Quantifying urban areas with multi-source data based on percolation theory

Wenpu Cao, Lei Dong, Lun Wu, and Yu Liu

1 Institute of Remote Sensing and Geographic Information Systems, School of Earth and Space Sciences, Peking University, Beijing 100871, China

(Dated: October 29, 2019)

How to define an urban area – the basic spatial unit for urban planning and studies – has been a long-standing problem for researchers and policymakers [1–3]. This problem has become more important in recent decades because of the emerging of a large number of fast-urbanizing regions around the world (e.g., China and India). However, due to the complexity of the urban system, especially the fuzzy urban-rural transition, consistent and robust measurement to quantify urban areas has remained elusive.

For a long time, governments have relied heavily on the administrative boundaries to address urban issues (e.g., environmental and sustainable problems); and many location-based policies are also implemented based on the administrative divisions. However, administrative divisions are mainly divided by historical, political, and geographical reasons, making it difficult to reflect the socio-economic dynamics of cities. Additionally, administrative divisions are incomparable across different countries and periods [4]. Therefore, some countries turn to employ socio-economic indicators (e.g., population, economic activity, commuting) to re-divide urban areas. For example, metropolitan areas (MAs), the most commonly used socio-economic boundaries, define the urban areas as closely related regions in terms of socio-economic connection [5]. However, the construction of MAs has three main shortcomings. First, the detailed data (e.g., census data, commuting survey data) to construct MAs are lacking in many developing countries. Second, the data collection process for MAs is time-consuming and expensive, making it unable to capture the rapid urbanization process in fast-growing regions. Third, the standard to define MAs is different in different countries. There is still a lack of a unified approach to obtain functional urban areas, which can be applicable to all countries.

Remote sensing data, especially the satellite-based data, provide continuous and consistent observations of urban activities on earth [6]. They are easily accessible for most countries and thus have been widely used to study urban dynamics at different spatial scales [7–11]. Based on different urban characteristics (e.g., multispectral information, light emissions, morphological structures), several global urban maps have been derived from the remote sensing data, such as the MODIS500m [12, 13], GHSL [14], GlobeLand30 [15], and GUF [16] products. Meanwhile, some emerging urban data with humans as sensors, such as volunteered geographical information (VGI, e.g., OpenStreetMap) [17] and social sensing data (e.g., mobile phone, social media) [18], have also shown great potential in revealing the socio-economic boundaries of cities [19–21]. In addition to these multi-source datasets, some new methods have also been developed to delimit urban areas [6, 11, 21–24]. Especially, the City Clustering Algorithm (CCA), which defined cities as the maximally connected populated areas, has attracted great attention due to its simplicity and efficiency [8, 20, 25–28]. Benefited from advanced computing techniques and easily accessed data sources, CCA or other data-driven methods can derive urban areas in a timely and simple manner. However, due to the complexity of the urban system and the fuzzy urban-rural transition, these methods still have difficulties in...
Complexity science of cities sheds some light on the optimal threshold problem. Urban systems, as typical self-organized systems, display some universal macroscopic patterns, such as Zipf’s law [28, 29], scaling laws [30], and fractal characteristics [31]. Previous studies have shown that these macroscopic patterns emerge at the critical point of the urban system [32], and several physical models have been adopted to study the critical phenomena of cities [33–35]. Notably, the percolation model, a typical model for studying complex systems [36], was used on the road network data of Britain to show that the urban system emerges at the critical point of the percolation process [37, 38]. These works inspire us to address the optimal threshold problem with the percolation model.

