

## Calibrating Spatial Stratified Heterogeneity for Heavy-Tailed Distributed Data

Bisong Hu, Tingting Wu, Qian Yin, Jinfeng Wang, Bin Jiang & Jin Luo

**To cite this article:** Bisong Hu, Tingting Wu, Qian Yin, Jinfeng Wang, Bin Jiang & Jin Luo (2024) Calibrating Spatial Stratified Heterogeneity for Heavy-Tailed Distributed Data, *Annals of the American Association of Geographers*, 114:7, 1568-1586, DOI: [10.1080/24694452.2024.2351002](https://doi.org/10.1080/24694452.2024.2351002)

**To link to this article:** <https://doi.org/10.1080/24694452.2024.2351002>



Published online: 18 Jun 2024.



Submit your article to this journal [↗](#)



Article views: 150






View related articles [↗](#)



View Crossmark data [↗](#)

# Calibrating Spatial Stratified Heterogeneity for Heavy-Tailed Distributed Data

Bisong Hu,<sup>a</sup>  Tingting Wu,<sup>a</sup> Qian Yin,<sup>b</sup> Jinfeng Wang,<sup>b</sup>  Bin Jiang,<sup>c</sup>  and Jin Luo<sup>a</sup>

<sup>a</sup>School of Geography and Environment, Jiangxi Normal University, China; <sup>b</sup>State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, China; <sup>c</sup>Urban Governance and Design Thrust, Society Hub, The Hong Kong University of Science and Technology (Guangzhou), China

The phenomena with within-strata characteristics that are more similar than between-strata characteristics are ubiquitous (e.g., land-use types and image classifications). It can be summarized as spatial stratified heterogeneity (SSH), which is measured and attributed using the geographical detector (Geodetector)  $q$ -statistic. SSH is typically calibrated by stratification and hundreds of algorithms have been developed. Little is discussed about the conditions of the methods. In this work, a novel stratification method based on head/tail breaks is introduced for the purpose of better capturing the SSH of geographical variables with a heavy-tailed distribution. Compared to conventional sample-based stratifications, the presented approach is a population-based optimized stratification that indicates an underlying scaling property in geographical spaces. It requires no prior knowledge or auxiliary variables and supports a naturally determined number of strata instead of being subjectively preset. In addition, our approach reveals the inherent hierarchical structure of geographical variables, characterizes its dominant components across all scales, and provides the potential to make the stratification meaningful and interpretable. The advantages were illustrated by several case studies in natural and social sciences. The proposed approach is versatile and flexible so that it can be applied for the stratification of both geographical and nongeographical variables and is conducive to advancing SSH-related studies as well. This study provides a new way of thinking for advocating spatial heterogeneity or scaling law and advances our understanding of geographical phenomena. *Key Words:* Geodetector  $q$ -statistic, head/tail breaks, spatial stratified heterogeneity, stratification, structural hierarchy.

The first law of geography, also known as Tobler's law (Tobler 1970), pertains to the term of spatial autocorrelation, which can be measured by the Moran's  $I$  statistic, semivariogram, or local indicators of spatial association (LISA; Moran 1950; Matheron 1963; Anselin 1995). The so-called second law of geography refers to another fundamental property of spatial data—spatial heterogeneity—which implies geographical variables exhibiting uncontrolled variances and refers to spatial nonstationarity (Goodchild 2004). Specifically, spatial heterogeneity can indicate spatial variability (local clustering or dispersing) or spatial stratification of heterogeneity significance at different spatial scales (Wang, Zhang, and Fu 2016). The latter refers to the term of spatial stratified heterogeneity (SSH), which is ubiquitous in social and natural phenomena; for example, the annual distribution of the

Normalized Difference Vegetation Index (NDVI) under different climate zones and the population distribution between two sides of the Hu Line (an important population dividing line in China). SSH describes the spatial heterogeneity between strata for spatial variables, where the within-strata variance is less than the between-strata variance. Note that SSH implies the heterogeneity by strata for observations of variables or for parameters of models, as well as the underlying distinct mechanisms, and is conducive to the enhancement of statistical inference if it is appropriately recognized (Wang, Zhang, and Fu 2016).

The Geodetector  $q$ -statistic was developed to measure SSH and to make attributions for or by SSH (Wang et al. 2010). In parallel to the semivariogram for kriging, stratification is the critical parameter for Geodetector and has been

## ARTICLE HISTORY

Initial submission, November 2023; revised submissions, January and February 2024; final acceptance, March 2024

CORRESPONDING AUTHOR Bin Jiang  [binjiang@hkust-gz.edu.cn](mailto:binjiang@hkust-gz.edu.cn)

© 2024 by American Association of Geographers

implemented by prior knowledge and numerous classification algorithms. Appropriate stratifications can help identify the inherent SSH of a geographical variable (Wang, Zhang, and Fu 2016; Li et al. 2019; Guo et al. 2022; Wang et al. 2024), characterize the potential determinants (Wang et al. 2010; Zhang et al. 2019; Hu et al. 2020; Hu, Fu, et al. 2023), evaluate the representativeness of alternatives (Yin et al. 2019; Hu, Ning, Li, et al. 2021), and overcome statistical confounding in modeling (Hu, Ning, Qiu, et al. 2021). Stratification also plays an extremely important role in constructing the zoning layer for the sandwich estimator, which is a method of spatial prediction for the SSH population (Wang et al. 2013; Liu et al. 2018; Liao et al. 2019). Moreover, it is beneficial to the SSH measure in the spatial statistic trinity, which is a generic framework for spatial statistics that integrates design-based and model-based statistical procedures (Wang, Gao, and Stein 2020).

There are plenty of stratification methods that can be classified into four categories: univariate, cluster-based, multicriteria, and supervised stratifications (Guo et al. 2022). Note that the strata of a target variable can be a partition either on its own or based on an explanatory variable. In other words, stratification can be implemented based on a categorical variable or by discretizing a numerical variable. Previous studies have implemented several optimizations for stratification (Li et al. 2008; Cao, Ge, and Wang 2013, 2014; B.-B. Gao et al. 2015; Song et al. 2020). Nevertheless, optimizing stratification for a geographical variable without prior knowledge or auxiliary variables is still challenging. Existing optimized stratifications are typically implemented by selecting the one with the highest  $q$ -statistic value from a range of classification or discretization methods (e.g., Cao, Ge, and Wang 2013; Song et al. 2020). There is no guarantee that the corresponding strata of a geographical variable are reasonable or meaningful. Furthermore, another key issue in stratification is determining the number of strata. Many previous classification or discretization methods, such as Jenks natural breaks (Fisher 1958; Jenks and Caspall 1971), require a subjectively preset parameter to determine the number of strata. Finally, it is surely understood that there should be different applicable conditions of various stratification methods. We

should investigate the population property of a geographical variable before conducting the stratification for the sample observations from its population.

On another aspect, heavy-tailed distributions are commonly found in numerous social and natural phenomena, such as the Earth's terrain surface, the distribution of residential areas across urban and rural regions, the Zipf's law of city sizes, and the Pareto's 80/20 rule of human incomes. The heavy-tailed distribution can be recognized as an underlying pattern in geographical spaces (Jiang 2013). Generally speaking, a heavy-tailed distribution is heavily or extremely right-skewed, with far more small values in the tail than large values in the head (e.g., power laws). The head/tail breaks method was initially developed as a classification scheme to recursively capture the underlying hierarchical structure of the data with a heavy-tailed distribution (Jiang 2013). It captures the essence of the data distribution (scaling or hierarchy) and provides the potential to reveal the structural hierarchy of the data for a geographical variable. This makes the classification result meaningful in geographical spaces. Meanwhile, the classification intervals are iteratively derived, and the number of classes is naturally determined during the process of head/tail breaks (i.e., a preset number of classes is not required). One more notable advantage is that the head/tail breaks process can derive meaningful dominant components of a geographical variable with a heavy-tailed distribution. For example, it can be used to identify natural cities that indicate urban areas from nighttime light (NTL) images (Jiang and Yin 2014; Jiang 2015b).

In view of these considerations, this work aims to introduce a population-based optimized stratification method based on head/tail breaks to better capture the SSH of geographical variables with a heavy-tailed distribution. Several real-world case studies were subsequently conducted to validate our approach and its advantages. The proposed stratification is versatile and flexible, and it can be applied in other studies for the stratification of both geographical and nongeographical variables that exhibit heavy-tailed distributions (for the purpose of advancing SSH-related studies; e.g., capturing SSH, characterizing determinant powers, supporting stratified modeling, and others). This article also provides novelty in terms of the way of

thinking for advocating spatial heterogeneity or scaling law and advances our understanding of geographical phenomena.

The remainder of this article is structured as follows. First, we briefly summarize the principles of SSH and  $q$ -statistic, especially the impact of its critical parameter (i.e., stratification). We then provide an overview of the methodology organized as three subsections, including heavy-tailed distribution detection, head/tail breaks and ht-index, and head/tail breaks-based stratification. The results of four real-world case studies are explicitly interpreted to verify the properties and advantages of the methodology, along with the evaluation of its performance for calibrating SSH. We then delve into a comprehensive discussion of several main contributions of this article, as well as the limitations and future work. Finally, we summarize our conclusions.

## Spatial Stratified Heterogeneity and $q$ -Statistic

Tobler's law refers to near things more related than distant things (Tobler 1970) and describes a global characteristic of the spatial distribution of geographical phenomena, known as spatial autocorrelation. In another aspect, SSH is a universal characteristic of nature at all scales, referring to the within-strata variance less than the between-strata variance (Wang, Zhang, and Fu 2016). Here, strata refer to the sampled subsets of the population or can be simply understood as zones, subregions, classes, and so on. When the observations are relatively homogeneous within each stratum but heterogeneous between strata, the SSH exists. More specifically, a perfect SSH can be recognized when the observations within each stratum are identical (i.e., the variance within each stratum is zero). Contrarily, the SSH disappears when the mean of the observations within each stratum is identical to the mean of all observations across the entire study area (i.e., there is no difference between strata).

