



OPEN

A raster-based spatial clustering method with robustness to spatial outliers

Haoyu Wang¹, Changqing Song^{1✉}, Jinfeng Wang² & Peichao Gao^{1,3}

Spatial clustering is an essential method for the comprehensive understanding of a region. Spatial clustering divides all spatial units into different clusters. The attributes of each cluster of the spatial units are similar, and simultaneously, they are as continuous as spatially possible. In spatial clustering, the handling of spatial outliers is important. It is necessary to improve spatial integration so that each cluster is connected as much as possible, while protecting spatial outliers can help avoid the excessive masking of attribute differences. This paper proposes a new spatial clustering method for raster data robust to spatial outliers. The method employs a sliding window to scan the entire region to determine spatial outliers. Additionally, a mechanism based on the range and standard deviation of the spatial units in each window is designed to judge whether the spatial integration should be further improved or the spatial outliers should be protected. To demonstrate the usefulness of the proposed method, we applied it in two case study areas, namely, Changping District and Pinggu District in Beijing. The results show that the proposed method can retain the spatial outliers while ensuring that the clusters are roughly contiguous. This method can be used as a simple but powerful and easy-to-interpret alternative to existing geographical spatial clustering methods.

Spatial clustering, also known as regionalization, is a traditional research topic in geography. Spatial clustering is when a set of spatial objects are grouped into several subsets, which are internally similar and as spatially contiguous as possible. These subsets are called clusters (or regions). Spatial clustering reveals the common processes and characteristics within the clusters and the differences between the clusters¹. The new clusters have a more comprehensive geographic extent than the original spatial objects², making clusters the foundation of further integrated research³. Spatial clustering has been used in many disciplines in geography^{4–7}. Based on the spatial differentiation of ecosystem functions and ecological environment, ecological function zoning provides a scientific foundation for comprehensive ecological management policies^{8–10}. Based on DEM and other morphometric attributes, geomorphometric regionalization plays a vital role in understanding the Earth's surface processes, natural hazards and landscape evolutions^{11,12}.

Spatial clustering encounters a significant challenge in adapting standard clustering techniques like K-means to capture spatial relationships¹³. To address the adaptation process, a careful balance between capturing spatial relationships and preserving attribute similarity is required. On the one hand, according to the first law of geography, nearby things are more closely related¹⁴, which makes nearby spatial units tend to be clustered into one cluster in spatial clustering. However, this process can result in the loss of similarity within the cluster. On the other hand, standard clustering methods disregard spatial relationships and focus solely on maximizing dissimilarity between clusters and maximizing similarity within clusters, which can lead to fragmented spatial clustering results.

To achieve a reasonable tradeoff between spatial and attribute relationships, the handling of spatial outliers is crucial. Spatial outliers refer to spatial units that are quite inconsistent with the attribute values of the surrounding units¹⁵. In practical geographical datasets, these spatial outliers can originate from various factors, such as errors in data collection, natural spatial variations, or the influences of human activities. These spatial outliers possess the capacity to severely disrupt the functionality of conventional clustering algorithms, resulting in clustering outcomes that lack the precision and reliability required for credible geospatial analysis. Specifically, if the spatial outliers are clustered into the same cluster with the surrounding units based on geographical location, the substantial attribute differences can severely damage the similarity within the cluster. Hence, enhancing

¹Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China. ²Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China. ³State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing 100875, China. ✉email: songcq@bnu.edu.cn

the robustness of spatial clustering methods against spatial outliers is of paramount importance. Improving the robustness to spatial outliers ensures that spatial clustering is less susceptible to being significantly affected by them, regardless of how the tradeoff between spatial and attribute relationships is managed. Nevertheless, despite the availability of numerous spatial outlier detection algorithms¹⁶, there is currently a shortage of spatial clustering algorithms designed to enhance robustness against spatial outliers.

