remote sensing

MDPI

Article

A Zipf's Law-Based Method for Mapping Urban Areas Using NPP-VIIRS Nighttime Light Data

Wenjia Wu [1,2], Hongrui Zhao [1,2,*] and Shulong Jiang [1,2]

[1] Institute of Geomatics, Department of Civil Engineering, Tsinghua University, Beijing 100084, China; wu-wj14@mails.tsinghua.edu.cn (W.W.); jsl15@mails.tsinghua.edu.cn (S.J.)

[2] 3S Center, Tsinghua University, Beijing 100084, China

* Correspondence: zhr@tsinghua.edu.cn; Tel.: +86-136-8154-1221

Received: 20 November 2017; Accepted: 15 January 2018; Published: 18 January 2018

Abstract: A significant difficulty in urban studies is obtaining urban areas. Nighttime light (NTL) data provide efficient approaches to map urban areas. Previous methods have utilized visual particularities of cities with ancillary data to obtain the optimal thresholds. How cities behave differently from rural areas should be considered. A Zipf's law-based method is proposed to bootstrap the optimal threshold based on the statistical properties of a Zipf's law model on continuous thresholds at the country scale. In our method, the Zipf's law model is utilized to quantify fractal, self-organized, and agglomeration behaviors of cities. The three-phase cluster dynamics are discovered and the abrupt transition between Phase 1 and Phase 2 denotes the rural-urban demarcation point. The urban areas are derived by the proposed method from the Suomi National Polar-Orbiting Partnership Visible Infrared Imaging Radiometer Suite (NPP-VIIRS) NTL data in 2013 in China. An accuracy assessment is conducted to compare it with the GlobeLand30-2010 data and the overall accuracy has directly confirmed the effectiveness of the method. The validation using point of interest (POI) data verifies that the urban areas show strong responses to social interactions and production with R^2 values of 0.91 and 0.92, respectively, implying that the city areas extracted by our method can be a proxy for human activities. Comparisons with existing methods validate the effectiveness and high degree of automation of the proposed method in mapping urban areas at the country level. According to our analyses, the Zipf's law-based method shows great potential to provide a universal criterion to map urban areas from the perspective of the behaviors of urban systems without ancillary data, and a valuable tool for spatial and temporal urban studies.

Keywords: urban areas; Zipf's law; social interactions; nighttime light (NTL) data; NPP-VIIRS

1. Introduction

The inexorable urbanization trend across the world presents an increasing need for a quantitative and predictive theory of urban organization [1], termed the second quantitative revolution by Michael Batty [2]. The availability of "new data"—big data—and "new science"—complex science—promotes the development of the new science of cities [2]. A major difficulty in the new science of cities is how a city can be defined properly [3]. Cities, regarded as typical self-organized systems [4], present general macroscopic features, such as urban scaling [5] and fractal geometry [2], which are all based on the size of the cities. Parameters in these urban models are sensitive to the units of the aggregated data [6]. For example, the scaling of transported-related CO_2 emissions with the population results in different conclusions at different aggregation levels [7].

However, recent administrative boundaries are not good candidates for urban studies due to the emphasis on political management and not on the city functions and human activities [8,9] and the inconsistencies across different countries and periods [10]. The current definitions of the morphological and functional views have not yet reached a consensus [11]. Concerning the mismatches between the

existing urban boundaries and quantitative urban studies, we look forward to functional and strongly interacting city definitions [12] from the view of the systems of cities consistently across the temporal and spatial dimension.

Nighttime light (NTL) data, especially the Visible Infrared Imaging Radiometer Suite (VIIRS) day/night band (DNB) data, can be applied to delineate accurate boundaries across the world [13–15]. The NTL data are positively correlated [16,17] with social interactions, which are regarded as the origin of why cities exist and thrive [5]. Face-to-face interactions on the road and non-face contacts through information systems in urban areas result in larger radiance values in the NTL data in the related pixels than in the neighboring dark rural areas [7]. In addition, extracting urban areas from the NTL data ensures urban research from both spatial and temporal perspectives [18]. Furthermore, the new generation of NTL data from the VIIRS began to produce more suitable data for the accurate extraction of urban areas [19]. In detail, the VIIRS provides a substantial number of improvements over the Defense Meteorological Satellite Program (DMSP) Operational Linescan System (OLS) data, including a higher spatial resolution [20], lower light imaging detection [21], finer quantization [21], and in-flight calibration [19].

However, challenges still exist when utilizing VIIRS NTL data to map urban areas [22]. The blooming effect in the NTL data refers to the fact that the radiance values of NTL data outside urban areas are still significantly above zero, which increases the difficulties of extracting urban areas from the surrounding non-urban regions [22,23]. Recent studies have been conducted to address this issue related to extracting urban areas, and these methods include thresholding methods [24–26], index methods including the enhanced built-up index [27] and vegetation-adjusted NTL urban index [28], and classification methods [18,29,30].

The thresholding technique has drawn more attention considering its simplicity and reasonable accuracy [24]. A number of studies have focused on empirical information; Elvidge et al. [13] estimated the threshold for a local window based on the background information, and Sutton et al. [31] proposed different thresholds for areas with different income levels after a comparison analysis between the NTL and population distributions. Then, more information including urban morphology [32] and brightness gradient [33] is considered; for example, the perimeter of an urban area polygon was applied as the objective function to obtain the optimal threshold to determine the "sudden jump" point through searching continuous thresholds [32]. Furthermore, ancillary data can help determine the thresholds (e.g., socioeconomic data, remote sensing data) [23,25,34]. Using this approach, statistical information of the urban areas of a region or city has been used to derive the optimal thresholds at different spatial levels; for example, He et al. [34] determined the optimal threshold for 1992, 1996, and 1998 in China using statistical information. In addition, remote sensing data at a medium- to high-resolution can help determine the threshold. MODIS data and TM data were used by Milesi et al. [35] and Henderson et al. [36] to search for the optimal threshold. Moreover, classified Land Use and Land Cover (LULC) data were utilized to obtain thresholds [37]; for instance, Liu et al. [38] extracted urban areas with the assistance of LULC data from Landsat TM images for 1995, 2000, and 2005 in China, and Zhou et al. [25] presented a method to derive optimal dynamic thresholds to map the urban area of each urban cluster, which was generated using a segmentation algorithm. This method was then extended to map urban extent at the global level [39].

These thresholding methods utilize empirical information or ancillary data about land use to determine the optimal threshold [22]. However, these methods lack a consideration of how cities behave differently from rural areas [2]. A high population density and hence bright nightlights make urban environments identifiable in remote sensing images [7,8]. These are obvious, visual particularities of cities, but some of their behaviors are subtler [40]. Cities are typical examples of complex systems; thus, city boundaries should have a meaning of "systems within systems of cities", as proposed by Berry [41]. In addition, cities can be the results of social interactions [5] at different spatial and temporal scales so that urban areas refer to a consistent definition gathering human

activities and social interactions [14]. Furthermore, geographic and economic incoherence [42] results in obvious boundaries between urban and non-urban areas.

It has been observed that systems of cities exhibit some consistent behaviors: the distribution of city sizes at the country or global level is well demonstrated by the Pareto distribution, which was first put forward by Auerbach and most notably refined by Zipf [43]. Zipf's law for cities is one of the most distinct empirical regulations in urban science [44]. Zipf's law states that the distribution of city sizes not only exhibits a power-law distribution but also has the exponent equal to 1 ($\pm$0.1) [45,46]. Recent works validate this empirical regularity for cities in both empirical [45] and theoretical [44] ways. Zipf's law for cities also reveals that cities with geographic and economic incoherence can be regarded as an interconnected whole [45]. Zipf's law for cities also reflects the fractal nature of city systems [2,47] and the self-organized behavior of a complex urban system, which are significant characteristics of urban dynamics. Thus, Zipf's law provides a universal measure of the "degree of urbanness" [14], which can be utilized to determine thresholds for urban area extraction.

Zipf's law is capable of quantifying changes in behaviors between non-urban areas and urban areas. To extract urban areas signifying the essential features of city systems, a Zipf's law-based method that does not depend on ancillary data is proposed. The optimal threshold is derived based on the three-phase cluster dynamics, termed the three-phase model. The method is performed on the NTL data from VIIRS that was collected by the Suomi National Polar-Orbiting Partnership (Suomi NPP). Furthermore, we validate the proposed method using an accuracy assessment, evaluating the response to social interactions by utilizing point of interest (POI) data and comparing it with existing methods.

The remainder of this paper is structured as follows. Section 2 displays the study area and the data. Section 3 introduces the proposed method for deriving urban areas. The results and discussions are presented in Section 4. Finally, Section 5 presents the conclusion and provides ideas for future work.

2. Study Area and Data

2.1. Study Area: China

China is utilized as the study area because of its more rapid urban development and more complex urbanization pattern than most other regions of the world. The complex urbanization pattern includes heterogenous urban expansion levels across regions [24] and an inconsistent land expansion pace and population growth. Thus, China is selected as an appropriate study area for evaluating the global applicability of the proposed urban area delineation method. Considering an absence of relevant statistical data, Taiwan is not included in this research.

There are five levels of administrative divisions in mainland China, namely, province, prefecture, county, township, and village. In Sections 4.4 and 4.5, we examine the responses of the urban areas to the variables related to social interactions and creations at the prefectural level.

2.2. NPP-VIIRS NTL Composite Data

The version 1 series of the NPP-VIIRS NTL data in 2013 are acquired from the NOAA-NCEI [48]. These NPP-VIIRS NTL data are monthly and cloud-free average radiance composite images for 2013. These products are produced in 15 arc-second grids that span from 180°E longitude to 180°N and from 75°N latitude to 65°S [48] under the WGS84 geographical coordinate system. In this study, we set out to obtain the monthly average data throughout the year, but data from May to August are removed due to the lack of good quality data in the high latitudes of China. After obtaining the NPP-VIIRS NTL data spanning months 1–4 and 8–12, the monthly NPP-VIIRS NTL data are first extracted by the administrative boundary of China and projected utilizing an Albers equal-area conic projection. Next, the extracted NPP-VIIRS NTL data for each month in China are corrected with data preprocessing, which is explained in detail in Section 3.1. Finally, the average pixel brightness values from the eight corrected images are calculated and we obtain the NPP-VIIRS NTL image for 2013 in China, as shown in Figure 1.

