Future Generation Computer Systems 112 (2020) 1038–1056

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

MSGC: Multi-scale grid clustering by fusing analytical granularity and visual cognition for detecting hierarchical spatial patterns

Zhipeng Gui [a,b,c,*], Dehua Peng [a,b,c,*], Huayi Wu [b,c,*], Xi Long [b,c]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China
[b] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
[c] Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China

ARTICLE INFO

Article history:
Received 4 February 2020
Received in revised form 16 June 2020
Accepted 26 June 2020
Available online 29 June 2020

Keywords:
Multi-scale spatial clustering
Scale dependence
Aggregation pattern
Noise curve
Visual cognition
Gestalt's law

ABSTRACT

Spatial clustering is a widely used data mining method for discovery of spatial aggregation pattern. However, existing methods often neglect scale dependence, impeding the full recognition of point patterns and the detection of hierarchical spatial structures. Spatial clustering is scale dependent and linked to the size of analysis unit as well as the hierarchy of visual cognition. Therefore, this paper proposes a novel multi-scale grid clustering (MSGC) algorithm, which fuses dual scale factors, i.e., analytical scale and visual scale that sequentially integrates multi-analytical-scale clustering (MASC) and multi-visual-scale clustering (MVSC). MASC generates multi-granularity grids to transform the analytical scales, and MVSC extracts multi-level clusters to express the hierarchy of visual cognition. Comparative experiments validated the proposed algorithm against the classical Density-based Spatial Clustering of Applications with Noise (DBSCAN) and WaveCluster algorithms on both synthetic and real-world geographic datasets. The results demonstrate that MSGC can generate multi-scale clusters for increased understanding of the spatial aggregation patterns and hierarchical structures of geographic entities. Moreover, it can eliminate noise adaptively and effectively identify clusters with arbitrary shapes. Due to the nature of grid clustering, the low computational complexity enables near real-time visual analytics and efficient point pattern mining on large spatial datasets.

1. Introduction

Spatial clustering is a powerful data mining method for exploring spatial aggregation patterns, which classifies objects into clusters based on their spatial distribution proximity [1,2]. This method is widely applied in many fields such as medicine [3], economics [4], sociology [5] and geology [6]. However, the scale effect of spatial clustering remains insufficiently explored, which results in the absence of unexplored macroscopic or microscopic aggregation patterns [7]. Especially in the era of big data, a valid multi-scale clustering (MSC) algorithm for fully detecting aggregation patterns and hierarchical spatial structures is highly desired.

Scale effects can be caused by the granularity of analysis unit and visual cognition, which have significant impacts on the clustering results. Spatial points present distinct aggregation patterns at different analytical granularities such as the neighboring radius in DBSCAN or the resolution in grid clustering as shown in Fig. A.1, and single-granularity parameters cannot detect the patterns fully. Coarse-granularity analysis units can detect macroscopic spatial distribution patterns, while fine-granularity analysis units excel in microscopic pattern discovery. Moreover, the hierarchy of visual cognition is another factor leading to scale effects. The neuroimaging research has provided solid evidence that the distribution of visual attention is controlled by both of the observing intentions and salience of the physical stimulus [8,9]. Visual neuroscience demonstrates that the visual cortex has a hierarchical structure [10], enabling the human brain to recognize massive amounts of information selectively in an orderly way. The human brain receives visual information with distinct visual characteristics such as clear structure and geometric features first, and then processes objects with vague contours or abstract structures [11]. Such a selective visual attention mechanism of human eyes [12], also causes the observation and recognition of massive clusters with intricate structures to be modeled as a staged process as shown in Fig. A.2. Besides, Gestalt's law of proximity holds that human perception of objects in the receptive field is based on the degree of closeness to each other. This theory has confirmed that close objects tend to be regarded as a whole [13].

Scale effects are particularly remarkable in large volume of spatial datasets. With the widespread use of GPS navigators,

* Corresponding authors.
E-mail addresses: zhipeng.gui@whu.edu.cn (Z. Gui), pengdh@whu.edu.cn (D. Peng), wuhuayi@whu.edu.cn (H. Wu).

https://doi.org/10.1016/j.future.2020.06.053

smartphones and other location devices, more types of social sensing data with high-precision location and exploding data volume becomes easily accessible [14,15]. Large spatial point datasets incorporate clusters with arbitrary shapes and massive noise that reflect intricate and hierarchical aggregation patterns. Meanwhile, the clusters in real-world datasets tend to present inhomogeneity distributions and vary greatly in density. These features also pose difficulties to obtain accurate and complete point patterns efficiently.

Aforementioned difficulties can be tackled by explicitly modeling the scale effects from the perspectives of analytical granularity and visual cognition. Different from the conventional single-scale clustering algorithms that rely on the prior domain knowledge of users, multi-granularity clustering avoids intensive parameter tunings and subjective errors. Meanwhile, it can explore both macro and micro aggregation patterns. As a hot topic of machine learning currently, explainable artificial intelligence (AI) works on discovering the interpretability and causality [16,17], for establishing the associated relationships between the algorithm mechanism and the thinking patterns of human brain. Simulating the cognitive process can reveal the spatial hierarchy and expansion patterns of geographic entities and assist the human eyes to recognize massive clusters efficiently. Therefore, this paper proposed a multi-scale grid clustering algorithm (MSGC), which integrates two innovative algorithms, i.e., multi-analytical-scale clustering (MASC) and multi-visual-scale clustering (MVSC), for the transformation of the analytical and visual scales. To evaluate the effectiveness, both synthetic and real-world large geographic datasets were exploited to test the performance of MSGC in noise disposal, clustering quality and time efficiency at different scales.

The rest of the paper is organized as follows: Section 2 briefly reviews classical clustering algorithms, MSC algorithms and their limitations. Section 3 proposes two types of scales, and details the workflow and key methods of MSGC algorithm. Section 4 describes the experimental datasets and analyzes the clustering results. Section 5 discusses several vital issues and Section 6 draws the conclusions and directions for future work.

2. Related work

2.1. State-of-the-art classical clustering algorithms

Classical clustering algorithms can be categorized into five types, including partitional, density-based, hierarchical, prototype-based, and grid clustering approaches [18]. Partitional clustering algorithms such as K-means and K-medoids group points into the nearest cluster spatially and modify the cluster centroids through continuous iterative computations. Partitional clustering however is unable to detect non-spherical clusters due to their reliance on the distance-based principle, which becomes an obstacle when handling clusters with arbitrary shapes. In density-based clustering, the number of points within the neighborhood of each point is defined as its density and the dense areas are identified as clusters. Typically, DBSCAN and Ordering Points to Identify Clustering Structure (OPTICS) excel in detecting arbitrarily shaped clusters, but low-density clusters are easily misidentified as noise. Hierarchical clustering attempts to find discrete groups with varying degrees of similarity in a dataset and generates a dendrogram. Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is a representative hierarchical clustering algorithm with a low space complexity, but the data input sequence affects its clustering quality. Prototype-based clustering aims to find an optimal probability model or prototype to depict the data distribution, such as Gaussian Mixed Model (GMM) [19] and Self Organizing Map (SOM) [20], which are capable to handle high-dimensional data but have a relatively high time complexity. Grid clustering converts the original data space to a grid space and thus reduces the computational expense dramatically. However, this approach might lose subtle boundary information and generate zigzag boundaries in clusters. The quality can be improved using optimizing strategies that balance accuracy and time efficiency.

Partitioning and noise disposal are two essential steps for grid clustering algorithm optimization. Statistical Information Grid (STING) uses a multi-resolution grid partitioning that the statistical information at a low-resolution grid, e.g., mean and standard deviation of attributes in cells, can be mapped from a high-resolution grid. Despite the huge improvement in query speed, it is still insufficient to handle high-dimensional data efficiently due to the exponential growth of the number of grid cells as the number of dimensions increases. To fill this gap, a Grid-Clustering algorithm for High-dimensional very Large spatial databases (GCHL) generates overlapped multi-resolution grid cells and uses information in lower dimensions to avoid unnecessary search in higher dimensions [21]. The diversity and complexity of application scenarios give birth to various grid-partitioning methods [22–24].

Noise disposal in grid clustering has a significant impact on the final clustering quality. WaveCluster takes advantage of low-pass filters in a wavelet transform to remove the noise, but boundaries tend to be destroyed at the same time [25]. To detect noise accurately, a local outlier factor (LOF) was developed to calculate the inverse of the average relative density for measuring the probability that each point is noise [26]. Manually setting of LOF threshold however is subjective potentially biasing the results. To address this issue, a noise curve analytic method was proposed that generates a curve by counting the quantity of noise under different thresholds [27]. The point where the noise curve grows sharply is considered the optimal threshold. However, this approach does not account for cluster density differences on the noise curve, which leads to inaccurate extraction when handling point data with uneven density distributions.

Most of the classical clustering algorithms however, proposed upon different working principles lack an explicit expression for the spatial scale effect, thus missing meaningful point patterns at some scales. Exploring the internal relations of point patterns at different scales is necessary for comprehensive geographic phenomenon analysis in the geosciences. Measuring the impact of the scale effect when pattern mining and specifying appropriate spatial scale factors for clustering algorithms, enables effective exploration of inhomogeneous point patterns [28]. It can be argued that developing scale factor driven clustering methods to discover inhomogeneous distribution patterns and hierarchical relations at multiple scales is essential.

2.2. Emerging of multi-scale clustering

In fact, the scale effect has received attention in some spatial clustering studies. Different scales have been proposed with various concepts, and we categorize these scale factors into three types, including data scale, analytical scale and visual scale.

Data scale depicts the sampling resolution of the sensors in data acquisition. In remote sensing, data scale measures the spatial resolution that represents the minimum identifiable distance. The number of detectable semantic classes is uncertain and depends on the spatial resolution of image. To overcome the uncertainty, a collaborative clustering algorithm that fuses two images at different data scales was designed to extract semantic classes completely [29]. However, this pixel-level algorithm would not integrate the information extracted from different levels. To fill this gap, multi-level analyses were considered, i.e., area

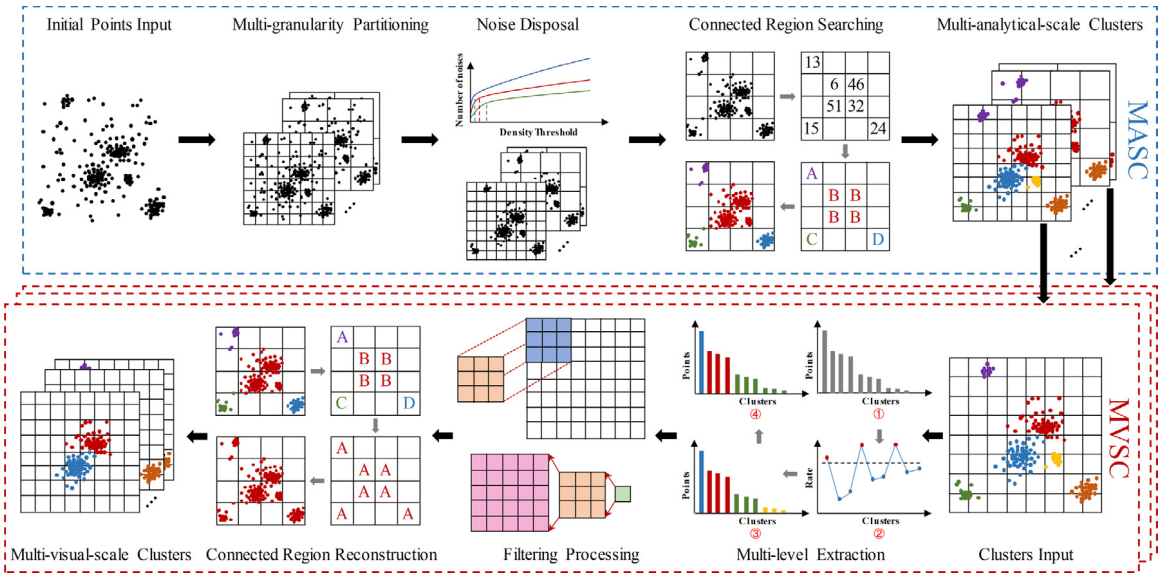


Fig. 1. The two-stage workflow of multi-scale grid clustering (MSGC).

level, block level and region level [30]. With the fusion of multi-resolution images, the method improves the clustering quality.

Analytical scale is usually defined as a controllable indicator that measures the size of analysis unit. In scale-space theory, analytical scale is modeled as the bandwidth of Gaussian root mean square [31]. Based on this theory, a MSC algorithm was developed to determine the number of clusters and prototype locations adaptively [32], but it cannot separate a weak cluster from a strong cluster nearby. To overcome this defect, a variable step size is adopted to modify the original method for extracting distinctive clusters adaptively in the color reduction [33]. DBSCAN can also be extended to a MSC algorithm by defining the neighboring radius as an analytical scale [34,35]. In addition, the partitioning granularity in grid clustering determines the size of grid cells. Both STING and WaveCluster exploit multi-granularity grid generation to explore multi-scale aggregation patterns and can be applied to large datasets due to the low time complexity.

Visual scale is leveraged to model the cognitive hierarchy caused by human visual system in clustering. The human information acquisition is a multi-staged complicated process which is affected by multiple biological factors [8–10]. Visual resolution, the size of receptive field and selective visual attention mechanism can all influence the process of information acquisition. They might cause a hierarchical cognitive process from fuzzy to clear or from local core areas to global scope [8,9]. Such hierarchy of visual cognition can reveal the evolutionary process of geographical elements with a polycentric distribution in spatial clustering. Two visual scales have been defined to model the impacts brought by visual resolution and size of receptive field [36]. The first scale measures the observation distance between the eyes and clusters. Because it affects the sampling resolution of the retina as a biological sensor. Another scale depicts the size of receptive field based on graph theory. This method can generate multi-scale spatial patterns and reduce the subjective impact on input parameters, but cannot handle massive number of points due to the time costs incurred in the construction of a Delaunay triangulation network. Selective visual attention mechanism also influences the cognitive process, controlling the cognitive prioritization of clusters with different sizes [37]. However, there are few studies that model visual scale on the basis of selective attention mechanism, which is the focus of our clustering algorithm.