In this paper, we propose a novel method to extract urban areas from multi-source urban data. We adopt a broader definition of urban areas as maximally connected areas that have more urban elements (i.e., population, infrastructure, economic activity) than non-urban areas, and these three urban elements are widely acknowledged in the urban geography and urban economics fields to measure the urbanization process [3, 26, 37]. We find the optimal urban/non-urban threshold solely through the input data themselves by considering the critical nature of urban systems. Specifically, we traverse all potential thresholds and aggregate the urban units into a cluster system under each threshold. Based on the percolation theory, we get the optimal urban areas when the whole system is at the critical point. To verify our method, we investigate the geographical layouts of urban areas derived by three datasets (population, road, and nighttime light). Despite the datasets of great difference, we find that: i) our method can capture the similar geographical distributions of urban areas; and ii) the rank-size distribution of urban areas fits well with Zipf’s law, a fundamental law of urban systems. We also compare our results with the Global Human Settlement Layer (GHSL). The derived urban areas by different datasets show good agreement with the reference data and can be further improved by data fusion. These findings demonstrate the effectiveness of our method and also deepen our understanding of cities. From the perspective of applications, the efficient, consistent, and low-cost properties of this method make it a good starting point for mapping urban areas around the world.

STUDY AREA AND DATA

Study area

We choose China as the main study country to demonstrate the effectiveness of our method. Then we apply our method to 28 countries to validate the universality of this approach. Despite the rapid urbanization experienced in the past few decades, many areas of China are still underdeveloped, and the rural-to-urban process is quite uneven across regions. More importantly, as the largest developing country, China is lacking in consistent urban area data, which highlights the meaningfulness of this study. For the remaining countries, we choose those largest countries in each continent, with an area of not less than 100,000 km², including both developed and developing countries (Table I). The national surface area information is from the 2016 United Nations Demographic Yearbook, available through United Nations Statistics Division. The administrative boundary data of all countries are available from GADM (www.gadm.org), an open-source database of global administrative areas.

| Study area       | Urban data          | Population        | Road                  |
|------------------|---------------------|--------------------|-----------------------|
| Africa           | Nighttime light     | China mobile phone | China road network    |
| Asia             | Global NPP-VIIRS    | estimated population dataset | dataset from the Ordnance Survey |
| Europe           |                     |                     | Global population distribution dataset from WorldPop |
| North America    |                     |                     |                       |
| Oceania          |                     |                     |                       |
| South America    |                     |                     |                       |

Data

We use three datasets – nighttime light (remote sensing data), population (social sensing data), and road networks (VGI data) – in this research. These datasets represent the three most important urban elements: eco-
nomic activity, population, and infrastructure, respectively.

a. Nighttime light. We use the new generation of nighttime light (NTL) data, the global NPP-VIIRS NTL data (available through [www.ngdc.noaa.gov/eog/viirs](http://www.ngdc.noaa.gov/eog/viirs)). The NPP-VIIRS NTL data is produced from the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB). Compared with the old DMSP/OLS data, the NPP-VIIRS data has a higher spatial resolution (15 arc-second) and partially relieves the saturation effects and blooming effects [39]. Before averaging the observations, the annual composites exclude any data impacted by stray light, lightning, lunar illumination, and cloud-cover. We collect the annual `vcm-orm-ntl` average radiance data for 2016, which has undergone the outlier removal process, with non-lights background set to zero. The NTL data is publicly accessible for most countries, which make it a useful data source to map socio-economic activities and functional urban areas [9, 11, 26, 27].

b. Population. We use two population data sources. The first one is the WorldPop dataset (available through [www.worldpop.org](http://www.worldpop.org)). WorldPop provides an open-access archive of high-resolution population distribution data [10]. Especially, the ‘Global per country 2000-2020’ datasets have been improved in terms of global consistency. We collect the 2016 data for all study countries. For China, we also collect the second population dataset, which is estimated by the anonymous mobile phone location data. Detailed information about this dataset is presented in [41]. Here we use the aggregated version with a resolution of $0.001^\circ \times 0.001^\circ$. Note that mobile phone estimated population is only a sample of the whole population; thus we scale up the data with a factor derived by (national population) / (number of mobile phone users in the sample).

c. Road. We collect the OpenStreetMap (OSM) road shapefiles (available through [geofabrik.de](http://geofabrik.de)) for all study countries. For China, we also collect the road network data from the Ordnance Survey, a more detailed dataset than the OSM data [12]. The raw ordnance survey data records every segment in road networks with two endpoints’ IDs and locations. We identify road intersections by endpoint’s ID and obtain about 21 million ones.