Raster data essentially represent a region as a matrix of grids, and each grid stores the value of a spatial attribute¹⁷. The location of a cell is defined by its row and column numbers. The spatial clustering of raster data divides each grid into different clusters according to its attribute value and location. The attributes of the same cluster are similar, and the spatial position is contiguous. A great deal of research in geography employs raster data, especially when expressing information that varies continuously in space. This information is often challenging to segment and represented as vector data, such as temperature, elevation, or spectral data. In addition, advances in remote sensing technology have enriched raster data. Satellite images or aerial photos are all raster data.

This paper proposes a spatial clustering of raster data that is robust to spatial outliers. It is easy to understand and convenient to implement. This method explores the range and standard deviation of the attribute values of each locality through a sliding window. Based on this, the classification results of the raster data are adapted to improve spatial integration while protecting spatial outliers. In addition, we designed a comparison experiment for spatial clustering methods to compare the clustering effects of different types of spatial clustering methods.

Related works

Applications of spatial clustering

Spatial Clustering serves two primary applications¹³. First, it plays a crucial role in data preprocessing. It serves as an automated, unsupervised method for data organization. This enhances the efficiency of search or query algorithms and contributes to the overall system's performance. Secondly, cluster analysis is employed in exploratory tasks, enabling knowledge discovery and stimulating the generation of new hypotheses by identifying underlying patterns. In the context of data preprocessing, spatial clustering can help organize geographical data, making it more convenient for further data mining operations¹⁸. Some studies utilize spatial clustering for image feature matching^{19,20}. Spatial clustering is also a significant method for outlier detection²¹. Additionally, research has shown that spatial clustering can enhance the efficiency of spatial query processing^{22,23}.

In the domain of exploratory research, spatial clustering finds extensive applications across various fields. For instance, in climatology, Wang et al.²⁴ employed spatial clustering to analyze urban signatures in Precipitation Extremes over Mainland China. In urban studies, Yu et al.²⁵ used nighttime light data for spatial clustering to delineate urban spatial clusters and assess urban landscape patterns. In public health, researchers employ spatial clustering to analyze factors influencing the spread of infectious diseases^{26,27}. In agriculture, Gao et al.²⁸ utilized spatial clustering for the analysis of agricultural soil data. In the field of remote sensing, spatial clustering serves as a crucial tool for processing hyperspectral images^{29,30}.

Existing methods for clustering spatial data

Spatial clustering methods cluster spatial units with similar attributes while imposing constraints of being spatially contiguous³¹. There are two ways to impose constraints, imposing strict constraints or non-strict constraints³². A strict constraint means that the spatial units of the same cluster must be contiguous. Non-strict constraints require that only most, but not all, spatial units in a cluster be contiguous. The different clustering results of the two types of constraints are shown in Supplementary Figure S1.

The strict spatially constrained clustering methods require that all of the spatial units of a cluster be contiguous. These spatial clustering methods are also referred to as regionalization methods. These methods can be divided into two types³³. The first type of algorithm achieves spatial integration by postprocessing the results of standard clustering or maximizing regional compactness. For example, ³⁴ regarded spatially contiguous parts of the clustering result as clusters. The second type of algorithm incorporates spatial integration into the clustering process as an explicit constraint. For example, adapted hierarchical clustering aggregates two contiguous and most similar clusters at each step³⁵.

Strict spatially constrained clustering methods have disadvantages. First, some spatial units that are very similar but far away will be divided into different clusters³². Second, there are some contiguous but not remarkably similar units in one cluster. Moreover, this type of method fails when it is allowed to sacrifice spatial integration to some extent (e.g., allowing enclaves) to improve attribute similarity. Non-strict spatially constrained clustering methods can avoid these disadvantages to a certain extent.