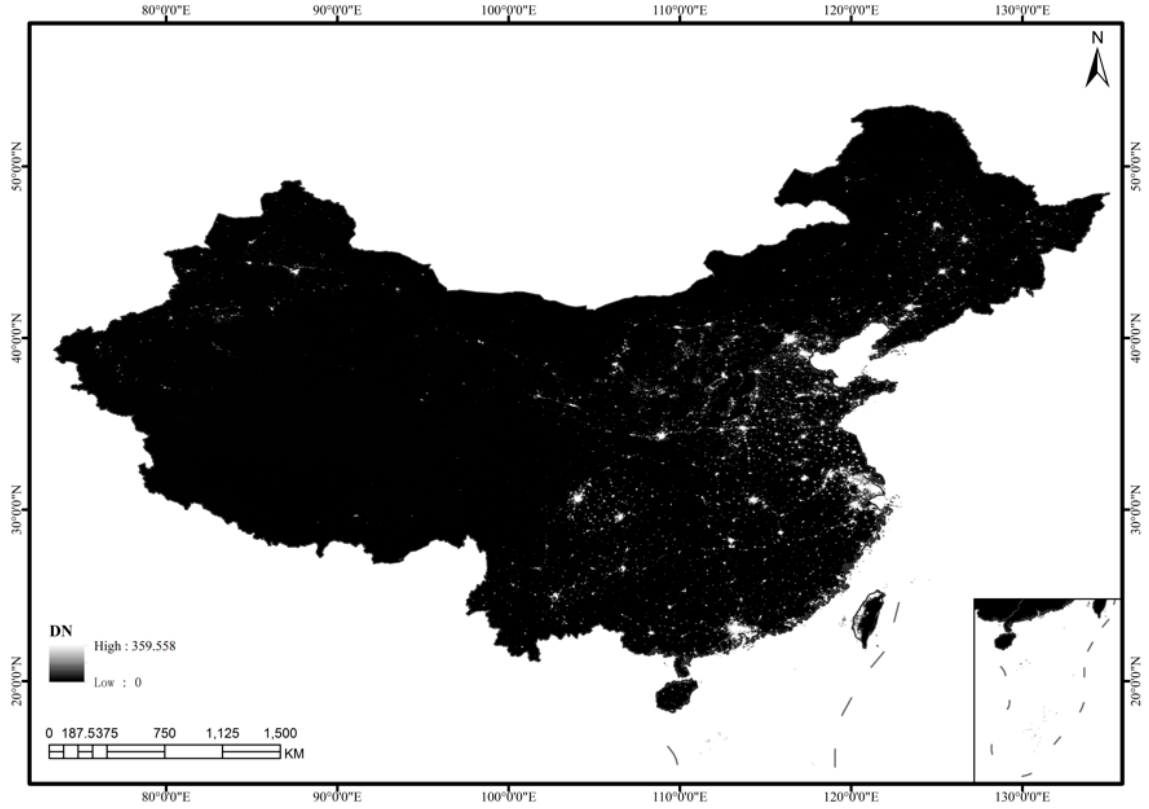

Figure 1. The corrected NPP-VIIRS NTL data of China in 2013.

2.3. Other Datasets

Since point of interest (POI) data, a new kind of location-based VGI (volunteered geographic information) data [49], can be utilized as a proxy for human mobility or social interactions such as understanding mobility based on GPS data, the POI data in China are utilized to explore the relationship between the extracted urban areas and social interactions for validation in Section 4.4. Due to the limits of data availability, POI data for the year 2014 are obtained and geo-coded from the business website maintained by GeoHey [50], as shown in Figure 2a. We assume that the POI data in two adjacent years were consistent, so the POI data from 2014 can be used in 2013. The POI data contain 45 million POIs categorized into 23 types, among which the shopping service POI, living service POI, and catering service POI are chosen to denote human activities. Thus, these three kinds of POI data are extracted and spatially and statistically analyzed at the prefectural level. There are 12.3 million shopping service POIs, 4.4 million living service POIs, and 5.3 million catering service POIs, as indicated in Figure 2b–d. These POIs are used to evaluate the response to social interactions.

For evaluating the proposed method, high spatial resolution LULC datasets, specifically GlobeLand30-2010 data, were acquired from the National Geomatics Center of China (NGCC) [51]. GlobeLand30 data represent the land cover of Earth between latitudes 80°N and 80°S with a spatial resolution of 30 m. Multispectral images [51,52] from the America Land Resources Satellite (Landsat TM5 and ETM+), the China Environmental Disaster Alleviation Satellite (HJ-1), and auxiliary data [51,52] are used during data production. The classification system consists of 10 land cover types, among which artificial surfaces represent the urban areas [51] in this study. However, the GlobeLand30 data for 2013 is not yet available; thus, the GlobeLand30 data for 2010 is adopted as an alternative.

The gross domestic product (GDP) data at the prefectural level in China are gathered from the China Statistical Yearbook to examine the responses of the urban areas to the variables related to production. Considering the lack of data in the western provinces, 289 out of 333 prefectures in China are evaluated.

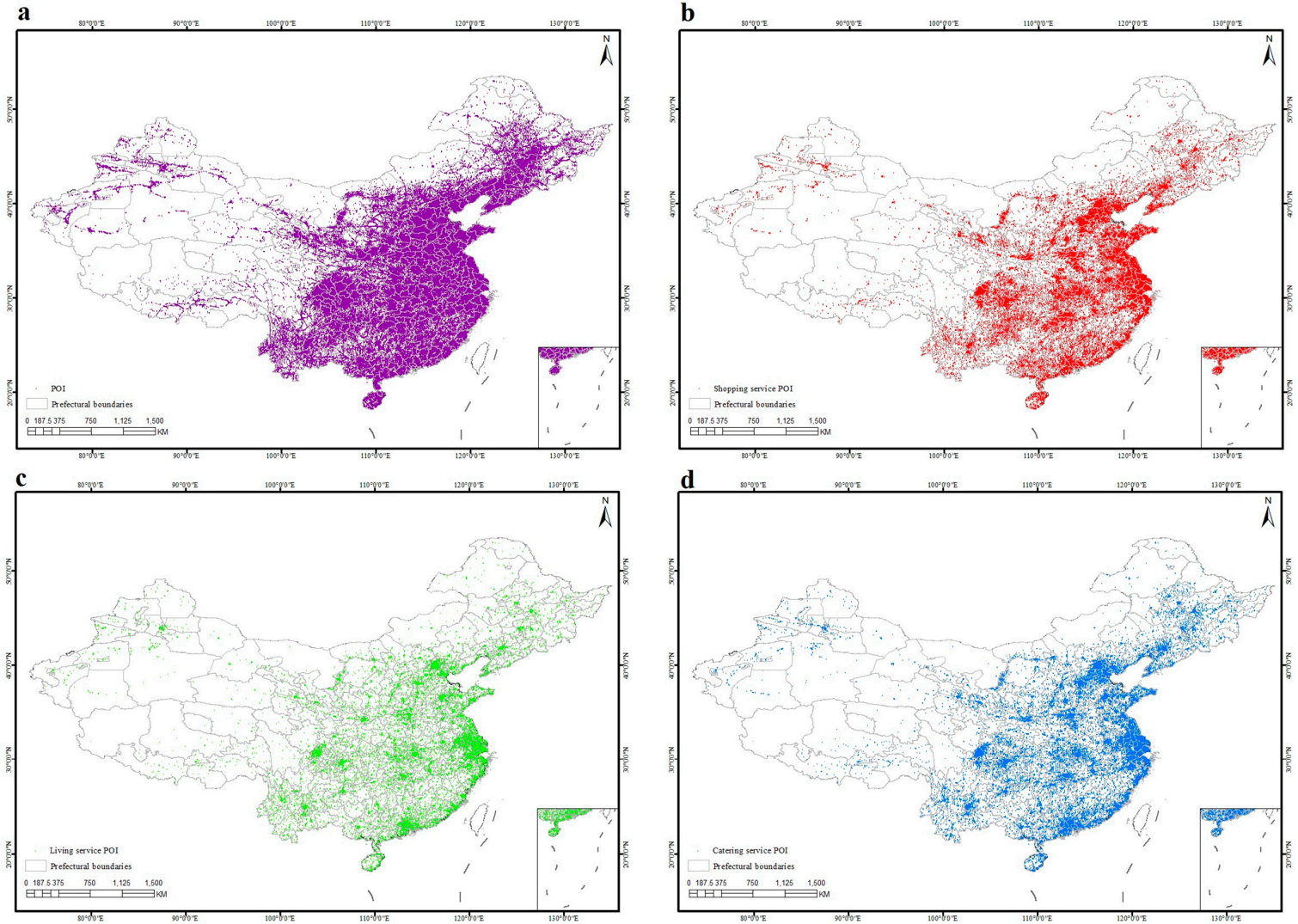

Figure 2. POIs of China. (**a**) Total POIs in China in 2014; (**b**) Shopping service POIs in China; (**c**) Living service POIs in China; (**d**) Catering service POIs in China.

3. Methodology

Human activities and social interactions in urban areas result in larger radiance values in NTL data in related pixels than the neighboring non-urban extents [53]. Previous thresholding methods utilize these obvious, visual particularities of cities with the assistance of empirical information or ancillary data about land use to obtain the optimal threshold [22,49]. How cities behave differently from rural areas should be considered.

Zipf's law reflects the consistent behaviors of cities. Thus, the proposed method, the Zipf's law-based method, is founded on the statistical properties of Zipf's law model on continuous thresholds at the country scale and allows for an automated and systematic way of finding the optimal threshold.

This method includes four major steps: correction of NTL data, estimation of Zipf's law of NTL clusters, estimation of Zipf's law-based threshold, and mapping of urban areas. First, NPP-VIIRS data are filtered by excluding the background noise and abrupt bright spots. Second, mathematical models are built to estimate Zipf's law model for the NTL clusters. Third, the statistical properties of Zipf's law model on continuous thresholds are explored and utilized for threshold estimation. Finally, the urban areas are extracted with the optimal threshold. Each step will be discussed in more detail in the following subsections.

3.1. Correction of the NPP-VIIRS NTL Data

In this step, we remove the background noise and temporal lights to generate the corrected NTL data. The utilized NTL data from monthly composites are preliminary. These data have not excluded lights from temporal lights and background noise [53], which are unrelated to human settlements and will reduce the accuracy of urban area extraction; therefore, the original NPP-VIIRS data in 2013 before correction cannot be used to derive the urban areas.

To correct the data for each month, we filter the original NTL data by removing the background noise and temporal lights. There are several pixels of negative radiance values on the images from monthly composites. These negative radiance values result from the system noise in the dark current correction process [20]. The average northern hemisphere dark night ocean data was utilized in the dark current correction process, but some system noise remained that made some pixels slightly negative in no-light areas [20]. Therefore, the background noise is screened out and set as no data. We utilize a data correction method [26] to exclude abrupt bright spots originating from temporal lights. Following such a method, Beijing, Shanghai, Guangzhou, and Shenzhen compose the most developed metropolitan regions, and their maximum light intensity values are selected as the maximum values for all pixels in China. Table 1 shows the maximum values of the first-tier cities in each month. Then, each pixel in the data is traversed, and the pixels with values greater than the maximum values are identified and set to the maximum value of the neighboring pixels.

Table 1. The maximum values for data correction of NPP-VIIRS NTL data in 2013.

Month	January	February	March	April	September	October	November	December
Maximum value	337.589	309.868	341.728	253.113	300.827	299.849	607.028	426.462

Thus, the NPP-VIIRS NTL image of 2013 in China is generated by averaging the pixel brightness values from eight corrected images, as shown in Figure 1.

3.2. Estimating Zipf's Law of NTL Clusters from NPP-VIIRS Data

Zipf's law can be utilized to assess urban behaviors for different low light thresholds in NPP-VIIRS data. In this step, a Zipf's law model for NTL clusters is built mathematically after the NTL clusters in the NPP-VIIRS data are first extracted using the potential threshold. Second, the method of estimating the parameters of the Zipf's law model for NTL clusters is described.

3.2.1. Zipf's Law Model for NTL Clusters

Because Zipf's law highlights the universal regularities of urban systems [14], the next step is to build the Zipf's law model for the NTL clusters at a potential threshold. First, if we have a potential threshold DN_P, the pixels with radiance values greater than the potential threshold will be set as NTL clusters $\{n_i, i = 1, 2, \ldots, n\}$, and four near neighbor pixels are required to have brighter DN values than the DN_P. Then, the Zipf's law model for the NTL clusters is founded to search for the optimal threshold.

Zipf's Law is widely considered as a ubiquitous rank-size regularity for complex systems [42]. Observations in different countries and across the world from various time periods provide empirical evidence supporting Zipf's law for cities. It has already been observed that the distribution of natural city sizes obeys Zipf's Law in DMSP-OLS data [45]. Researchers have proposed theoretical explanations using mainly probabilistic models based on multiplicative stochastic processes, including Simon's Model, Gabaix's Model [44], Cordoba's Model, and Krugman's Model, to validate the universality of Zipf's law for cities.

