philadelphia and Houston have higher IoU values (0.34 and 0.32, respectively) than Chicago and Los Angeles. For Chicago, the left center has less overlapped area than the right center, with IoU values of 0.25 and 0.30, respectively. The three datasets in Los Angeles do not overlap each other like the other cities, with the lowest IoU value of 0.13. In four cities, the downtown areas are mostly covered by NTL data. The OSM data takes larger proportions in Philadelphia and Chicago. The NTL data dominates in Los Angeles and Houston. In addition, the building centers are relatively far from the downtown center, which implies that they could be popular residential areas for people who work in the CBD. We roughly estimated the distances between the centroids of building centers and downtown areas. Houston has the nearest residential center of 3.9 km. The distances in Philadelphia and Los Angeles are 8.1 km and 8.3 km. Chicago has the longest distance, 11.15 km, implying a longer commute distance than other cities.

5. Further discussions on the data and the topological representation

There are two laws rooted in the topological representations: Scaling law and Tobler's law. Scaling law characterizes the far more smaller hotspots than larger ones across all scales in a city. Tobler's law, is also known as first law of geography (Tobler, 1970), characterizing the hotspots are similar in size within the same scale. The links across all scales and links within same levels are created based on the Scaling law and Tobler's law respectively. These two laws complement each other to depict the homogeneity and heterogeneity phenomena in our urban environment. This study not only identified the spatial extents of urban centers, but also characterized the hierarchical levels of those centers. Those centers with higher hierarchies are likely to form the image of the city in human mind (Lynch, 1960). A well-designed urban center can be well-perceived by human beings. In this regard, the topological representation is a remarkable leap that not only helps us understand how city looks currently, but also gives us guidance about how to design a sustainable and livable city in the future.

In this study, the three datasets complement each other to characterize the urban centers from both functional and morphological perspectives. The nightlight image is a good proxy of human activities during night-time (Ma et al., 2021; Sutton, 1997; Wu, Zhao, & Jiang, 2018). The building footprints are a permanent proxy of human movement destinations, which is suitable for predicting human activities (Ren et al., 2019) and characterizing urban forms (Ma, Seipel, Brandt, & Ma, 2022). A street network is a good data source for studying urban form since streets constitute a backbone of urban environment (Hillier & Hanson, 1984; Ma et al., 2019). The key to identifying a true urban center is determining how to combine three datasets together. In the present study, we assumed that urban centers are supposed to be covered by the high degree of wholeness value hotspots of three datasets together. We first spatial intersected the hotspots of three datasets and took out the hotspots that are intersected three times. We then selected those intersected hotspots with the highest wholeness values in three