Data quality assessment

For the nighttime light, the VIIRS data has been greatly improved with in-flight calibration, finer quantization, and lower light detection [43]. Moreover, the annual `vcm-orm-ntl` data is obtained through massive cloud-free observations and eliminates the background noise and ephemeral lights, thereby enhancing the radiance stability across the world [44]. Besides, our method relies only on the relative brightness values of different areas within a country, which can be reflected by the VI-IRS data. For the population, the WorldPop dataset has been proved to have high accuracy of population distribution as shown in [45]. The mobile phone estimated dataset comes from our previous work and also has high accuracy in measuring population distribution in China [11]. For example, at the district (county) level, the $R^2$s of the regression between mobile phone inferred population and census population are 0.97 and 0.98 for Beijing and Shanghai, respectively. For the road, the quality of OSM data varies greatly across countries regarding completeness and accuracy. However, previous studies have shown that OSM data can still be a reliable data source for mapping urban areas [20, 46].

**METHODS**

Our percolation-based city clustering algorithm (PCCA) includes three main steps. First, we aggregate the fine-scale urban data by $0.5^\circ \times 0.5^\circ$ grid cells to unify different datasets. Second, we apply the CCA to merge grid cells into urban clusters under each potential threshold. Third, we perform percolation analysis on the detected clusters to find the optimal threshold and then map the urban areas at the optimal threshold. Fig. 1 shows the schematic of the PCCA. All steps will be discussed in detail below.

Aggregate multi-source data by grid cells

Since multi-source urban data differ in granularity, data preprocessing is required to make the results comparable. For the nighttime light and population datasets, we directly downsample the data into $0.5^\circ \times 0.5^\circ$ grids. For the road datasets, we need more processing as raw data are vector files. For the Chinese ordnance survey data, we count the number of road intersections located in each grid cell and thus derive the road intersection grids. For the OSM data, to speed up the calculation, we divide the road lines into small segments and count the total length of road networks in each grid cell. Then we obtain the road length grids. Note that we further divide the cell values of the population and road data by cell’s spherical area to derive the consistent density maps.

City clustering algorithm (CCA)

At the grid cell level, we set each value of the above-mentioned datasets as a potential urban density threshold. Cells with a greater value (more urban elements) than the threshold will be marked as urban units. Then we apply the CCA to aggregate urban units into urban clusters under each potential threshold. The CCA originally uses fine-grained grid data of population and defines
Fig. 1. Schematic of the percolation-based city clustering method. (a) Multi-source urban data are aggregated by grid cells. (b) We use the CCA to merge urban units into a cluster system under each potential threshold. CCA: i) Cells with a greater value than the potential threshold are marked as urban units (light). ii) An unprocessed urban cell is selected to form a new urban cluster. iii) This urban cluster recursively adds the nearest urban cells until all nearest neighbors have been processed. We use the eight nearest neighbors in our research. iv) An urban cluster is formed (dark). Then another process begins until all urban cells belong to an urban cluster. (c) Two-dimensional site percolation model. For a $L \times L$ lattice, each site could be occupied with probability $p$, and adjacent occupied sites form a cluster (light). As $p$ increases, the largest cluster (dark) remains stable at first, but quickly becomes a giant spanning one at the critical point. We regard each potential threshold as probability $p$ in percolation and find the optimal threshold $D^*$ at this critical point. (d) Urban areas with optimal threshold $D^*$. 

Percolation of the cluster systems

To find the optimal threshold of the CCA, we apply the percolation theory to analyze the properties of the extracted clusters. The percolation theory was originally developed in statistical physics and mathematics to study the emergent structures of the clusters on a random graph. Since the percolation can lead to some critical phenomena, urban researchers have then used the percolation theory to model urban growth and to understand the critical phenomena of cities [3, 24, 25]. The two-dimensional site percolation is a simple and intuitive model to explain the percolation theory and explore the critical phenomena (Fig. [a]). As for a $L \times L$ lattice, each site can be occupied with probability $p$, and adjacent occupied sites form a cluster (light). When $p$ is small, there are only a few small clusters. As $p$ becomes larger, the size of the largest cluster (dark) remains stable, despite more occupied sites. When $p$ reaches a certain point, a giant cluster quickly forms and spans the whole lattice. This point is called the critical point or the continuous phase transition. Around the critical point, the cluster system exhibits some critical phenomena (e.g., size distribution follows power-law), which are also found in urban systems. Therefore, analogous to the two-dimensional lattice, we regard each potential threshold as the occupation probability $p$ in percolation, the optimal threshold $D^*$ can be found when the largest cluster of each cluster system becomes a giant one with a continuous phase transition. We consider the threshold at this critical point as the optimal threshold. After determining the optimal threshold, we obtain the final results of urban areas.