The non-strict spatially constrained clustering methods trade off between spatial integration and attribute similarity³⁶ divided spatial clustering methods into four categories: strict spatially constrained clustering and the other three are all non-strictly constrained clustering³⁷ held the same opinion. That is, non-strict spatially constrained clustering methods can be divided into three categories. The first category considers the geographical coordinates as additional attributes. The trade-off between attribute similarity and locational similarity is reflected by adjusting the weights of the geographical coordinates^{38,39}. For example⁴⁰, proposed a bicriterion median clustering problem by adding geographic coordinates to median clustering for spatial units with a single attribute. The second category modifies the dissimilarity measure between different spatial units of the non-spatial clustering methods. The modified dissimilarity measure combines attribute dissimilarity and locational dissimilarity⁴¹ first proposed a spatial clustering method for this category, and the modified dissimilarity measure was applied to K-means clustering. Some other nonspatial clustering methods have been modified to this category more recently³² proposed a hierarchical clustering algorithm including spatial constraints and implemented it with the R package ClustGeo⁴² proposed an improved version ClustGeo, Bootstrap ClustGeo. All the methods

employed a weighted sum of the attribute and locational dissimilarity. The third category is based on statistical models³⁶. Ambroise et al.³⁶ proposed a spatial clustering method based on Markov random field, which was first used for image segmentation⁴³ proposed a software SpaCEM3 in which spatial clustering using Markov models is implemented. This software applied the model for gene clustering based on the Markov models proposed by⁴⁴.

These categories of methods have limitations in handling spatial outliers. For the first two categories, spatial integration and attribute similarity are controlled by location weights (namely, the weights of geographical coordinates and locational dissimilarity). Suppose the weight of the location is too large. In that case, the spatial outliers will be clustered with closer spatial units, resulting in excellent attribute dissimilarity of this cluster. If the weight of the location is too small, although the spatial outliers can be correctly delineated, the resulting clusters are spatially scattered and have poor spatial integration. The third category of methods requires the spatial units involved in clustering to satisfy certain statistical assumptions, such as the independence assumption³⁷. For geographic problems, these assumptions cannot often be satisfied, and the methods will perform poorly, so this type of category will not be considered in this paper. To reasonably delineate spatial outliers, this paper proposes a new non-strict spatially constrained clustering method for univariate spatial data (spatial data with a single attribute).

A new method for clustering raster data

Overview of the strategy

Our method is based on two fundamental principles. The first is the contiguousness principle. Spatial units of the same cluster should be connected as much as possible. The second is the reservation principle. The difference should be preserved when the spatial unit differs too much from the surrounding attributes. This principle makes it possible to characterize spatial outliers in contiguous areas. The new method proposed in this paper performs spatial clustering by improving nonspatial clustering. At present, for the spatial clustering of raster data, the method of data classification is mainly used, and the raster is classified according to an attribute value^{11,45–47}. We use data classification schemes for nonspatial clustering for univariate data. Then the method adapts the nonspatial clustering results of some spatial units to improve spatial integration. As stated before, the spatial outliers deviate considerably from the values of the attributes of their spatial neighborhood. Hence, it is important to define how ‘considerable deviation’ is in the method⁴⁸. In statistics, range and standard deviation are two commonly used measures of difference. So, the range and standard deviation of the values of the attributes of the spatial unit and its neighbors are used in the method to portray the spatial outliers. In general, the strategy of the proposed method consists of three steps. The first step is to classify spatial units based on attribute values to achieve nonspatial clustering. The second step is to delineate the spatial outliers using the range. The third step is to delineate the spatial outliers using the standard deviation. The framework is shown in Fig. 1.

In the proposed method, the principle of spatial outlier protection ensures the attribute similarity within the cluster, while the principle of contiguousness improves the spatial integration of the clusters of units other than spatial outliers. This is the way the proposed method as a non-strictly constrained clustering tradeoffs spatial integration and attribute similarity. Based on this, the proposed method can reasonably cluster spatial outliers under various degrees of spatial and attribute trade-offs, i.e., it is robust to spatial outliers.