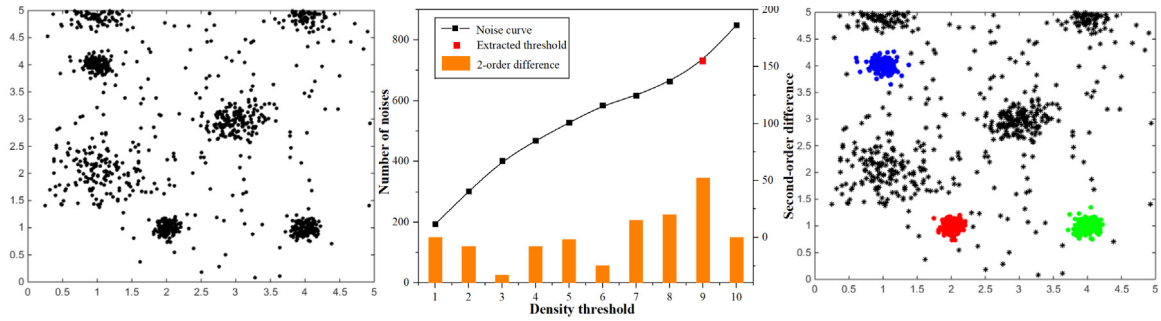
2.3. Limitations of existing work

Although there have been studies focusing on multi-scale clustering, they lack unified scale definitions and take no account of the scale effects caused by biological factors. Recently, explainable AI has become a hot research branch in machine learning, which aims to reveal the causality between algorithms and human brain [38]. It can be argued that humans are an indispensable part of the machine learning methods or the control loop of systems [39]. Comprehending the thinking and recognizing patterns of human brain facilitates the control and improvement of algorithms, further discovering explainable laws [40,41]. Scale based on biological modeling connects the algorithms and human brain. Inspiration from the biological process is beneficial for explaining the mechanism and recognizing the patterns that fit to the real world. As a biological sensor for information acquisition, human eyes would produce cognitive hierarchy of brain when observing aggregation patterns. However, such hierarchical patterns are rarely discussed in existing studies. In addition, the calculation of MSC is a computation-intensive task, especially when handling the large spatial datasets with intricate structures and massive noise, which hinders the scalability of algorithms.

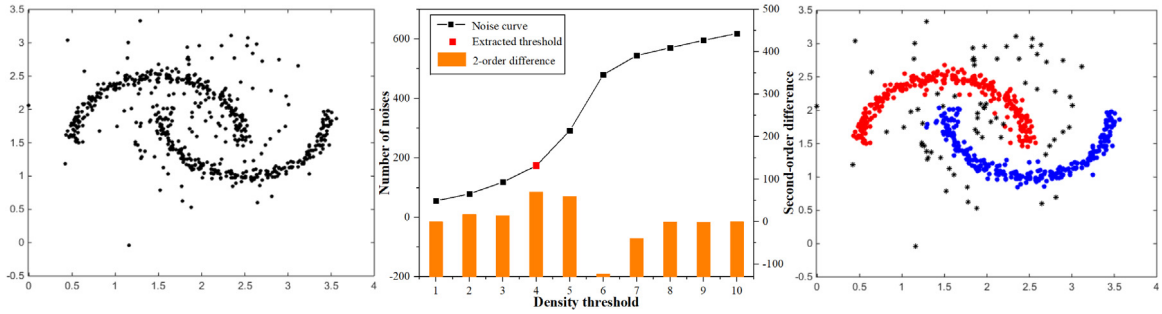
Considering the high time-efficiency of grid clustering on large datasets [42], we propose a novel MSGC algorithm that fuses analytical and visual scales. Specifically, the analytical scale is defined as the grid granularity for revealing aggregation patterns of geographic entities with different sizes of grid cells. MASC performs clustering in a multi-granularity partitioning and incorporates a refined noise curve method for adaptive noise elimination. While, in order to depict the selection attention mechanism, the visual scale is introduced to simulate multiple stages of cognitive process when exploring potential spatial hierarchies. MVSC generates the cognitive priority using a criteria-based extraction strategy that categorizes all clusters into multiple levels. In addition, MVSC merges nearby clusters and removes isolated tiny clusters at each visual scale by filtering with variable-length templates to conform the Gestalt's law of proximity.

3. Methodology

The overall workflow of MSGC is illustrated in Fig. 1, this method integrates the procedure of MASC and MVSC. As shown, MASC contains four steps:



(a) Noise curve and the clustering result of the first dataset



(b) Noise curve and clustering result of the second dataset

Fig. 2. The noise disposal results of two synthetic datasets based on original noise curve analytic method using second-order difference to measure the change of growth rate of the noise quantity as density threshold grows.

(1) Partition the initial data space into grid space with multiple grid granularities, and each grid granularity corresponds to an analytical scale.

(2) Extract optimal noise threshold based on refined noise curve and eliminate noise points at each analytical scale.

(3) Search the connected regions and assign the same cluster labels to each cell of a connected region.

(4) Assign the cell labels to the points inside and obtain the clustering results of MASC.

MVSC takes multi-analytical-scale clusters generated by MASC as input and conducts the following four steps for the result at each analytical scale:

(1) Group all the clusters into multiple levels based on the size difference of clusters and the total number of points in each level.

(2) Merge the close clusters at each cluster level by filtering with variable-length templates.

(3) Reconstruct connected regions from the filtering results, and assign the cell labels to the points and generate multi-visual-scale clusters.

(4) Repeat steps 1 through 3 till all analytical scales are processed.

In the following sections, based on the working mechanisms of the two scale factors, analytical scale and visual scale, we introduce the details of the algorithm and MASC and MVSC scale transformations.

3.1. Multi-analysis-scale clustering (MASC)

In grid clustering, the results are explicitly affected by the partitioning granularity. Spatial point datasets usually contain clusters of different point sizes and densities. However, loose clusters are separated by a fine-granularity grid, and tiny clusters are detected as noise using a coarse-granularity grid. To overcome this problem and explore point patterns comprehensively, we employed multi-granularity partitioning in MASC, and each granularity corresponds to an analytical scale.

Whereas, setting a noise threshold that can adapt to different grid granularity is a challenging task. Most of traditional noise disposal methods rely on expert knowledge and intensive parameter tuning, which adds labor cost when determining the optimal noise threshold. Therefore, we designed a modified noise curve method based on the convexity index to improve threshold extraction.

3.1.1. Noise disposal based on convexity index of noise curve

The original noise curve is an effective adaptive noise disposal method but cannot handle clusters with significant density differences. The model assumption is that a noise curve grows gently at first, but then rises dramatically. However, significant density differences between clusters might violate this assumption. As shown in Fig. 2(a), the extracted threshold in the first dataset is much higher than its optimal value that misidentified four clusters as noise, because there exists a significant density difference between the dense and loose spherical clusters. While the two banded clusters are well-found in the second dataset since the points in the banded clusters are almost evenly distributed (Fig. 2(b)).

To address this problem, we created a convexity index (CI) as shown in Eq. (1) to measure the relative change of the growth rate, and the point with maximum CI value represents the optimal threshold. The original method tends to extract the maximum absolute change of the growth rate between the high-density and low-medium-density clusters, which leads to the misidentification of the low-medium-density clusters. While the use of relative change in CI considers cluster density, which ensures the extracted threshold is between the noise points and clusters. In addition, the differences of cluster density may cause the shape of noise curve varied. Employing absolute value in CI facilitates to handle both concave and convex noise curve.

$$CI = \left| \frac{f''(x)}{2f(x)} \right| \tag{1}$$

where $f(x)$ denotes the number of noises with the density threshold x, and $f''(x)$ refers to the second-order derivative of the noise curve. Since the density threshold is discrete with an equal

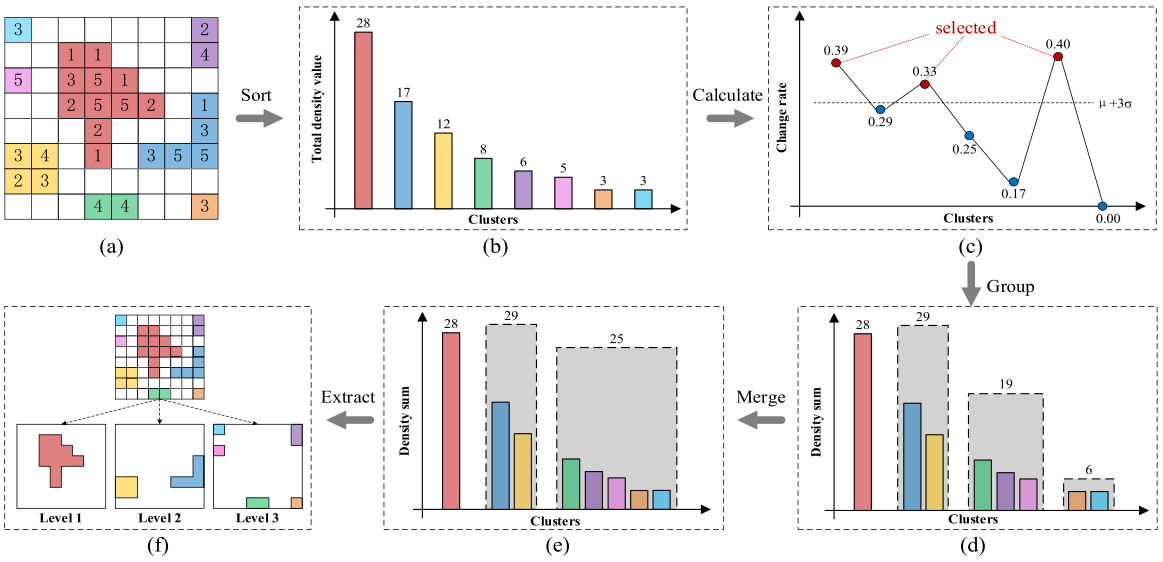


Fig. 3. Illustration of multi-level cluster extraction workflow.

interval one, the second derivative $f''(x)$ can be converted into a second-order difference as Eq. (2).

$$CI = \left|\frac{f''(x)}{2f(x)}\right| = \left|\frac{\frac{1}{2}(f(x+1)+f(x-1))-f(x)}{f(x)}\right| \tag{2}$$

The modified noise curve method can adapt to different spatial distribution of points and grid granularities, thereby extracting the optimal noise threshold adaptively. In order to preserve the details of the cluster boundaries, we set the cell whose density and that of its eight neighbors are all lower than the threshold as a noise grid cell.

3.2. Multi-visual-scale clustering (MVSC)

Simulating the cognitive process facilitates to reveal hierarchical structures and spatial expansion patterns of geographic entities. Moreover, modeling the stages from core clusters to secondary clusters eliminates visual interference of massive tiny clusters when focusing on the core clusters. MVSC categorizes all clusters into multiple levels with significant visual differences to generate multiple cognitive stages using a criteria-based extraction method. For each visual scale, MVSC filters the clusters with variable-length templates to conform to the Gestalt's law of proximity.

3.2.1. Criteria-based multi-level cluster extraction

The first step of MVSC is to extract multiple cluster levels that are quite different visually. The extraction criteria are proposed to guarantee the representiveness of extracted multiple cognitive stages and the consistency in the amount of information from all visual scales.

Criterion 1: the clusters in different levels should be significantly different in the number of points contained.

Criterion 2: the total number of points contained in each level should be approximately the same.

The general workflow categorizes all clusters into several initial levels and then merges the adjacent levels constantly until it reaches the terminal condition according to the two criteria, as illustrated in Fig. 3. Clusters are sorted in a descending order according to the number of points contained (Fig. 5(b)). The relative change rates of every two point-amount-adjacent clusters are calculated and compared with a threshold to ensure the clusters belonging to different levels varies significantly in size (Fig. 3(c)).

The initial levels are generated by grouping clusters using the change rates larger than the threshold (Fig. 3(d)). Here, we select μ+3σ as the threshold, where μ denotes the mean of change rates and σ denotes the standard deviation. Actually, this threshold can be customized in a certain range since the impacts caused by lower thresholds could be eliminated by merge operations. Merge operations are conducted on each two adjacent levels only if it can reduce the coefficient of variation of the total number of points in each level until reach the maximum iterations or the coefficient of variation remains unchanged (Fig. 3(e)). The coefficient of variation measures the dispersion of the sample data and its decline indicates the total number of points in each level is closer. Finally, multiple cluster levels are extracted (Fig. 3(f)). Algorithm 1 shows the pseudocode of multi-level cluster extraction.

Algorithm 1. Multi-level cluster extraction algorithm.

Input the number of points in each cluster at an analytical scale $\{x_1, x_2, \ldots, x_n\}$
Output the extracted multiple cluster levels $L = \{l_1, l_2, \ldots, l_m\}$

Sort $\{x_1, x_2, \ldots, x_n\}$ in a descending order to $\{y_1, y_2, \ldots, y_n\}$;
Calculate the change rates of each two adjacent clusters $R = \{r_1, r_2, \ldots, r_{n-1}\}$;
Calculate the mean μ and variance σ of R;
Assign $t = 1$;
for each r_i in $\{r_1, r_2, \ldots, r_{n-1}\}$ **do**
  if $r_i > \mu + 3\sigma$ **then**
    Put $\{y_{j+1}, y_2, \ldots, y_i\}$ into level list L as a new level l_t;
    Assign $j = i$ and t++;
  end if
end for
Calculate the coefficient of variation c_0 of initial levels L;
for iterations reach the maximum number or c_0 remains unchanged **do**
  for each level l_i in L **do**
    Try to merge l_i and l_{i+1} into one level;
    Calculate the new coefficient of variation as c_1;
    if $c_1 < c_0$ **then**
      Confirm merge operation and assign $c_0 = c_1$;
    end if
  end for
end for
return the extracted levels $L = \{l_1, l_2, \ldots, l_m\}$;

3.2.2. Cluster merging using variable-length filter templates

Gestalt's law of proximity leads to the visual effect that close clusters tend to be viewed as a whole, and this effect becomes more remarkable with the expansion of focus area [13]. Filtering is widely applied in digital image and signal processing for smoothing and denoising [43]. Research indicates that filtering is capable to connect neighboring connected components in multi-scale image processing using variable-length operators [44]. Since the continuous space is expressed as grid cells in grid clustering, filtering can be adopted in MVSC to depict Gestalt's law. Therefore, we propose a filtering processing with variable-length templates to simulate this visual effect by merging close clusters and removing isolated clusters.