The concept of SSH was first introduced by the power of determinant in geographical detector (Wang et al. 2010), which was later renamed the  $q$ -statistic (Wang, Zhang, and Fu 2016). Given a target geographical variable, a specific stratification is implemented to stratify all observations into several strata. It is worth noticing that stratification can occur through geographical divisions (i.e.,

subregions), categorical explanatory variables (i.e., classes), or the discretization of a numerical variable (i.e., intervals). To quantify the SSH under such a stratification, the geographical detector  $q$ -statistic is defined as follows (Wang, Zhang, and Fu 2016):

$$q = 1 - \frac{\sum_{h=1}^L N_h \sigma_h^2}{N \sigma^2} \quad (1)$$

where  $N$  is the number of all observations and  $\sigma^2$  denotes the corresponding variance. The observations of the target variable are stratified into  $L$  strata, denoted by  $h = 1, 2, \dots, L$ .  $N_h$  is the number of observations within stratum  $h$  and  $\sigma_h^2$  denotes the corresponding variance. The  $q$ -statistic value ranges from zero to one and indicates the stratification-specific strength of the SSH or the determinant power of an explanatory variable to the target variable. It is a monotonically increasing function of the SSH strength, and it increases as the SSH strength increases from the weakest to the strongest. A simple transformation of the  $q$ -statistic can satisfy a noncentral  $F$  distribution with the first degrees of freedom ( $df$ ) of  $L - 1$  and the second  $df$  of  $N - L$  (Wang, Zhang, and Fu 2016). Subsequently, the statistical significance of the  $q$ -statistic can be tested, with the null and alternative hypotheses as the absence and presence of SSH, respectively.

In general, the SSH measure of a target variable, without prior knowledge or auxiliary variables, depends on both the stratification method and the number of strata. Let us generate a toy data set with sixty-four numbers distributed in  $8 \times 8$  grids,  $[1, 1/2, 1/3, \dots, 1/64]$ , to demonstrate how stratification influences the SSH measure (Figure 1). It can be simply stratified into two strata shown in Figures 1A and 1B. Based on Equation 1, the former receives a  $q$ -statistic value of 0.1391 ( $p = 0.0059$ ), indicating a statistically significant SSH to some extent. In contrast, the latter achieves a  $q$ -statistic value of 0.0331 ( $p = 0.1631$ ), indicating an extremely weak SSH without statistical significance. By overlaying these two stratifications, as shown in Figure 1C, a new stratification with four strata is generated. The SSH measure increases to a  $q$ -statistic value of 0.2030 ( $p = 0.0100$ ). At last, with a consistent number of strata, an alternative in Figure 1D achieves an extremely high  $q$ -statistic value of 0.9557 ( $p = 4.18 \times 10^{-10}$ ), indicating a nearly perfect SSH.



Figure 1. An illustration of the spatial stratified heterogeneity (SSH) measures varying by stratification.

Note that there is a stratum with only one number in Figure 1D and the within-stratum variance is set as zero to calculate the  $q$ -statistic.

The majority of conventional classification or discretization methods (e.g., equal interval, quantile, geometrical interval, standard deviation,  $k$ -means, and Jenks natural breaks) require a preset number of intervals or classes to implement the stratification.

This preset parameter introduces subjective influence when measuring the SSH for a geographical variable. We intend to address this limitation and make the stratification more “natural” (i.e., the number of strata is naturally determined). The optimized stratification is expected to reveal the essence of the data distribution underlying geographical spaces, to



characterize the structural hierarchy and dominant components, and to better capture the SSH for geographical variables.

## Methodology

### Heavy-Tailed Distribution Detection

The imbalance between the head and tail in a heavy-tailed distribution (e.g., a power law distribution, a log-normal distribution, or an exponential distribution) can be expressed as a notion of far more small values than large ones. A heavy-tailed distribution can be easily observed using a rank-size plot of the raw data (Jiang 2013). Alternatively, we can detect whether a data set follows the heavy-tailed distribution by testing it against a hypothesized specific heavy-tailed distribution, such as power laws. Taking the power law test as an example, a variable of interest  $x$  is drawn from a probability distribution as follows (Clauset, Shalizi, and Newman 2009):

$$p(x) \propto x^{-\alpha} \quad (2)$$

where  $\alpha$  denotes the power law exponent (or scaling parameter), which typically lies in the range of  $2 < \alpha < 3$ .

There must be a lower bound to the power law behavior in the majority of empirical phenomena (Clauset, Shalizi, and Newman 2009), and the power law applies only for values above a minimum threshold (denoted as  $x_{\min}$ ). Then, provided  $\alpha > 1$ , the probability density  $p(x)$  can be described as

$$p(x) = \frac{\alpha - 1}{x_{\min}} \left( \frac{x}{x_{\min}} \right)^{-\alpha}. \quad (3)$$

Next, the parameters  $\alpha$  and  $x_{\min}$  of the power law model can be estimated using the robust maximum likelihood method (Clauset, Shalizi, and Newman 2009). Furthermore, a *goodness-of-fit* test can be applied to determine the plausibility of the hypothesized power law distribution. The null and alternative hypotheses are defined as the absence and presence, respectively, of the plausibility of power laws. This test uses the Kolmogorov–Smirnov (KS) statistic to quantify the “distance” between the empirical data (observations) and the synthetic data sets repetitively drawn from the hypothesized power law model. By counting the number of times the synthetic distance is larger than the empirical

distance, a  $p$  value indicator is defined as the proportion of this number to the total number of repetitions (Clauset, Shalizi, and Newman 2009). Therefore, with large  $p$  values the difference between the data and the power law model can be attributed to statistical fluctuations. In other words, the resulting  $p$  value indicates the plausibility of the power law hypothesis. If  $p > 0.1$ , reject the null hypothesis and accept the alternative; that is, one can accept the power law as a plausible hypothesis for the data; otherwise, accept the null hypothesis that the power law is not plausible.

Furthermore, alternative hypotheses (e.g., log-normal/exponential distributions) can be compared with the power law via a likelihood ratio test to determine whether an alternative is favored over the power law or not (more details can be found in Clauset, Shalizi, and Newman 2009).

### Head/tail Breaks and *ht-Index*

Head/tail breaks is a powerful approach first developed to implement classification for data with a heavy-tailed distribution (Jiang 2013). The arithmetic mean is used to partition the data into two parts: the head, which consists of values greater than the mean, and the tail, which consists of values smaller than the mean. Specifically, if the observations (values) of a given variable follow a heavy-tailed distribution, then the mean can divide all the values into a high percentage of small values (the tail) and a low percentage of large ones (the head). In brief, the head/tail breaks method is a recursive function used to iteratively derive the head and tail parts across all levels (scales) and recursively reveal the inherent hierarchical structure of a data set (Jiang 2013).

For a data set of observations of a specific geographical variable that follows a heavy-tailed distribution, the fundamental process of head/tail breaks is shown in Figure 2A. It is worth noticing that a heavy-tailed distribution of the data is a prerequisite for the process of head/tail breaks. Based on the descending sorting of the data, we can generate a head with values above the mean and a tail with values below the mean. While the head part derived through the first partitioning is still heavy-tailed distributed (i.e., it satisfies the notion of far more small values than large ones), it can be further partitioned and the head of the head is derived through the

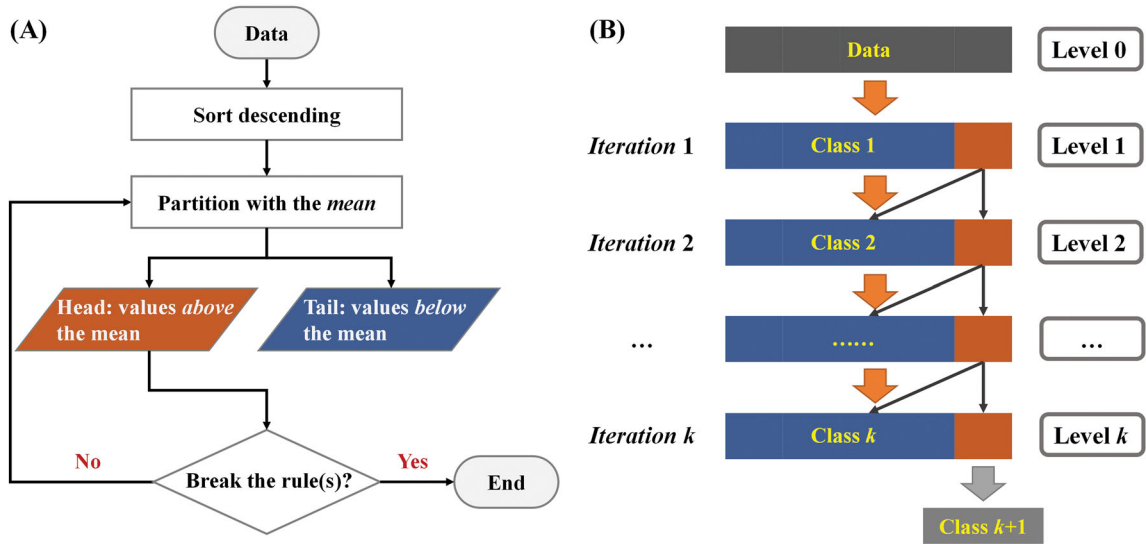


Figure 2. The head/tail breaks-based stratification: (A) the fundamental process; (B) the derived hierarchical structure.

second partitioning. The partitioning continues iteratively for the head until it breaks the rule (i.e., it violates the notion of far more small values than large ones), and the process of head/tail breaks terminates. Note that the notion of far more small values than large ones can be judged by a threshold (e.g., 40 percent) for the head proportion during each iteration (Jiang 2013, 2015c). The judgment rule can be modified and the threshold can be flexible. For instance, in head/tail breaks version 2.0 (Jiang 2019), this threshold can be relaxed for the average of all head proportions instead of each individual head proportion during iterations. In other words, although several head proportions are allowed to exceed a specific threshold (e.g., 40 percent) during iterations, the average of all head proportions is restricted to be smaller than it. It should also be noted that there is no need to examine whether the data strictly obey a power law distribution, a log-normal distribution, or an exponential distribution to implement the head/tail breaks process. Instead, it is recommended to use a straightforward way to determine whether the data are heavy-tailed distributed or not: if the notion of far more small values than large ones recurs at least twice during the head/tail breaks process, the data are heavy-tailed distributed; otherwise, they are not (Jiang 2019).