RESULTS

Urban areas extracted by PCCA

We first apply the PCCA method to the datasets of China. Following the Methods section, we obtain the density maps of population, road intersections, and nighttime light of China. Then, we extract the cluster systems under all potential thresholds and apply the percolation analysis to the cluster systems to find the optimal threshold.

Fig. 2a-c present the size of the largest cluster of different thresholds, and the size has been normalized by the total area of all clusters. At each threshold, there are some fragmented areas, possibly due to data noise or some special land use (e.g., oil fields, scenic areas). We
FIG. 2. The normalized size of the largest cluster for different thresholds in the datasets of population (a), road (b), and nighttime light (c), and the distribution entropy of the cluster system for different thresholds in the datasets of population (d), road (e), and nighttime light (f). The solid blue lines and dotted red lines represent the results of clusters larger than 20 km$^2$ and all clusters, respectively. The critical points of the continuous phase transition are marked with the solid vertical lines. The values of these points are 550 people per km$^2$ in population data, 20 intersections per km$^2$ in road data, 3.0 DN value in nighttime light data. Maximum entropies are marked with the dotted horizontal lines, which are also around the critical points. Insets: Similar to the main figures but with a log-scale for x-axis.

set a minimum size – 20 km$^2$ – to filter those fragmented areas. This value is set because the smallest land area of city in China is approximately 20 km$^2$. We also test the sensitivity of our method to this parameter by setting the minimum size to 10, 15, 25 km$^2$, and the result shows that our method is robust (Fig. A2). In Fig. 2 solid blue lines present the results of clusters larger than 20 km$^2$, and dotted red lines show the results of all clusters. For all datasets, as we lower the threshold, a giant spanning cluster quickly forms when the threshold reaches a critical point (vertical lines) – indicating a continuous phase transition. This phenomenon reflects the characteristics of the urban system as an interconnected complex system. Since the intra-city connectivities are much stronger than the inter-city connectivities, weak inter-city connections break up as we increase the threshold. When the threshold reaches a certain point, all weak inter-city connections do not exist, while the intra-city connections can still be tied closely. As a result, the size of the largest cluster goes through a critical point, where we consider as the optimal threshold to quantify urban area.

Besides the largest cluster, we also calculate Shannon’s entropy $H$ of the size distribution for each cluster system:

$$H = - \sum_{i=1}^{N} p_i \log p_i \quad (1)$$

where $N$ is the number of clusters in the system, $p_i$ is the proportion of the area of cluster $i$ in all clusters. In Fig. 2d-f, we find that the entropy also reaches the maximum (horizontal lines) around the critical point (vertical lines). Moreover, for each dataset, the entropies at the critical point are close for the clusters larger than 20 km$^2$, which are 5.88 (population), 5.32 (road), and 5.34 (nighttime light), indicating the similar size distributions of urban areas extracted from different data sources.

Furthermore, we map the urban areas at the critical point (Fig. 3). Strikingly, different data yield similar results. Especially, those larger clusters are well-developed cities, such as Chengdu, Xi’an, and Xiamen. Moreover, the maps also echo the uneven regional development in China. The urbanization level is much higher in the southern and eastern regions, and the coastal areas are more developed than the inland areas. The top three largest areas delineated by our method are the Beijing-Tianjin-Hebei, the Yangtze River Delta, and the Pearl River Delta economic zones (enlarged view of Fig 3), which are the most urbanized megalopolis areas in China. These areas break the geographic constraints of administrative boundaries and have highly integrated connections. Our results confirm their high integration, regardless of population, transportation infrastructure, or socio-economic activities.