Fig. 4 illustrates the workflow of variable-length filtering processing. According to the intermediate clustering result generated at a certain analytical scale (Fig. 4(a)), multiple visual scales are initialized by integrating multiple cluster levels based on extraction criteria defined in Section 3.2.1 (Fig. 4(b)). Variable-length filtering templates merge close clusters and eliminate isolated clusters at each visual scale. The connected regions are reconstructed based on the filtering results (Fig. 4(c)). Grid labels are mapped to the original points at each visual scale to generate the ultimate clustering results (Fig. 4(d)).

In this paper, we adopt mean filtering in MVSC, assuming that each neighboring cells has the same impact on the centered cell. Although the weights of filter templates influence the density of each cell, density does not affect the construction of connected regions, and thus has no impact on the ultimate results. Hence, the weights of the filter template can be customized. Meanwhile, we linearly increase the length of filter template to simulate the merging visual effect of the human eyes, which simplifies parameter setting for scale transformation. Actually, the decrease in visual resolution that occurs with the expansion of focus area presents a complicated nonlinear trend, which is hard to be modeled precisely. A series of continuous linear increases at a fine interval however, can approximate a nonlinear expansion of the focus area.

4. Experiments

MSGC is an integration of MASC and MVSC, which models the scale effects of analytical and visual scales respectively. In order to make comprehensive analysis of the algorithm, we designed three experiments to validate the effectiveness of the two components of the MSGC as well as the integrated algorithm respectively. All three experiments were conducted on a desktop computer with a 4-core Intel i7 processor and 16 GB RAM. The operating system was Windows 10 and algorithms were implemented in Java.

4.1. Experimental data and design

Both synthetic and real-world spatial datasets were selected to test the performance of the proposed algorithm. The data descriptions for each experiment are illustrated in Table 1, including data type, cluster shape and number of points.

In Table 1, the real-world POI data used in experiment 1 is the enterprise registration data of Hubei Province of China, and taxi trajectory point dataset of Wuhan City of China from 0:00 am to 1:00 am on May 17, 2014. The enterprise registration data has a polycentric industrial distribution with intricate cluster shapes and reflects distinctive differences in regional development patterns. The taxi trajectory data however, presents a network distribution. The trajectory points form clusters on busy roads but are sparsely located on less traveled roads, which results in density differences along different sections. While the POI data in experiment 3 is from the enterprise registration data of mainland China. The dataset presents a sharp economic gap between Southeast China and Northwest China. Massive tiny clusters and noise surround the central clusters with a polycentric distribution.

Table 1
Descriptions of the experimental data.

Experiments	Data type		Cluster shape	Number of points
Experiment 1	Synthetic		Spindle	1000
			Banded	619
			Combined	2971
			Ringed	1774
	Real-world	POI	Combined	0.37 million
		Trajectory	Network	0.57 million
Experiment 2	Synthetic		Combined	2109
Experiment 3	Real-world	POI	Combined	2.35 million

Experiment 1 aims to verify the capability of noise disposal, clustering quality and time-efficiency of MASC. In order to validate the feasibility of the modified noise curve analytic method in-depth, we designed four synthetic datasets containing clusters with different densities, diverse shapes, and many noise points, and compared MASC clustering results with DBSCAN, WaveCluster, as well as clustering results with fine-tuned thresholds visually. Meanwhile, we applied MASC to two large real-world datasets with different cluster shapes and adopted seven validity metrics to evaluate clustering quality and time efficiency quantitatively. The features of the enterprise registration data facilitate evaluation of the ability of MASC to detect clusters with different shapes and sizes at different scales. While regional difference in the trajectory data reflects the connectivity and integrity of the network, which could validate the applicability of MASC to complex network data.

Experiment 2 is designed to verify whether MVSC can detect hierarchical patterns that conform to visual cognition and Gestalt's law. A synthetic dataset is selected with manually tagged multi-level labels which were tagged manually following Gestalt's law (Fig. 11). This dataset contains clusters with different shapes and hierarchical point patterns. Meanwhile, we also analyzed the influence of MASC on MVSC furtherly under different analysis scales.

Experiment 3 assesses the overall performance of MSGC by evaluating its capability for exploring industrial spatial agglomeration and expansion trends on the mainland China enterprise registration dataset. A statistical analysis is also performed for the clustering results of the mainland China enterprises to study the changing tends of cluster properties as visual scale increases and explore the size distribution of clusters.

4.2. Experiments of multi-analytical-scale clustering (MASC)

4.2.1. Effectiveness of noise disposal on synthetic datasets

In this experiment, we visually compared the clustering results of MASC, DBSCAN, and WaveCluster as shown in Fig. 5. In general, all the three algorithms can yield relative high clustering quality but have subtle differences in the boundary areas with appropriate manual noise threshold settings for DBSCAN and WaveCluster. However, the results of the last two algorithms may be unsatisfactory without cautious noise threshold settings, e.g., the WaveCluster clustering results with a fixed noise threshold on the fourth column in Fig. 5. More specifically, the clustering results of MASC are approaching to that of DBSCAN and better than WaveCluster. MASC can identify all clusters accurately and detect the complete shapes of clusters. For DS1 and DS3, the density of the clusters is inhomogeneous. Most clusters have a high

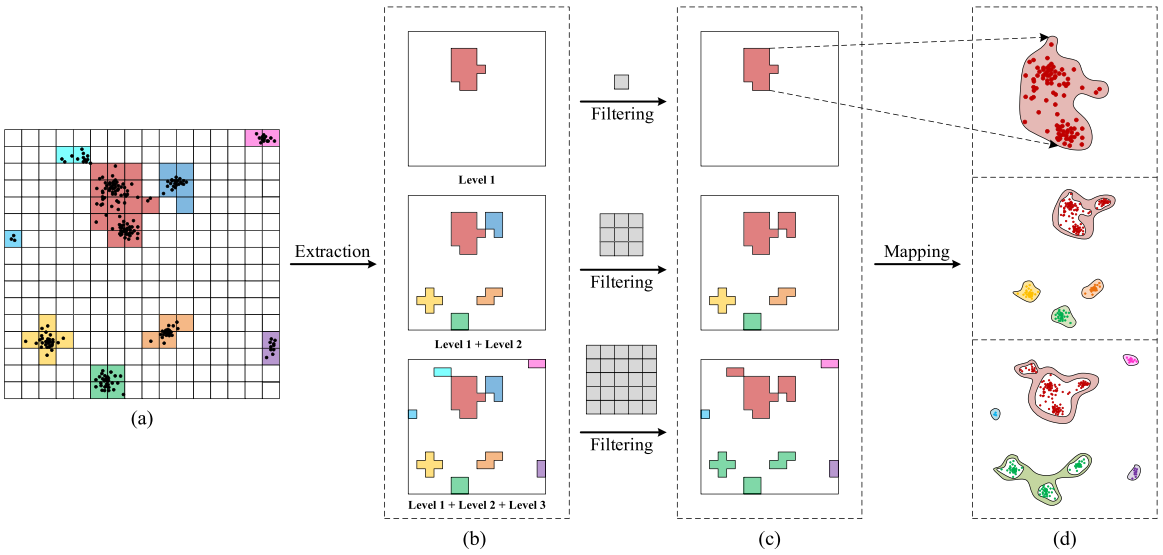


Fig. 4. Illustration of multi-visual-scale clustering with variable-length filter templates.

density in the central areas and low density in the boundary areas (Fig. 5(a) and (c)), which causes to MASC perform worse than DBSCAN in the identification of cluster boundaries since grid-based clustering coarsens the analysis unit from point to grid cell. Theoretically, the clustering label of each point is determined by the cell it belongs to, but zigzag structures occur at the boundaries of clusters usually, which leads to the loss of boundary detail. But the strategy of noise disposal in MASC performs better than WaveCluster and retains more boundary points of clusters. While, the clusters in DS2 and DS4 present banded or ringed distribution (Fig. 5(b) and (d)), with much denser point densities than the noisy areas. The clear boundaries of clusters are beneficial for the performance of MASC even close to that of DBSCAN.

In order to further verify the accuracy of noise disposal, we manually tuned the thresholds extracted from the modified noise curve method and compared the clustering results as shown in Fig. 6. Grid clustering is sensitive to noise threshold selection. A subtle change of noise threshold will bring a significant variation on the results. We found that a threshold lower than the extracted value fails to remove the noise and leads to the merge of different clusters, while a relatively higher threshold tends to misidentify the points belonging to clusters as noise, thus splitting a complete cluster into multiple discrete parts. Therefore, the modified noise curve method can extract relatively reasonable noise threshold and avoid labor-intensive parameter tuning. As the result, MASC can yield promising clustering accuracy comparable to DBSCAN with appropriate granularity.

4.2.2. Clustering quality and time efficiency on large geographical datasets

In this experiment, we designed four analytical scales, corresponding to partitions of 4096 × 4096, 2048 × 2048, 1024 × 1024 and 512 × 512 grid cells. In the 4096 × 4096 grid, the size of each cell is approximately 180 m in the enterprise registration data and 30 m in the taxi trajectory data. We picked DBSCAN as the comparative algorithm since it can handle clusters with arbitrary shapes and can be extended to a multi-analytical-scale algorithm, simply by adjusting the neighboring radius. In order to ensure that the parameter settings in DBSCAN were consistent as possible with MASC, we set the input parameter *MinPts* equal to the noise threshold extracted by MASC and the neighboring radius *Eps* as Eq. (3). This ensured that the area of circular analysis unit in DBSCAN was approximately equal to that of a grid cell in MASC.

$$Eps = \sqrt{\frac{(Lng_{max} - Lng_{min}) \cdot (Lat_{max} - Lat_{min})}{\pi m}} \tag{3}$$

where $Lng_{max} - Lng_{min}$ and $Lat_{max} - Lat_{min}$ represent the longitude range and latitude range of the dataset respectively, and m refers to the total number of cells in the grid.

We used seven metrics to compare MASC and DBSCAN quantitatively, including number of clusters, number of noise points, running time, compactness index (CPI), separation index (SPI), Davies–Bouldin index (DBI) and the connectivity index (CNI). The first three metrics depict basic clustering attributes and the remaining four evaluate the validity of the algorithm for datasets without multi-scale reference labels. CPI measures the compactness by calculating the average inner-cluster distance from the centroid to other points. While SPI measures the separation by calculating the mean distance of all pairs of clusters. DBI leverages both compactness and separation by combining CPI with SPI [45]. CNI considers the impact of the cluster size and introduces the number of points in the clusters as weights. In addition, it calculates the furthest distance between all point pairs in a cluster to measure the compactness rather than the average distance [46]. The lower CPI and DBI and the higher SPI and CNI reflect better clustering quality.

(1) Clustering analysis on enterprise registration data

The clustering results of MASC and DBSCAN for enterprise registration data of Hubei Province are illustrated in Fig. 7. With the increase of analytical scale (the grid granularity becomes coarser), these two algorithms present the same tendency that dispersed tiny clusters at a fine granularity gradually merge to larger clusters and show macroscopic spatial aggregation patterns. As shown in the amplified regions (Wuhan City, China), enterprises are separated into various different clusters. The increase in analytical scale causes them to merge. In general, clusters from DBSCAN are more fragmented and present more details than those of MASC at each analytical scale. Moreover, MASC merges faster than DBSCAN, because DBSCAN has a stricter spatial proximity rule than MASC. Regardless of the impact of other neighboring points, two points could be grouped only if their distance is lower than *Eps* in DBSCAN, while any point pair belonging to neighboring cells whose distance is even larger than *Eps* might be clustered in MASC.