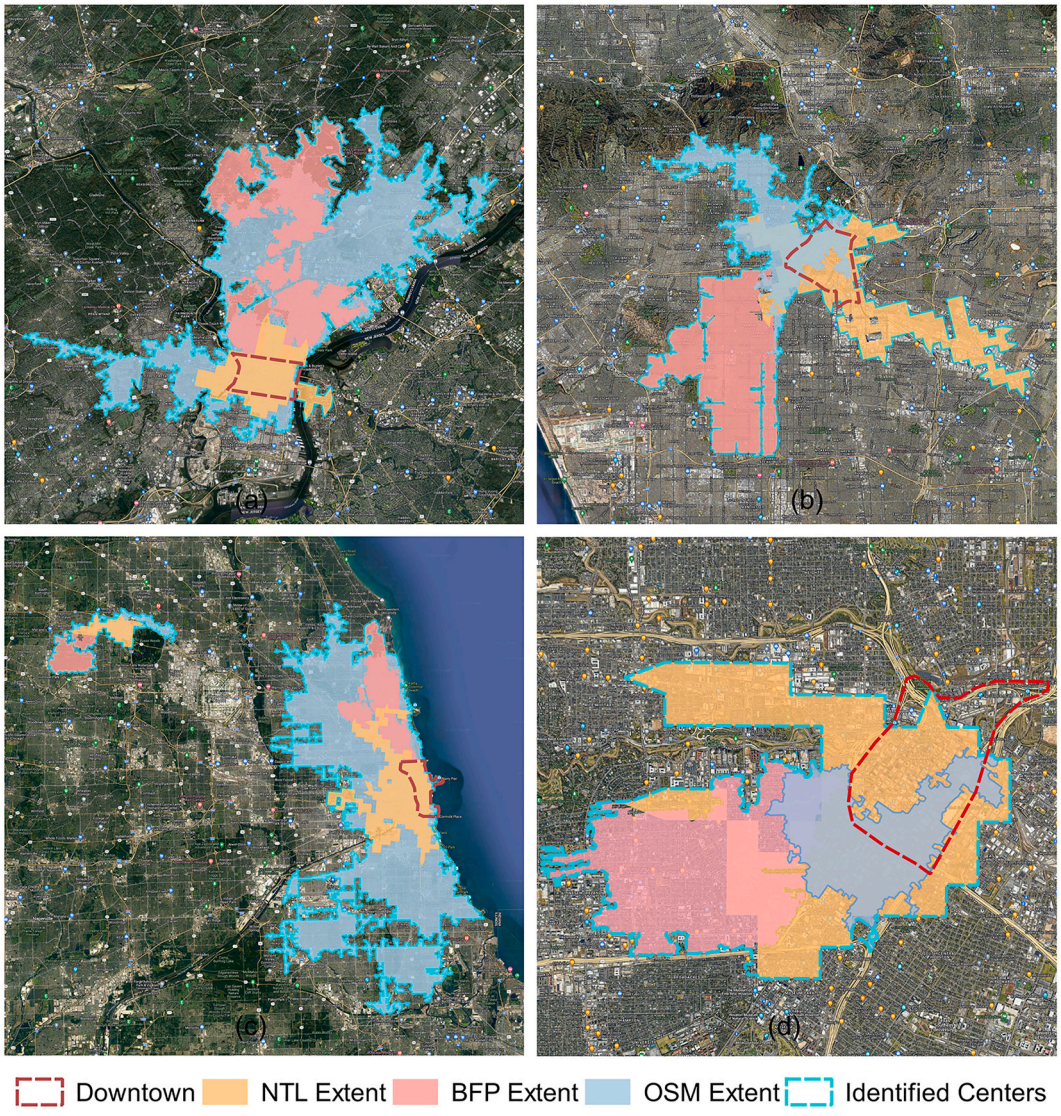


Fig. 8. Identified urban centers in four largest US cities. (Note: (a), (b), (c), and (d) are Philadelphia in the New York polycentric area, Los Angeles, Chicago and Houston, respectively. The basemap is Google Satellite Hybrid.) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

datasets as the urban centers. In this way, three datasets are fused together to ensure the reliability of our results.

In order to further illustrate the necessity of topological representation and compare the difference between our approach to the previous approaches, we generated the hotspots from three data sources respectively using without topological representation and data fusion. The downtown areas in four cities are used to make comparisons. Fig. 9 shows that three datasets led to three different results. The highest intensity value for nighttime light image data is located at the bottom of Los Angeles which cannot be regarded at main urban centers. Three datasets generated different urban centers. The uncertainty of multi-source geospatial data makes the data fusion using the proposed centric index in formula (1) useful and necessary.

Furthermore, by comparing the hotspots before and after the topological representation, we found that some hotspots expand and some shrink, while some remain unchanged in terms of sizes and degrees of wholeness. The expanding centers have lots of surrounding supporters, while shrinking centers lack supporters. Some unchanged hotspots reached a steady phase, indicating that their sizes and locations are well adapted to their surrounding centers. The shrinking and expanding hotspots show the real urban development. Some newly built areas in big cities continuously attract people, while some old urban centers are on the wane. In addition, the shrinking and expanding phenomena indicate that our method can effectively and objectively avoid the misjudgments of urban centers that have high values in size, but fewer surrounding hotspots. For example, Fig. 10(a) shows that before topological representation, the identified urban center is located at the northwest of the downtown Los Angeles, which is a mis-detected center. After topological representation shown in Fig. 10(b), the error center shrinks and the identified centers located at the central expand to a larger size. In this way, our approach can automatically detect the mis-detected centers and make an adjustment. In addition, the identified centers can help us design the future urban centers with their possible locations.