The head/tail breaks process implements the classification for a data set with a heavy-tailed distribution and simultaneously reveals its inherent hierarchical structure. As shown in Figure 2B, the

head at a certain level consists of the head and tail at the lower level, whereas the head and tail at a certain level constitute the head at the upper level. There are  $k$  iterations during the process, which are naturally determined by the heads satisfying the notion of far more small values than large ones. In other words, the pattern of far more small values than large ones recurs multiple times at different levels (scales) of the structural hierarchy. The hierarchy can be quantified by an index called ht-index, which is calculated as the number of recurring times plus one; that is,  $\text{ht-index} = k + 1$  (Jiang and Yin 2014). The ht-index value also indicates the total number of levels in the structural hierarchy (levels are denoted from 0 to  $k$  with level 0 indicating the raw data). The classification result of the head/tail breaks process consists of the tail parts from level 1 to level  $k$  and the head part at the last level  $k$  (Figure 2B). Note that the number of classes is identical to the ht-index value and is naturally determined, i.e., it is not a subjectively preset parameter. Taking the previous data set with sixty-four numbers as an example (Figure 1), it is heavy-tailed distributed and achieves an inherent structural hierarchy of four levels through the head/tail breaks process ( $\text{ht-index} = 4$ ). This toy data set is classified into four classes ([1/14, 1/15, ..., 1/64], [1/5, 1/6, ..., 1/13], [1/2, 1/3, 1/4], and [1]), which almost perfectly capture the SSH of the data set with an extremely high  $q$ -statistic value of 0.9557 (Figure 1D).

## Head/tail Breaks-Based Stratification

We should first investigate the heavy-tailed distributions of the data, either qualitatively or quantitatively, e.g., using a rank-size plot of the raw data or employing a heavy-tailed distribution detection (e.g., the power law test). In fact, a more straightforward way to be recommended is judging the number of recurring times (more than twice) during the process of head/tail breaks for the data (or  $ht\text{-index} \geq 3$ ; Jiang 2019).

To better capture the SSH of geographical variables with a heavy-tailed distribution, we modified the head/tail breaks process to implement an optimized stratification. As shown in Figure 2A, the partitioning iteration for the optimized stratification continues until it violates any of the following rules: (1) the head proportion during each iteration is smaller than a relaxed threshold of 50 percent (satisfying the notion of far more small values than large ones); (2) the average of all head proportions is smaller than 40 percent (consistent to the criterion in head/tail breaks version 2.0); (3) each head has at least two observations (supporting the  $q$ -statistic calculation for the SSH measure). In brief, the partitioning terminates once the head proportion exceeds 50 percent, or the average of the head proportions is greater than 40 percent, or the head part has only two observations.

It should be noted that this proposed approach investigates the population property of geographical variables and then conducts optimized stratification. It is population-based, as opposed to conventional sample-based stratifications. Besides, the head/tail breaks-based stratification simultaneously generates the stratification result and derives the inherent hierarchical structure for data with a heavy-tailed distribution (Figure 2B). It can indicate an underlying scaling pattern in geographical spaces. Based on the theory and principle of head/tail breaks, no prior knowledge or auxiliary variables are required to implement the stratification and the number of strata (classes) is naturally determined instead of being subjectively preset. The stratification captures the inherent hierarchical structure of geographical features and reveals the spatial hierarchy and scaling property.

In addition, the head/tail breaks process is effective in generating dominant components of a geographical variable with a heavy-tailed distribution, which can be meaningful and interpretable. For instance, when using head/tail breaks for NTL images, the head at the first level of the hierarchy

represents a dominant component of the raw data, and it can be recognized as natural cities indicating urban areas (Jiang 2015b). Similarly, when using head/tail breaks for population-density data, the heads at the first and second levels are expected to represent the primary and secondary components, respectively, indicating settlements in general and settlements in urban areas. In other words, we can also generate component-based stratifications in addition to head/tail breaks-based stratification; for example, a two-strata stratification consisting of a primary component and the rest, or a three-strata stratification consisting of a primary component, a secondary component, and the rest. In this regard, this approach characterizes meaningful and interpretable dominant components of geographical variables across all scales (levels).

The properties and advantages of the proposed approach, mentioned earlier, are summarized based on the theory and principle of head/tail breaks, which in essence surpasses other conventional stratification methods. We will further use several case studies to verify these properties and advantages, as well as the performance of calibrating the SSH of geographical variables with a heavy-tailed distribution.

In summary, the technical process of the methodology is as follows. First, the heavy-tailed distributions of the raw data should be investigated using a rank-size plot, a heavy-tailed distribution detection, or by judging the recurring times or  $ht\text{-index}$ . Next, if the data are heavy-tailed distributed, the head/tail breaks-based stratification can be implemented to derive the inherent hierarchical structure, generate the dominant components, and achieve optimized stratification. Finally, the geographical detector  $q$ -statistic can be used to assess the performance of our approach in calibrating SSH or capturing the SSH of geographical variables that exhibit a heavy-tailed distribution. Note that our approach can generate multiple stratification results including the optimized stratification and the component-based stratifications (two-strata or three-strata). Thus, we can further evaluate the performance of our approach for calibrating SSH even with the least two or three strata.

## Case Studies

We applied the proposed approach to four real-world case studies to verify the properties and advantages, as well as to evaluate the performance for



calibrating SSH. Both social and natural features in geographical spaces were considered in the case studies, and the raw data consisted of vector features or raster images.

### City-Size Hierarchy Based on Urban Populations

We collected the data of China's urban populations in cities from the China Population Census Yearbook 2020. There are a total of 683 administrative cities in mainland China, including 4 municipalities, 292 prefecture-level cities, and 387 county-level cities. We can see from the rank-size plot (Figure 3A) that the urban populations are heavy-tailed distributed and there are far more small-sized cities than large-sized ones. Alternatively, we can conduct heavy-tailed distribution detection. The power law test received an estimated power law exponent of  $\alpha = 2.3080$ , with an estimated lower bound of  $x_{\min} = 74.23 \times 10^4$  over which the power law behavior holds. The  $p$  value for the hypothesized power law distribution was detected as  $p = 0.9310$  (with a minimum goodness-of-fit KS statistic of 0.0317). Thus, the hypothesized power law distribution for the urban population data set is statistically trustworthy (Figure 3B).

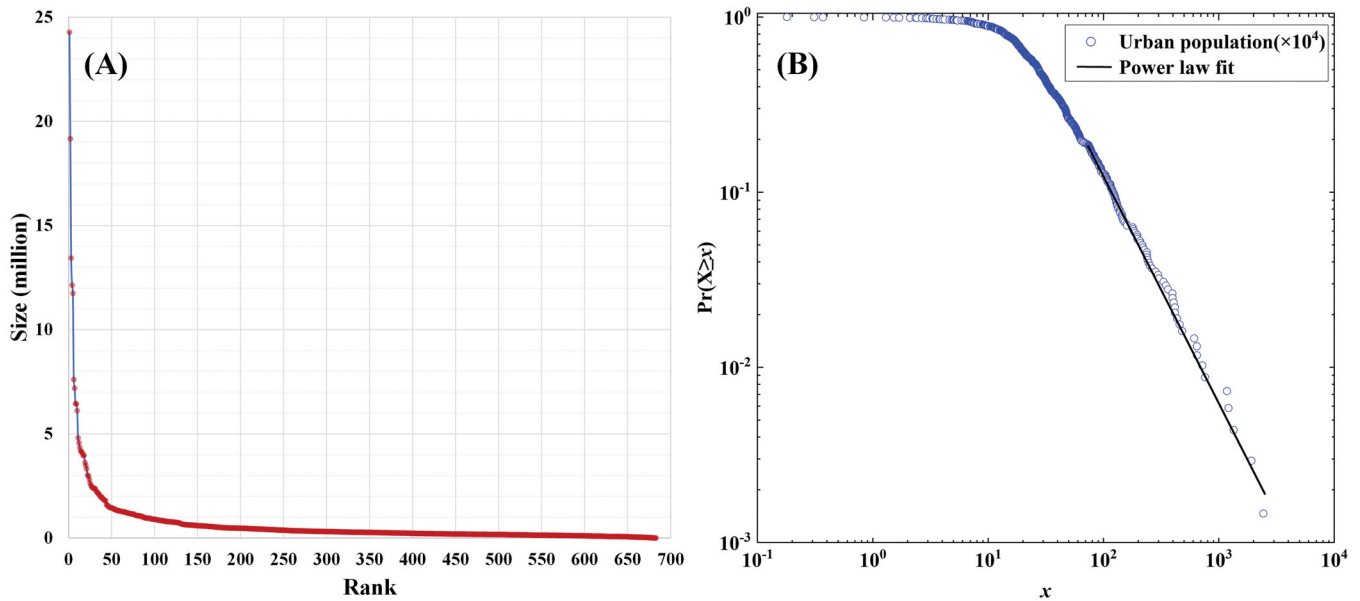
By applying the proposed head/tail breaks-based stratification, we characterized the city-size hierarchy and subsequently measured the SSH of urban

populations. The partitioning recurred five times during the head/tail breaks process (i.e.,  $ht\text{-index} = 6$ ) and the average of the head proportions across all levels is 36 percent. Table 1 shows the five levels (scales) of the city-size hierarchy, excluding level 0, which indicates the raw data. We can see that the pattern of far more small-sized cities than large-sized ones recurs at different levels of the structural hierarchy (scaling law). It demonstrates a hierarchical structure similar to the pyramid structure of global city-size distribution (Fang, Pang, and Liu 2017).