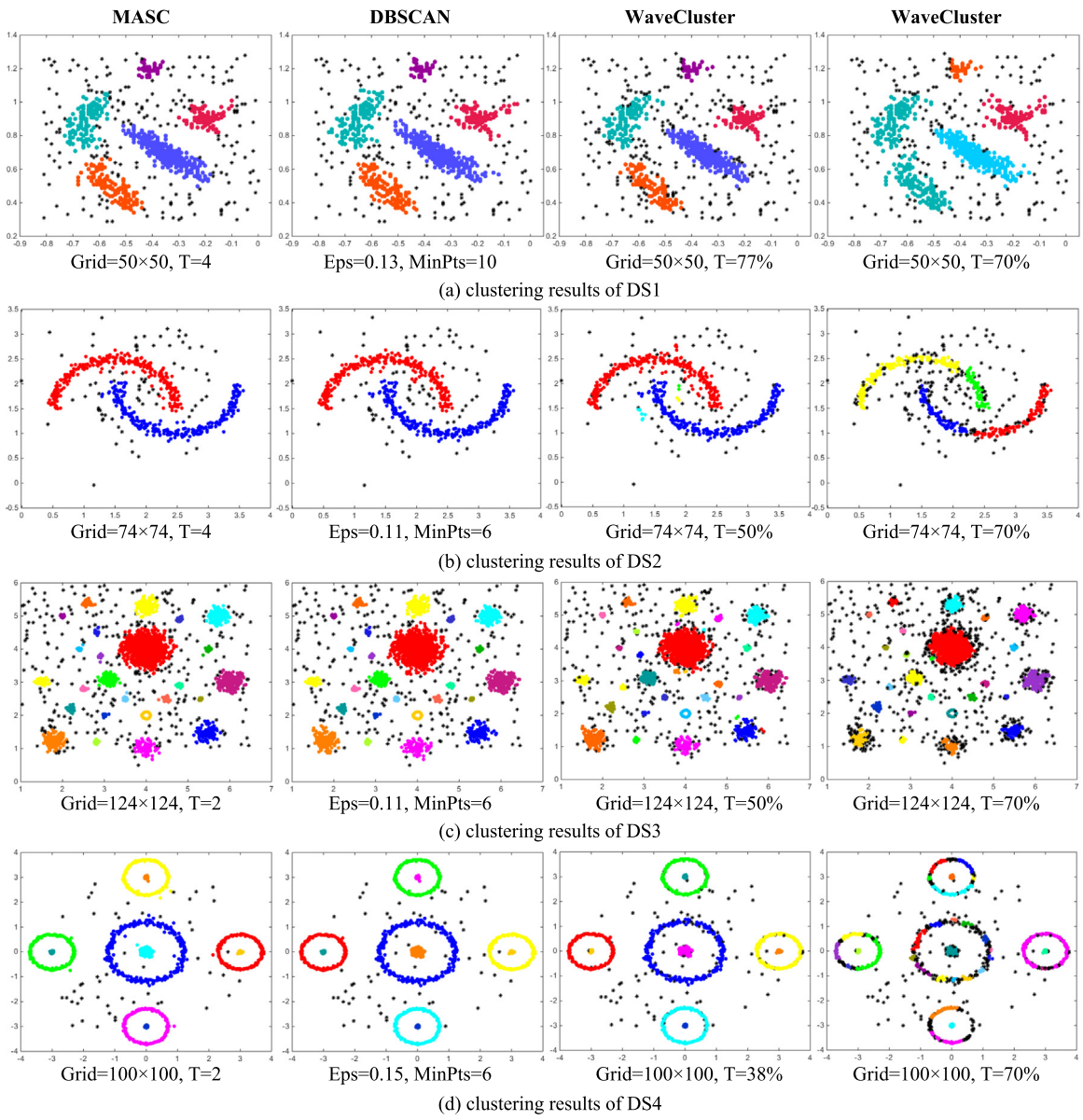


Fig. 5. The results of MASC, DBSCAN and WaveCluster on four synthetic datasets.

Table 2
Basic metrics of MASC and DBCAN on enterprise registration data.

Scale	Number of clusters		Number of noises		Running time (s)	
	MASC	DBSCAN	MASC	DBSCAN	MASC	DBSCAN
1	5,651	9,582	12,184	11,183	3.89	769.83
2	4,223	6,391	9,151	8,724	1.35	767.95
3	3,282	4,691	7,466	7,198	1.08	785.36
4	2,107	3,807	6,104	5,887	0.98	774.81

The first three basic metrics are presented in Table 2. The number of clusters and noise decreases for both MASC and DBSCAN with the increase of analytical scale, but fewer clusters and more noise are identified in MASC since MASC destroy the boundary points of clusters. However, MASC is more efficient than DBSCAN. The running time of MASC presents a decreasing trend give the drop in the number of cells. The running time of DBSCAN remains stable because it depends on the number of points.

To evaluate the clustering quality of MASC, the four validity metrics are shown in Fig. 8. The CPI and CNI values indicate that these two algorithms have their own strengths in clustering quality at different scales. The compactness of DBSCAN decreases with the increase of analytical scale, so the CPI displays an upward trend, while MASC has better compactness at large scales. Because MASC tends to lose the boundary details at large scales, which causes the areas of clusters to be smaller than that of DBSCAN. CNI shows that the clustering quality of MASC is much better than DBSCAN in the 4096 × 4096 grid, but is close at other scales. SPI and DBI illustrate that MASC performs better than DBSCAN in separation and the balance between compactness and separation. The tiny isolated clusters surrounding the large clusters are detected as noise in MASC, which makes the between-cluster distances larger than DBSCAN.

(2) Clustering analysis on taxi trajectory data

The comparisons of MASC and DBSCAN on taxi trajectory data of Wuhan City are illustrated in Fig. 9. At small scales, both of these algorithms can identify the roads in different degrees of traffic congestion especially for heavily traveled sections, and preserve abundant cluster details, e.g., the connectivity between different road sections. As the increase of scale, massive tiny

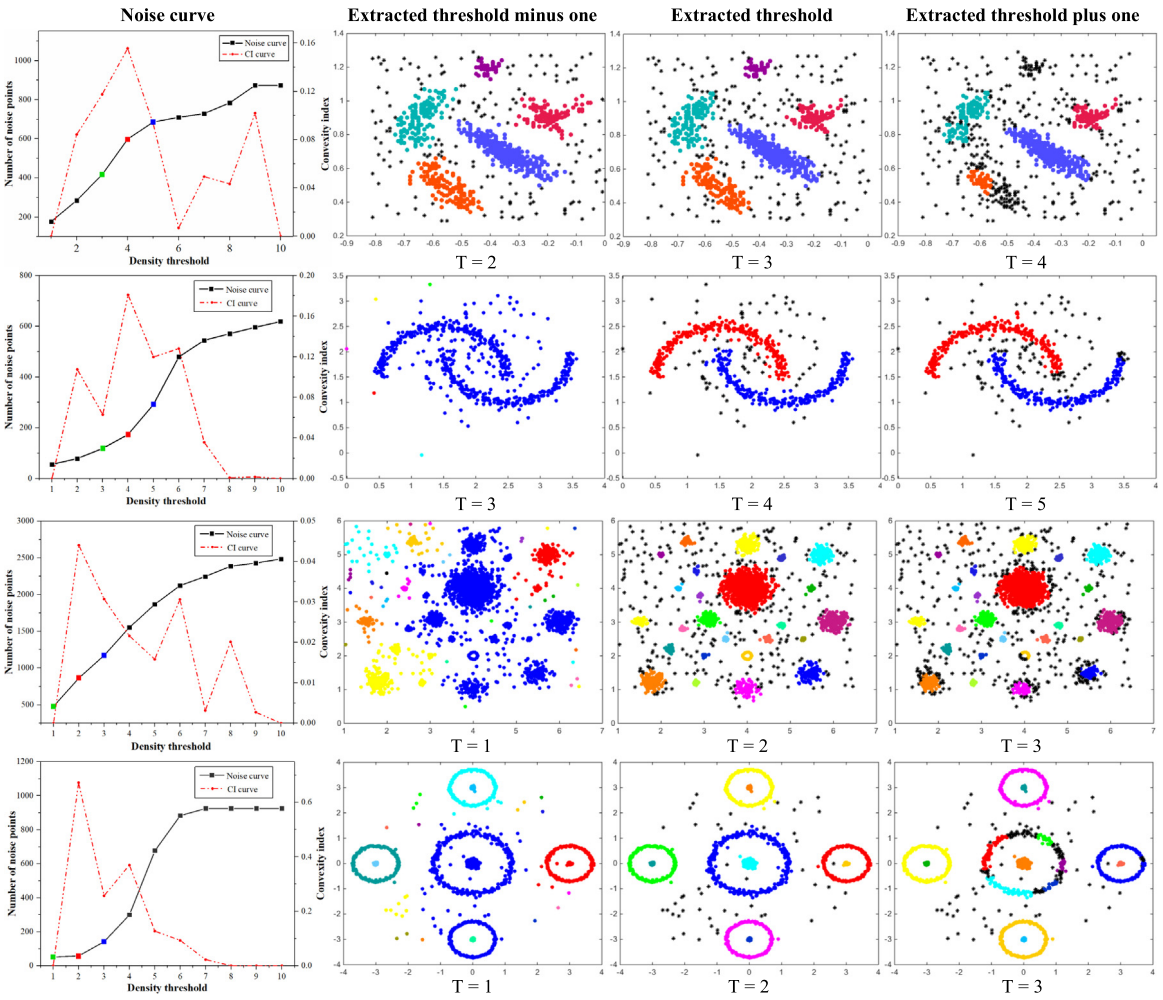


Fig. 6. The results of MASC on four synthetic datasets with the extracted optimal thresholds and neighboring values.

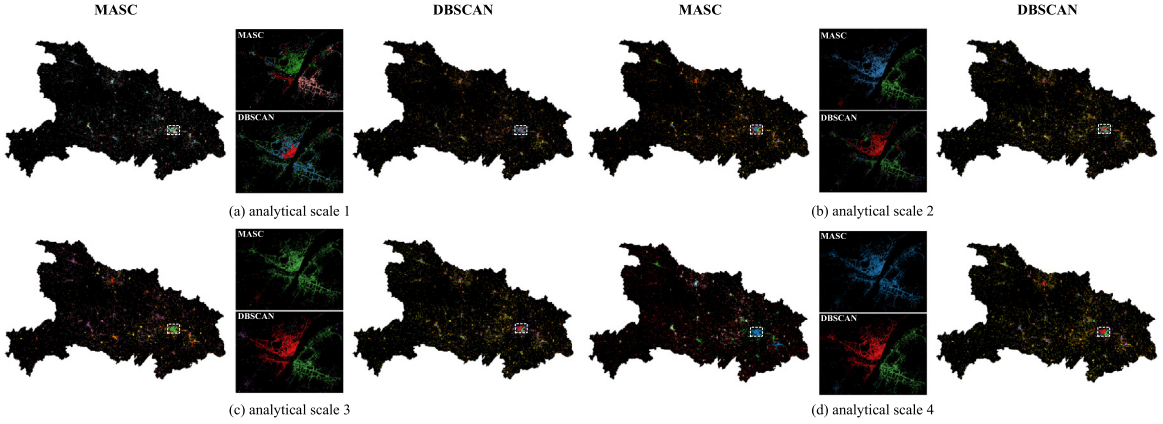


Fig. 7. The results of MASC and DBSCAN on enterprise registration data of Hubei Province.

clusters on road sections are connected into larger networks and the amount of clusters declines dramatically. In general, the clusters of DBSCAN have stronger network connectivity and integrity, e.g., scale 2 and 3. Despite that the largest cluster has a smaller spatial coverage than that of MASC at scale 4, DBSCAN preserves detailed and complete network structures on tiny road sections. While many gaps exist within MASC clusters resulting in a weaker connectivity due to the weakness of noise disposal method.

Table 3 illustrates three basic metrics for the clustering results on taxi trajectory data. Similar to the changing pattern presented in Table 2, MASC identifies fewer clusters and more noise than DBSCAN at each scale. However, the numbers of clusters are much smaller than those in the enterprise registration dataset and decrease faster as the scale increases. This might be attributed to the connectivity of the network architecture, which causes the clusters to connect spatially. We also found that the running time of MASC was higher on the enterprise registration dataset

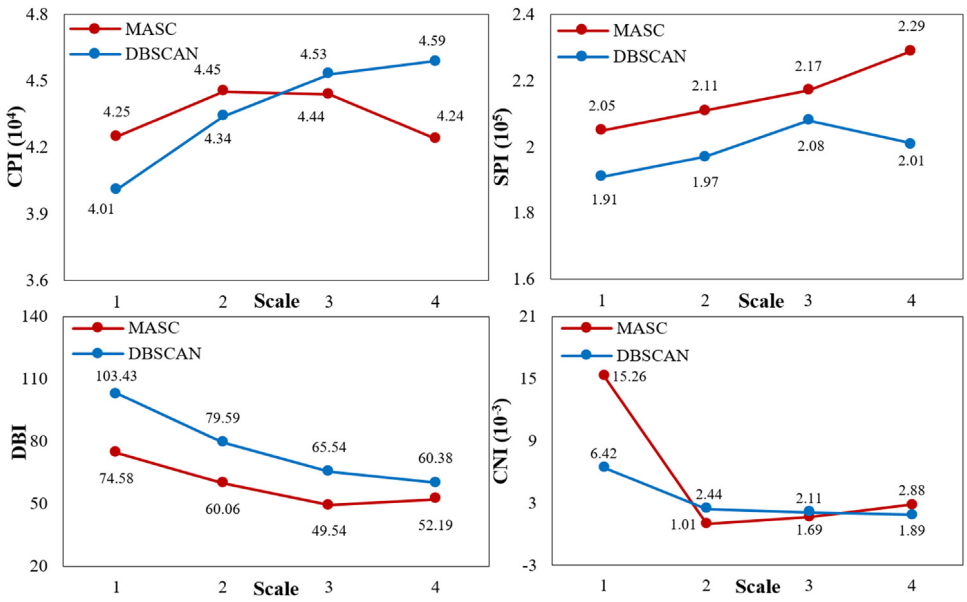


Fig. 8. Validity metrics of MASC and DBSCAN on enterprise registration data.

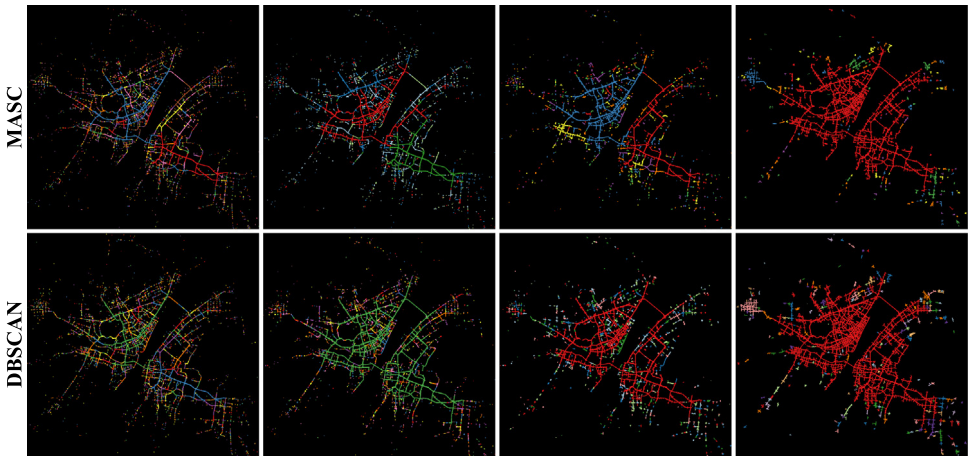


Fig. 9. The results of MASC and DBSCAN on taxi trajectory data of Wuhan City.