Zipf's law is often formulated as a function of the distribution of city sizes. All cities are ranked in descending order according to size; for instance, the first refers to the largest city, the second to the second largest, and so on [44]. Then, the size of a city $s(r)$ decreases with its rank r as a power-law formation,

$$s(r) = Ar^{-\alpha}, \tag{1}$$

where A is a constant and α is a constant parameter known as the Zipf's exponent. The Zipf's exponent for city sizes is equal to 1 (± 0.1), which is validated by empirical [45] and theoretical [44] studies. Therefore, the Zipf's law model for the NTL clusters $\{n_i, i = 1, 2 \ldots n\}$ can be written as:

$$a(r) = Ar^{-\alpha}, \tag{2}$$

where $a(r)$ denotes the area of a nightlight cluster n_i with rank r. Mathematically, we usually utilize the probability distribution function (PDF) $p(x)$ of areas for nightlight clusters to describe Zipf's law,

$$p(x) = Cx^{-\beta}, \tag{3}$$

where C and β are the normalization constant and exponent, respectively.

Aiming at computing β, we utilize the cumulative density function (CDF) to establish the relation between Zipf's law and the probability distribution function. Since the inverse relation $r(x)$ for $a(r)$ is:

$$r(x) = a(r)^{-1} = \frac{x^{-\frac{1}{\alpha}}}{A}, \tag{4}$$

which is a rank/frequency plot equivalent to the CDF [54], Zipf's law model $a(r)$ has a mathematical relationship with the PDF as:

$$P(x) = \Pr(a(r) > x) = \frac{r(x)}{R} \propto r(x) \propto x^{-\frac{1}{\alpha}}, \tag{5}$$

where R is the maximum rank. The relationship between the CDF and the PDF can be described as:

$$P(x) = \int_x^{+\infty} p(x')dx' \propto x^{1-\beta}. \tag{6}$$

Thus, we can obtain β as:

$$\beta = 1 + \frac{1}{\alpha}. \tag{7}$$

3.2.2. Estimating the Parameters in the Zipf's Law Model

In this step, the central question that we will consider is as follows: given a set of NTL clusters from one potential threshold DN_P, how can you determine if those NTL cluster areas are well described by Zipf's Law? And if so, what is the best way to estimate the exponent? Extensive discussions by Newman [54] and Clauset [55] stated that these questions can be surprisingly subtle and misunderstood.

There are some challenges in estimating the probability distribution function, such as that in Equation (3). The first problem originates from the bin width setting. When forming a histogram, you should choose a suitable bin width to make the numbers of points within each bin equal. However, for this power-law distribution, when small numbers compose the highest portion of the distribution, normal bin width methods do not work. Logarithmic binning was proposed for power-law distribution [54]. The second problem results from the large fluctuations in the tail [55], which represents large but rare events. The cumulative density function is utilized to solve this problem. It can be verified that the cumulative density function is statistically stable with great robustness compared to the probability distribution function [54]. This cumulative density function enables us to smooth the large fluctuations in the tail [54,55]. The third problem centers on the least-squares fitting method (LS), the most common method. It has been suggested that the LS method will introduce systematic bias [56]. Some methods based on maximum-likelihood fitting methods and goodness-of-fit tests based on the Kolmogorov-Smirnov statistic and likelihood ratios can solve this problem [54–56].

Thus, to detect Zipf's distribution and obtain the component, we follow the technical method proposed by Clauset et al. [55]. Here, we provide a brief introduction to this method.

According to the maximum-likelihood fitting method, given a set of NTL cluster areas $x = \{a_i, i = 1, 2, \ldots, n\}$, we maximize the likelihood function $\mathcal{L}$ as:

$$\mathcal{L} = p(x|\beta) = p(a_1|\beta)p(a_2|\beta)\ldots p(a_n|\beta). \tag{8}$$

Then, we can obtain the estimating exponent β for the probability distribution function $p(x)$ for continuous x,

$$\beta = 1 + n\left(\sum_{i=1}^{n} \ln \frac{a_i}{x_{\min}}\right)^{-1}, \tag{9}$$

where $x_{\min}$ is the lower bound of where the power-law behaviors hold. According to Equation (7), the estimating exponent α for Zipf's law is:

$$\alpha = \frac{1}{n}\left(\sum_{i=1}^{n} \ln \frac{a_i}{x_{\min}}\right). \tag{10}$$

Next, we explain how to identify the power-law region, that is, we estimate the value of $x_{\min}$. Aiming to ensure a good balance between discarding fewer data and obtaining better power-law behavior, for each $x_{\min}$, we calculate the estimating exponent β and then select the $x_{\min}$ that leads to a model that minimizes the distance between the data and $p(x)$. For distance, Kolmogorov-Smirnoff (KS) distances are utilized.

Furthermore, we will demonstrate how we test the goodness-of-fit. First, synthetic data are generated by sampling from $p(x)$ with β and $x_{\min}$; then, the estimating exponent and $x_{\min}$ are calculated to determine the model for synthetic data and the distance d_i between the synthetic data and the model $p_i(x)$ is calculated. Thus, we can obtain the goodness-of-fit index (GFI) *p-value* using:

$$p\text{-}value = \frac{\{number\ of\ d_i \geq d\}}{total\ number}, \tag{11}$$

where d is the distance between the given data and the model $p(x)$ with β and $x_{\min}$. The larger the p-value, the better the fit. Generally, in the statistical field, if the p-value is greater than 0.05, or more strictly, than 0.1, Zipf's law and the estimating exponents are accepted. In practice, the particular conditions and priori knowledge should be considered to decide the adopted rule [55]. For geographical features like the NTL clusters, the exponents with a more lenient rule are meaningful for Zipf's law [45], but the p-value can reflect the goodness-of-fit of Zipf's law for geographical features.

3.3. Zipf's Law-Based Threshold Estimation Method

The Zipf's-based model provides a continuous measure of changes in behaviors between non-urban areas and urban areas. Aiming at optimal threshold estimation, Zipf's law will be systematically investigated for NTL clusters by varying the potential threshold. Then, the optimal thresholds will be obtained by utilizing the statistical properties of the Zipf's law model on continuous thresholds. The statistical properties of the Zipf's law model on continuous thresholds demonstrate the three-phase cluster dynamics, termed the three phases model, and the optimal threshold corresponds to the potential threshold where there is an abrupt transition of the parameters of Zipf's Law between Phase 1 and Phase 2.

3.3.1. Estimating the Three-Phase Model Based on the Statistical Properties of the Zipf's Law Model on Continuous Thresholds

This section is concerned with whether there are apparent differences in the power-law distributions outside or within the urban areas. Thus, we examine the transformations of the power-law exponent β results from the change in the threshold value from low to high values.

In this step, various DN values are traversed as potential thresholds using certain step sizes in certain intervals, and then we computed two parameters for each set of clusters at different thresholds: the power-law exponent β according to Eqution (9) and the GFI $p\text{-}value$ according to Equation (11). Both the statistical properties of the exponents with a lenient rule and the exponent with a strict rule, $p\text{-}value \geq 0.05$, will be explored. In detail, we change the threshold from dark to light, and consistently examine the power-law size distributions—the mathematical model of Zipf's law in Section 3.2—of the NTL clusters on the NPP-VIIRS data in China in 2013 with the step size set to 1 within the interval [1,70].

As the potential threshold increases from the dimmest to the brightest, the parameters describing Zipf's law for each set of clusters change due to the cluster dynamics [57]. Four change patterns in the visual view can be summarized as: (1) Shrinkage outside: the size of each spatially contiguous cluster shrinks around the perimeter; (2) Shrinkage inside: the size of each spatially contiguous cluster shrinks inside and results in holes; (3) Separation: large clusters break up internally since the connecting dim lights are exceeded by the threshold [14]; and (4) Vanish: small clusters disappear. In the mechanism view, the parameters change for two reasons when the threshold value increases from low to high: (1) the fractal behavior and self-organized behavior—"far more small things than large ones" proposed by Jiang [8]—vary; and (2) different thresholds generate areas with different degrees of connectivity [14]. These visible and invisible cluster dynamics are the performances of different urban-nonurban patterns with varying thresholds. The cluster dynamics can be categorized into three phases according to the urban-nonurban patterns, termed the three-phase model: (1) Phase 1: the NTL clusters are the combination of the urban areas and non-urban areas; (2) Phase 2: the NTL clusters consist of the urban areas, and the urban areas can be regarded as a whole made up of several interconnected sub-areas; and (3) Phase 3: the NTL clusters are the combination of several isolated and broken sub-areas. The cluster dynamics in three phrases result in the transformation of the β value of the NTL clusters on continuous thresholds. The first sudden point between Phase 1 and Phase 2 corresponds to the rural-urban demarcation point. The second sudden point between Phase 2 and Phase 3 corresponds to the urban violation point. Therefore, the optimal thresholds can be obtained by utilizing the three-phase model based on the statistical properties of the Zipf's law model on continuous thresholds.

3.3.2. Threshold Estimation

The abrupt transition of the exponent β emerges for the NTL clusters on continuous thresholds, which reflects the diverse fractal, self-organized, and agglomeration behaviors of cities that are different from rural areas. The optimal threshold corresponds to the potential threshold where an abrupt transition of the parameters of Zipf's Law appears.

The proposed threshold estimation method is described by the following recursive function, Algorithm 1. Given the corrected NPP-VIIRS NTL data and step size i, the power-law exponent β and the GFI p-value are computed on the continuous potential threshold DN_P. The three-phase model is developed. The optimal threshold equals the potential threshold DN_P when there is an abrupt transition of the parameters of Zipf's Law between Phase 1 and Phase 2 in the three-phase model.

The three-phase model and the abrupt transition between Phase 1 and Phase 2 can be described quantitatively in detail. When the potential threshold DN_P increases in Phase 1, the non-urban areas with low DN values shrink. Since the non-urban areas are fragmented and inconsistent, the fractal behaviors and scale-free behaviors of the NTL cluster systems fluctuate greatly. Thus, the power-law exponent β fluctuates in Phase 1 and the GFI p-value fluctuates with a low value. When the potential threshold DN_P varies in Phase 2, the urban areas reduce. Since the urban areas can be regarded as a whole with the geographic and economic coherences, the behaviors of the NTL cluster systems are stable and continuous. The power-law exponent β holds steadily and continuously around the theoretical value, that is $\beta = 2$, corresponding to $\alpha = 1$ in Zipf's law according to empirical [45] and theoretical [44] studies, in Phase 2, and the GFI p-value is relatively high. When the potential threshold DN_P increases in Phase 3, the whole urban area separates into several isolated sub-areas. The fractal, self-organized, and agglomeration behaviors are violated, and the power-law exponent β begins to increase to a very large value greater than the theoretical value and fluctuates greatly. These conditions provide criteria to estimate the abrupt change between Phase 1 and Phase 2.

Algorithm 1 Algorithm for optimizing the threshold

Inputs:
The corrected NTL data on the NPP-VIIRS data.
The step size i.
Outputs:
1. Set the step size i as the potential threshold DN_P.
2. **While** $DN_P \leq 70$ **do**
3. Set the pixels—values less than the threshold DN_P—to zero.
4. Segment the data into extracted extents and non-urban extents.
5. Calculate the power-law exponent β and the GFI p-value of the size distribution of the extracted area.
6. $DN_P + i$ as the new potential threshold DN_P
7. End while
8. Develop the three-phase model.
9. Obtain the optimized threshold $DN_T = DN_P$

3.4. Urban Areas Mapping

With the optimal threshold DN_T, the pixels with radiance values greater than DN_T will be set as urban areas, and all other areas will be set as non-urban areas. The final urban map products in China are then generated.