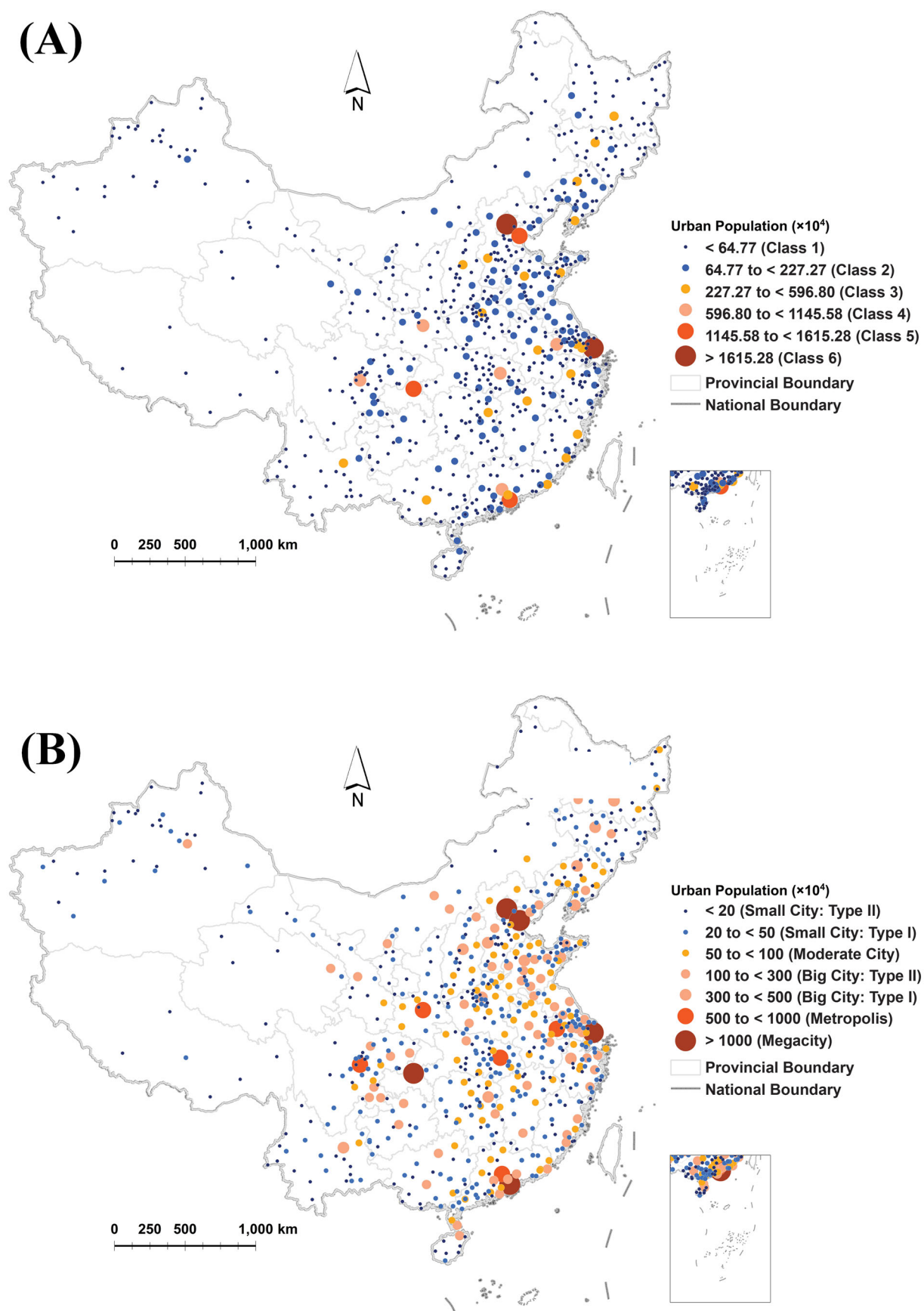
Our approach achieved the stratification result, which included a structural hierarchy of city size consisting of six strata (classes) as shown in Figure 4A. Meanwhile, according to a city-size criterion established by the State Council of China, cities can be categorized into five classes or seven grades (Figure 4B), including megacities (with an urban population over 10 million), metropolises (5–10

**Table 1.** Head/tail breaks statistics for urban populations and the city-size hierarchy

Level	Data	Head	Tail	% head	% tail	M (millions)
1	683	133	550	19	81	0.65
2	133	31	102	23	77	2.27
3	31	10	21	32	68	5.97
4	10	5	5	50	50	11.46
5	5	2	3	40	60	16.15



**Figure 3.** The power law distribution of the data set of urban populations in China's 683 cities: (A) the rank-size plot; (B) the power law fitting in double-logarithm coordinates.



**Figure 4.** China's city-size distributions: (A) the structural hierarchy derived by the head/tail breaks-based stratification; (B) the strata based on a governmental criterion.

million), big cities (Type I, 3–5 million; Type II, 1–3 million), moderate cities (0.5–1 million), and small cities (Type I, 0.2–0.5 million; Type II, less than 0.2 million). Note that our results also demonstrate an inherent hierarchical structure of city sizes in addition to the stratification, which is not found from the classification based on the governmental criterion. Furthermore, we can derive the dominant components from the structural hierarchy of city sizes. Specifically, a primary component consists of 133 cities (the head at the first level or strata or classes 2–6 in Figure 4A), indicating the major cities in China (municipalities and the majority of key prefecture-level cities); a secondary one consists of thirty-one cities (the head at the second level or strata or classes 3–6), indicating municipalities, the majority of provincial capitals, and several key prefecture-level cities (e.g., Shenzhen and Suzhou).

As shown in Table 2, the governmental city-size criterion has already demonstrated a strong SSH for urban populations, as indicated by the  $q$ -statistics of 0.8866 for five classes and 0.9207 for seven grades, respectively. Nevertheless, the presented stratification still achieved a higher value of  $q=0.9681$  and captures a nearly perfect SSH for urban populations. Based on the components of the structural hierarchy, the SSH of urban populations can be characterized with  $q$ -statistics of 0.2403 and 0.5445 by a two-strata stratification and a three-strata stratification, respectively (Table 2). That is to say, our stratification still explains 24.03 percent of the SSH of urban populations even with two strata, whereas the explanatory power can increase to 54.45 percent with just three strata. In brief, this case study verified the aforementioned properties and advantages of the presented approach. It was implemented based on the essential distribution property of urban populations, requiring no prior knowledge or auxiliary variables, and naturally determining the number of strata. Meanwhile,

the stratification revealed the inherent hierarchical structure of city sizes, characterized its dominant components, and better captured the SSH of urban populations in China.

### NTL-Based Urban-Area Hierarchy

The head/tail breaks-based stratification was further applied to the 2020 NTL image of the urban agglomeration in the middle reaches of the Yangtze River (UAMRYR), China. The annual NTL image came from an extended time series of annual composite NTL data (Chen et al. 2021), which were generated by a cross-sensor calibration using the Defense Meteorological Satellite Program Operational Linescan System (DMSP-OLS) NTL data and monthly composite Suomi National Polar-orbiting Partnership Visible Infrared Imaging Radiometer Suite (NPP-VIIRS) NTL data. There were a total of 1,183,232 pixels in the 2020 NTL image of UAMRYR, with values ranging from 0 to 328. The pixel values exhibited a typical heavy-tailed distribution, with far more dark pixels than light ones.

We characterized the urban-area hierarchy from the NTL image using the proposed stratification. The partitioning recurred ten times (i.e., ht-index = 11), and the average of the head proportions was 31 percent. Table 3 illustrates the levels of the urban-area hierarchy in the UAMRYR region. The pattern of far more dark pixels than light ones can be observed across all levels of the structural hierarchy (scaling law). Similarly, the dominant components can be derived from the NTL-based urban-area hierarchy. Specifically, as shown in Figure 5, the head at the first level (115,740 pixels) was identified as a

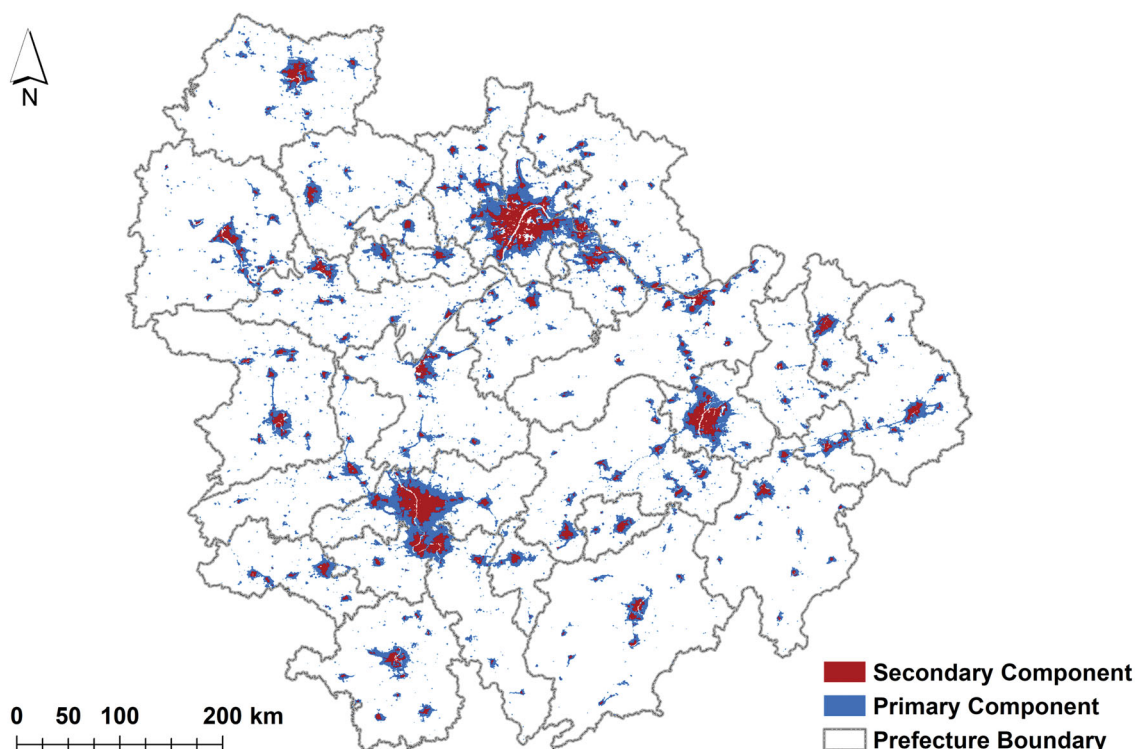
**Table 2.**  $q$ -Statistic calculations of the stratification for the city-size hierarchy

Stratification	Number of strata	$q$ -statistic <sup>a</sup>
Head/tail breaks-based stratification	6	0.9681
Component-based stratifications	2	0.2403
	3	0.5445
Governmental criteria	5	0.8866
	7	0.9207

<sup>a</sup>For all listed  $q$ -statistics,  $p < 0.001$ .