To measure the similarities of urban areas delineated by different datasets, we use the Dice similarity coefficient (DSC). DSC is a similarity measure over sets and ranges from 1, with the same sets, to 0, with two com-
FIG. 3. Urban areas in China delimited by the PCCA in the datasets of nighttime light (a), population (b), and road (c). The numbers of clusters (≥ 20 km²) are 1085, 1260, and 931 for the nighttime light, population, and road datasets, respectively. Urban clusters are colored according to their geographical areas. Note that for simplicity, the South China Sea Islands are not shown in the maps.

FIG. 4. Differences (grid cells marked as urban areas in only one dataset) between urban areas delineated by different dataset (red). ‘2-Synthetic’ corresponds to urban cells identified in two datasets. ‘3-Synthetic’ corresponds to urban cells identified in all datasets. Note that for simplicity, the South China Sea Islands are not shown in the maps.

\[DSC = \frac{2|X \cap Y|}{|X| + |Y|}\] (2)

where \(X\) and \(Y\) are the sets of grid cells of urban areas. We remove the clusters smaller than 20 km². The DSCs are 0.62 between population and road, 0.68 between road and nighttime light, and 0.59 between population and nighttime light, which indicates that the spatial distributions of the urban areas obtained by different datasets are similar. Besides, we find the differences, those grid cells marked as urban areas in only one dataset, are mainly from the different distributions of intra-city ‘holes’ and the peripheries of each urban cluster (Fig. 4). For example, the Olympic Park in Beijing, with few people but dense roads, is a ‘hole’ in the population map but urbanized in the road map. The blooming effects in nighttime light data lead to less intra-city variations, which is consistent with the result found by [26].

Zipf’s law of urban areas

To further verify our results, we investigate whether Zipf’s law holds for the urban areas delimited by our method. Zipf’s law is one of the most important laws for the size distribution of the urban system. It reflects the self-organized nature of urban systems, and has been found in most countries [20, 29, 47–49]. Mathematically, it characterizes the size distribution of the urban system...
FIG. 5. Zipf’s law of urban areas in China. (a) Size indicators (geographical area, population, road intersections number, and nighttime light power) of an urban cluster. For each cluster in each dataset, we calculate the geographical area and the amount of population, road intersections, and nighttime light that fall within the cluster, through raw data. (b-d) PDFs of urban cluster sizes in the datasets of population (b), road (c), and nighttime light (d). Solid lines show the fitting results. All follow a power law with the exponent close to 2.

by the city size and the city rank, that is, if we order the cities of a country by their sizes, the city size is inversely proportional to the rank of the city:

$$S_i = S_1/R_i$$  \hspace{1cm} (3)

where $S_i$ and $R_i$ are the size and rank of the city $i$, respectively; and $S_1$ is the size of the largest city. This is equivalent to the statement that the probability of a city larger than size $S$ is inversely proportional to $S$:

$$P(S_i > S) = kS^{-\alpha}$$  \hspace{1cm} (4)

where $k$ is a constant and $\alpha = 1$. The corresponding probability distribution of size $S$ is:

$$p(S_i) = kS^{-(\alpha+1)}$$  \hspace{1cm} (5)

which means the probability distribution function (PDF) of the city size in a country follows a power law and the exponent is close to 2.

We obtain the urban clusters at the critical point for each dataset in China. Then we calculate four size indicators (geographical area, population, road intersections number, and nighttime light power) for each cluster and investigate the size distribution of urban clusters (Fig. 5). To fit a power-law distribution, we use the method proposed by [50]. The PDFs and fitting lines of cluster sizes in each dataset are presented in Fig. 5b-d. We find that the size distributions of urban clusters follow a power law and the exponents are close to 2 with standard errors less than 0.06. This indicates that Zipf’s law holds well for the urban areas delimited by our method in China.