4. Results and Discussion

4.1. Threshold Estimation from the Statistical Properties of Zipf's Law Model on Continuous Thresholds

The power-law distributions of the NTL cluster sizes with varying DN values from the NPP-VIIRS data in China in 2013 are demonstrated in Figure 3. Following Algorithm 1 in Section 3.3.2,

we transverse the various DN values as the potential threshold with step size $i = 1$. Both the results of exponents on continuous thresholds and the results of exponents with a *p-value* greater than 0.05 are shown in Figure 3 with the red line and the blue line. As shown in Figure 3, there are three phases according to the statistical properties of Zipf's law model on continuous thresholds, which has been marked in the figure.

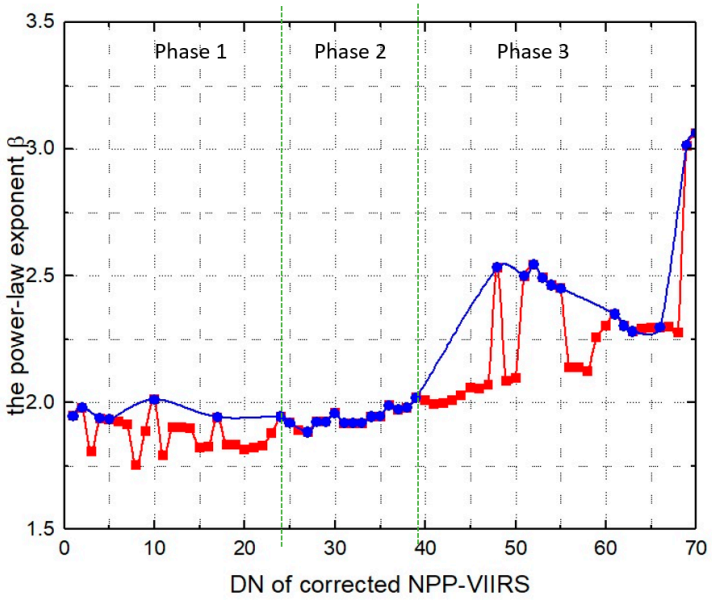

Figure 3. The transition of the power-law distribution exponent β on continuous thresholds. The blue line with circle symbols and the red line with square symbols, respectively, show the curve of the exponent β on continuous thresholds with and without a *p*-value limit.

In Phase 1, power-law exponent β without a *p-value* limit fluctuates considerably between 1.75 and 2.01 within the interval [1,24], suggesting a Zipf's exponent between 1.33 to 0.99. Both the local maximum and the local minimum fluctuate violently in this interval. The goodness-of-fit test is not accepted for quantities of power-law distribution models at DN values within the interval [1,24]. For the 24 points within the interval, only six points are accepted with a $p\text{-}value \geq 0.05$ and are illustrated in the blue line. Thus, most of the red lines are below the blue lines within the interval. The statistical variation of the exponent and *p-value* is consistent with the cluster dynamics in Phase 1, and reflects the violent fluctuation of the behaviors of the NTL clusters in Phase 1 in Section 3.3.1.

There is a rather sharp transition that begins at the value of 24. In Phase 2, the exponent β without a *p-value* limit remains markedly stable in the range from 1.88 to 2.02 within the interval [24,39], suggesting a Zipf's exponent between 1.14 to 0.98. As a result, β is approximately 2.00 within this interval. The goodness-of-fit test is accepted for quantities of power-law distribution models at DN values within the interval [24,39]. For the 16 points within the interval, 15 points are accepted with a $p\text{-}value \geq 0.05$ and are illustrated in the blue line. Thus, most of the red lines coincide with the blue lines within the interval, as shown in Figure 3. The statistical characteristics of the exponent and *p-value* are consistent with the cluster dynamics in Phase 2, and illustrate the stable behaviors of the NTL clusters in Phase 2. This result suggests that Zipf's exponent α equals approximately 1.00 and there are fractal and self-similar behaviors when the DN value is between 24 and 39.

There is another transition that begins at the value of 39. In Phase 3, the power-law exponent β increases and becomes greater than 2, and this suggests that α is less than 1.00 beyond the sudden point, which is inconsistent with Zipf's law. Large fluctuations between 1.99 and 3.06 occur within this interval, indicating a Zipf's exponent between 1.01 to 0.49. Both the local maximum and the local minimum fluctuate violently in this interval. The goodness-of-fit test is not accepted for quantities of

power-law distribution models at DN values beyond 39, and most of the red lines are below the blue lines within the interval, as shown in Figure 3. The abrupt transition indicates the disappearance of the fractal and self-organized behavior for cities as a whole compared to the power-law distribution within the interval [24,39]. This reflects the violent fluctuation of the behaviors and the violation of the urban behaviors for the clusters in Phase 3, which is comparable to the violation of the self-similarity in Jiang [8] and the transition observed in the percolation models with census tracts [57] and road network data [6].

Thus, we can obtain the urban areas using the definition of DN set to 24; that is $DN_T = 24$. Three different threshold intervals can be discovered according to the statistical properties of Zipf's law model on continuous thresholds, including $DN_P < DN_T$, $DN_T \leq DN_P < DN_S$ (DN_S is the second demarcation point between Phase 2 and Phase 3), and $DN_P \geq DN_S$, and these different threshold intervals correspond to the three phases in Section 3.3.1.

4.2. The Cluster Dynamics in Three Phases

The visual influences of the thresholds on the NTL clusters are illustrated to validate that the cluster dynamics vary at the DN values in three different phases.

The NTL clusters maps at the varying thresholds from NPP-VIIRS data at the DN values of the three categories—$DN_P < DN_T$ ($DN_P = 2$ and $DN_P = 15$), $DN_T \leq DN_P < DN_S$ ($DN_P = 24$ and $DN_P = 30$), and $DN_P \geq DN_S$ ($DN_P = 52$)—are illustrated in Figure 4. Four change patterns, including shrinkage inside, shrinkage outside, separation, and vanish, all take place in three phases but with different probability distributions, which is related to the urban-nonurban patterns. For Phase 1, when $DN_P < DN_T$, large clusters separate since the connecting non-urban areas are exceeded by the threshold. Some clusters shrink outside due to the surrounding non-urban areas, and some clusters, made up of non-urban areas, vanish. For Phase 2, when $DN_T \leq DN_P < DN_S$, the NTL clusters capture the whole urban areas made up of several interconnected sub-areas, and the sizes of each spatially contiguous urban area shrink around the perimeter. For Phase 3, when $DN_P \geq DN_S$, the whole urban clusters break up internally, and separate into several isolated sub areas. The NTL clusters cannot capture the core of the urban areas, for example, Beijing divided into several broken parts.

Furthermore, the generation (or definition) of fractals proposed by Jiang [8], and applied in many fields, including urban planning [58], transportation [59], and biodiversity [60]—a feature is thought to be fractal if the scaling pattern of "far more small things than large ones" can be observed multiple times—is adopted in this research. Since both the first generation fractals like Koch curve and Sierpinski carpet, and the second generation fractals in a statistical sense—a power-law relationship can be found between the measurement scale—are strict [61], and both definitions cannot directly work for complex geographical features. The fractal indexes have been computed for the NTL clusters in the previous paragraph. When $DN_P < DN_T$, the fractal indexes at the DN values of 2 and 15 are 6 and 5, respectively. When $DN_T \leq DN_P < DN_S$, both the fractal indexes at the DN values of 24 and 30 are 7. When $DN_P \geq DN_S$, the fractal index at the DN value of 52 is 4. Thus, the fractal index changes in the three phases.

The influences of the thresholds on the rank-size distributions of the NTL clusters are evaluated to validate that the clusters at the DN values on the opposite side of the optimal threshold DN_T have different system behaviors and urban-rural behaviors.

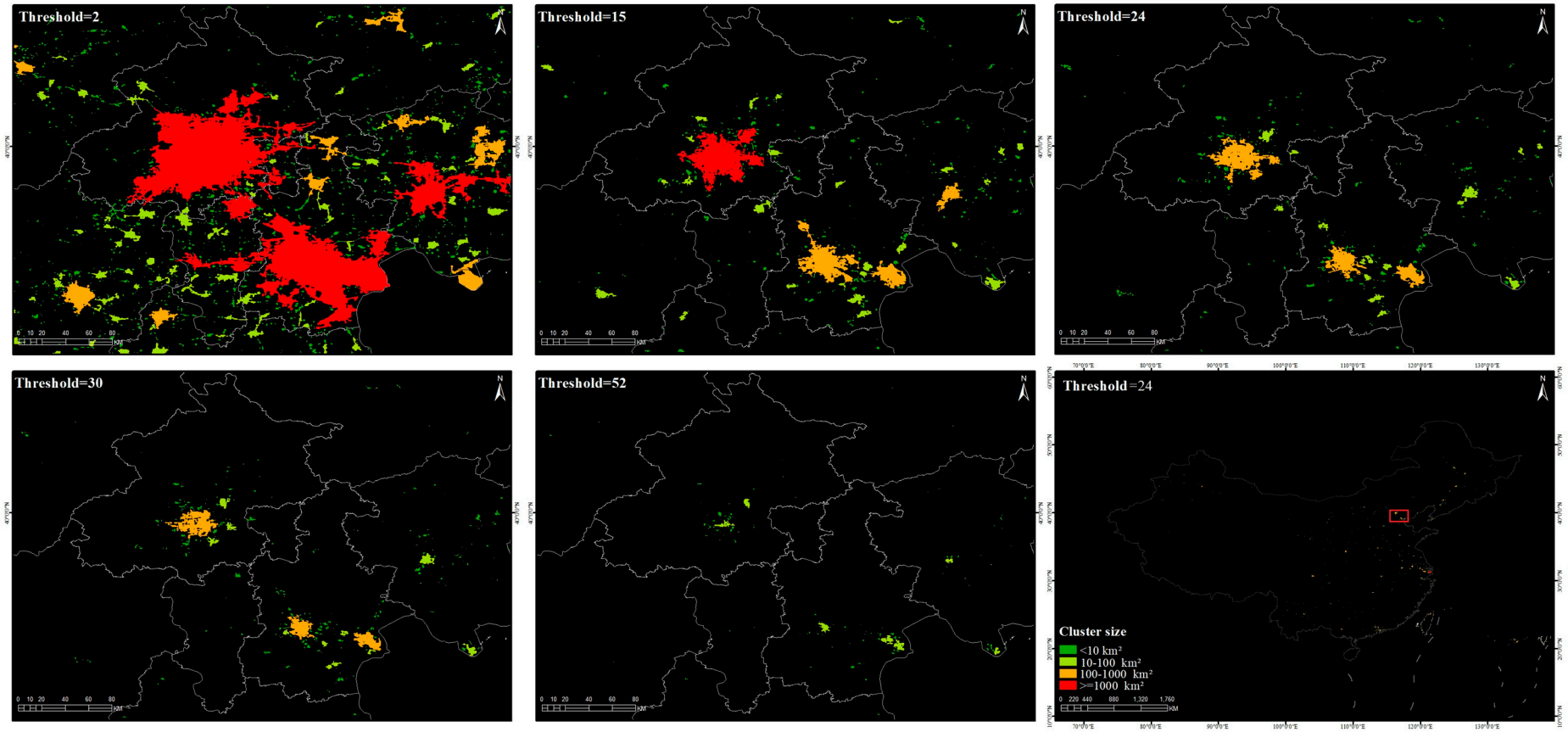

Figure 4. NTL clusters at varying thresholds from NPP-VIIRS data. The first map in the top row shows the NTL clusters at the optimal threshold in China, and the remaining five maps show the enlargement of the Beijing-Tianjin-Hebei metropolitan region when the potential threshold DN_P is set to thresholds in different categories, $DN_P < DN_T$ ($DN_P = 2$ and $DN_P = 15$), the optimal threshold ($DN_P = DN_T = 24$), $DN_T \leq DN_P < DN_S$ ($DN_P = 30$), and $DN_P \geq DN_S$ ($DN_P = 52$). The NTL clusters with different sizes are illustrated in a different color with 0–10 km^2 (green), 10–100 km^2 (peridot green), 100–1000 km^2 (orange), and >1000 km^2 (red).