**Table 3.** Head/tail breaks statistics for nighttime lights (NTL) pixels and the urban-area hierarchy

Level	Data	Head	Tail	% head	% tail	M
1	1,183,232	115,740	1,067,492	10	90	0.56
2	115,740	32,409	83,331	28	72	5.69
3	32,409	12,145	20,264	37	63	15.67
4	12,145	4,647	7,498	38	62	25.90
5	4,647	1,726	2,921	37	63	35.89
6	1,726	553	1,173	32	68	46.41
7	553	152	401	27	73	59.84
8	152	40	112	26	74	82.50
9	40	10	30	25	75	125.70
10	10	5	5	50	50	205.70



**Figure 5.** Primary and secondary components of the urban-area hierarchy derived from the nighttime lights (NTL) image in urban agglomeration in the middle reaches of the Yangtze River (UAMRYR) region.

**Table 4.**  $q$ -Statistic calculations of the stratification for the urban-area hierarchy

Stratification	Number of strata	$q$ -statistic <sup>a</sup>
Head/tail breaks-based stratification	11	0.9664
Component-based stratifications	2	0.2863
	3	0.6652

<sup>a</sup>For all listed  $q$ -statistics,  $p < 0.001$ .

primary component, which were termed natural cities (Jiang and Yin 2014; Jiang 2015b), indicating urban areas in general; the head at the second level (32,409 pixels) was identified as a secondary component, indicating core urban areas.

It should be noted that the stratification was implemented based on the distribution property of pixels in the NTL image, without prior knowledge or auxiliary information. The number of strata was naturally determined as eleven. Our approach achieved an extremely high value of  $q=0.9664$  (Table 4), which captures a nearly perfect SSH for urban areas. Besides, the dominant components derived from the urban-area hierarchy characterized

a satisfactory SSH as well. A two-strata stratification consisting of a primary component and the rest explained 28.63 percent of the SSH of urban areas in the UAMRYR region, whereas the explanatory power increased to 66.52 percent with a three-strata stratification consisting of a primary component, a secondary component, and the rest (Table 4). Thus, we concluded that the head/tail breaks-based stratification better captures the SSH of urban areas in the UAMRYR region.

### NDVI-Based Greenness Hierarchy

We subsequently used the optimized stratification for the NDVI to derive the hierarchy of greenness in Jiangxi Province, China. The image of NDVI for June 2020, with a spatial resolution of 1 km, was obtained from the Resource and Environment Science and Data Center at the Chinese Academy of Sciences. It had a total of 166,357 pixels with values ranging from 0 to 0.92 (negative values were removed from the raw image). In general, pixels in the NDVI with values approaching zero indicate water bodies, urban areas, or bare soil, whereas low-value pixels indicate barren areas with rocks, sand,



**Table 5.** Head/tail breaks statistics for Normalized Difference Vegetation Index (NDVI) pixels and the greenness hierarchy

Level	Data	Head	Tail	% head	% tail	M
1	166,357	69,086	97,271	42	58	0.70
2	69,086	24,513	44,573	35	65	0.57
3	24,513	8,940	15,573	36	64	0.43
4	8,940	3,302	5,638	37	63	0.28

or others. Moderate-value greenness pixels are expected to represent grasslands or others, whereas high-value greenness pixels indicate forests. Note that there are far more greenness pixels (high values) than nongreenness ones (low values) in the NDVI of Jiangxi province. The NDVI pixels are still heavy-tailed distributed when the concepts of the head and tail are exchanged. More specifically, during the head/tail breaks process in this case study, the pixels with values below the mean were expected as the head, whereas those with values above the mean were considered as the tail.

The partitioning recurred four times during the head/tail breaks process (i.e., ht-index = 5) and the average of the head proportions was 38 percent. The greenness hierarchy derived from the NDVI was revealed to have a total of five levels and there were far more greenness pixels in the tail than nongreenness pixels in the head across all levels (Table 5). The dominant components of the current greenness hierarchy should be characterized starting from the tail parts, as a result of the modification of the head/tail breaks process. More specifically, a primary component consisted of the tails at the first and second levels of the NDVI-based greenness hierarchy (141,844 pixels), whereas a secondary component was expected to be the tail at the first level (97,271 pixels). As shown in Figure 6, the primary component (greenness pixels) indicates the greenness areas in general (e.g., forests and grasslands), whereas the secondary component (dark greenness pixels) indicates forests to a great extent.

The aforementioned properties and advantages of the proposed stratification were verified in this case study as well. We can also conclude that it better captured the SSH of greenness areas in Jiangxi Province. The stratification introduced a highly satisfactory SSH with a  $q$ -statistic value of 0.8803 (Table 6). More important, the component-based stratification also achieved unexpectedly good performance in capturing the SSH of greenness areas. It can explain 57.76 percent of the SSH of greenness

areas even with the least two strata and the explanatory power increases to a very high value of 77.49 percent with just three strata (Table 6).

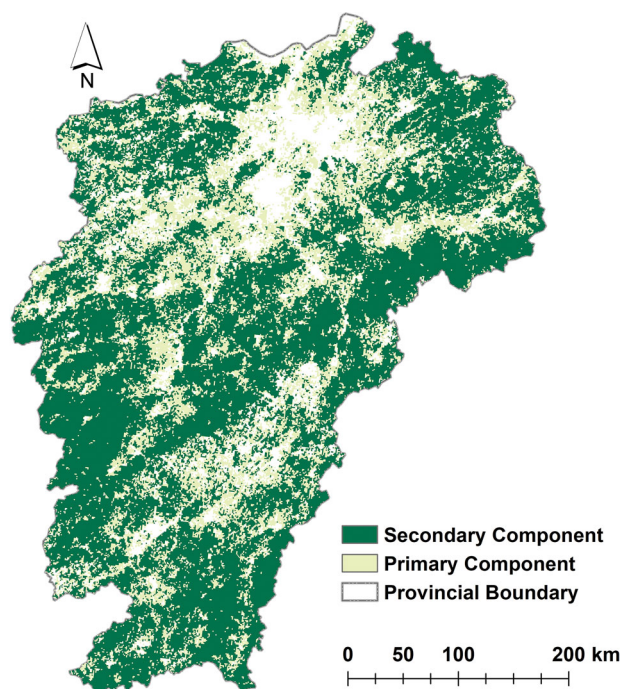
### Habitat Hierarchy of Bamboo Forests

At last, the head/tail breaks-based stratification was applied to the bamboo-forest density image to characterize the habitat hierarchy of bamboo forests. We collected the image of China's bamboo-forest density in 2020, with a spatial resolution of approximately 1 km, from the Resource and Environment Science and Data Center. There were a total of 9,222,895 pixels with values ranging from zero to one, and the pixel values exhibited a typical heavy-tailed distribution, with far more low-density pixels than high-density ones.

The partitioning recurred nine times (i.e., ht-index = 10) and the average of the head proportions was 38 percent. The habitat hierarchy of bamboo forests was identified to have a total of ten levels, shown in Table 7. The scaling law can also be found across all levels of the hierarchy, and at each level, there were far more low-density pixels than high-density ones. We further derived the dominant components of the bamboo-forest habitat hierarchy. Specifically, as shown in Figure 7, the head at the first level was expected to be a primary component (664,774 pixels), indicating suitable bamboo-forest habitats in general; the head at the second level was considered a secondary one (168,246 pixels), indicating core bamboo-forest habitats.

The presented stratification captured a nearly perfect SSH for bamboo-forest habitats with an extremely high value of  $q = 0.9643$  (Table 8). Furthermore, the dominant components derived from the bamboo-forest habitat hierarchy also demonstrated a satisfactory performance in capturing the SSH for bamboo-forest habitats. A two-strata stratification explained 28.17 percent of the SSH for bamboo-forest habitats, whereas the explanatory power increased to a very high value of 69.64 percent with just three strata (Table 8). The optimized stratification better captures the SSH of bamboo-forest habitats in China, also revealing the inherent hierarchical structure of bamboo-forest habitats and characterizing its meaningful dominant components. Besides, it was a population-based stratification based on the essential distribution property of bamboo forests and the number of strata was naturally determined without any prior knowledge or auxiliary information.





**Figure 6.** Primary and secondary components of the greenness hierarchy derived from the Normalized Difference Vegetation Index (NDVI) in Jiangxi Province.

**Table 6.**  $q$ -Statistic calculations of the stratification for the greenness hierarchy

Stratification	Number of strata	$q$ -statistic <sup>a</sup>
Head/tail breaks-based stratification	5	0.8803
Component-based stratifications	2	0.5776
	3	0.7749

<sup>a</sup>For all listed  $q$ -statistics,  $p < 0.001$ .

## Discussion

This work introduces a novel head/tail breaks-based stratification to address the primary key issue (i.e., stratification) in the geographical detector method. Different from existing stratifications, the proposed stratification captures the essence of the data distributions in various geographical phenomena and reveals the underlying scaling pattern in geographical spaces. In essence, it is a population-based stratification, whereas previous stratifications are mainly sample-based. Meanwhile, our approach addresses the limitation of requiring a predetermined number of strata that is common in most existing stratifications. It also significantly reduces subjective influence and makes the stratification more

**Table 7.** Head/tail breaks statistics for bamboo-forest density pixels and the bamboo-habitat hierarchy

Level	Data	Head	Tail	% head	% tail	M
1	9,222,895	664,774	8,558,121	7	93	0.01
2	664,774	168,246	496,528	25	75	0.11
3	168,246	63,418	104,828	38	62	0.33
4	63,418	27,539	35,879	43	57	0.56
5	27,539	12,230	15,309	44	56	0.72
6	12,230	5,474	6,756	45	55	0.83
7	5,474	2,509	2,965	46	54	0.91
8	2,509	1,167	1,342	47	53	0.95
9	1,167	578	589	50	50	0.98

“natural.” Furthermore, our approach requires no prior knowledge or auxiliary variables and is straightforward to implement. We conducted four case studies of different social and natural features in geographical spaces to illustrate the properties of our approach.

It is worth noting that the presented approach can implement stratification for a geographical variable with a heavy-tailed distribution, and simultaneously reveal the inherent hierarchical structure, which provides the potential to make the stratification result meaningful and interpretable. The scaling law (the pattern of far more small values than large ones recurring across all levels of the structural hierarchy) enables us to identify meaningful dominant components from the hierarchy; for example, the primary component of the city-size hierarchy indicates the major cities and the primary component of the NTL-based urban-area hierarchy indicates natural cities of urban areas (Jiang and Yin 2014; Jiang 2015b). The proposed stratification introduces more reasonable information compared to other conventional stratifications.