The famous architect Christopher Alexander stated that the city is not a tree but a semilattice (Alexander, 1965). A tree is a graph-theoretical structure with nodes and edges that has no overlapped links between each node, while a semilattice is a complex network that encloses some intertwined links. It is such overlapping substructures that make a structure complex, beautiful and sustainable (Jacobs, 1961, Alexander, 2002–2005, Jiang & Huang, 2021). In the present study, we added the links across all scales to make it more complex than the

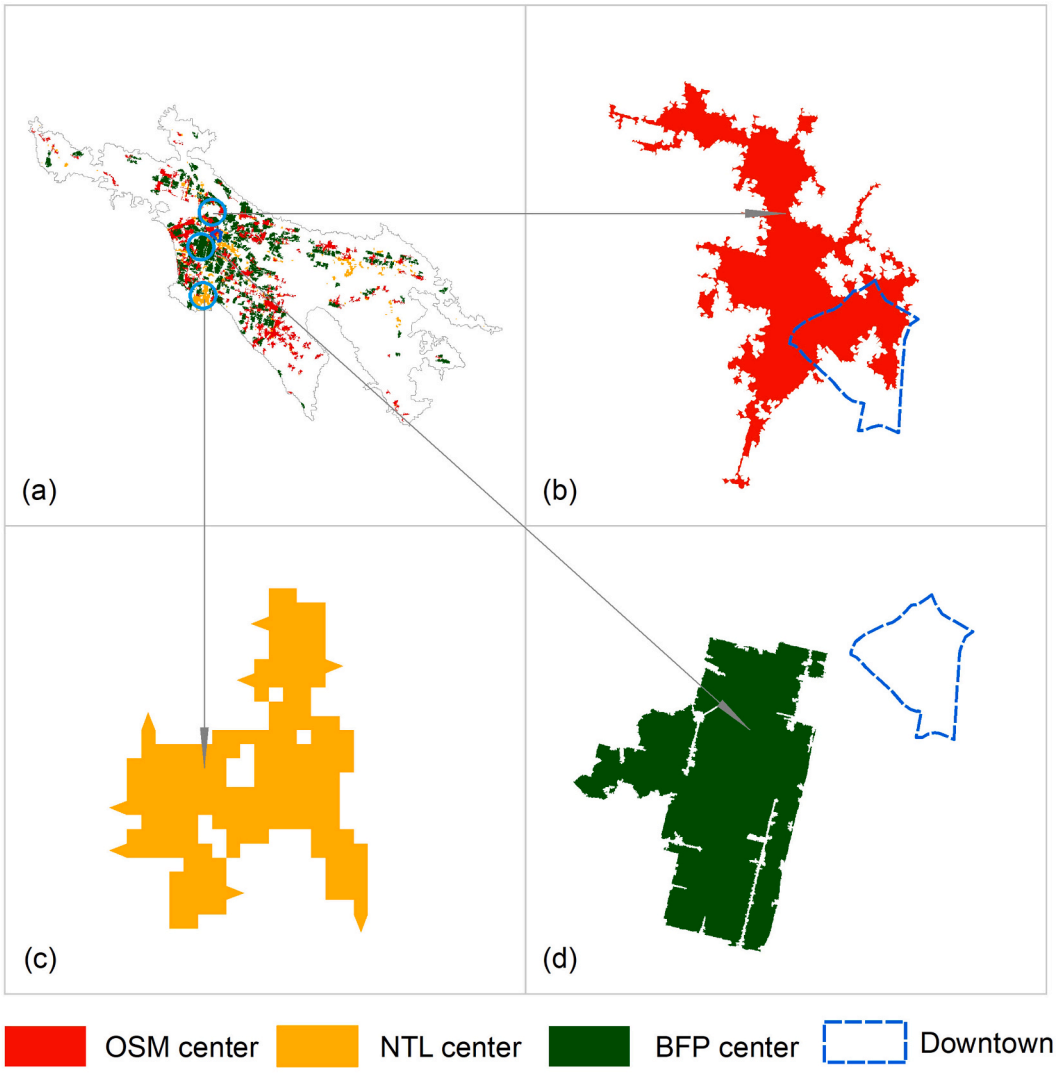


Fig. 9. The top-ranked hotspots in Los Angeles without topological representation (Note: (a) (b) (c) (d) are respectively all hotspots of three datasets, the largest OSM center, NTL center and BFP center). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

network that only has relationships at two consecutive scales. There are pervasive interactions of people, goods, and services in our urban environment. The interactions refer to the human's presence, traffic flows and other socio-economic activities. The hotspots are the 'hubs' of those interactions, where human activities are more concentrated than other places. The interactions are complex but organized rather than chaotic, which makes cities livable (Jacobs, 1961). The three kinds of links of the topological representation vividly and accurately characterize these interactions. In order to evaluate how well the topological representation can character the interactions or relationships between hotspots, more data sources from social economic perspectives will be needed in the future such as logistics data, taxi routes and cell phone signaling data. The present study identifies urban centers from a structural perspective to reveal the wholeness property of urban structure. In this regard, the results could be different when considering more socio-economic data such as employment rate and GDP data. Moreover, although this study gives the future location of the urban centers, due to the lack of historical data, it is difficult to evaluate how well the prediction is. The future study can be focused on the aforementioned perspectives.

6. Conclusion

In this paper, we identified urban centers using a topological approach with three different sources of geographic big data. With the topological representation, we can see the underlying scaling structure of hotspots within a city. The NTL hotspots exhibit Zipf's laws in four cities. The latent notion of wholeness can be depicted by the topological representations using centers as nodes and spatial relationships as links. All four cities possess living structures with higher ht-indices afterwards. We combined three datasets together for urban center identification both spatially and statistically. The three datasets complement each other to reinforce the reliability of the results. The high identification accuracy suggests the reliability of our approach. The Chicago case indicates that if we take the larger area as consideration, the future city center could be located at the west of the original center.

Our urban environment is more living with more recursive substructures. Through topological representation, we observe that urban centers exhibit a higher sense of coherence and degree of wholeness. Rather than isolated entities, these urban hotspots form a semilattice structure, akin to the complex nature of cities themselves. This proposed topological framework offers a valuable lens through which to comprehend our urban environment, promoting a holistic understanding through topological analysis. The geospatial data utilized in this