The log-log plots of the power-law distributions of the NTL cluster sizes at the DN values in three phases—$DN_P < DN_T$ ($DN_P = 2$ and $DN_P = 15$), $DN_T \leq DN_P < DN_S$ ($DN_P = 24$ and $DN_P = 30$), and $DN_P \geq DN_S$ ($DN_P = 52$)—are illustrated in Figure 5. When $DN_P < DN_T$, the power-law distributions are dominated by the large segments, and the large segments at the DN values of 2 and 15 differ greatly due to the change patterns in Phase 1. Therefore, the exponents and the goodness of fit at the DN values of 2 and 15 differ greatly, indicating the fluctuations of the system behaviors. When $DN_T \leq DN_P < DN_S$, the exponents at the DN values of 24 and 30 are similar, approximately 2, with a *p-value* greater than 0.05, which implies Zipf's law and the self-organized behavior of urban systems. When $DN_P \geq DN_S$, larger agglomerations connected by dimmer regions break up internally so that the preferential attachment is violated, and the power-law distributions are dominated by medium segments. The slopes increase abruptly when $DN_P \geq DN_S$ with violent fluctuations, which deviates from Zipf's law and indicates that the violation of the scale-free, self-organized, and agglomeration behaviors of cities shrinks.

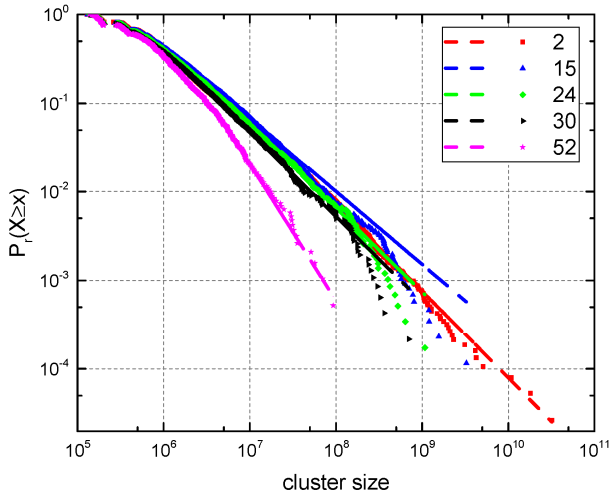

Figure 5. The log-log plot of the power-law distributions of the NTL cluster sizes at varying thresholds. When the potential threshold DN_P is set to 2 (red), 15 (blue), 24 (green), 30 (black), and 52 (magenta), the cumulative density distribution of the cluster sizes at each threshold is shown in the corresponding color, and its statistical power-law distribution function is displayed as the dotted line in the same color.

Most importantly, the Zipf's-based model is capable of quantifying changes in fractal behavior, self-organized behavior, and agglomeration behavior between non-urban and urban areas, so that it can be utilized to map urban extents and study the underlying structures and dynamics.

4.3. Comparison with the Head/Tail Breaks Method [8]

To assess the performance of the Zipf's law-based method, the head/tail breaks method established by Jiang [8] is conducted to obtain the universal optimal threshold, and this value is compared with the results of our method. Jiang [8] found that the NTL data followed a heavy-tailed distribution, where large values above the average were the minority, termed the head, and small values were the majority, termed the tail. This head/tail breaks method [62] iteratively divides the NTL data around the average into the head and tail parts until the head no longer represents a long-tailed distribution. This method has been verified as a universal and powerful method to delineate urban areas, namely, natural cities by Jiang [8].

We utilize the same NPP-VIIRS data in China in 2013 to determine a threshold using head/tail breaks methods, as shown in Table 2. The value of 33.14 is the optimal threshold with the head/tail breaks method. The optimal threshold from the head/tail breaks method is greater than that from the Zipf's law method. In detail, the value of 33.14 is greater than 24 (DN_T in the Zipf's law method)

and less than 39 (DN_S in the Zipf's law method). Thus, the optimal threshold from the head/tail breaks method locates in Phase 2 ($DN_T \leq DN_P < DN_S$), where the NTL clusters capture the whole urban areas made up of several interconnected sub-areas, and the size of each spatially contiguous urban areas shrinks around the perimeter. For the urban areas derived from the head/tail breaks method, since the optimal threshold is in Phase 2, the power-law exponent is around 2 with a high goodness-of-fit, indicating the Zipf's law behavior. However, the value of 33.14 is between the two endpoints of the interval in Phase 2 and closed to the right endpoint DN_S. Therefore, the change patterns of the urban areas are closed to breaking up. Figure 6 shows the urban areas around Beijing extracted from both methods. Although the urban areas capture the core in Beijing, separation patterns can be discovered in this stage and the sub-areas begin to appear.

Table 2. Statistics for the head/tail breaks method and power-law distribution exponent on NTL data.

Light	Number of Pixels	Mean Value	Count of Head	Percentage of Head	Count of Tail	Percentage of Tail	fi	*p*-Value
0.00–359.59	55,548,465	0.26	9,171,200	16.51	46,377,265	83.49		
0.26–359.59	9,171,200	1.73	1,685,716	18.38	7,485,484	81.62	2.07	0.00
1.73–359.59	1,685,716	8.23	519,035	30.79	1,166,681	69.21	1.98	0.48
8.23–359.59	519,035	19.87	191,648	36.92	327,387	63.08	1.76	0.00
19.87–359.59	191,648	33.14	70,202	36.63	121,446	63.37	1.83	0.01
33.14–359.59	70,202	47.71	29,265	41.69	40,937	58.31	1.94	0.26
47.71–359.59	29,265	66.13	23,514	79.37	5751	20.63	2.08	0.00

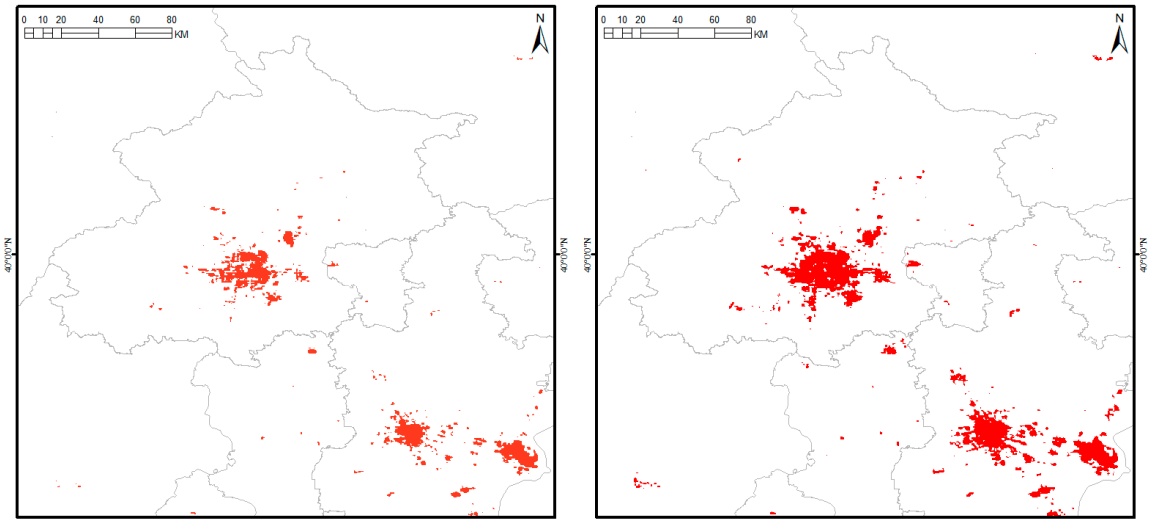

Figure 6. The urban areas extracted from the head/tail breaks method and those derived by the Zipf's law-based method.

Thus, the optimal threshold from the head/tail breaks method is lower than the nonurban-urban demarcation point. Both methods consider the fractal and self-similar behaviors of cities. The head/tail breaks method utilizes the percentage of the head and tail parts to describe the self-similarity and explores each mean value. Our proposed method utilized the Zipf's law model or power-law distribution model and calculates continuous values. For both methods, utilizing one threshold derives urban areas without the help of ancillary data, but the Zipf's law base method utilizes the model which reflects more information and calculates the values with a finer precision.

Furthermore, our method can solve the sensitivity issue in the head/tail breaks method. This sensitivity issue refers to the fact that objectivity is required for the condition to determine when the recursion process terminates in the method. Following the head/breaks method, the count of the tail is

less than that of the head [8]. Thus, the percentage of the tail, as seen in the seventh column in Table 2, can be regarded as a sensitive variable for the method. Previous studies [8] have suggested that this sensitive variable is approximately set to 60%. However, Jiang [62] noted that the sensitive variable could be modified in many cases, such as to 50% or even more. Therefore, this sensitive variable in the head/tail breaks method needs more objectivity. The proposed Zipf's law-based method provides the criterion to determine the sensitive variable. Taking the results in Table 2 as an example, the power-law distribution exponent β and the goodness-of-fit is computed at each break, as shown in the eighth and ninth column in Table 2. According to our method, the threshold is set to 33.14. This optimal threshold is the same as the results when the sensitive variable is assumed to be 60%. If we relax the threshold to 50% in this case, the exponent locates in Phase 3; moreover, the exponent then suggests the disappearance of the self-organized behavior. It can be suggested that the stopping threshold in the head/tail breaks method is better to set at 40% of the NPP-VIIRS NTL data.

Therefore, compared with the head/tail breaks method, the Zipf's law method can capture the change of the nonurban-urban pattern better and automatically terminate the recursion process.

4.4. Urban Area and Accuracy Evaluation

We apply this method to map the urban extents from the NPP-VIIRS NTL data in 2013. The optimal threshold in 2013 is 24 according to Section 4.1. Figure 7 displays the urban areas extracted by the proposed method. The urban areas vary significantly in different cities in China, and the majority are located in the southeast coastal cities and the northern areas, while the western inland cities are a relative minority.

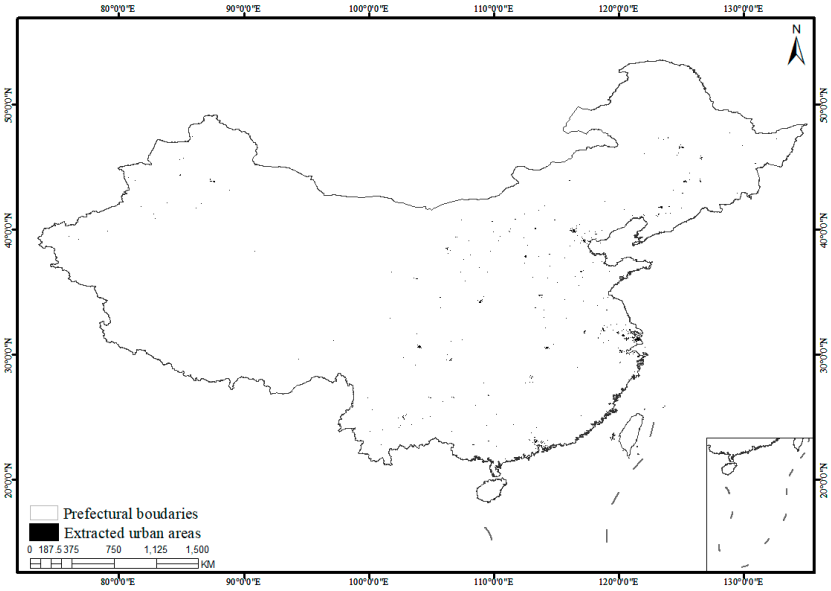

Figure 7. The urban areas in 2013 extracted from the NPP-VIIRS NTL data via the proposed method.