Classification of spatial units

In our method, two data classification schemes, the natural breaks and head/tail breaks, are provided to perform nonspatial clustering. The natural breaks scheme follows the clustering principle, minimizing the variance within the class while maximizing the variance between classes. At the same time, the number of classes must be specified in advance. In contrast, the head/tail breaks can adaptively determine the number of classes. Nevertheless, this scheme is specifically designed for data with heavy-tailed distributions. Users can choose an appropriate data classification scheme according to their needs. For univariate raster data, each grid is taken as a spatial unit, and the attribute values of all the spatial units are classified.

The natural breaks scheme is the most popular data classification scheme and has been widely used in geography^{49,50}. The natural breaks scheme classifies data into several classes in light of the breaks or gaps in the data⁵¹, which is done by maximizing the between-class variance and minimizing the within-class variance⁵². This data classification scheme follows the principle of data clustering⁵³. It should be noted that when performing natural breaks, the user needs to specify the number of classes.

The head/tail breaks scheme is a new data classification method^{49,51}. It is used to classify data with a heavy-tailed distribution. In a heavy-tailed distribution, the histogram of the data values is right-skewed. A majority of data values are small, and a minority of data values are large, which can also be stated as ‘there are far more small things than large ones’⁵¹. Heavy-tailed distributions have been detected in many social and natural phenomena. The head/tail breaks scheme first divides the data into two parts based on the mean. The part above the mean is called the head, which is the minority. The part below the average is called the tail, which is the majority. The division is continued for the head iteratively until the head part is no longer heavy-tailed⁵⁴. Unlike the natural breaks, the head/tail breaks scheme does not require the user to specify the number of classes.

Delineation of spatial outliers based on the range

Before the following two steps, the shape and size of the neighbourhood need to be determined. Since, in raster data, all the spatial units are uniform in size and neatly arranged, a fixed-size square is selected as the neighbourhood. If a square with side length n is set as a neighbourhood, each group of n rows and n columns of the spatial units (grids) forms a neighbourhood. Then, the second and third steps can be performed for each neighbourhood. Users decide the fixed side length. The fixed-size square slides across the whole research area so that the classes of all the spatial units can be adapted. We also call the fixed-size square a ‘sliding window’, and the sliding step is