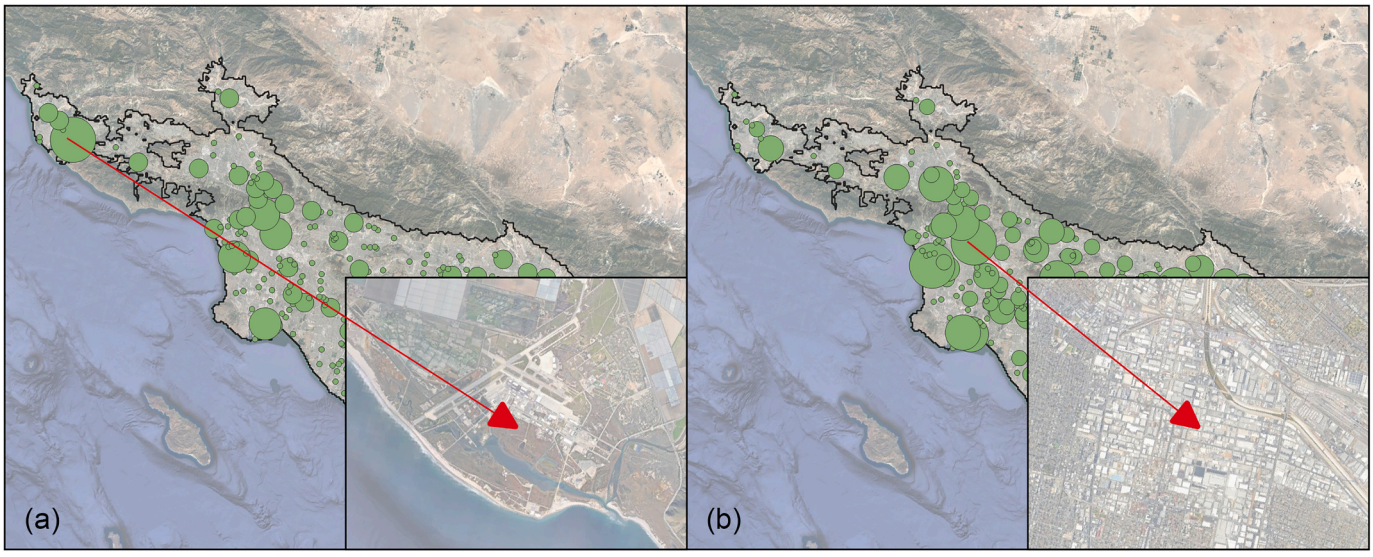


Fig. 10. The hotspots shrinkage after topological representation due to few neighbors in Los Angeles, (a) the greed dots show the original hotspots based on intensity values, (b) the green dots show the hotspots levels based on PageRank value after topological representation. Note that the zoom-in frame in (a) shows the mis-represented center of airport, the zoom-in frame in (b) shows the adjusted center of Los Angeles. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

study provide critical insights into the underlying urban structures and spatial patterns. It's important to note that our definition of urban centers in this study aligns with Alexander's theory of centers, highlighting their distinctive spatial attributes. Furthermore, in future research, the integration of socio-economic data holds the potential to enhance our ability to identify conventional urban centers and provide practical guidance for urban planning and related studies.

CRediT authorship contribution statement

Zheng Ren: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Stefan Seipel:** Supervision, Validation, Writing - review & editing. **Bin Jiang:** Conceptualization, Supervision, Funding acquisition, Investigation, Methodology, Project administration, Validation, Writing - review & editing.

Acknowledgements

We would like to thank the anonymous referees for their insightful and deep comments, but any shortcomings are ours. The paper was partially supported by the Swedish Research Council FORMAS through the ALEXANDER project with grant number FR-2017/0009 (2017-00824).

Appendix A

The power law statistics for the chosen hotspots of building footprints data and OSM street nodes data are included. And the ht-indices of building hotspots and street nodes hotspots are included in the Appendix A.

Table A1
Power law statistics for the chosen building hotspots in four cities.

City	α	Xmin	P	NC#
New York	2.30	3012	0.026	13,105
Los Angeles	2.03	477	0.028	5164
Chicago	1.92	315	0.0018	6098
Houston	2.69	896	0.28	4250

Table A2
Power law statistics for the chosen OSM hotspots in four cities.

City	α	Xmin	P	NC#
New York	2.05	23	0.078	24,072
Los Angeles	2.02	75	0.024	5663
Chicago	1.99	36	0.22	5999
Houston	2.15	32	0.82	3439

Table A3
Ht-index for the ABF hotspots in four cities.

City	HT hotspots	HT city centers	HT variation
New York	9	10	1
Los Angeles	6	7	1
Chicago	7	8	1
Houston	8	9	1

Table A4
Ht-index for the OSM hotspots in four cities.

City	HT hotspots	HT city centers	HT variation
New York	6	7	1
Los Angeles	5	7	2
Chicago	6	8	2
Houston	5	7	2