We evaluated the extracted urban area results utilizing the GlobeLand30-2010 data [51] at the pixel level. Now that the GlobeLand30 data, the high-resolution (30-m spatial resolution) LULC data, utilized multispectral images and auxiliary data in the classification process, the overall accuracy of the GlobeLand30-2010 product arrives at 80.33% [52] and the kappa indicator reaches 0.75 [52]. Therefore, these data can be assumed to be the ground truth reference. The GlobeLand30 classification data contain 10 land cover types including artificial surfaces. Therefore, the artificial surfaces, where pixel values equal eight, were segmented as the urban areas [51] in this study. After we extracted and calculated the non-urban areas derived from both the proposed method and the GlobeLand30-2010

data, the urban areas delineated from both our method and the ground truth reference, and for the total area, the overall accuracy has been calculated and reached 98.45%.

Previous studies generate varying thresholds for separated clusters or different cities, whilst the proposed method, which adopts the universal regularity—Zipf's law—for all cities, utilizes one unique threshold throughout China. With assistance from urban science theory, this method does not require ancillary data, which makes this method less data-dependent and more convenient. The accuracy assessment directly confirmed the effectiveness of our proposed method. Direct relationships between derived urban areas and administrative cities [45] cannot be found in this study since Zipf's law-based method regard cities with geographic and economic incoherence as an interconnected whole [45]. As a consequence, this proposed method provides a universal model to define urban behaviors [45] at the country scale, and the study in the future will discuss the applicability of the proposed method at a global and regional scale in detail.

4.5. Validation of Urban Area with POI Data

Previous studies have widely confirmed the utility of NTL data for extracting urban areas at the country level by examining quantitative connections between NTL intensity and demographic and economic factors [18]. The "Big Data Era" provides better clues to the properties of cities. The point of interest (POI) data, a new kind of location-based VGI data [49] with strong timeliness, can be utilized for validation as a proxy for human mobility or social interactions.

To detect the relationships between the urban areas extracted via our method and social interactions or human activities, three human activity variables covering most aspects of daily human life are extracted: shopping service POI, living service POI, and catering service POI, and the spatial patterns are statistically analyzed at the prefectural level, as indicated in Figure 8a–c. It is noticeable in these figures that most of the derived urban areas (the black polygons) are located in the prefectures with the bigger circles (indicating bigger counts of POI). It indicates that the delineated cities have strong connections with social interactions in the visual view. The further study explores the statistical correlations between the urban areas delineated by the proposed method and the counts of these three kinds of POI across the cities in China.

Figure 9 indicates the linear regression correlation results. Since the extracted areas fit the heavy-tailed distribution—small numbers compost the most portion of the distribution, the influence of large valued data is underrated for the conventional Ordinary Least Square (OLS) method. The Weighted Least Square (WLS) method is utilized in this scenario. Facing the problem caused by the heavy-tailed distribution data, the weighted matrix is used to make small data less overrated, and big data less underrated. Thus, we build the weighted matrix, which is the reciprocal for the probability distribution function (PDF) of extracted urban areas. The correlation results imply that the extracted areas have a clear positive linear correlation with the counts of shopping service POI ($R^2 = 0.90$, $p < 0.001$), living service POI ($R^2 = 0.93$, $p < 0.001$), and catering service POI ($R^2 = 0.91$, $p < 0.001$). Since these three categories of POIs are regarded as indicative of human activities, the clear linear relationships suggest that the cities extracted by the proposed method are functional, strongly interacting, and collocated social networks. Further statistical analysis of the response of the extracted urban areas to social production was carried out to establish the statistic correlations between the extracted urban areas and socioeconomic variables. For the NPP-VIIRS NTL data, our study confirmed a clear linear correlation between the derived urban extents and the social production variable: GDP (Figures 8d and 9d). These linear correlations connections with human and economic activity variables verify that the derived urban areas can be a proxy for city functions and human activities.

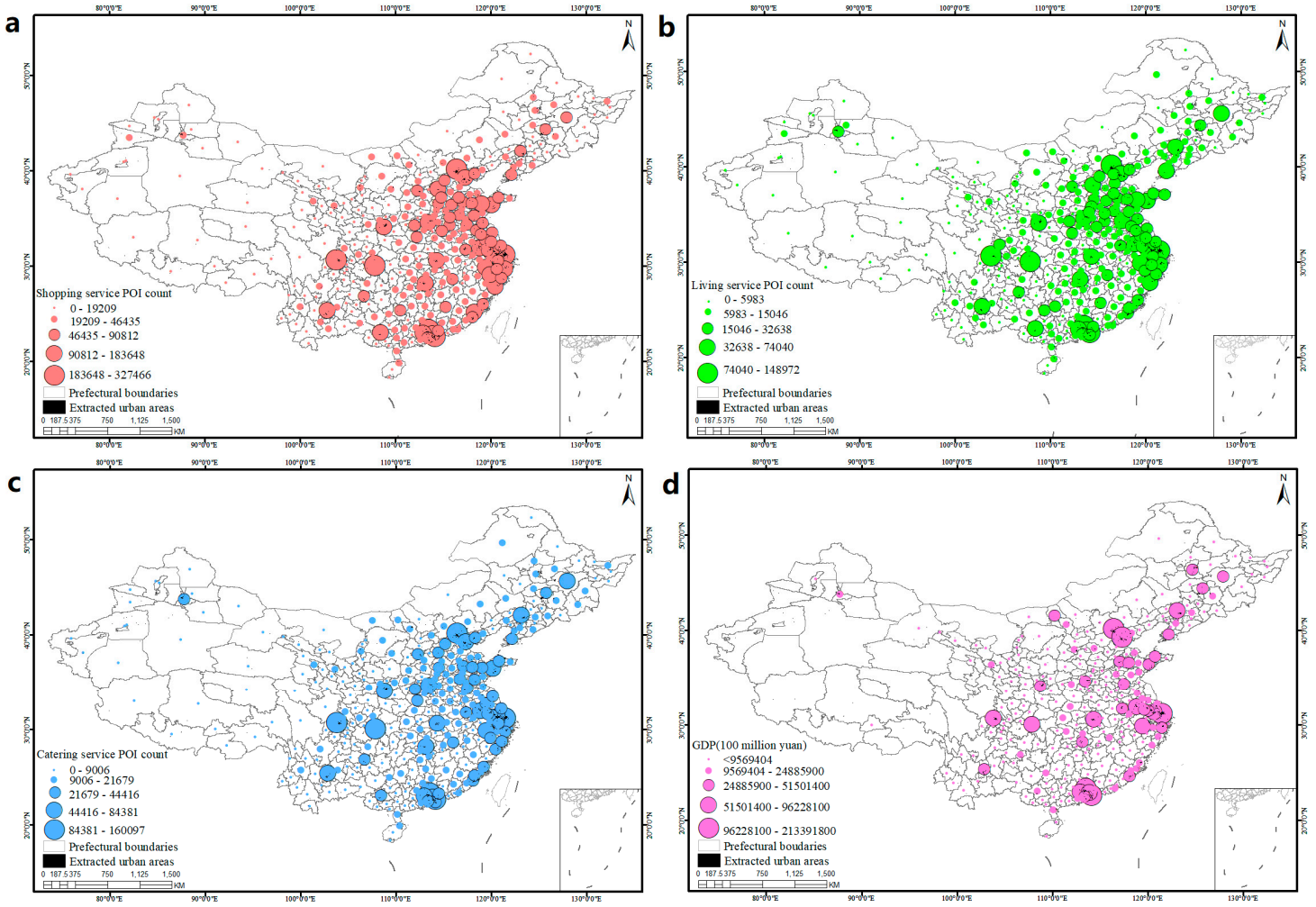

Figure 8. The statistical POI count and extracted urban areas in 2013 in NPP-VIIRS. (**a**) The statistical shopping service POI count and extracted urban areas; (**b**) The statistical living service POI count and extracted urban areas; (**c**) The statistical catering service POI count and extracted urban areas; (**d**) The GDP and extracted urban areas.

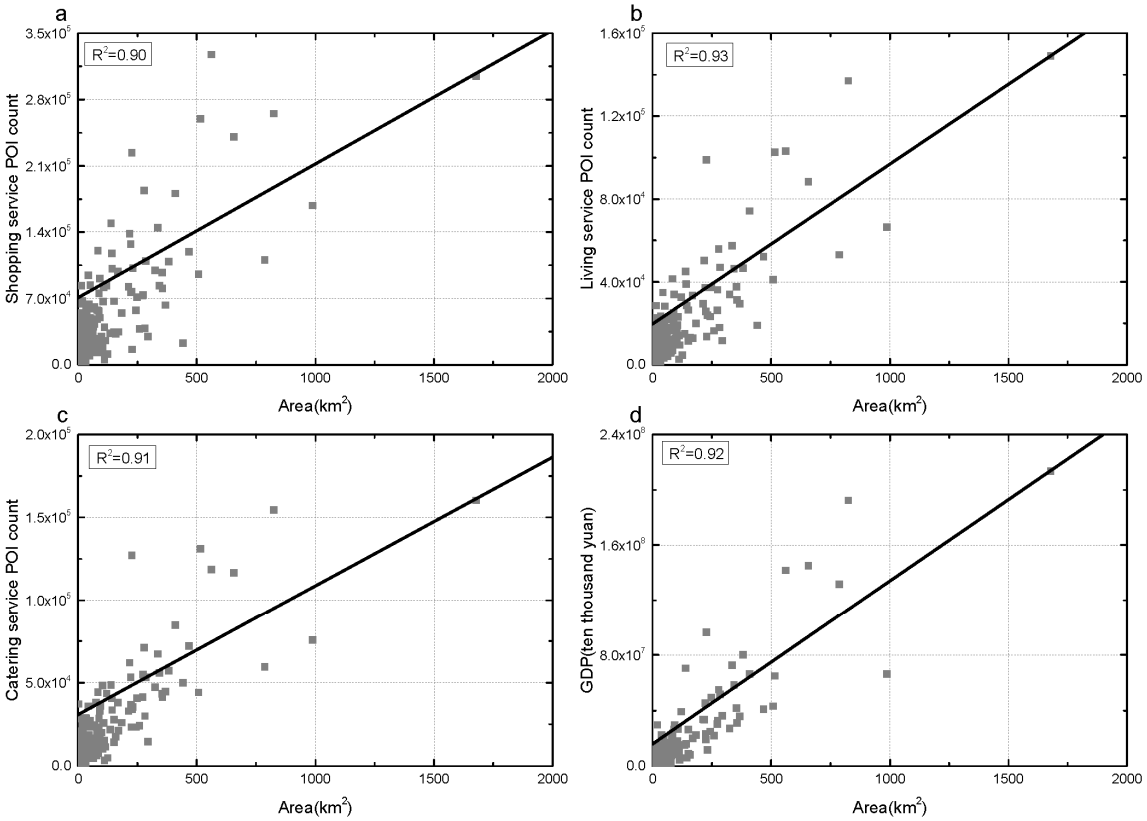

Figure 9. Statistical relationships between urban areas and social interactions and production. (**a**) The relationship between urban areas and shopping service POI count; (**b**) The relationship between urban areas and living service POI count; (**c**) The relationship between urban areas and catering service POI count; (**d**) The relationship between urban areas and GDP.