In addition, four case studies have verified that our approach always introduces large  $q$ -statistic values for geographical variables with a heavy-tailed distribution; for example, the proposed stratification achieved an average value of  $q = 0.9448$ , indicating a nearly perfect SSH. In particular, the stratification based on the dominant components of the structural hierarchy only has two or three strata but still demonstrates very satisfactory performance; for example, the two-strata and three-strata stratifications achieved average  $q$ -statistic values of 0.3465 and 0.6703, respectively, in this study. We believe that a meaningful stratification with a high  $q$ -statistic value is preferable to a meaningless one with a maximum

$q$ -statistic value (approaching 1). Maximizing the  $q$ -statistic should not be regarded as the sole criterion for optimizing stratification. Our approach implements meaningful and interpretable stratification, indicating a nearly perfect SSH. It also achieves satisfactory  $q$ -statistic values, even with two or three strata. That is the reason why we conclude that the head/tail breaks-based stratification better captures the SSH of geographical variables with a heavy-tailed distribution.

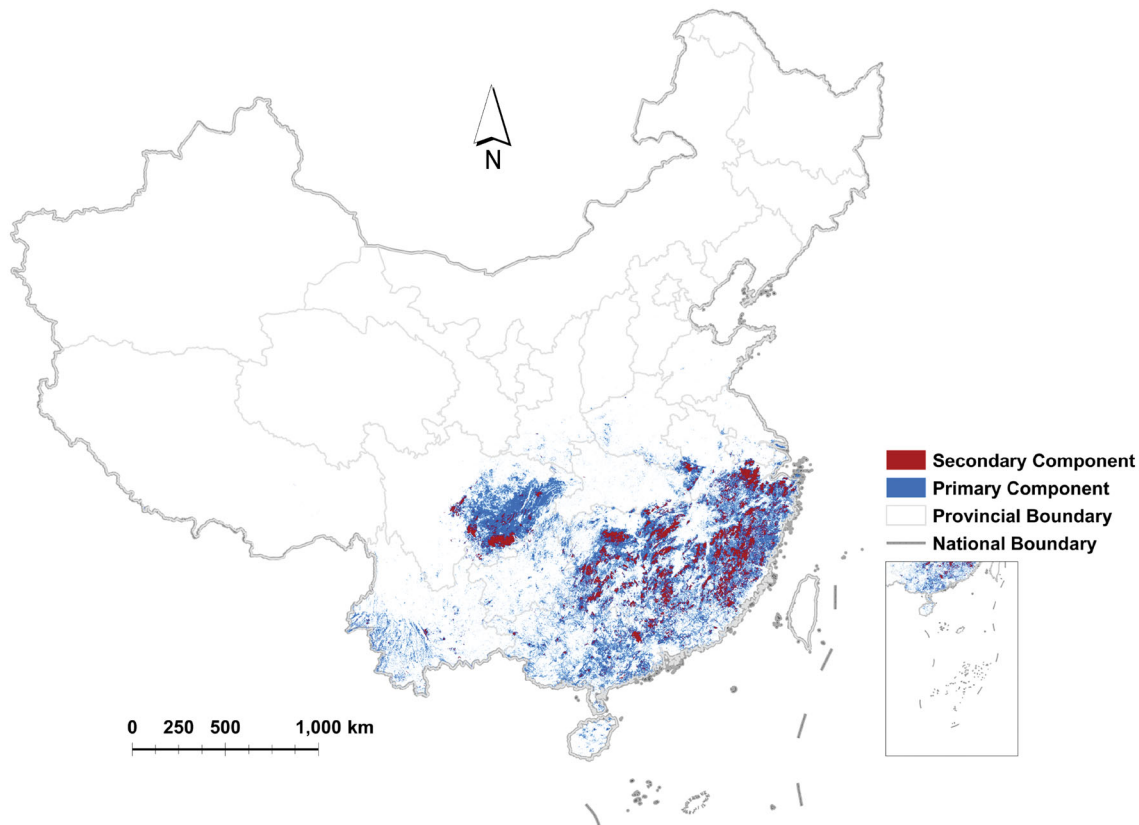
We intend to further compare the  $q$ -statistic in geographical detector and the ht-index in head/tail breaks, as well as several specific terms associated with them. Table 9 lists a detailed comparison between  $q$ -statistic and ht-index. We can reveal both spatial dependence and spatial variability (as two spatial properties) from the principles of  $q$ -statistic and ht-index. As mentioned before, spatial autocorrelation described in the first law of geography refers to the property of spatial dependence in geographical phenomena (Tobler 1970). It can be observed across the structural hierarchy of a geographical variable with a heavy-tailed distribution;

that is, there are more or less similar things at each level of the hierarchy. Meanwhile, there are far more small values than large ones across various levels of the structural hierarchy (scaling law). Tobler's law and scaling law are expected to be two governing laws underlying the structural hierarchy of a geographical variable with a heavy-tailed distribution, which complement each other and recur at different levels of the hierarchy (Jiang and de Rijke 2021, 2022; Jiang and Huang 2021). Furthermore, spatial heterogeneity, mentioned in the controversial second law of geography, refers to the property of spatial variation or variability (Goodchild 2004). It

**Table 8.**  $q$ -Statistic calculations of the stratification for the bamboo-habitat hierarchy

Stratification	Number of strata	$q$ -statistic <sup>a</sup>
Head/tail breaks-based stratification	10	0.9643
Component-based stratifications	2	0.2817
	3	0.6964

<sup>a</sup>For all listed  $q$ -statistics,  $p < 0.001$ .



**Figure 7.** Primary and secondary components of the habitat hierarchy derived from the bamboo-forest density image.

**Table 9.** A detailed comparison between  $q$ -statistic and ht-index

	Wang's $q$ (Wang et al. 2010, 2024; Wang, Zhang, and Fu 2016)	Jiang's ht (Jiang 2013; Jiang and Yin 2014)
Type	A statistic	An index
Alternative	Geographical detector	Scaling law, scale-free or hierarchy
Principle	Within-strata variance less than between-strata variance	The recursive pattern of far more small values than large ones
Property	SSH	Spatial heterogeneity (or scaling law)
Spatial dependence	Observations are more or less homogeneous (similar or related) within each stratum	There are more or less similar observations at each level of the hierarchy
Spatial variability	Observations are heterogeneous from stratum to stratum	Observations are heterogeneous from level to level in the structural hierarchy
Target	Numerical $y$	Numerical $y$
Constraint	Geographical divisions or stratifications determined by categorical or numerical $x$	Judgment rule(s) to hold the scaling law across different levels of the hierarchy
Calculation	The $q$ -statistic equals one minus the ratio of the within-strata sums of squares of deviations (SSW) to the total sum of squares of deviations (SST)	The ht-index equals one plus the recurring times of far more small values than large ones at different levels (scales)
Measure	$0 \leq q \leq 1$ and $q = 0$ indicates no SSH while $q = 1$ indicates a perfect SSH	$ht \geq 3$ and larger ht values indicate stronger spatial heterogeneity
Implication	$q(y x)$ indicates the spatial/nonspatial heterogeneity between strata determined by $x$ , each of which is composed of a few $y$ observations	$ht(y)$ indicates the fractal complexity, scaling hierarchy, or spatial/nonspatial heterogeneity of the $y$ observations
Example (generalized)	Q: Is individual wealth heterogeneous between strata? A: Yes. Wealth observations might be homogeneous within a stratum but might be not between strata (e.g., occupations or educations).	Q: Is individual wealth heterogeneous across scales? A: Yes. Wealth observations satisfy a heavy-tailed distribution and the pattern of far more poor ones than rich ones recurs across scales.
Example (geographical)	Q: Is NDVI heterogeneous between climate zones? A: Yes. NDVI observations are similar or related to some extent within a climate zone but the within-zones variance is less than the between-zones variance. The $q$ -statistic is large and significant.	Q: Is terrain surface heterogeneous across scales? A: Yes. The pattern of far more low locations than high ones recurs across scales and the ht-index is extremely high (natural phenomena are usually more complicated than human-made phenomena).
Application	SSH measure, attribution analyses, determinant power, interactive detection, representativeness, nonlinear association, stratified modeling, etc.	Spatial heterogeneity measure, natural cities, fractal structure, map generalization, cognitive mapping, structural beauty, living structure, degree of order, etc.

Note: SSH = spatial stratified heterogeneity; NDVI = Normalized Difference Vegetation Index.

should be formulated as a scaling law of geography (Jiang 2015a; Jiang and Brandt 2016). More levels of the structural hierarchy introduce stronger spatial heterogeneity for geographical variables. In this regard, ht-index serves as an index to express hierarchical levels for heterogeneous scales, and provides a practicable measure to quantify the spatial

heterogeneity for geographical variables (Jiang and Yin 2014). Specifically, spatial heterogeneity exists when ht-index  $\geq 3$  (head/tail breaks induced value), and larger ht-index values indicate stronger spatial heterogeneity (Jiang 2019). On the other hand, spatial dependence and spatial variability can also be observed in the stratification of a geographical

variable with a significant SSH; that is, observations are more or less homogeneous (similar or related) within each stratum but are heterogeneous between strata. Note that spatial heterogeneity can manifest in both the spatial stratification of heterogeneity significance and across the inherent structural hierarchy. Meanwhile,  $q$ -statistic is a powerful approach for measuring the SSH of geographical variables, whereas ht-index is conducive to quantifying the scaling structure and spatial heterogeneity or scaling law. In other words,  $q$ -statistic and ht-index are quantitative measures of SSH and spatial heterogeneity, respectively. They complement each other and together provide a comprehensive characterization of spatial heterogeneity.

This work first integrates the principles of  $q$ -statistic and ht-index, which can be recognized as a theoretical advancement in the field of spatial statistics, and introduces an optimized stratification for geographical variables with a heavy-tailed distribution. We recommend using head/tail breaks to investigate heavy-tailed distributions for geographical variables, applying our approach to stratify the heavy-tailed distributed data, and further implementing SSH-related studies, such as attribution analyses (Hu, Zou, et al. 2022), stratified modeling (Luo et al. 2022; Yang et al. 2022), and examination of the modifiable areal unit problem effect (F. Gao et al. 2021; Hu, Fu, et al. 2023). Note that  $q$ -statistic and ht-index are not limited in geographical phenomena when measuring the SSH and spatial heterogeneity (see the examples in Table 9), and thus, our approach is versatile and flexible for the stratification of both geographical and nongeographical variables.