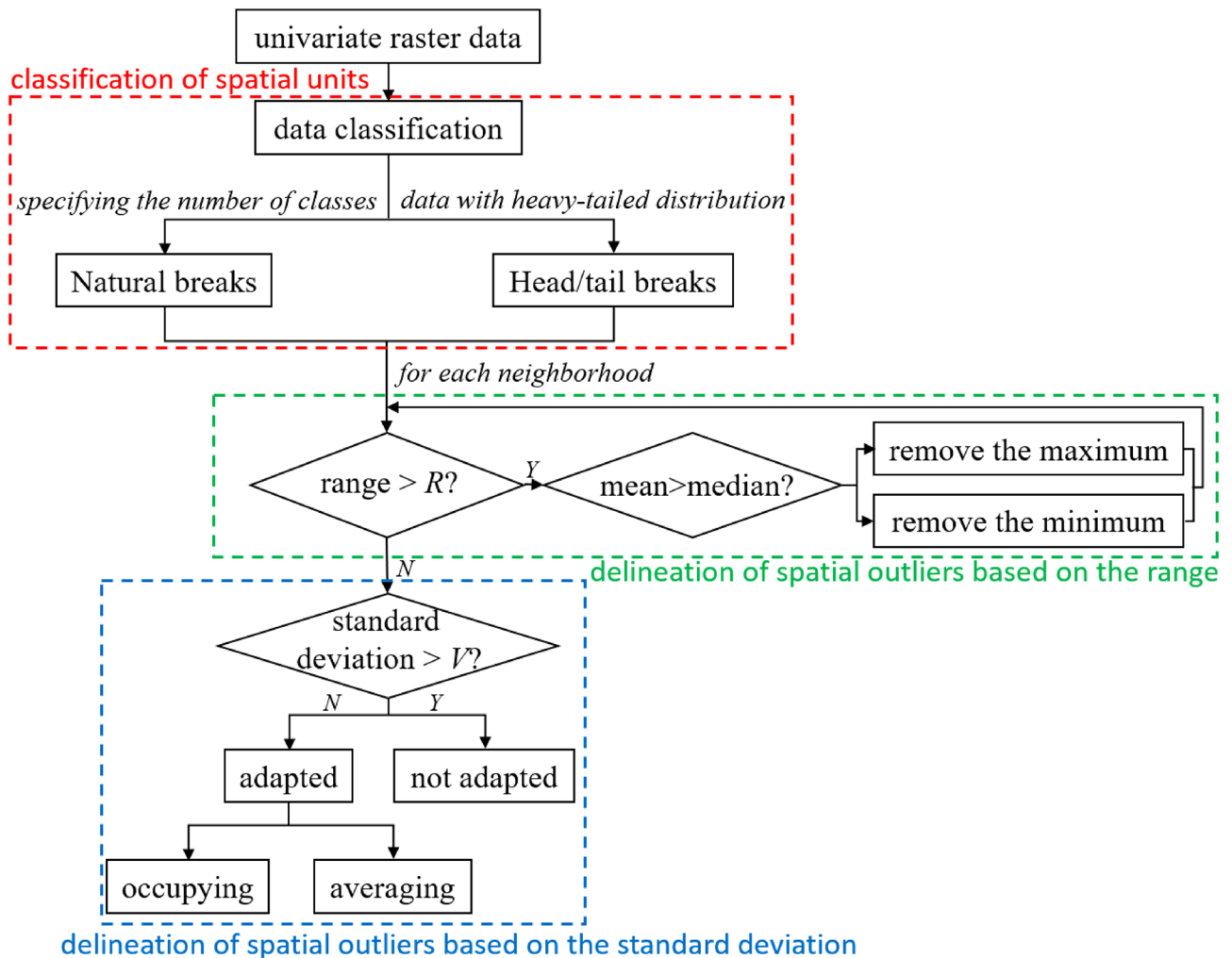


Figure 1. The framework of the proposed method.

one grid. An example of a sliding window is shown in Fig. 2. The whole figure represents the classification of the research area, and the different colours represent different classes. The side length of the sliding window is three in this example, as the black box is shown in the top left corner of the area. Each area consisting of grids of three rows and three columns is a sliding window, such as the grey dashed box and the solid grey box.

Spatial outliers are delineated and preserved based on the range of the attribute values of the spatial units in the sliding window. If the range is too large, it is considered that there are spatial outliers in the sliding window; otherwise, there are no spatial outliers. Spatial outliers will be preserved. The criteria for judging whether the range is large or not is through the threshold R given by the user. If the range is greater than R , the grid with an extreme attribute value, which means either a maximum or minimum, will be removed from the sliding window. If the mean attribute value in the sliding window is greater than the median, the grid with the maximum attribute value should be removed; otherwise, the grid with the minimum value should be removed. Then, the above process is repeated until the range is less than R . The grids removed from the sliding window are spatial outliers whose classes will be preserved; they will not be adapted anymore. The relationship between the number of grids removed from the window and the range threshold is shown in Supplementary Figure S2. Examples of the preservation of spatial outliers are shown in the left half of Fig. 3 and Fig. 4. Two sliding windows are shown in the left subfigures with black borders in Figs. 3 and 4, and the threshold R was set to 70. The numbers in the grids represent the attribute values. The range in the window on the left in Fig. 3 is 85, which is greater than 70. Therefore, there will be a grid removed from the window. Since the mean of the window is greater than the median, the grid with the largest attribute value, 100, should be removed. Then, the range of the remaining attribute values is 54 less than 70, so the process ends. While the range of the window in Fig. 4 is 25, which is less than 70, no grids will be removed. The results are shown in the middle subfigures.

Delineation of spatial outliers based on the standard deviation

Before introducing this step, it is necessary to clarify the concepts. Since we chose data classification schemes as nonspatial clustering methods, to avoid confusion, the results of the data classification (i.e., nonspatial clustering) are called *classes*, and the adapted results (i.e., the results of spatial clustering) are called *clusters*.