Robustness check

To check the robustness of our method, we expand the analysis to 28 countries and present the results of France (Fig. 6) and India (Fig. 7). In the Appendix, we show the results of the remaining countries. In most countries, we obtain similar results as in China: (1) A giant spanning cluster quickly forms when the threshold reaches the critical point. (2) The distribution entropy reaches...
the maximum around the critical point. (3) The spatial distributions of urban areas delineated by three datasets are similar. In France, the largest cluster is the Paris metropolitan area, the political and economic capital of France. Other larger clusters also correspond to those well-developed regions, such as Marseille, Lyon, Toulouse (Fig. 6). In India, our method can also capture those important urban clusters, such as New Delhi, Mumbai, Bengaluru (Fig. 7). The critical points are 100 population per km$^2$, 6 km per km$^2$, 1.0 DN value in France, and 1800 population per km$^2$, 2 km per km$^2$, 2.0 DN value in India. These findings further validate the robustness and generalization of our method.

However, in some countries, we fail to observe a continuous phase transition; and entropy is the maximum at the minimum threshold, especially in population and nighttime light datasets. There may be two reasons. On the one hand, limited to resolution, these urban data cannot sense lower urban intensity than the minimum threshold, like weaker nighttime light, which mainly occurs in some underdeveloped regions, such as Kazakhstan (Fig. B18) and Chad (Fig. B19). On the other hand, geographical barriers could block the geographical proximity. For example, in Australia, almost all population clusters are distributed in isolated areas along the coast, due to the complex geography (e.g., deserts, mountains, rainforests). These geographical barriers break the weak inter-city connections (e.g., population distribution) even at the minimum threshold, while road networks can still connect the cities. Thus, in Australia (Fig. B2), we find the continuous phase transition only in the road dataset.

Comparison with global urban maps

To evaluate the accuracy of our method, we compare our derived urban areas with one well-known global urban map – the Global Human Settlement Layer (GHSL) [14], which is produced by the European Commission (EC) using Landsat images. We use the GHSL as reference data since EC has released the 2015 version recently, which is closer to the date of our results (2016) than most
FIG. 7. PCCA in India. Same as Fig. 6.

of the other global urban maps (2010 and before). We obtain the GHSL built-up 250m dataset, which measures the global built-up area density. We first coarse-grain the raw GHSL data into $0.5\degree \times 0.5\degree$ grids to make the resolution comparable to our results. Then we convert the built-up density data to urban/non-urban binary data. Based on previous works [24], we define the areas with the built-up density larger than 20% as urban areas. Finally, we apply the CCA to merge urban units into urban clusters, and remove clusters smaller than $20\text{km}^2$.

We compare our results with the GHSL data pixel by pixel ($0.5\degree \times 0.5\degree$) in China, Mexico, and the United States. Due to the large number of non-urban pixels, the overall accuracy (OA) is high (> 90%) in all countries. The Kappa coefficients ($\kappa$) are 0.56 (population), 0.57 (road), 0.56 (nighttime light) in China, and 0.73 (population), 0.61 (road), 0.45 (nighttime light) in the United States, which is better or similar to the performance of the cluster-based method [24] and the object-based thresholding method [51]. In Mexico, our results also show good agreement with the GHSL data. However, due to the poor quality of OSM data in Mexico, the results by road dataset ($\kappa = 0.32$) are not as good as results by population ($\kappa = 0.62$) and nighttime light ($\kappa = 0.55$) dataset.

We also sample some urban areas, from small to big cities in China, to investigate the spatial patterns of differences between our results and the GHSL. Fig. 8 presents the comparison of Beijing, Changsha, Guangzhou, Lanzhou, and Taizhou. Visually, our method captures the accurate urban extent for all sampled cities, and the results match well with the GHSL data. In addition, our method is superior to the GHSL in extracting small-scale urban areas beyond the urban center, especially in some medium-developed cities. In Lanzhou, our method captures the fast-growing Lanzhou New Area where the airport is located, while this area is ‘missing’ in the GHSL; in Taizhou, our method captures the urban areas of all three counties (Sanmen, Tiantai, Xianju) under the jurisdiction of Taizhou, while only Tiantai county is found in the GHSL. Besides, the differences between our results by three datasets reflect the variety of urban development. For example, in Beijing, the construction of road networks grows faster near the
FIG. 8. Comparison of our results (nighttime light, population, road) with GHSL reference data in the Beijing, Changchun, Guangdong, Lanzhou, Shanghai, Zhengzhou.

newly built Daxing international airport, while the population of that area has not grown up.