On the other hand, this article provides novelty in terms of the way of thinking for advocating spatial heterogeneity or scaling law. In other words, we can see how  $q$ -statistic and ht-index reciprocally support each other in promoting spatial heterogeneity, under the Paretian way of thinking rather than the Gaussian way of thinking. In fact, conventional literature always states that spatial homogeneity is the primary effect, whereas spatial heterogeneity is the secondary effect. This is the typical Gaussian way of thinking. We would like to advocate the opposite, however; that is, under the Paretian way of thinking, spatial heterogeneity (scaling law) is ubiquitous as the dominant effect rather than spatial homogeneity (Tobler's law) being the primary effect.

Several further analyses have been identified for this study. The SSH  $q$ -statistic can identify the interactions between two or more explanatory variables to a target variable (Wang et al. 2010). It can examine their interactions and quantify the interactive power of determinants of multiple variables. The presented approach should be further improved to address the stratification of multivariate geographical features, which is a theoretical optimization task for future work. In addition, we should consider how to integrate our approach into the framework of the spatial statistic trinity (Wang, Gao, and Stein 2020). Besides, further empirical analyses supported by our approach are required. In fact, one of our ongoing works is the development of an optimized method for spatial stratified interpolation based on the presented approach. Furthermore, the scale effect of this approach requires investigation to clarify its universality at different spatial scales. Finally, several theoretical issues need to be addressed, such as the robustness of the presented approach to the outliers of the heavy-tailed distributed data, and the nonlinear association between the ht-index measures identified by head/tail breaks and the  $q$ -statistic values introduced by the proposed stratification.

## Conclusions

This work introduces a novel approach to optimal stratification for geographical variables with a heavy-tailed distribution, which is a critical parameter for the SSH  $q$ -statistic. The properties and advantages of our approach can be summarized as follows: (1) it is a population-based optimized stratification that indicates an underlying scaling pattern in geographical spaces; (2) it is straightforward to implement and requires no prior knowledge or auxiliary variables; (3) the number of strata (classes) is naturally determined instead of being subjectively preset; (4) the stratification captures the inherent hierarchical structure of geographical features and reveals the spatial hierarchy and scaling property; (5) it characterizes meaningful and interpretable dominant components of geographical features across all scales (levels); and (6) the stratification usually introduces large  $q$ -statistic values and better captures the SSH of geographical variables with a heavy-tailed distribution. We verified these properties and advantages and evaluated the performance for calibrating SSH using four case studies. This approach can implement reasonable stratification and better capture the SSH



of both social and natural features in geographical spaces. It is versatile and straightforward to be applied in other studies for the stratification of either geographical or nongeographical variables.

This article first integrates the principles of  $q$ -statistic and ht-index, which can be recognized as a theoretical advancement for spatial statistics. In this regard, this study provides fundamental improvements of advancing SSH-related studies (e.g., calibrating SSH and supporting stratified modeling). Furthermore, this article advocates novelty in terms of advocating spatial heterogeneity or scaling law. This new insight might have a broad impact on the way of thinking and enhance the changes of understanding geographical phenomena.

## Disclosure Statement

No potential conflict of interest was reported by the authors.

## Funding

This study was supported by the National Natural Science Foundation of China (Grant Nos. 42061075, 42071375), the Startup Fund of The Hong Kong University of Science and Technology (Guangzhou), and the City-University Joint Fund of the Science and Technology Project of Guangzhou.

## ORCID

Bisong Hu  <http://orcid.org/0000-0003-3875-8792>

Jinfeng Wang  <http://orcid.org/0000-0002-6687-9420>

Bin Jiang  <http://orcid.org/0000-0002-2337-2486>

## Data Availability Statement

The data and codes that support the proposed approach and case studies of this article are available at figshare. com at the identifier <https://doi.org/10.6084/m9.figshare.21786014>. The file package includes (1) the codes in R to support the head/tail breaks-based stratification and the SSH calculation in the case studies, and the codes in MATLAB to support the power law detection for urban populations; and (2) the raw data used in the case studies. Alternatively, the geographical detector and head/tail breaks software can be found at [http://](http://www.geodetector.cn/)

[www.geodetector.cn/](http://www.geodetector.cn/) and [https://en.wikipedia.org/wiki/head/tail\\_breaks](https://en.wikipedia.org/wiki/head/tail_breaks), respectively. The codes to support the power law detection and the detections of other heavy-tailed distributions can be found at <https://aaronclauset.github.io/>. The data used in the case studies are publicly available. The data set of China's urban populations is owned by China Population Census Yearbook 2020 (<http://www.stats.gov.cn/tjsj/pcsj/rkpc/7rp/zk/indexce.htm>). The annual NTL image comes from an extended time series of global NPP-VIIRS-like NTL data (<https://doi.org/10.7910/dvn/ygivcd>). The monthly image of NDVI and annual bamboo-forest density image are owned by the Resource and Environment Science and Data Center, Chinese Academy of Sciences (<https://www.resdc.cn/>).

## References

- Anselin, L. 1995. Local indicators of spatial association—LISA. *Geographical Analysis* 27 (2):93–115. doi: [10.1111/j.1538-4632.1995.tb00338.x](https://doi.org/10.1111/j.1538-4632.1995.tb00338.x).
- Cao, F., Y. Ge, and J. Wang. 2013. Optimal discretization for geographical detectors-based risk assessment. *GIScience and Remote Sensing* 50:78–92. doi: [10.1080/15481603.2013.778562](https://doi.org/10.1080/15481603.2013.778562).
- Cao, F., Y. Ge, and J. Wang. 2014. Spatial data discretization methods for geocomputation. *International Journal of Applied Earth Observation and Geoinformation* 26:432–40. doi: [10.1016/j.jag.2013.09.005](https://doi.org/10.1016/j.jag.2013.09.005).
- Chen, Z., B. Yu, C. Yang, Y. Zhou, S. Yao, X. Qian, C. Wang, B. Wu, and J. Wu. 2021. An extended time series (2000–2018) of global NPP-VIIRS-like nighttime light data from a cross-sensor calibration. *Earth System Science Data* 13 (3):889–906. doi: [10.5194/essd-13-889-2021](https://doi.org/10.5194/essd-13-889-2021).
- Clauset, A., C. R. Shalizi, and M. E. J. Newman. 2009. Power-law distributions in empirical data. *SIAM Review* 51 (4):661–703. doi: [10.1137/070710111](https://doi.org/10.1137/070710111).
- Fang, C., B. Pang, and H. Liu. 2017. Global city size hierarchy: Spatial patterns, regional features, and implications for China. *Habitat International* 66:149–62. doi: [10.1016/j.habitatint.2017.06.002](https://doi.org/10.1016/j.habitatint.2017.06.002).
- Fisher, W. D. 1958. On grouping for maximum homogeneity. *Journal of the American Statistical Association* 53 (284):789–98. doi: [10.1080/01621459.1958.10501479](https://doi.org/10.1080/01621459.1958.10501479).
- Gao, B.-B., J.-F. Wang, H.-M. Fan, K. Xu, M.-G. Hu, and Z.-Y. Chen. 2015. A stratified optimization method for a multivariate marine environmental monitoring network in the Yangtze River estuary and its adjacent sea. *International Journal of Geographical Information Science* 29 (8):1332–49. doi: [10.1080/13658816.2015.1024254](https://doi.org/10.1080/13658816.2015.1024254).
- Gao, F., S. Li, Z. Tan, Z. Wu, X. Zhang, G. Huang, and Z. Huang. 2021. Understanding the modifiable areal unit problem in dockless bike sharing usage and exploring the interactive effects of built environment