References

Alexander, C. (1965). A city is not a tree. *Architectural Forum, 122*(1,2), 58–62.
Alexander, C. (2002–2005). *The Nature of Order: An essay on the art of building and the nature of the universe*. Berkeley, CA: Center for Environmental Structure.
Alonso, W. (1964). *Location and land use: Toward a general theory of land rent*. Cambridge, MA: Harvard University Press.
Alstott, J., Bullmore, E., & Plenz, D. (2014). Powerlaw: A Python package for analysis of heavy-tailed distributions. *PLoS ONE, 9*(4). e95816.
Anas, A., Arnott, R., & Small, K. (1998). Urban spatial structure. *Journal of Economic Literature, 36*, 1426–1464.
Batty, M. (2008). The size, scale, and shape of cities. *Science, 319*, 769–771.
Batty, M., & Longley, P. (1994). *Fractal cities: A geometry of form and function*. London: Academic Press.
Bohm, D. (1980). *Wholeness and the implicate order*. London and New York: Routledge.
Burger, M., & Meijers, E. (2012). Form follows function? Linking morphological and functional polycentricity. *Urban Studies, 49*(5), 1127–1149.
Chen, Y. (2012). The mathematical relationship between Zipf`s law and the hierarchical scaling law. *Physica A, 391*(11), 3285–3299.
Chen, Z., Yu, B., Song, W., Liu, H., Wu, Q., Shi, K., & Wu, J. (2017). A new approach for detecting urban centers and their spatial structure with nighttime light remote sensing. *IEEE Transactions on Geoscience and Remote Sensing, 55*(11), 6305–6319.
Christaller, W. (1933). *Central places in southern Germany*. Prentice Hall: Englewood Cliffs, N.J [1966].
Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review, 51*, 661–703.
Elvidge, C. D., Baugh, K., Zhizhin, M., Hsu, F. C., & Ghosh, T. (2017). VIIRS night-time lights. *International Journal of Remote Sensing, 38*, 5860–5879.
Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal, 69*, 211–221.
Goodchild, M. F., & Mark, D. M. (1987). The fractal nature of geographic phenomena. *Annals of the Association of American Geographers, 77*(2), 265–278.
Hillier, B., & Hanson, J. (1984). *The social logic of space*. Cambridge: Cambridge University Press.
Jacobs, J. (1961). *The death and life of great American cities*. New York: Random House.
Jiang, B. (2013). Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution. *The Professional Geographer, 65*(3), 482–494.
Jiang, B. (2015). Wholeness as a hierarchical graph to capture the nature of space. *International Journal of Geographical Information Science, 29*(9), 1632–1648.
Jiang, B. (2018). A topological representation for taking cities as a coherent whole. *Geographical Analysis, 50*(3), 298–313.
Jiang, B., & Huang, J. (2021). A new approach to detecting and designing living structure of urban environments. *Computers, Environment and Urban Systems, 88*, 1–10.
Jiang, B., & Liu, X. (2012). Scaling of geographic space from the perspective of city and field blocks and using volunteered geographic information. *International Journal of Geographical Information Science, 26*(2), 215–229.
Jiang, B., & Yin, J. (2014). Ht-index for quantifying the fractal or scaling structure of geographic features. *Annals of the Association of American Geographers, 104*(3), 530–541.
Jiang, B., Yin, J., & Liu, Q. (2015). Zipf`s Law for all the natural cities around the world. *International Journal of Geographical Information Science, 29*, 498–522.
Langville, A. N., & Meyer, C. D. (2006). *Google's PageRank and beyond: The science of search engine rankings*. Princeton, N.J: Princeton University Press.
Liu, X., Huang, J., Lai, J., Zhang, J., Senousi, A. M., & Zhao, P. (2021). Analysis of urban agglomeration structure through spatial network and mobile phone data. *Transactions in GIS, 00*, 1–21.
Lynch, K. (1960). *The image of the City*. Cambridge, MA: The MIT Press.
Ma, D., Guo, R., Jing, Y., Zheng, Y., Zhao, Z., & Yang, J. (2021). Intra-urban scaling properties examined by automatically extracted city hotspots from street data and nighttime light imagery. *Remote Sensing, 13*(7), 1322.