The differences among three datasets allow us to synthesize a more accurate urban area map. Specifically, we merge the urban areas delimited by population, nighttime light, and road datasets; and extract urban areas identified in at least two datasets. In Fig. 8, we also show the synthesized results for the sampled cities. We compare this synthesized urban map with the GHSL data in China. The result shows a significant improvement, with the Kappa coefficient increasing from 0.56 to 0.60. This finding demonstrates that multi-source data can not only delineate similar urban extents, but also can be further fused to derive more accurate urban area maps.

**Zipf’s law, optimal threshold, and socio-economic development**

We also investigate Zipf’s law of the delineated urban areas for each country. We set the optimal threshold as the minimum threshold when there is no continuous phase transition and measure the size of the urban cluster by its geographical area. In Table III, we present the optimal threshold \((D_{\text{pop}}, D_{\text{road}}, D_{\text{ntl}})\), entropy at the threshold \((E_{\text{pop}}, E_{\text{road}}, E_{\text{ntl}})\), Zipf exponent with one standard error \((\alpha_{\text{pop}}, \alpha_{\text{road}}, \alpha_{\text{ntl}})\) for each country and each dataset. We find that \(D_{\text{pop}}\) varies greatly from country to country, while \(D_{\text{road}}\) and \(D_{\text{ntl}}\) change less since the development of road networks and nighttime light is limited by geospatial space, especially for nighttime light. For each country, the entropies at the critical point in three datasets are similar, indicating the similar size distributions of the delineated urban areas. Meanwhile, the Zipf exponents of size distributions fall within [1.75, 2.25] in 24/28 (population), 20/24 (road), and 24/28 (nighttime light) countries, which means that Zipf’s law holds well in most countries.

Furthermore, we explore the relationship between the optimal thresholds and countries’ socio-economic indicators. Here, we use the urban population density as the proxy for socio-economic development. We calculate the urban population density by dividing the total population (WorldPop data) that fall within urban clusters for each country. Intuitively, the population threshold \(D_{\text{pop}}\)
is highly correlated with urban population density (the $R^2$ is 0.81, Fig. 9a), and $D_{pop}$ of each country is about 1/3 of the country’s urban population density. However, thresholds of road ($D_{road}$) and nighttime light ($D_{ntt}$) have weak positive correlations with urban population density (Fig. 9b,c). This may result from the country’s slow development of transportation infrastructure, which mainly occurs in some developing countries with large population size, such as India, Algeria, and Mexico. Besides, the OSM data quality varies greatly across different countries, and many road lines are not captured by OSM in some developing countries. Therefore, $D_{road}$ is smaller than the actual urban road density threshold in some countries. (The $R^2$ between $D_{road}$ and urban population density become 0.71 if removing some special cases, as shown in Fig. 9b.)

Finally, we apply the method to the entire world using nighttime light data and present the delineated urban areas in Fig. 10. Similar to the country level findings, the largest cluster size goes through a critical point, and the maximum distribution entropy is exactly at this point. Then we obtain the optimal threshold $\sim 1.0$ DN value. Through this world map (Fig. 10), we find that the urbanization level is much higher in North America, Europe, and East Asia. The top six urban clusters correspond to the Manchester - Milan (Europe), the Greater Cairo (Egypt), the Yangtze River Delta (China), the Boston - Washington (USA), the Delhi National Capital Region (India), and the Taiheiyo Belt (Japan) megalopolises (Fig. 10). The size distribution of urban areas of the world also fits well with Zipf’s law, with an exponent of 1.97 ± 0.01. We note that with our proposed method, we capture different dimensions of urban areas at different spatial scales. At the country level (such as Fig. 9), we delimit the metropolitan areas; while at the world level (Fig. 10), due to the differences in the economic basis of each country, we delimit those large mega-regions (or urban corridors). For example, develop-
FIG. 9. Relationship between the optimal threshold and urban population density. All countries are included in the datasets of population (a), road (b), and NTL (c). Some countries are removed in the datasets of population (d), road (e), and NTL (f). In population and NTL dataset, we remove the countries without a continuous phase transition; in the road dataset, we remove those with large population size and poor transportation infrastructure (India, Algeria, Indonesia, Mexico, Colombia, and Argentina).

opment level of the capitals of some African countries is even far less than that of rural areas in some developed countries. Therefore, the meanings of our extractions are different at different spatial level, and they depend on applications.