- factors. *International Journal of Geographical Information Science* 35 (9):1905–25. doi: [10.1080/13658816.2020.1863410](https://doi.org/10.1080/13658816.2020.1863410).
- Goodchild, M. F. 2004. The validity and usefulness of laws in geographic information science and geography. *Annals of the Association of American Geographers* 94 (2):300–03. doi: [10.1111/j.1467-8306.2004.09402008.x](https://doi.org/10.1111/j.1467-8306.2004.09402008.x).
- Guo, J., J. Wang, C. Xu, and Y. Song. 2022. Modeling of spatial stratified heterogeneity. *GIScience & Remote Sensing* 59 (1):1660–77. doi: [10.1080/15481603.2022.2126375](https://doi.org/10.1080/15481603.2022.2126375).
- Hu, B., S. Fu, J. Luo, H. Lin, Q. Yin, V. Tao, B. Jiang, L. Zuo, and Y. Meng. 2023. Geographical detector-based assessment of multi-level explanatory powers of determinants on China's medical-service resumption during the COVID-19 epidemic. *Environment and Planning B: Urban Analytics and City Science* 50 (7):1739–58. doi: [10.1177/23998083221143122](https://doi.org/10.1177/23998083221143122).
- Hu, B., P. Ning, Y. Li, C. Xu, G. Christakos, and J. Wang. 2021. Space-time disease mapping by combining Bayesian maximum entropy and Kalman filter: The BME-Kalman approach. *International Journal of Geographical Information Science* 35 (3):466–89. doi: [10.1080/13658816.2020.1795177](https://doi.org/10.1080/13658816.2020.1795177).
- Hu, B., P. Ning, J. Qiu, V. Tao, A. T. Devlin, H. Chen, J. Wang, and H. Lin. 2021. Modeling the complete spatiotemporal spread of the COVID-19 epidemic in mainland China. *International Journal of Infectious Diseases* 110:247–57. doi: [10.1016/j.ijid.2021.04.021](https://doi.org/10.1016/j.ijid.2021.04.021).
- Hu, B., J. Qiu, H. Chen, V. Tao, J. Wang, and H. Lin. 2020. First, second and potential third generation spreads of the COVID-19 epidemic in mainland China: An early exploratory study incorporating location-based service data of mobile devices. *International Journal of Infectious Diseases* 96:489–95. doi: [10.1016/j.ijid.2020.05.048](https://doi.org/10.1016/j.ijid.2020.05.048).
- Hu, B., L. Zou, S. Qi, Q. Yin, J. Luo, L. Zuo, and Y. Meng. 2022. Evaluating the vulnerability of Siberian crane habitats and the influences of water level intervals in Poyang Lake Wetland, China. *Remote Sensing* 14 (12):2774. doi: [10.3390/rs14122774](https://doi.org/10.3390/rs14122774).
- Jenks, G. F., and F. C. Caspall. 1971. Error on choroplethic maps: Definition, measurement, reduction. *Annals of the Association of American Geographers* 61 (2):217–44. doi: [10.1111/j.1467-8306.1971.tb00779.x](https://doi.org/10.1111/j.1467-8306.1971.tb00779.x).
- Jiang, B. 2013. Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution. *The Professional Geographer* 65 (3):482–94. doi: [10.1080/00330124.2012.700499](https://doi.org/10.1080/00330124.2012.700499).
- Jiang, B. 2015a. Geospatial analysis requires a different way of thinking: The problem of spatial heterogeneity. *GeoJournal* 80 (1):1–13. doi: [10.1007/s10708-014-9537-y](https://doi.org/10.1007/s10708-014-9537-y).
- Jiang, B. 2015b. Head/tail breaks for visualization of city structure and dynamics. *Cities* 43:69–77. doi: [10.1016/j.cities.2014.11.013](https://doi.org/10.1016/j.cities.2014.11.013).
- Jiang, B. 2015c. Wholeness as a hierarchical graph to capture the nature of space. *International Journal of Geographical Information Science* 29 (9):1632–48. doi: [10.1080/13658816.2015.1038542](https://doi.org/10.1080/13658816.2015.1038542).
- Jiang, B. 2019. A recursive definition of goodness of space for bridging the concepts of space and place for sustainability. *Sustainability* 11 (15):4091. doi: [10.3390/su11154091](https://doi.org/10.3390/su11154091).
- Jiang, B., and S. A. Brandt. 2016. A fractal perspective on scale in geography. *ISPRS International Journal of Geo-Information* 5 (6):95. doi: [10.3390/ijgi5060095](https://doi.org/10.3390/ijgi5060095).
- Jiang, B., and C. de Rijcke. 2021. Structural beauty: A structure-based computational approach to quantifying the beauty of an image. *Journal of Imaging* 7 (5):78. doi: [10.3390/jimaging7050078](https://doi.org/10.3390/jimaging7050078).
- Jiang, B., and C. de Rijcke. 2022. Representing geographic space as a hierarchy of recursively defined subspaces for computing the degree of order. *Computers, Environment and Urban Systems* 92:101750. doi: [10.1016/j.compenvurbsys.2021.101750](https://doi.org/10.1016/j.compenvurbsys.2021.101750).
- Jiang, B., and J.-T. Huang. 2021. A new approach to detecting and designing living structure of urban environments. *Computers, Environment and Urban Systems* 88:101646. doi: [10.1016/j.compenvurbsys.2021.101646](https://doi.org/10.1016/j.compenvurbsys.2021.101646).
- Jiang, B., and J. Yin. 2014. Ht-index for quantifying the fractal or scaling structure of geographic features. *Annals of the Association of American Geographers* 104 (3):530–40. doi: [10.1080/00045608.2013.834239](https://doi.org/10.1080/00045608.2013.834239).
- Li, J., C. Xu, M. Chen, and W. Sun. 2019. Balanced development: Nature environment and economic and social power in China. *Journal of Cleaner Production* 210:181–89. doi: [10.1016/j.jclepro.2018.10.293](https://doi.org/10.1016/j.jclepro.2018.10.293).
- Li, L., J. Wang, Z. Cao, and E. Zhong. 2008. An information-fusion method to identify pattern of spatial heterogeneity for improving the accuracy of estimation. *Stochastic Environmental Research and Risk Assessment* 22 (6):689–704. doi: [10.1007/s00477-007-0179-1](https://doi.org/10.1007/s00477-007-0179-1).
- Liao, Y., D. Li, N. Zhang, C. Xia, R. Zheng, H. Zeng, S. Zhang, J. Wang, and W. Chen. 2019. Application of sandwich spatial estimation method in cancer mapping: A case study for breast cancer mortality in the Chinese mainland, 2005. *Statistical Methods in Medical Research* 28 (12):3609–26. doi: [10.1177/0962280218811344](https://doi.org/10.1177/0962280218811344).
- Liu, T., J. Wang, C. Xu, J. Ma, H. Zhang, and C. Xu. 2018. Sandwich mapping of rodent density in Jilin Province, China. *Journal of Geographical Sciences* 28 (4):445–58. doi: [10.1007/s11442-018-1483-z](https://doi.org/10.1007/s11442-018-1483-z).
- Luo, P., Y. Song, D. Zhu, J. Cheng, and L. Meng. 2022. A generalized heterogeneity model for spatial interpolation. *International Journal of Geographical Information Science* 37 (3):1–26. doi: [10.1080/13658816.2022.2147530](https://doi.org/10.1080/13658816.2022.2147530).
- Matheron, G. 1963. Principles of geostatistics. *Economic Geology* 58 (8):1246–66. doi: [10.2113/gsecongeo.58.8.1246](https://doi.org/10.2113/gsecongeo.58.8.1246).
- Moran, P. A. P. 1950. Notes on continuous stochastic phenomena. *Biometrika* 37 (1–2):17–23. doi: [10.1093/biomet/37.1-2.17](https://doi.org/10.1093/biomet/37.1-2.17).
- Song, Y., J. Wang, Y. Ge, and C. Xu. 2020. An optimal parameters-based geographical detector model enhances geographic characteristics of explanatory variables for spatial heterogeneity analysis: Cases with different types of spatial data. *GIScience & Remote Sensing* 57 (5):593–610. doi: [10.1080/15481603.2020.1760434](https://doi.org/10.1080/15481603.2020.1760434).

- Tobler, W. R. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46:234. doi: [10.2307/143141](https://doi.org/10.2307/143141).
- Wang, J., B. Gao, and A. Stein. 2020. The spatial statistic trinity: A generic framework for spatial sampling and inference. *Environmental Modelling & Software* 134:104835. doi: [10.1016/j.envsoft.2020.104835](https://doi.org/10.1016/j.envsoft.2020.104835).
- Wang, J.-F., R. Haining, T.-J. Liu, L.-F. Li, and C.-S. Jiang. 2013. Sandwich estimation for multi-unit reporting on a stratified heterogeneous surface. *Environment and Planning A: Economy and Space* 45 (10):2515–34. doi: [10.1068/a44710](https://doi.org/10.1068/a44710).
- Wang, J., R. Haining, T. Zhang, C. Xu, M. Hu, Q. Yin, L. Li, C. Zhou, G. Li, and H. Chen. 2024. Statistical modeling of spatially stratified heterogeneous data. *Annals of the American Association of Geographers* 114 (3):499–519. doi: [10.1080/24694452.2023.2289982](https://doi.org/10.1080/24694452.2023.2289982).
- Wang, J., X. Li, G. Christakos, Y. Liao, T. Zhang, X. Gu, and X. Zheng. 2010. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region, China. *International Journal of Geographical Information Science* 24 (1):107–27. doi: [10.1080/13658810802443457](https://doi.org/10.1080/13658810802443457).
- Wang, J., T. Zhang, and B. Fu. 2016. A measure of spatial stratified heterogeneity. *Ecological Indicators* 67:250–56. doi: [10.1016/j.ecolind.2016.02.052](https://doi.org/10.1016/j.ecolind.2016.02.052).
- Yang, J., J. Wang, X. Liao, H. Tao, and Y. Li. 2022. Chain modeling for the biogeochemical nexus of cadmium in soil–rice–human health system. *Environment International* 167:107424. doi: [10.1016/j.envint.2022.107424](https://doi.org/10.1016/j.envint.2022.107424).
- Yin, Q., J. Wang, Z. Ren, J. Li, and Y. Guo. 2019. Mapping the increased minimum mortality temperatures in the context of global climate change. *Nature Communications* 10 (1):4640. doi: [10.1038/s41467-019-12663-y](https://doi.org/10.1038/s41467-019-12663-y).
- Zhang, L., W. Liu, K. Hou, J. Lin, C. Song, C. Zhou, B. Huang, X. Tong, J. Wang, W. Rhine, et al. 2019. Air pollution exposure associates with increased risk of neonatal jaundice. *Nature Communications* 10 (1):3741. doi: [10.1038/s41467-019-11387-3](https://doi.org/10.1038/s41467-019-11387-3).

BISONG HU is a Full Professor at the School of Geography and Environment, Jiangxi Normal University, Nanchang, China. E-mail: [hubisong@jxnu.edu.cn](mailto:hubisong@jxnu.edu.cn). His

research interests center on spatial statistics, spatial epidemiology, epidemic spread simulation, and spatial-temporal big data analysis.

QIAN YIN is an Associate Professor at the State Key Laboratory of Resources and Environmental Information Systems, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China. E-mail: [yinq@reis.ac.cn](mailto:yinq@reis.ac.cn). Her research interests include spatial statistics and environmental health.

JINFENG WANG is a Distinguished Professor at the State Key Laboratory of Resources and Environmental Information Systems, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing, China. E-mail: [wangjf@reis.ac.cn](mailto:wangjf@reis.ac.cn). His research interests include spatial statistics and their application in geoscience and population health.

BIN JIANG is a Full Professor at Urban Governance and Design Thrust, Society Hub, The Hong Kong University of Science and Technology (Guangzhou), China. E-mail: [binjiang@hkust-gz.edu.cn](mailto:binjiang@hkust-gz.edu.cn). His research interests center on geospatial analysis of urban structure and dynamics, or geospatial big data in general.

TINGTING WU is a Master's Candidate at the School of Geography and Environment, Jiangxi Normal University, Nanchang, China. E-mail: [202140100105@jxnu.edu.cn](mailto:202140100105@jxnu.edu.cn). Her research interests include spatial analysis and urban systems.

JIN LUO is an Associate Professor and the Executive Dean of the School of Geography and Environment, Jiangxi Normal University, Nanchang, China. E-mail: [luojin@jxnu.edu.cn](mailto:luojin@jxnu.edu.cn). He specializes in spatial databases and data mining.