Ma, D., Omer, I., Osaragi, T., Sandberg, M., & Jiang, B. (2019). Why topology matters in predicting human activities. *Environment and Planning B: Urban Analytics and City Science, 46*(7), 1297–1313.
Ma, L., Seipel, S., Brandt, S. A., & Ma, D. (2022). A new graph-based fractality index to characterize complexity of urban form. *International Journal of Geo-Information, 11*, 287.
Ma, M., Lang, Q., Yang, H., Shi, K., & Ge, W. (2020). Identification of polycentric cities in China based on NPP-VIIRS nighttime light data. *Remote Sensing, 12*(19), 3248.
Mandelbrot, B. (1982). *The fractal geometry of nature*. New York: W. H. Freeman and Co.
McDonald, J. F., & Prather, P. J. (1994). Suburban employment centers: The case of Chicago. *Urban Studies, 31*, 201–218.
McMillen, D. P. (2003). Identifying sub-centers using contiguity matrices. *Urban Studies, 40*, 57–69.
Meijers, E. (2005). Polycentric urban regions and the quest for synergy: Is a network of cities more than the sum of the parts? *Urban Studies, 42*, 765–781.
Microsoft. (2022). available online: https://github.com/Microsoft/USBuildingFootprints (Data accessed on September 2022).
Murphy, R. E., & Vance, J. E. (1954). Delimiting the CBD. *Economic Geography, 30*, 189–222.
NOAA/NCEI. (2021). Available online (accessed on 30th December 2021) https://eogdata.mines.edu/nighttime_light/monthly/v10/2021/202101/vcmcfg/.
Page, L., & Brin, S. (1998). The anatomy of a large-scale hypertextual web search engine. *Proceedings of the Seventh International Conference on World Wide Web, 107–117*.
Ren, Z. (2016). Extracting the street nodes data using Natural Cities Model in ArcGIS. http://www.arcgis.com/home/item.html?id=47b1d6fdd1984a6fae916af389cdc57d.
Ren, Z., Jiang, B., & Seipel, S. (2019). Capturing and characterizing human activities using building locations in America. *International Journal of Geo-Information, 8*(5), 200.
Roth, C., Kang, S. M., Batty, M., et al. (2011). Structure of urban movements: Polycentric activity and entangled hierarchical flows. *PLoS One, 6*(1), Article e15923.
Salingaros, N. (1997). Life and complexity in architecture from a thermodynamic analogy. *Physics Essays, 10*(1), 165–173.
Schmitt, A., Uth, P., Standfuß, I., Heider, B., Siedentop, S., & Taubenböck, H. (2023). Quantitative assessment and comparison of urban patterns in Germany and the United States. *Computers, Environment and Urban Systems, 100*, 101920.
Sun, M., & Fan, H. (2021). Detecting and analyzing urban centers based on the localized contour tree method using taxi trajectory data: A case study of Shanghai. *ISPRS International Journal of Geo-Information, 10*(4), 220.
Sun, Y., Fan, H., Li, M., & And, Z. A. (2016). Identifying the city center using human travel flows generated from location-based social networking data. *Environment and Planning. B, Planning & Design, 43*(3), 480–498.
Sutton, P. C. (1997). Modeling population density with night-time satellite imagery and GIS. *Computers, Environment and Urban Systems, 21*, 227–244.
Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography, 46*(2), 234–240.
Wu, W., Zhao, H., & Jiang, S. (2018). A Zipf`s law-based method for mapping urban areas using NPP-VIIRS nighttime light data. *Remote Sensing, 10*, 130.
Xie, Z., Yuan, M., Zhang, F., Chen, M., Shan, J., Sun, L., & Liu, X. (2023). Using remote sensing data and graph theory to identify polycentric urban structure. *IEEE Geoscience and Remote Sensing Letters, 20*, 1–5.

Yang, Z., Chen, Y., Guo, G., Zheng, Z., & Wu, Z. (2021). Using nighttime light data to identify the structure of polycentric cities and evaluate urban centers. *Science of the Total Environment, 780*, 146586.

Zhou, Y., He, X., & Zhu, Y. (2022). Identification and evaluation of the polycentric urban structure: An empirical analysis based on multi-source big data fusion. *Remote Sensing, 14*, 2705.

Zipf, G. K. (1949). *Human behaviour and the principles of least effort*. Cambridge, MA: Addison Wesley.