4.6. Comparison with the INN-SVM Method [30]

To further verify the method described in this study, we make a comparative analysis of the urban areas extracted via our method and the Integrated NTLs and normalized difference vegetation index (NDVI) support the vector machine (SVM) classification (INN-SVM) method from Yang [30], as shown in Figure 10. Yang produced a stratified SVM-based method at the country level using NTL data and NDVI data to map the urban land. A similar method has been validated to have the best performance compared to the local-optimized thresholding (LOT) and vegetation-adjusted NTL urban index (VANUI) methods, which are popular methods to derive urban areas utilizing NTL data [19].

As shown in Figure 10, we can see that the extracted areas by the proposed method are slightly smaller than those utilizing the INN-SVM method. However, the locations of the extracted areas by both methods are consistent. This may be attributed to the fact that the proposed method is from the perspective of the behaviors of urban systems resulting from social interactions, while the INN-SVM method utilizes ancillary data about NDVI from the perspective of land use.

Comparing the overall accuracy of both methods, the overall accuracy of the proposed method is 98.45% and the overall accuracy of the INN-SVM method is 98.47%. The result confirmed that the Zipf's law-based method has as good a performance as the INN-SVM method.

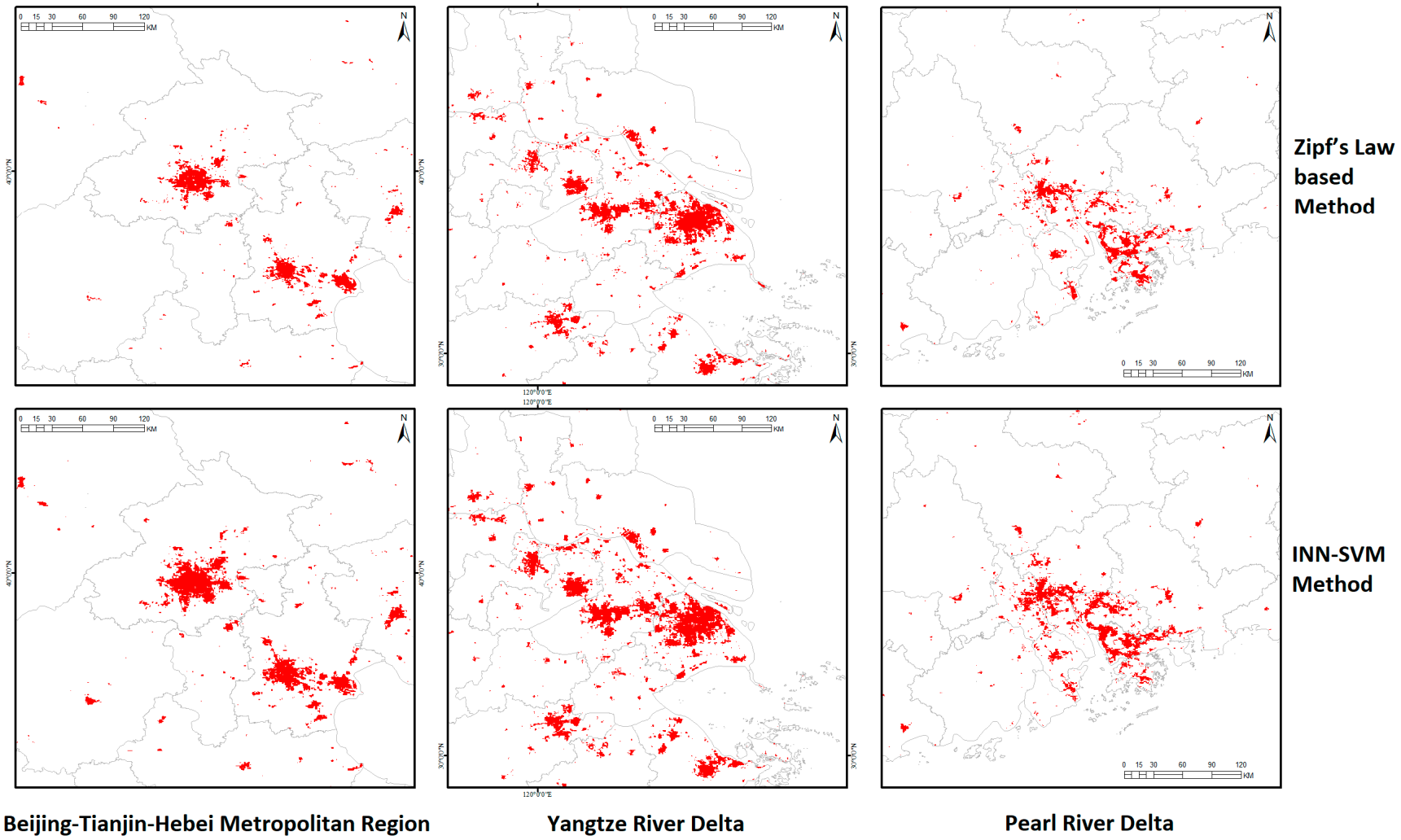

Figure 10. The urban areas extracted using the INN-SVM method [30] and those derived by the Zipf's Law-based method.

The comparative analysis has been made to calculate the quantitative correlations between social interactions and the urban areas extracted via our method and the INN-SVM method. It is noteworthy that urban areas extracted via these two methods show similar significant responses to the three kinds of POI variables denoting social interactions. The urban areas extracted by our method show slightly stronger responses to the shopping service POI ($R^2 = 0.90$ vs. $R^2 = 0.90$), living service POI ($R^2 = 0.91$ vs. $R^2 = 0.93$), and catering service POI ($R^2 = 0.93$ vs. $R^2 = 0.91$). This result suggests that the city area extracted by the proposed method can be a proxy for human activities.

Following the evaluation system in Shu et al. [63], we summarize the two methods from five perspectives: convenience, the degree of automation, data dependency, accuracy, and response to human activity. Similar to the INN-SVM method, the proposed method has directly shown a high overall accuracy and strong response to human activity, thus validating the effectiveness. With the assistance of urban science theory, the proposed Zipf's law method does not require ancillary data, indicating a high degree of automation. Therefore, the proposed method is not only effective but also more convenient and achievable.

5. Conclusions

To extract urban areas from NTL data from the perspective of the behaviors of urban systems, this paper proposes a Zipf's law-based method to generate the optimal threshold and extract urban areas without ancillary data, referring to human activities. The optimal threshold at the country scale is estimated through the three-phase cluster dynamics based on the statistical properties of the Zipf's law model of NTL clusters on continuous potential thresholds. The three-phase model is discovered through the power-law distribution investigated on clusters by varying the thresholds. The experiment on the NPP-VIIRS NTL data in 2013 in China shows the three-phase cluster dynamics. The exponents and the goodness-of-fit fluctuate violently in Phase 1 and Phase 3, and there are consistent statistically significant power-law exponents in the range of 1.88 to 2.02 in Phase 2. The optimal threshold corresponds to the abrupt transition between Phase 1 and Phase 2. The optimal threshold based on this abrupt transition is compared with the head/tail breaks method, indicating that the proposed method can capture the change of the nonurban-urban pattern better and shows a higher degree of automation. The urban areas derived from the optimal threshold obtain a high overall accuracy. The validation using POI data verifies that the proposed method is not only effective but also more convenient and achievable compared to the INN-SVM method.

Our evaluation of the cluster dynamics in three phases suggests that the abrupt transition between Phase 1 and Phase 2 denotes the rural-urban demarcation point, indicating the change in fractal behavior, self-organized behavior, and agglomeration behavior on both sides. The validation of the mapped urban extent at the country level in China with POI data confirms that it can be a proxy of human activities. A comparison with existing methods validates the effectiveness of mapping urban areas at the universal scale and the dynamics using the NPP-VIIRS NTL data in a convenient and automated way. We conclude that a Zipf's law-based optimal threshold allows for an automated and systematic method to derive human activities in general at the country scale, from the perspective of behaviors of urban systems without ancillary data. The proposed method is less data dependent; thus, it can define and delineate the whole world utilizing big data. In consequence of the universality, this method provides a valuable tool for urban studies from both spatial and temporal perspectives.

Zipf's law, the urban science theory from the bottom up, provides information about how cities behave differently from rural areas. The Zipf's law-based method shows great potential to provide a universal criterion to map urban areas from the perspective of the behaviors of urban systems without ancillary data. In this paper, the statistical properties of a Zipf's law model on continuous thresholds provide an effective and efficient manner to map urban areas at the country scale. More analysis should be conducted to explore the global and regional applicability of the proposed method in detail.

Acknowledgments: This research was supported by the Fund of National Natural Science Foundation of China [grant number 41571414] and the Tsinghua University Initiative Scientific Research Program [grant number

2015THZ01]. We thank anonymous reviewers for their constructive comments, which greatly improved this article. We also thank Xiaoli Ding and Peichao Gao at the Hong Kong Polytechnic University for helpful discussions.