Table 3
Basic metrics of MASC and DBSCAN on taxi trajectory data.

Scale	Number of clusters		Number of noises		Running time (s)	
	MASC	DBSCAN	MASC	DBSCAN	MASC	DBSCAN
1	3,173	4,373	160,636	105,216	16.12	1,525.26
2	1,534	1,973	154,317	90,038	5.23	1,663.11
3	675	853	173,833	103,050	3.31	1,615.46
4	188	273	119,989	66,505	2.45	1,734.83

although the number of cells in these two sub-experiments was same at each scale. This occurred because the clusters of road networks have a larger grid coverage, increasing the iterations when reassigning cell labels in the connected region search.

For the validity metrics in Fig. 10, CPI, SPI and DBI show the same trends of change with the increase in scale, as the two algorithms capture similar spatial aggregation patterns. Meanwhile, the CNI results show a big difference at scale 3 and 4, as CNI takes the size of clusters into account, and the smaller number of clusters in MASC makes the influence of large clusters dominant. SPI, DBI and CNI indicate that MASC yielded higher quality results. In contrast, the CPI indicates that DBSCAN achived higher quality because MASC identifies more tiny clusters as noise, so these clusters had larger average inner-cluster and between-cluster distances.

In summary, through the noise disposal analysis on the typical synthetic datasets and the quantitative comparisons of clustering quality and time-efficiency on two real-world datasets, we found that MASC and DBSCAN have their own strengths. Although DBSCAN delivers higher integrity and preserves precise boundary and cluster details, MASC performs better than DBSCAN on some validity metrics at different scales. Meanwhile, parameter tuning requires prior knowledge and a large number of experiments, thus limiting the application of DBSACN when handling large datasets. While, MASC has an adaptive strategy of noise disposal and performs in a much more time efficient manner than DBSCAN due to the nature of grid clustering. High time-efficiency makes MASC more appropriate for near-real-time exploration of hierarchical point pattern on large spatial datasets.

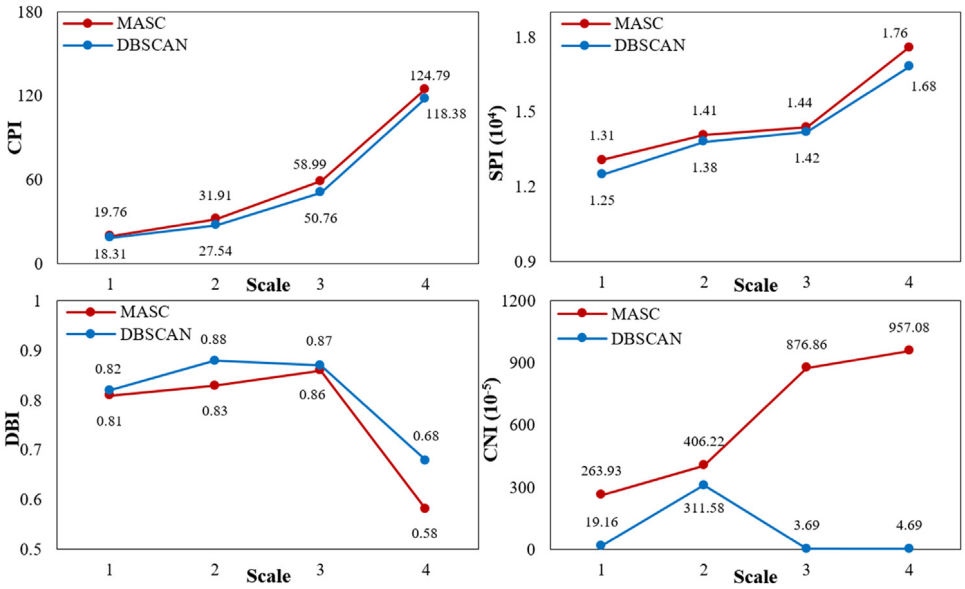


Fig. 10. Validity metrics of MASC and DBSCAN on taxi trajectory data.

4.3. Experiment of multi-visual scale clustering (MVSC)

(1) Effectiveness of MVSC

In order to verify whether MVSC can detect hierarchical patterns that conform to visual cognition and Gestalt's law, we applied MVSC to the MASC results from a synthetic dataset generated in a 120 × 120 grid. The initial labeled clusters in Fig. 11(a) present a hierarchy of visual cognition. Specifically, large clusters, e.g., C1 and C2, tend to be first caught by human eyes. Smaller clusters are gradually captured with eye movements and expansion of the focus area, e.g., C3–C13. Meanwhile, nearby clusters at a previous scale are merged when the visual resolution decreases. Our experimental results show that extracting multi-level clusters can simulate this human cognitive process.

As shown in Fig. 11, MVSC extracted three distinct levels as seen in Fig. 11(b). The corresponding multi-visual-scale clusters were obtained as in Fig. 11(c)–(e) based on the criteria proposed in Section 3.2.1. At visual scale 1, C1 and C2 were detected separately, due to their large size and the clearly visible spatial gap between them at high visual resolution. Eight smaller clusters C3–C10 appear when C1 and C2 were merged as the visual resolution declined with the expansion of the focus area at scale 2. Three tiny clusters C11–C13 were identified at scale 3. At this scale, neighboring clusters at scale 2 were merged i.e., C3–C5, C6–C8 and C9–C10. The merging operation establishes inclusion relations between clusters at adjacent visual scales when clusters at the current visual scale contain one or more clusters at a previous scale.

(2) Influence of analytical scale setting

Since MVSC inputs the clusters obtained by MASC rather than the original points, to furtherly analyze the influence of analytical scale setting on MVSC, we analyzed the results of MVSC quantitatively upon the MASC results at five different grid granularities, including the 160 × 160, 140 × 140, 120 × 120, 100 × 100 and 80 × 80 scales. We manually assigned multi-level labels to all points as benchmarks, which are consistent with the multi-visual-scale results in the 120 × 120 grid in Fig. 11(c–e) except one misidentified point marked in Fig. 11(a). The MVSC results as shown in Fig. B.1 were compared with the benchmarks using five validity metrics including Precision, Recall, F1-score, Rand index, and Jaccard index. The F1-score takes both precision and recall into account. The Rand index calculates the proportion of correctly clustered point-pairs, while the Jaccard index measures the similarity between the results and labels [47]. All five metrics range from 0 to 1, and the higher value, the more precise the results.

The validity metrics are presented in Table 4. Different grid granularities generated different numbers of visual scales. MASC with a fine-granularity grid tends to identify more clusters and enrich the spatial hierarchy of aggregation patterns, thus it gives birth to more visual scales. While the coarsening of grid granularity reduces the number of clusters and simplifies the process of visual cognition. Variation in the grid granularity changes the clustering quality of MASC and further alters the MVSC results. In this case, MASC obtains the most precise result in the 120 × 120 grid, and all five validity metrics are the highest comparing to the other four grid granularities. Moreover, the clustering qualities have no significant association between the adjacent visual scales, e.g., the precisions of visual scale 1 and 2 in the 160 × 160 and 100 × 100 grids. Nonetheless, the clustering qualities of the last visual scales for all analytical scales are relatively high, since the filtering processing connects dispersed patterns and makes the results similar to the final macroscopic cognitive stage.

The experiments proved the effectiveness of MVSC on the base of quantitative analysis. MVSC is capable to detect the hierarchical patterns by adopting multi-level process of clustering information reception, which accords with visual cognition. Specifically, multi-level cluster extraction generates visually distinct levels according to the size of clusters. Filtering can achieve the evolutionary process from local core clusters to global scope by merging spatially nearby clusters.

4.4. Experiment of integrated multi-scale grid clustering (MSGC)

In order to assess the overall performance of MSGC, we use the enterprise registration POI data of mainland China to analyze the clustering results at different analytical scales, regions and visual scales, which demonstrates the capability of MSGC for exploring industrial spatial agglomeration and expansion trends. Exploring the aggregation patterns of enterprises contributes to an understanding of uneven regional economic development, and a more nuanced interpretation of the influence of policies, population, and other socioeconomic factors on the industrial spatial distribution.

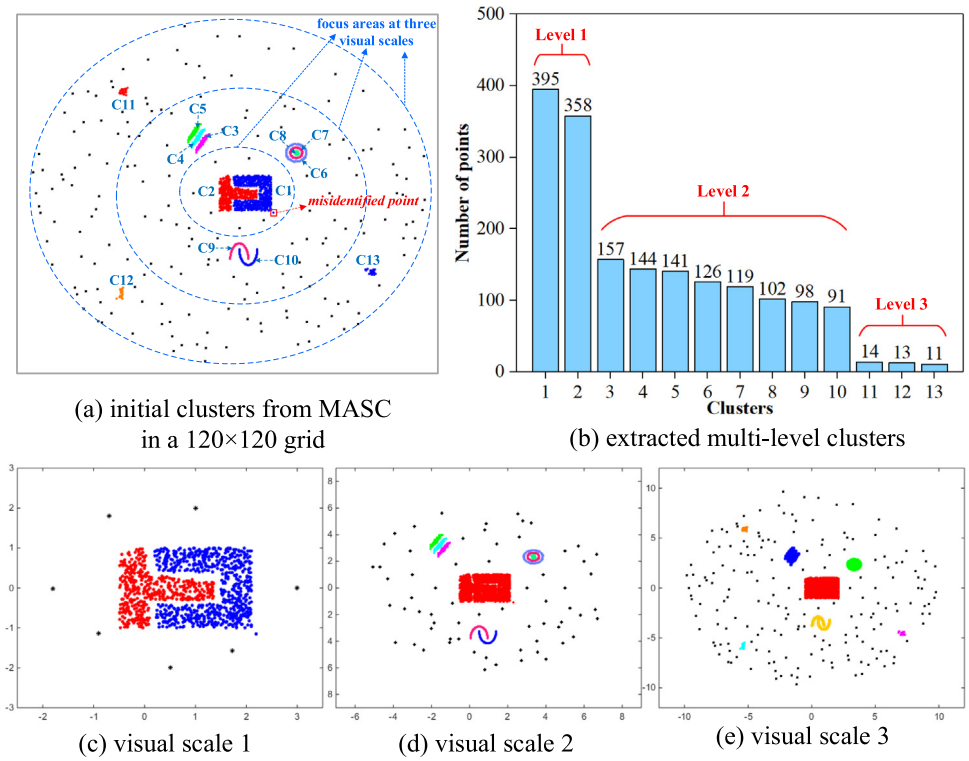


Fig. 11. The clustering result of MVSC for the initial clusters of a synthetic dataset by MASC with manual-tagged multi-level cluster labels in 120 × 120 grid.

Table 4
Validity metrics of MVSC at five analytical scales.

Analytical Scale	Visual Scale	Precision	Recall	F1-score	Rand Index	Jaccard Index
160 × 160	1	0.9959	0.9020	0.9467	0.9477	0.8987
	2	0.5271	0.9039	0.6659	0.8099	0.4991
	3	0.9679	0.9074	0.9366	0.9567	0.8808
140 × 140	1	0.9999	0.9921	0.9960	0.9959	0.9920
	2	0.9989	0.9922	0.9955	0.9966	0.9911
	3	0.9946	0.9920	0.9933	0.9953	0.9867
120 × 120	1	**1.0000**	0.9980	0.9990	0.9990	0.9980
	2	0.9999	0.9982	0.9990	**0.9996**	0.9980
	3	0.9998	**0.9995**	**0.9991**	0.9994	**0.9983**
100 × 100	1	0.5232	0.9896	0.6838	0.5306	0.5195
	2	0.9971	0.9985	0.9983	0.9988	0.9966
80 × 80	1	0.5223	0.9845	0.6833	0.5291	0.5190
	2	0.9951	0.9982	0.9966	0.9976	0.9933

(1) Clustering results on multiple analytical scales

Through the visual results in Fig. 12 and quantitative results in Table C.1, we can find a trend that massive tiny clusters are merging into larger clusters with the transformation of analytical scales. Specifically, at smaller scales, i.e., in Fig. 12(a)–(b), large clusters located in economically important regions are surrounded by massive tiny clusters. This pattern shows the details spatial distribution of industrial clusters corresponding to the location of major cities and counties across the country. While, at large scales, i.e., in Fig. 12(c)–(d), the maps present a sharp industrial spatial distribution gap between Northwest China and Southeast China that reflects the unbalance of regional economic development [48]. The clusters in the southeast coast tend to be connected as a whole, which reveals that the industrial spatial agglomeration evolved from the pole-axis model to the network development model [49].

In the 4096 × 4096 grid, detailed industrial spatial distributions of four major urban agglomerations are amplified in Fig. 13. Each region contains a central cluster surrounded by massive tiny clusters, reflecting the radial distribution in these regions. Central clusters in Yangtze River Delta and Pearl River Delta have higher densities and wider ranges than that in Beijing-Tianjin region and Chengdu City, and tend to connect to the surrounding clusters. In addition, the central clusters of inland regions, e.g., Beijing and Chengdu, are spherically shaped, but the coastlines lead to non-spherical cluster shapes in Yangtze River Delta and Pearl River Delta. Therefore, MASC recognizes clusters with arbitrary shapes and captures the diverse structures found in urban agglomerations.