DISCUSSION AND CONCLUSIONS

In this study, we propose a ‘percolation-based city clustering algorithm’ to extract urban areas from multi-source urban data (nighttime light, population, road networks). Our method only needs one parameter (urban/non-urban threshold), which can be derived solely through the input data themselves by considering the critical nature of urban systems. Specifically, we form urban clusters under each potential threshold and perform the percolation analysis to locate the optimal threshold when a spanning cluster begins to form. We assume that the intra-city connections are much stronger and more stable than the inter-city connections so that at the optimal threshold, inter-city connections all break up while intra-city connections still tie closely. We evaluate the effectiveness and the robustness of our method by mapping the urban areas of 28 countries from three datasets. The observations of continuous phase transitions validate our assumption. Results show that our method can capture the similar geographical layouts of urban areas from multi-source data, and Zipf’s law holds well in most countries. Besides, the derived urban areas by our method show good agreement with the GHSL data and they can be further improved by multi-source data fusion.

The contributions of this study can be summarized in three aspects. First, we bridge the gap between remote sensing and emerging urban data in the task of delimiting urban areas. Our study has demonstrated that despite of great differences, different urban data can reflect the similar socio-economic dynamics of cities. Second, our method provides a consistent measurement of urban areas since the optimal threshold is derived automatically and under the same criteria. With our method, urban development can be measured under a unified standard, which allows comparisons across different countries and periods. Third, we show the potential of open-source data in delimiting urban areas. With the proposed method, we can produce reliable urban area maps from these publicly available data, which is especially helpful for those developing regions with limited survey data. Our study is also an attempt for applying complexity science to solve traditional urban problems and could deepen our understanding of urban systems.

There are still some limitations in this study, and several improvements can be explored in future work. First, due to the limited availability of temporal urban data, we have not been able to track the changes of urban areas over time. Such analysis could be possible with more detailed spatio-temporal data in the future. Second, it is meaningful to study the factors that influence the values of optimal thresholds. Possible explanations can be
complicated for geographical, social, or economic reasons. For example, environmental awareness can cause a decrease in brightness of nighttime light in some regions of Europe [52]. Third, the differences of the urban areas delineated by multi-source data are also worth exploring, as they reflect the inconsistent configuration of urban elements. However, how to explain these differences is full of challenges and requires more in-depth study.

DATA AVAILABILITY

All open-source datasets are available through the websites described in the data section. The PCCA codes and the maps of the delineated urban areas can be obtained through https://github.com/caowenpu56/PCCA.

ACKNOWLEDGEMENTS

This research was supported by the National Natural Science Foundation of China (Grant No. 41625003, 41830645, 41801299) and the China Postdoctoral Science Foundation (2018M630026).

[1] M. Batty, Nature 444, 592 (2006).
[2] X. Gabaix and Y. M. Ioannides, in Handbook of Regional and Urban Economics, Vol. 4 (Elsevier, 2004) pp. 2341–2378.
[3] H. D. Rozenfeld, D. Rybski, J. S. Andrade Jr., M. Batty, H. E. Stanley, and H. A. Makse, Proceedings of the National Academy of Sciences 105, 18702 (2008).
[4] W. Wu, H. Zhao, and S. Jiang, Remote Sensing 10, 130 (2018).
[5] B. Berry, P. Goheen, and H. Goldstein, Metropolitan Area Definition: A Re-evaluation of Concept and Statistical Practice, Vol. 28 (Washington, D.C.: U.S. Bureau of the Census, 1969).