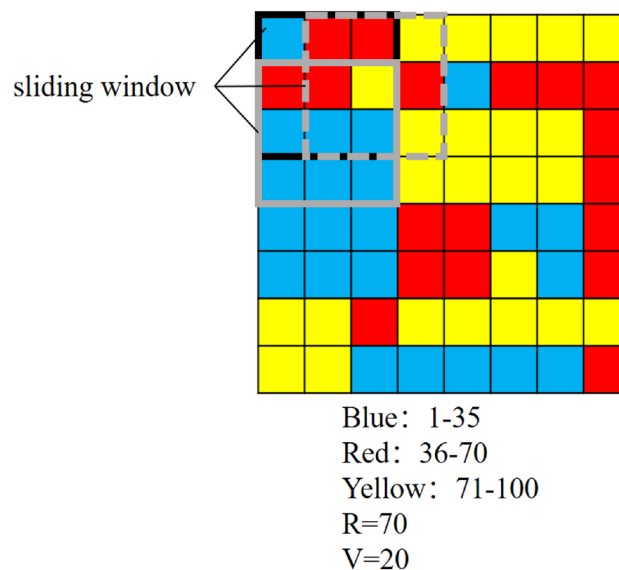


Figure 2. An example of a sliding window.

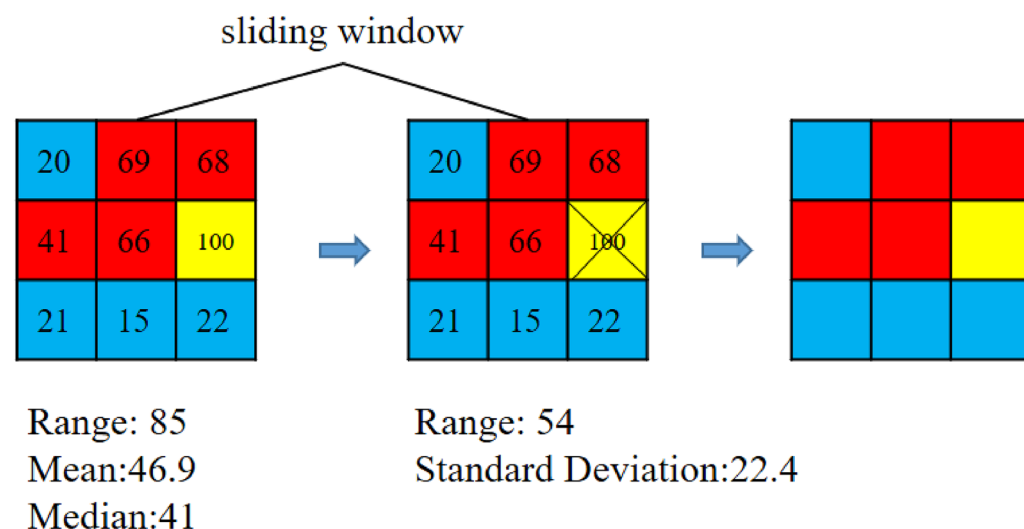


Figure 3. An example of delineation of spatial outliers based on the range.