(2) Clustering results on multiple visual scales

We applied MVSC to the results of MASC at all analytical scales (shown in Fig. C.1), and present three visual scales at the analytical scale 3 (1024 × 1024) in Fig. 14. MVSC can extract the core areas of industrial spatial agglomeration at small visual scales and macroscopic spatial distribution patterns at large scales. When human eyes observe core areas, interference of peripheral small clusters or low-density areas are removed. At visual scale 1, only one large cluster was found, located along the Yangtze River Delta. Clusters in secondary agglomeration regions were captured with the increase of visual scale, and clusters detected at the

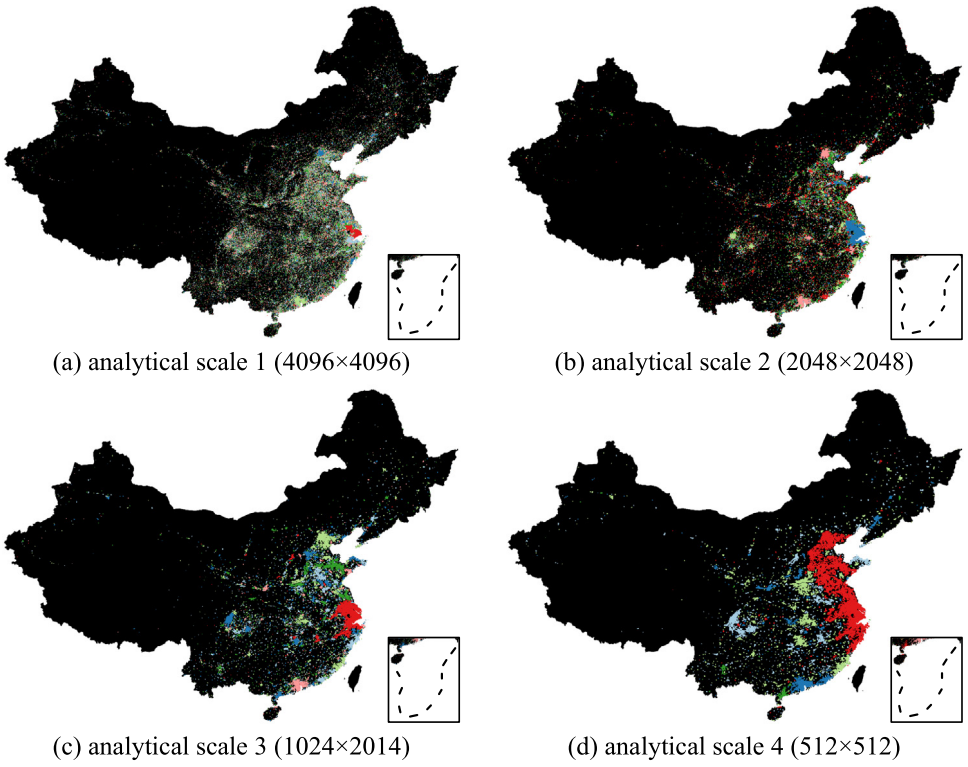


Fig. 12. The results of MASC on enterprise registration data of mainland China.

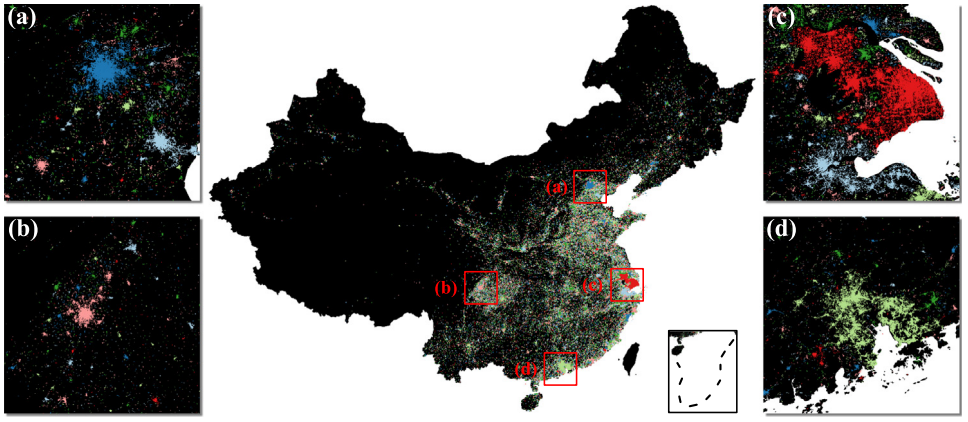


Fig. 13. The clustering result of MASC at scale 1 (4096 × 4096): (a) Beijing-Tianjin region; (b) Chengdu City; (c) Yangtze River Delta; (d) Pearl River Delta.

previous scale were merged with surrounding clusters to form larger clusters. MVSC simulates the visual cognitive processes to reveal hierarchies in the spatial structure when exploring spatial–temporal patterns in industrial agglomerations.

(3) Statistical analysis on clustering properties

The clustering properties of the three visual scales in Fig. 14 are furtherly analyzed from statistical perspective. The number of clusters and the total number of points contained in these clusters grow with the appearance of small clusters and the merging of previous clusters as the visual scale increases (Fig. 15(a)). Linear growth in the number of points contained conforms to the criterion of informational equality, while the number of clusters grows nonlinearly. The sizes of extracted clusters are skewed such that a few core clusters contain large number of points, while a massive number of tiny clusters contain few points. For example, the only cluster at scale 1 and twenty-one clusters at scale 2 contain more than 1/5 and 1/2 of the total points respectively. A log–log plot is drawn by ranking all MASC clusters in the 1024 × 1024 grid according to the number of points, as shown in Fig. 15(b).

A Kolmogorov–Smirnov test (p-value > 0.05) indicates that clusters follow a power law distribution in the number of points contained [50]. Hence, there are significant differences in the importance of clusters. The existence of massive tiny clusters will generate unnecessary information that interferes with human eyes cognition of core clusters. Multi-level cluster extraction simulates the expansion process of visual focus area from local core clusters to secondary global clusters, thereby weakening the cognitive interference.

5. Discussion

5.1. Relations between MASC and MVSC

MASC and MVSC have a sequential relation in the workflow of MSGC that MVSC takes the clusters detected by MASC as inputs.

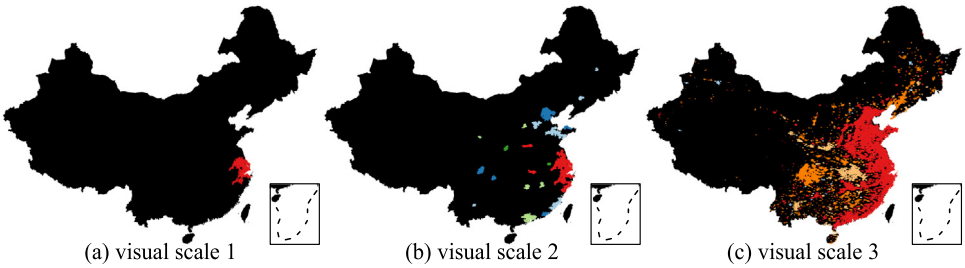


Fig. 14. The results of MVSC on the clusters generated by MVSC in the 1024 × 1024 grid.

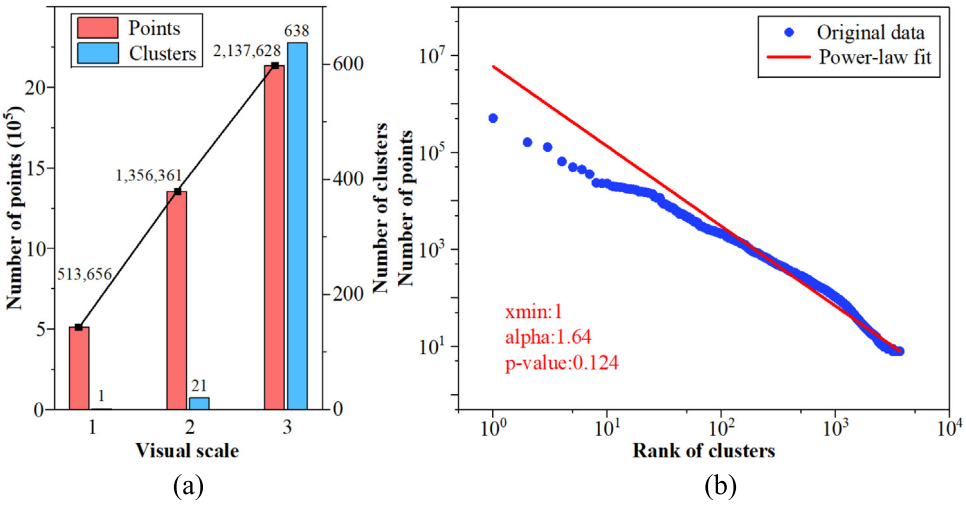


Fig. 15. Quantitative analysis of the growth trends of the number of clusters and points contained as the visual scale increases. (a) the growth trend of the number of clusters and points contained; (b) log–log plot of the number of points contained in the clusters from MASC in the 1024 × 1024 grid according to a discrete power law distribution with $\alpha = 1.64$ and $x_{min} = 1$.

Variations in the analytical scale affect both quantity and quality of the clusters identified in MASC, changing the spatial hierarchy in the patterns of aggregation as well as the visual cognition process in MVSC. The clustering results of MASC therefore determine the number of extracted visual scales and the MVSC clustering quality.

Multi-scale clusters obtained by MASC and MVSC display different trends when the scale changes. A coarsening of grid granularity leads to a decrease of the number of clusters extracted by MASC, coarsening the boundaries of clusters and expanding their spatial distribution areas. In MVSC, similar to the human eyes, more tiny clusters are captured and the receptive field becomes larger as the cognitive process evolves. The movement and expansion of the focus area from core clusters to secondary clusters reveals the hierarchies inherent in the patterns of spatial aggregation in geographic entities.

Clusters extracted by MASC might be similar with the MVSC results at a finer grid granularity, e.g. clusters at analytical scale 2 (2048 × 2048) in Fig. 12(b) and clusters at analytical scale 1 (4096 × 4096) and visual scale 7 in Fig. C.1, clusters at analytical scale 3 (1024 × 1024) in Fig. 12(c) and clusters at analytical scale 2 (2048 × 2048) and visual scale 2 in Fig. C.1. That is because the coarsening of grid granularity in MASC achieves the similar effect as filtering in MVSC, and both of them lose the details of the extracted clusters and generate macroscopic patterns. At different analytical scales, clusters extracted by MVSC might also show similar spatial patterns, e.g. the four clustering results at visual scale 3 in Fig. B.1(a)–(b) and at visual scale 2 in Fig. B.1(c)–(d). This indicates that transformation of analytical scale with fine interval may generate certain spatial patterns stably, although the corresponding visual scales for these similar clustering results are different. These similar clustering results reveal the significant spatial patterns in a way that is consistent with visual cognition, further verifying the effectiveness of MVSC. However, these similar clustering results show subtle differences and varying clustering quality. Therefore, the selection of appropriate analytical scales for exploring stable and significant patterns could be further investigated.

5.2. Grid granularity and boundary processing in MASC

Grid granularity affects the grid clustering results. Grid granularity determines the size of the analysis unit and the degree of cluster details. A fine-granularity grid preserves cluster details and exposes microscopic spatial distribution patterns, while a coarse-granularity grid loses subtle details but generates patterns of aggregation at a macroscopic level. Grid granularity has impact on the noise threshold selection, since different sizes of grid cell tend to generate different spatial and density distributions from gridded data. Setting parameters in single-granularity grid clustering for massive large point datasets requires intensive tuning tests, and cannot easily handle clusters with different densities. MASC enhances adaptability by employing multi-granularity partitioning for different cluster densities, to capture spatial patterns fully, unlike single scale clustering algorithms. In real-word large dataset applications, clustering must meet specific application requirements, so we leave space for users to specify grid granularities.

A refined method is highly desired to preserve detailed boundaries of clusters. The points in boundary areas are relatively low-density rather than centroid areas, and tend to be detected as noise by a single global threshold approach. In contrast to

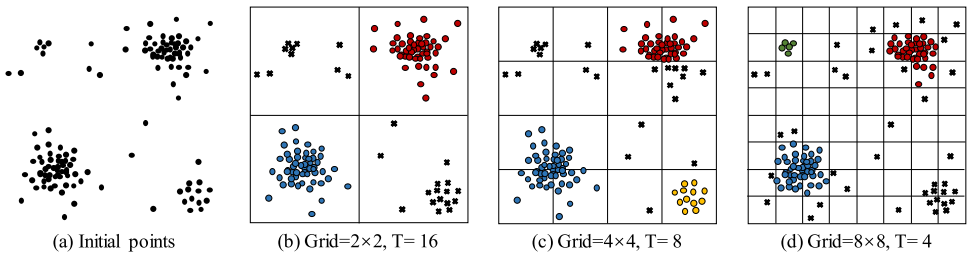


Fig. A.1. An exemplary grid granularity partitioning and its influence on clustering results.

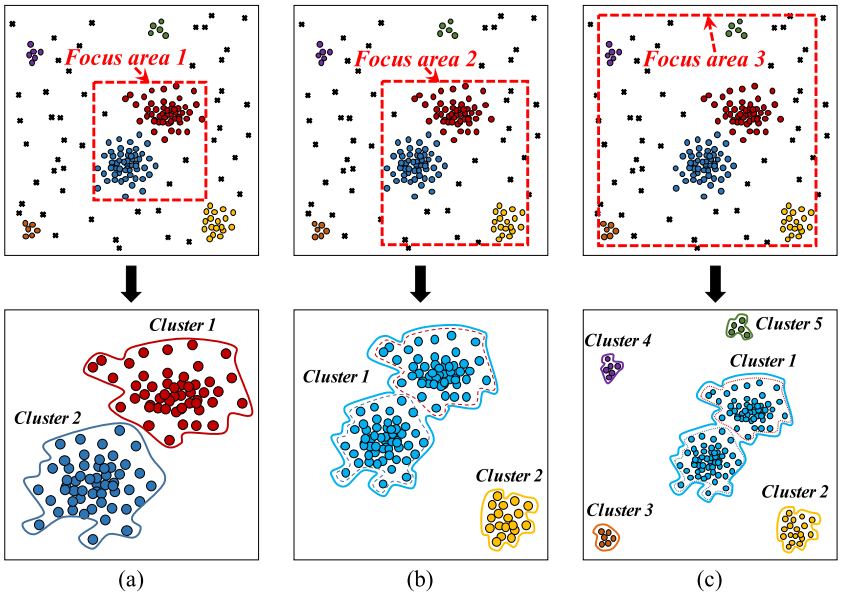


Fig. A.2. An example of the staged process of visual cognition when human eyes observe the clusters.

conventional grid clustering algorithms, MASC identifies noise cells according to the densities of the cell and its eight-neighbors. When the densities of these cells are lower than the threshold they are identified as noise, thus preserving the integrity of boundary areas. Nonetheless, MASC misidentifies more noise at the boundaries of clusters than DBSCAN when handling large spatial point datasets. Moreover, MASC is weak to preserve the integrity and connectivity of clusters with a network distribution.