Author Contributions: Wenjia Wu and Hongrui Zhao conceived and designed the experiments; Wenjia Wu performed the experiments; Wenjia Wu analyzed the data; Shulong Jiang contributed analysis tools; Wenjia Wu and Hongrui Zhao wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bettencourt, L.M.; Lobo, J.; Helbing, D.; Kühnert, C.; West, G.B. Growth, innovation, scaling, and the pace of life in cities. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 7301–7306. [CrossRef] [PubMed]
2. Batty, M. *The New Science of Cities*; The MIT Press: Cambridge, MA, USA, 2013; pp. 123–126.
3. Pumain, D. *Scaling Laws and Urban Systems*; Santa Fe Institute: Santa Fe, NM, USA, 2004.
4. Portugali, J. *Self-Organization and the City*; Springer Science & Business Media: Berlin, Germany, 2012.
5. Bettencourt, L.M. The origins of scaling in cities. *Science* **2013**, *340*, 1438–1441. [CrossRef] [PubMed]
6. Arcaute, E.; Hatna, E.; Ferguson, P.; Youn, H.; Johansson, A.; Batty, M. Constructing cities, deconstructing scaling laws. *J. R. Soc. Interface* **2015**, *12*, 20140745. [CrossRef] [PubMed]
7. Louf, R. Wandering in Cities: A Statistical Physics Approach to Urban Theory. Ph.D. Thesis, Cornell University, Ithaca, New York, 25 November 2015.
8. Jiang, B. Head/tail breaks for visualization of city structure and dynamics. *Cities* **2015**, *43*, 69–77. [CrossRef]
9. Jiang, B.; Miao, Y. The evolution of natural cities from the perspective of location-based social media. *Prof. Geogr.* **2014**, *67*, 295–306. [CrossRef]
10. Long, Y. Redefining chinese city system with emerging new data. *Appl. Geogr.* **2016**, *75*, 36–48. [CrossRef]
11. Berry, B.J.L.; Okulicz-Kozaryn, A. The city size distribution debate: Resolution for us urban regions and megalopolitan areas. *Cities* **2012**, *29*, S17–S23. [CrossRef]
12. Li, X.; Wang, X.; Zhang, J.; Wu, L. Allometric scaling, size distribution and pattern formation of natural cities. *Palgrave Commun.* **2015**, *1*, 15017. [CrossRef]
13. Elvidge, C.D.; Baugh, K.E.; Kihn, E.A.; Kroehl, H.W.; Davis, E.R. Mapping city lights with nighttime data from the dmsp operational linescan system. *Photogramm. Eng. Remote Sens.* **1997**, *63*, 727–734.
14. Small, C.; Elvidge, C.D.; Balk, D.; Montgomery, M. Spatial scaling of stable night lights. *Remote Sens. Environ.* **2011**, *115*, 269–280. [CrossRef]
15. Yu, B.; Shu, S.; Liu, H.; Song, W.; Wu, J.; Wang, L.; Chen, Z. Object-based spatial cluster analysis of urban landscape pattern using nighttime light satellite images: A case study of china. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 2328–2355. [CrossRef]
16. Ma, T.; Yin, Z.; Li, B.; Zhou, C.; Haynie, S. Quantitative estimation of the velocity of urbanization in China using nighttime luminosity data. *Remote Sens.* **2016**, *8*, 94. [CrossRef]
17. Yu, B.; Shi, K.; Hu, Y.; Huang, C.; Chen, Z.; Wu, J. Poverty evaluation using npp-viirs nighttime light composite data at the county level in china. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *8*, 1217–1229. [CrossRef]
18. He, C.; Liu, Z.; Tian, J.; Ma, Q. Urban expansion dynamics and natural habitat loss in china: A multiscale landscape perspective. *Glob. Chang. Biol.* **2014**, *20*, 2886–2902. [CrossRef] [PubMed]
19. Dou, Y.; Liu, Z.; He, C.; Yue, H. Urban land extraction using viirs nighttime light data: An evaluation of three popular methods. *Remote Sens.* **2017**, *9*, 175. [CrossRef]
20. Elvidge, C.D.; Baugh, K.; Zhizhin, M.; Hsu, F.C.; Ghosh, T. Viirs night-time lights. *Int. J. Remote Sens.* **2017**, *38*, 5860–5879. [CrossRef]
21. Elvidge, C.D.; Baugh, K.E.; Zhizhin, M.; Hsu, F.-C. Why viirs data are superior to DMSP for mapping nighttime lights. *Proc. Asia-Pac. Adv. Netw.* **2013**, *35*, 62–69. [CrossRef]
22. Li, X.; Zhou, Y. Urban mapping using DMSP/OLS stable night-time light: A review. *Int. J. Remote Sens.* **2017**, 1–17. [CrossRef]
23. Liu, X.; Hu, G.; Ai, B.; Li, X.; Shi, Q. A normalized urban areas composite index (nuaci) based on combination of dmsp-ols and modis for mapping impervious surface area. *Remote Sens.* **2015**, *7*, 17168–17189. [CrossRef]
24. Xie, Y.; Weng, Q. Updating urban extents with nighttime light imagery by using an object-based thresholding method. *Remote Sens. Environ.* **2016**, *187*, 1–13. [CrossRef]

25. Zhou, Y.; Smith, S.J.; Elvidge, C.D.; Zhao, K.; Thomson, A.; Imhoff, M. A cluster-based method to map urban area from DMSP/OLS nightlights. *Remote Sens. Environ.* **2014**, *147*, 173–185. [CrossRef]
26. Shi, K.; Yu, B.; Huang, Y.; Hu, Y.; Yin, B.; Chen, Z.; Chen, L.; Wu, J. Evaluating the ability of npp-viirs nighttime light data to estimate the gross domestic product and the electric power consumption of china at multiple scales: A comparison with dmsp-ols data. *Remote Sens.* **2014**, *6*, 1705–1724. [CrossRef]
27. Sharma, R.C.; Tateishi, R.; Hara, K.; Gharechelou, S.; Iizuka, K. Global mapping of urban built-up areas of year 2014 by combining modis multispectral data with viirs nighttime light data. *Int. J. Dig. Earth* **2016**, *9*, 1004–1020. [CrossRef]
28. Li, Q.; Lu, L.; Weng, Q.; Xie, Y.; Guo, H. Monitoring urban dynamics in the southeast USA using time-series dmsp/ols nightlight imagery. *Remote Sens.* **2016**, *8*, 578. [CrossRef]
29. Jing, W.; Yang, Y.; Yue, X.; Zhao, X. Mapping urban areas with integration of DMSP/OLS nighttime light and modis data using machine learning techniques. *Remote Sens.* **2015**, *7*, 12419–12439. [CrossRef]
30. Yang, Y.; He, C.Y.; Zhang, Q.F.; Han, L.J.; Du, S.Q. Timely and accurate national-scale mapping of urban land in china using defense meteorological satellite program's operational linescan system nighttime stable light data. *J. Appl. Remote Sens.* **2013**, *7*, 073535. [CrossRef]
31. Sutton, P. Modeling population density with night-time satellite imagery and gis. *Comput. Environ. Urban Syst.* **1997**, *21*, 227–244. [CrossRef]
32. Imhoff, M.L.; Lawrence, W.T.; Stutzer, D.C.; Elvidge, C.D. A technique for using composite dmsp/ols "city lights" satellite data to map urban area. *Remote Sens. Environ.* **1997**, *61*, 361–370. [CrossRef]
33. Ma, T.; Zhou, Y.; Zhou, C.; Haynie, S.; Pei, T.; Xu, T. Night-time light derived estimation of spatio-temporal characteristics of urbanization dynamics using DMSP/OLS satellite data. *Remote Sens. Environ.* **2015**, *158*, 453–464. [CrossRef]
34. He, C.; Shi, P.; Li, J.; Chen, J.; Pan, Y.; Li, J.; Zhuo, L.; Ichinose, T. Restoring urbanization process in china in the 1990s by using non-radiance-calibrated DMSP/OLS nighttime light imagery and statistical data. *Chin. Sci. Bull.* **2006**, *51*, 1614–1620. [CrossRef]
35. Milesi, C.; Elvidge, C.D.; Nemani, R.R.; Running, S.W. Assessing the impact of urban land development on net primary productivity in the southeastern united states. *Remote Sens. Environ.* **2003**, *86*, 401–410. [CrossRef]
36. Henderson, M.; Yeh, E.T.; Gong, P.; Elvidge, C.; Baugh, K. Validation of urban boundaries derived from global night-time satellite imagery. *Int. J. Remote Sens.* **2003**, *24*, 595–609. [CrossRef]
37. Li, X.; Gong, P.; Liang, L. A 30-year (1984–2013) record of annual urban dynamics of beijing city derived from landsat data. *Remote Sens. Environ.* **2015**, *166*, 78–90. [CrossRef]
38. Liu, Z.; He, C.; Zhang, Q.; Huang, Q.; Yang, Y. Extracting the dynamics of urban expansion in china using dmsp-ols nighttime light data from 1992 to 2008. *Landsc. Urban Plan.* **2012**, *106*, 62–72. [CrossRef]
39. Zhou, Y.; Smith, S.J.; Zhao, K.; Imhoff, M.; Thomson, A.; Bond-Lamberty, B.; Asrar, G.R.; Zhang, X.; He, C.; Elvidge, C.D. A global map of urban extent from nightlights. *Environ. Res. Lett.* **2015**, *10*, 054011. [CrossRef]
40. Dennett, D.C. Real patterns. *J. Philos.* **1991**, *88*, 27–51. [CrossRef]
41. Berry, B.J.L. Cities as systems within systems of cities. *Pap. Reg. Sci.* **1964**, *13*, 146–163. [CrossRef]
42. Cristelli, M.; Batty, M.; Pietronero, L. There is more than a power law in zipf. *Sci. Rep.* **2012**, *2*, 812. [CrossRef] [PubMed]
43. Zipf, G.K. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*; Martino Publishing: Eastford, CT, USA, 2012.
44. Gabaix, X. Zipf's law for cities: An explanation. *Q. J. Econ.* **1999**, *114*, 739–767. [CrossRef]
45. Jiang, B.; Yin, J.; Liu, Q. Zipf's law for all the natural cities around the world. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 498–522. [CrossRef]
46. Soo, K.T. Zipf's law for cities: A cross-country investigation. *Reg. Sci. Urban Econ.* **2005**, *35*, 239–263. [CrossRef]
47. Batty, M.; Longley, P.A. *Fractal Cities—A Geometry of Form and Function*; Academic Press Professional, Inc.: Cambridge, MA, USA, 1994.
48. Version 1 VIIRS Day/Night Band Nighttime Lights. Available online: https://www.ngdc.noaa.gov/eog/viirs/download_dnb_composites.html (accessed on 17 May 2017).
49. Goodchild, M.F. Citizens as voluntary sensors: Spatial data infrastructure in the world of web 2.0. *Int. J. Spat. Data Infrastruct. Res.* **2007**, *2*, 24–32.

50. Point of Interest (POI) Data in China in 2014. Available online: https://geohey.com/ (accessed on 15 June 2016).
51. National Geomatics Center of China. Available online: http://www.globallandcover.com (accessed on 17 May 2017).
52. Chen, J.; Chen, J.; Liao, A.; Cao, X.; Chen, L.; Chen, X.; He, C.; Han, G.; Peng, S.; Lu, M.; et al. Global land cover mapping at 30 m resolution: A pok-based operational approach. *ISPRS J. Photogramm. Remote Sens.* **2015**, *103*, 7–27. [CrossRef]
53. Shi, K.; Huang, C.; Yu, B.; Yin, B.; Huang, Y.; Wu, J. Evaluation of npp-viirs night-time light composite data for extracting built-up urban areas. *Remote Sens. Lett.* **2014**, *5*, 358–366. [CrossRef]
54. Newman, M.E. Power laws, pareto distributions and Zipf's law. *Contemp. Phys.* **2005**, *46*, 323–351. [CrossRef]
55. Clauset, A.; Shalizi, C.R.; Newman, M.E.J. Power-law distributions in empirical data. *Siam Rev.* **2009**, *51*, 661–703. [CrossRef]
56. Goldstein, M.L.; Morris, S.A.; Yen, G.G. Problems with fitting to the power-law distribution. *Eur. Phys. J. B-Condens. Matter Complex Syst.* **2004**, *41*, 255–258. [CrossRef]
57. Rozenfeld, H.D.; Rybski, D.; Andrade, J.S., Jr.; Batty, M.; Stanley, H.E.; Makse, H.A. Laws of population growth. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 18702. [CrossRef] [PubMed]
58. Long, Y.; Shen, Y.; Jin, X. Mapping block-level urban areas for all chinese cities. *Ann. Am. Assoc. Geogr.* **2016**, *106*, 96–113. [CrossRef]
59. Ma, D.; Sandberg, M.; Jiang, B. Characterizing the heterogeneity of the openstreetmap data and community. *ISPRS Int. J. Geo-Inf.* **2015**, *4*, 535–550. [CrossRef]
60. Ontoy, D.S.; Padua, R.N. Measuring species diversity for conservation biology: Incorpo-rating social and ecological importance of species. *Biodivers. J.* **2014**, *5*, 387–390.
61. Gao, P.; Liu, Z.; Liu, G.; Zhao, H.; Xie, X. Unified metrics for characterizing the fractal nature of geographic features. *Ann. Am. Assoc. Geogr.* **2017**, 1315–1331. [CrossRef]
62. Jiang, B.; Yin, J. Ht-index for quantifying the fractal or scaling structure of geographic features. *Ann. Assoc. Am. Geogr.* **2014**, *104*, 530–540. [CrossRef]
63. Shu, S.; Yu, B.; Wu, J.; Liu, H. Methods for deriving urban built-up area using night-light data: Assessment and application. *Remote Sens. Technol. Appl.* **2011**, *26*, 169–176.