5.3. Filter templates with variable-length in MVSC

Filtering with variable-length templates merges nearby clusters into larger clusters with the expansion of focus area following Gestalt's law. The templates are extended with a continuous fine-interval linear increase. The expansion of focus areas is nonlinear however, and determined by the spatial distribution of hierarchical clusters. Although continuous fine-interval linear increases in the filter template can fit the nonlinear patterns to some extent, it might slow down the merging speed of close clusters or lose macroscopic cognitive stages when the length of the largest template is lower than the distance of neighboring clusters. Thus, domain knowledge in visual neuroscience should be further investigated, and the mapping relation between the length of filter template and the size of focus area more closely approximate the biological mechanism.

6. Conclusion and outlook

This paper proposes a novel multi-scale grid clustering (MSGC) algorithm to model the scale effects from the granularity of analysis unit and the hierarchy of visual cognition, thereby exploring multi-level aggregation patterns. Specifically, MASC adopts multi-granularity partitioning to achieve the transformation of analytical scale, which can preserve both microscopic and macroscopic aggregation patterns for better understanding the multi-scale spatial distribution of geographic entities. The modified noise curve method achieves adaptive noise threshold extraction and handles the clusters with significant density differences by using a convexity index. MVSC simulates the cognitive process using a multi-level cluster extraction and variable-length filtering. It regards human as an essential factor in MSC by considering the selective attention mechanism, which is capable of exploring geographical patterns that conform to human cognition. Experiments validated the effectiveness of MSGC in noise disposal, clustering quality, and time efficiency on both synthetic and real-world geographic datasets. MSGC can detect multi-scale aggregation patterns with for large spatial point datasets time-efficiently and identify arbitrary shaped clusters with a relative high precision. It could be used to support near real-time visual analytics or as the input and reference for other compute-intensive pattern analytical methods, such as assisting the classification of remote sensing by providing a multi-granularity reference. In turn, it could be a potential use case of explainable AI by make algorithms explainable and adjusted reasonably.

Nonetheless, MSGC has three main limitations and can be further improved. Due to the nature of grid clustering, the fineness of the cluster boundaries is determined by the granularity of the grid, i.e., the analytical scale. Especially at coarse analytical scales, the grid would incorporate more noise points into the boundaries of clusters, thus reducing the clustering quality. Hence, the noise

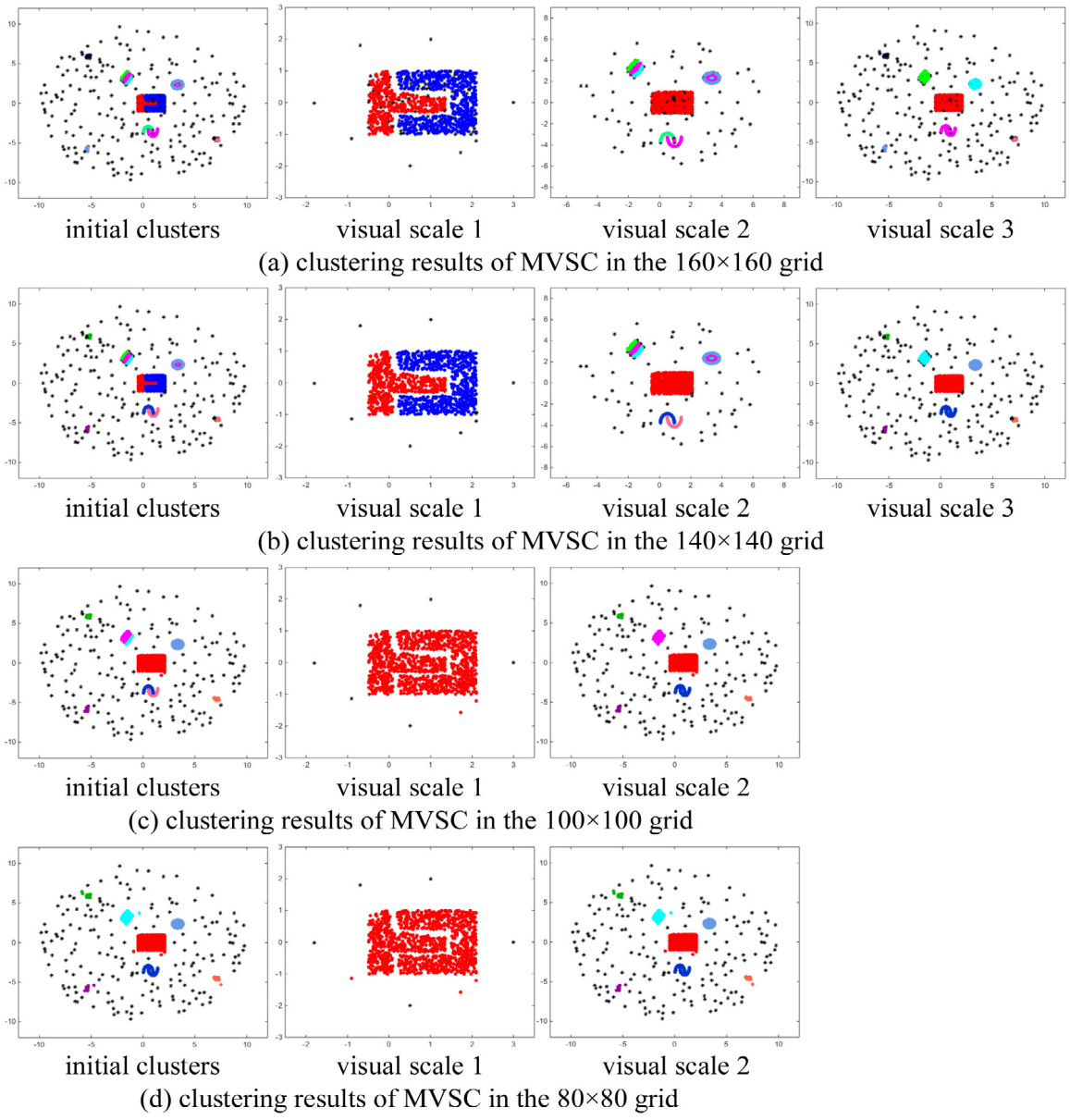


Fig. B.1. The clustering results of MVSC for the clusters generated at four analytical scales by MASC.

Table C.1
Basic clustering attributes, number of extracted visual scales and running time of each part of the integrated MSGC for enterprise registration data of mainland China.

Clustering attributes & Running time	Analytical scale			
	4096 × 4096	2048 × 2048	1024 × 1024	512 × 512
Number of clusters	40,753	10,539	3,367	1,425
Number of noise points	258,405	289,712	251,081	216,916
Number of visual scales	7	2	3	1
Total running time (s)	48.355	7.933	6.301	3.964
Input time (s)	0.233	0.231	0.249	0.258
MASC time (s)	16.539	3.273	1.935	1.772
MVSC time (s)	30.409	3.906	3.506	1.638
Output time (s)	1.174	0.527	0.611	0.296

detection method in MASC can be refined by analyzing the distribution of the internal points in boundary cells and modeling the impact of neighboring cells. Meanwhile, MVSC is insufficient to simulate the cognitive process of human brain exactly. The focus area of human eyes might present a nonlinear expanding pattern when observing clusters. However, our algorithm MVSC adopts a fine-gained linear increase to imitate this pattern, which might produce biases from the real cognitive process. Sufficient knowledge of visual neuroscience to formulate accurate quantitative models of the selective visual attention mechanism should be considered to assist the extension of the filter templates. Moreover, the adopted analytical scales are manually specified as

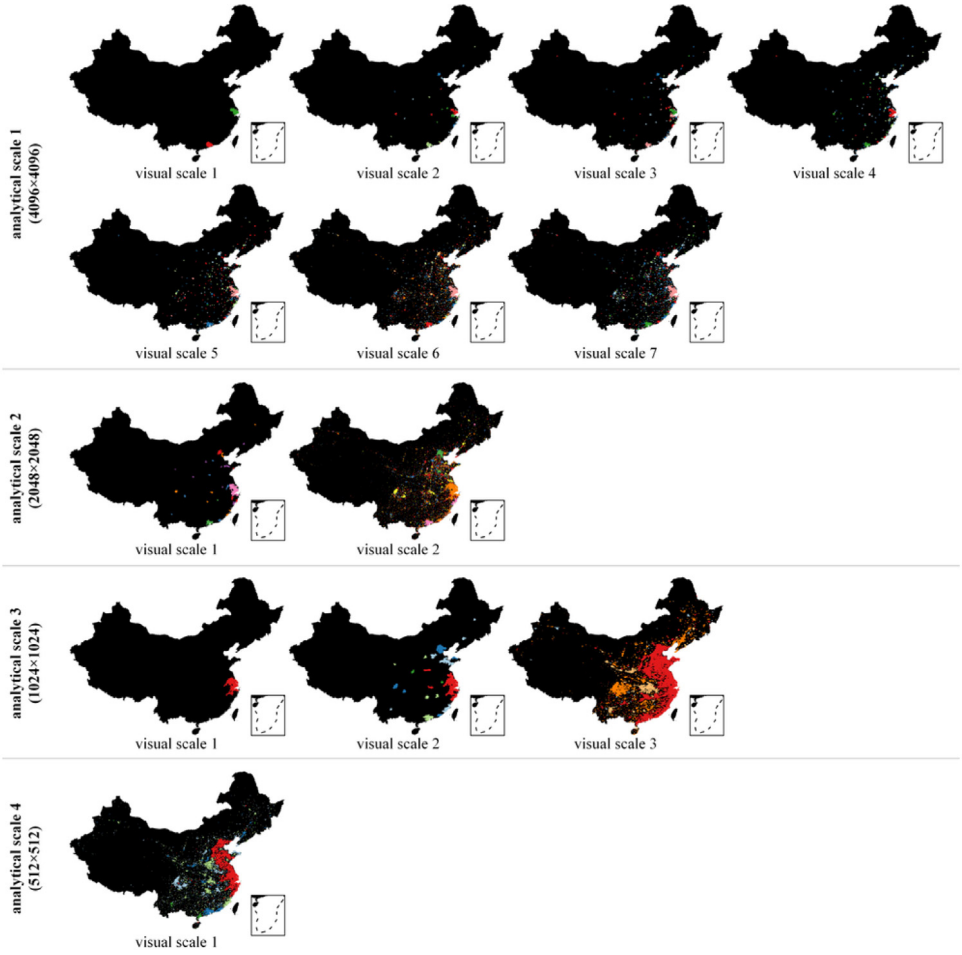


Fig. C.1. The clustering results of MSGC on enterprise registration data of mainland China.

input parameters currently. An adaptive analytical scale selection method can be further investigated by considering the spatial extent, point density, as well as significance of visual patterns at different scales.

CRediT authorship contribution statement

Zhipeng Gui: Conceptualization, Methodology, Data curation, Writing - original draft, Writing - review & editing, Validation, Supervision, Project administration. **Dehua Peng:** Methodology, Software, Writing - original draft, Writing - review & editing, Data curation, Investigation, Validation, Formal analysis. **Huayi Wu:** Supervision, Project administration. **Xi Long:** Data curation, Investigation, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This paper is supported by National Key R&D Program of China (No. 2017YFB0503704 and No. 2018YFC0809806) and National Natural Science Foundation of China (No. 41971349, No. 41930107 and No. 41501434).

Appendix A

Fig. A.1 gives an example of analytical scale. The initial points are shown in Fig. A.1(a) and the clustering results with three grid granularities are represented in Fig. A.1(b)–(d). At the coarsest granularity, two clusters on the bottom left and upper right are found in Fig. A.1(b). As the granularity becomes finer, a yellow-colored cluster emerges on the bottom right as shown Fig. A.1(c) and partial points in the red cluster in previous granularity are identified as noise points. In Fig. A.1(d), previous yellow-colored cluster is treated as noise, meanwhile a new green-colored cluster is detected in the upper left.

Fig. A.2 illustrates a staged process when human eyes observe the clusters, which generates different clustering results with the movement of focus area. According to the principle of visual neuroscience, the human eyes will firstly pay attention to the clusters with dense distribution, clear structure and large area like cluster 1 and 2 in Fig. A.2(a). The visual resolution is high at this stage due to the small range of the focus area, and these two clusters can be clearly distinguished. While as the range of focus area expands in Fig. A.2(b)–(c), more clusters emerge and previous close clusters are merged since the visual resolution declines.

Appendix B

Fig. B.1 illustrates a series of results of MVSC upon the MASC results at five different grid granularities, including the 160 × 160, 140 × 140, 120 × 120, 100 × 100 and 80 × 80 scales. As shown

in the figure, the quality of MASC directly affects that of MVSC. The number of visual scales obtained at different analytical scales is different. Generally speaking, it increases as the analytical scale gets more refined. Coarsening of analytical scale makes visual scale converges more quickly.

Appendix C

Table C.1 shows the statistical properties of the clustering results of MSGC. The total running time includes input time, MASC time, MVSC time and output time. Input time refers to the time the data was read from files and output time is the time the results were written to files. Even with millions of grid cells, the overall procedure of MSGC can be completed in less than a minute. It can be argued that MSGC can support efficient pattern analysis of large spatial datasets.

Fig. C.1 illustrates the complete clustering results of MSGC on mainland China enterprise registration data. Both the number of clusters and visual scales are influenced by the analytical scale. In the fine-granularity grid, more clusters, visual scales and cognitive stages were extracted. The refined results can reveal micro aggregation patterns at local regions. While the coarse-granularity grids produced the emergence of urban agglomeration and core economic regions. The evolutionary patterns of mainland China enterprises were explored and revealed.

References

[1] K. Ericson, S. Pallickara, On the performance of high dimensional data clustering and classification algorithms, Future Gener. Comput. Syst. 29 (4) (2013) 1024–1034, http://dx.doi.org/10.1016/j.future.2012.05.026.

[2] A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks, Science 344 (6191) (2014) 1492–1496, http://dx.doi.org/10.1126/science.1242072.

[3] E.B. Pathak, S. Reader, J.P. Tanner, M.L. Casper, Spatial clustering of non-transported cardiac decedents: the results of a point pattern analysis and an inquiry into social environmental correlates, Int. J. Health Geogr. 10 (2011) 46, http://dx.doi.org/10.1186/1476-072X-10-46.

[4] R. Boschma, M. Hartog, Merger and acquisition activity as driver of spatial clustering: the spatial evolution of the Dutch banking industry, 1850–1993, Econ. Geogr. 90 (3) (2014) 247–266, http://dx.doi.org/10.1111/ecge.12054.

[5] Z. Wang, L. Tu, Z. Guo, L.T. Yang, B. Huang, Analysis of user behaviors by mining large network data sets, Future Gener. Comput. Syst. 37 (2014) 429–437, http://dx.doi.org/10.1016/j.future.2014.02.015.

[6] B. Zhao, J. Wu, F. Yang, J. Pilz, D. Zhang, A novel approach for extraction of Gaoshanhe-group outcrops using Landsat Operational Land Imager (OLI) data in the heavily loess-covered Baoji District, Western China, Ore. Geol. Rev. 108 (2019) 88–100, http://dx.doi.org/10.1016/j.oregeorev.2018.01.034.

[7] Z. Li, Q. Liu, J. Tang, Towards a scale-driven theory for spatial clustering, Acta Geod. Cartogr. Sin. 46 (10) (2017) 1534–1548, http://dx.doi.org/10.11947/j.AGCS.2017.20170275.

[8] S. Shomstein, G.L. Malcolm, J.C. Nah, Intrusive effects of task-irrelevant information on visual selective attention: semantics and size, Curr. Opin. Psychol. 29 (2019) 153–159, http://dx.doi.org/10.1016/j.copsyc.2019.02.008.

[9] E. Awh, A.V. Belopolsky, J. Theeuwes, Top-down versus bottom-up attentional control: a failed theoretical dichotomy, Trends Cogn. Sci. 16 (2012) 437–443, http://dx.doi.org/10.1016/j.tics.2012.06.010.

[10] D.C. Vanessen, J.H.R. Maunsell, Hierarchical organization and functional streams in the visual-cortex, Trends Neurosci. 6 (9) (1983) 370–375, http://dx.doi.org/10.1016/0166-2236(83)90167-4.

[11] A. Holzinger, M. Kickmeier-Rust, H. Müller, KANDINSKY Patterns As IQ-Test for Machine Learning, in: Springer Lecture Notes in Computer Science LNCS, vol. 11713, 2019, pp. 1–14, http://dx.doi.org/10.1007/978-3-030-29726-8_1.

[12] T. Moore, M. Zirnsak, Neural mechanisms of selective visual attention, Annu. Rev. Psychol. 68 (2017) 47–72, http://dx.doi.org/10.1146/annurev-psych-122414-033400.

[13] S. Kim, K.J. Yoon, I.S. Kweon, Object recognition using a generalized robust invariant feature and Gestalt's law of proximity and similarity, Pattern Recognit. 41 (2) (2008) 726–741, http://dx.doi.org/10.1016/j.patcog.2007.05.014.

[14] Y. Wang, Z. Gui, H. Wu, D. Peng, J. Wu, Z. Cui, Optimizing and accelerating space–time Ripley's K function based on Apache Spark for distributed spatiotemporal point pattern analysis, Future Gener. Comput. Syst. 105 (2020) 96–118, http://dx.doi.org/10.1016/j.future.2019.11.036.

[15] M. Ghahramani, M.C. Zhou, C.T. Hon, Mobile phone data analysis: a spatial exploration toward hotspot detection, IEEE Trans. Autom. Sci. Eng. 16 (1) (2019) 351–362, http://dx.doi.org/10.1109/TASE.2018.2795241.

[16] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Muller, Causability and explainability of artificial intelligence in medicine, Wires Data Min. Knowl. 9 (2019) http://dx.doi.org/10.1002/widm.1312.

[17] H. Hagras, Toward human-understandable, explainable AI, Computer 51 (2018) 28–36, http://dx.doi.org/10.1109/MC.2018.3620965.

[18] M. Abdullah, H.N. Eldin, T. Al-Moshadak, R. Alshaik, I. Al-Anesi, Density grid-based clustering for wireless sensors networks, Proc. Comput. Sci. 65 (2015) 35–47, http://dx.doi.org/10.1016/j.procs.2015.09.074.

[19] J. Zhang, Z. Yin, R. Wang, Pattern classification of instantaneous cognitive task-load through GMM clustering, Laplacian eigenmap, and ensemble SVMs, IEEE ACM Trans. Comput. Biol. 14 (4) (2017) 947–965, http://dx.doi.org/10.1109/TCBB.2016.2561927.

[20] J. Vesanto, E. Alhoniemi, Clustering of the self-organizing map, IEEE Trans. Neural Netw. 11 (3) (2000) 586–600, http://dx.doi.org/10.1109/72.846731.

[21] A.H. Pilevar, M. Sukumar, GCHL: A grid-clustering algorithm for high-dimensional very large spatial data bases, Pattern Recognit. Lett. 26 (7) (2005) 999–1010, http://dx.doi.org/10.1016/j.patrec.2004.09.052.

[22] J. Cao, Y. Zhou, M. Wu, Adaptive grid-based k-median clustering of streaming data with accuracy guarantee, Database Syst. Adv. Appl. 9049 (2015) 75–91, http://dx.doi.org/10.1007/978-3-319-18120-2_5.

[23] J. Montalvao, J. Canuto, Clustering ensembles and space discretization - A new regard toward diversity and consensus, Pattern Recognit. Lett. 31 (15) (2010) 2415–2424, http://dx.doi.org/10.1016/j.patrec.2010.07.018.

[24] J. Zhang, X. Feng, Z. Liu, A grid-based clustering algorithm via load analysis for industrial internet of things, IEEE Access 6 (2018) 13117–13128, http://dx.doi.org/10.1109/ACCESS.2018.2797885.

[25] C. Ling, T. Yu, R. Chirkova, Wavecluster with differential privacy, Comput. Sci. 11 (2) (2015) 191–198, http://dx.doi.org/10.1145/2806416.2806546.

[26] D.R. Edla, P.K. Jana, A grid clustering algorithm using cluster boundaries, in: Proceedings of the 2012 World Congress on Information and Communication Technologies, 2012, pp. 254–259, http://dx.doi.org/10.1109/WICT.2012.6409084.

[27] C. Zheng, Y. Cao, Self-adaptive based on grid density clustering algorithm, Appl. Res. Comput. 36 (11) (2019) 1–7.

[28] S.Y. Kang, J. McGree, K. Mengersen, The impact of spatial scales and spatial smoothing on the outcome of Bayesian spatial model, PLoS One 8 (10) (2013) http://dx.doi.org/10.1371/journal.pone.0075957.

[29] C. Wemmert, A. Puissant, G. Forestier, P. Gancarski, Multiresolution remote sensing image clustering, IEEE Geosci. Remote Sens. 6 (3) (2009) 533–537, http://dx.doi.org/10.1109/LGRS.2009.2020825.

[30] C. Kurtz, N. Passat, P. Gancarski, A. Puissant, Multi-resolution region-based clustering for urban analysis, Int. J. Remote Sens. 31 (22) (2010) 5941–5973, http://dx.doi.org/10.1080/01431161.2010.512312.

[31] J. Babaud, A.P. Witkin, M. Baudin, R.O. Duda, Uniqueness of the Gaussian kernel for scale-space filtering, IEEE Trans. Pattern Anal. 8 (1) (1986) 26–33, http://dx.doi.org/10.1109/TPAMI.1986.4767749.

[32] E. Nakamura, N. Kehtarnavaz, Determining number of clusters and prototype locations via multi-scale clustering, Pattern Recogn. Lett. 19 (14) (1998) 1265–1283.

[33] H.J. Oh, N. Kehtarnavaz, P DSP-based automatic color reduction using multiscale clustering hotonics West-electronic Imaging, 2001, http://dx.doi.org/10.1117/12.424959.

[34] J. Capdevila, G. Pericacho, J. Torres, J. Cerquides, Scaling DBSCAN-Like Algorithms for Event Detection Systems in Twitter, in: Lect. Notes Comput. Sc., vol. 10048, 2016, pp. 356–373, http://dx.doi.org/10.1007/978-3-319-49583-5_27.

[35] X. Liu, Q. Huang, S. Gao, Exploring the uncertainty of activity zone detection using digital footprints with multi-scaled DBSCAN, Int. J. Geogr. Inf. Sci. 33 (6) (2019) 1196–1223, http://dx.doi.org/10.1080/13658816.2018.1563301.

[36] Q. Liu, Z. Li, M. Deng, J. Tang, X. Mei, Modeling the effect of scale on clustering of spatial points, Comput. Environ. Urban 52 (2015) 81–92, http://dx.doi.org/10.1016/j.compenvurbsys.2015.03.006.

[37] L. Wang, R.J. Krauzlis, Visual selective attention in mice, Curr. Biol. 28 (2018) 676–685, http://dx.doi.org/10.1016/j.cub.2018.01.038.

[38] A. Holzinger, P. Kieseberg, E. Weippl, A.M. Tjoa, Current advances, trends and challenges of machine learning and knowledge extraction: From machine learning to explainable AI, in: Machine Learning and Knowledge Extraction, Vol. 11015, CD-MAKE 2018, 2018, pp. 1–8, http://dx.doi.org/10.1007/978-3-319-99740-7_1.

[39] D.S. Nunes, P. Zhang, J.S. Silva, A survey on human-in-the-loop applications towards an internet of all, IEEE Commun Surv Tut 17 (2015) 944–965, http://dx.doi.org/10.1109/COMST.2015.2398816.

[40] A. Holzinger, M. Plass, M. Kickmeier-Rust, K. Holzinger, G.C. Crisan, C.M. Pintea, V. Palade, Interactive machine learning: experimental evidence for the human in the algorithmic loop: A case study on ant colony optimization, Appl. Intell. 49 (2019) 2401–2414, http://dx.doi.org/10.1007/s10489-018-1361-5.

[41] B. Heinrichs, S.B. Eickhoff, Your evidence? Machine learning algorithms for medical diagnosis and prediction, Hum. Brain Mapp. 41 (2020) 1435–1444, http://dx.doi.org/10.1002/hbm.24886.
[42] B. Wu, B.M. Wilamowski, A fast density and grid based clustering method for data with arbitrary shapes and noise, IEEE Trans. Ind. Inform. 13 (4) (2017) 1620–1628, http://dx.doi.org/10.1109/TII.2016.2628747.
[43] P. Nair, K.N. Chaudhury, Fast high-dimensional kernel filtering, IEEE Signal Proc. Lett. 26 (2) (2019) 377–381, http://dx.doi.org/10.1109/LSP.2019.2891879.
[44] J. Debayle, J.C. Pinoli, General adaptive neighborhood image processing: Part II: Practical application examples, J. Math. Imaging Vision 25 (2) (2006) 267–284, http://dx.doi.org/10.1007/s10851-006-7452-7.
[45] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A.Y. Zomaya, S. Foufou, A. Bouras, A survey of clustering algorithms for big data: taxonomy and empirical analysis, IEEE Trans. Emerg. Top. Comput. 2 (3) (2014) 267–279, http://dx.doi.org/10.1109/TETC.2014.2330519.
[46] C. Cai, A cluster validity evaluation index based on connectivity, Comput. Appl. Softw. 32 (11) (2015) 285–288.
[47] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, J. Intell. Inf. Syst. 17 (2001) 107–145, http://dx.doi.org/10.1023/A:1012801612483.
[48] F. Li, Z. Gui, H. Wu, J. Gong, Y. Wang, S. Tian, J. Zhang, Big enterprise registration data imputation: Supporting spatiotemporal analysis of industries in China, Comput. Environ. Urban. 70 (2018) 9–23, http://dx.doi.org/10.1016/j.compenvurbsys.2018.01.010.
[49] J. Gao, F. Yuan, Economic. transition, firm. dynamics, Economic transition firm dynamics and restructuring of manufacturing spaces in urban China: empirical evidence from Nanjing, Prof. Geogr. 69 (3) (2017) 504–519, http://dx.doi.org/10.1080/00330124.2016.1268059.
[50] M.E.J. Newman, Power. laws, Power laws Pareto distributions and Zipf's law, Contemp. Phys. 46 (5) (2005) 323–351, http://dx.doi.org/10.1080/00107510500052444.

Zhipeng Gui is an Associate Professor of Geographic Information Science in the School of Remote Sensing and Information Engineering, Wuhan University. His research interests include high-performance spatiotemporal data mining and geovisual analytics and Distributed Geographic Information Processing (DGIP), especially on 1) Spatiotemporal point pattern analysis and GeoAI; 2) High-performance geocomputation and spatial cloud computing; 3) Geospatial service chain modeling and optimization; 4) QoGIS-aware monitoring and evaluation of geospatial web services.

Dehua Peng is a master student in the School of Remote Sensing and Information Engineering, Wuhan University. He will become a doctoral student in the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing in Sep., 2020. His research interests include clustering algorithms and point pattern mining.

Huayi Wu is now a full professor in the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. His research interests include high-performance geospatial computing and intelligent geospatial web services.

Xi Long has received his master degree in the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, in 2019. His research interest is spatiotemporal data mining and database technology.