Spatial outliers are further delineated and preserved based on standard deviation. Similarly, a threshold (denoted as V) is needed to determine whether the standard deviation is large. If the standard deviation of the attribute values in the sliding window is greater than V, which means the attribute values fluctuate, the areas in the window are also considered spatial outliers. Grids in the window cannot be adapted at this point. If the standard deviation is less than V, the grids in the window will be adapted to the same cluster to improve the spatial integration. There are two adaptation mechanisms. The first one is called 'occupying'. Grids will turn into the class most resident in the window in this mechanism. The other is called 'averaging'. In this mechanism, grids will turn into the class according to the mean of the attribute values in the window. This process is illustrated in the right half of Figs. 3 and 4. The threshold V is set to 20. The standard deviation in the sliding window in the middle part of Fig. 3 is 22.4, so these grids cannot be adapted. Since the standard deviation in the sliding window in the middle part of Fig. 4 is 8.1, this sliding window should be adapted. The most resident class is yellow by the occupying mechanism, so the whole window turns yellow. By the averaging mechanism, the mean of the attribute values in the modified window is 70.3, so the corresponding class is red. The results are shown in the right parts of Figs. 3 and 4. The whole process can be likened to an 'infection' and 'quarantine'. The spatial units that are quite different from the surrounding area are 'quarantined' and are not affected by the outside. The rest of the spatial units will be 'infected' with the same type.

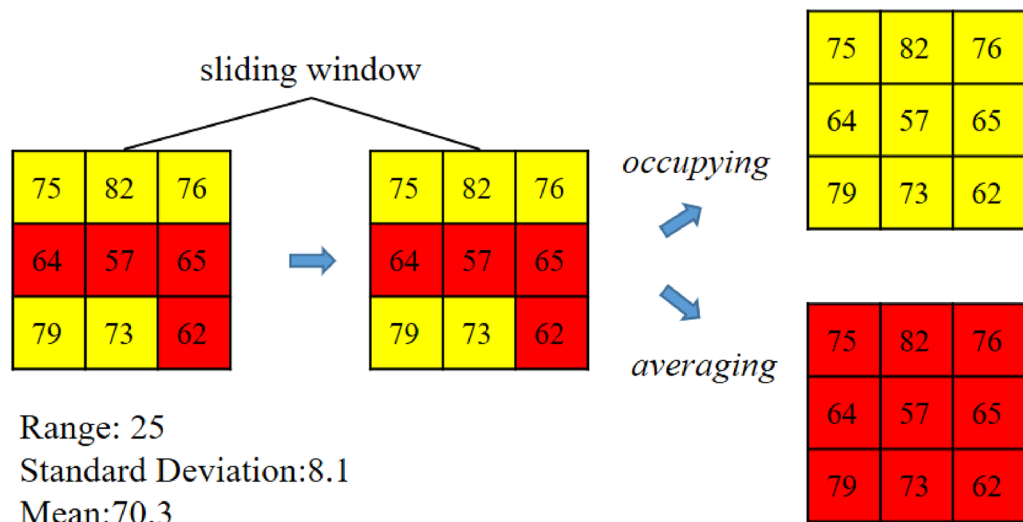


Figure 4. An example of delineation of spatial outliers based on the standard deviation.

Experimental design

To evaluate the robustness to spatial outliers of the proposed method, we designed two experiments on two sets of raster data. Then, the results obtained by the proposed method are compared with the baseline methods.

Experimental data

We selected the DEM data of two districts in Beijing, Changping District and Pinggu District, to evaluate the algorithm; each grid has an attribute value of DEM. The DEM data in Pinggu District are heavy-tailed distributions, while those in Changping District are not heavy-tailed distributions. It can be noted that both districts have large areas of low value (southeastern Changping District and southwestern Pinggu District). To test the robustness of the methods to spatial outliers, we used the method to cluster two representative types of outlier areas that are difficult to delineate, extreme areas and volatility areas, which refer to areas with large and fluctuating differences in attribute values compared to surroundings, respectively. To test the clustering of extreme and volatility areas, the attribute values of several grids in the low-value region were changed. The attribute values of several grids were changed from low values to high values to construct the extreme areas. The attribute values of some grids were changed to random values to construct the volatility areas. The maps after changing the attribute values of the two districts are shown in Figs. 5 and 6.

Baseline method selection

We compared the proposed method with two baseline methods. The first, spatial K-means (SKM for short), is commonly used among the first category of non-strict spatially constrained clustering methods. The second, a Ward-like hierarchical clustering algorithm including spatial/geographical constraints (WHS for short), is a recently proposed method in the second category of non-strict spatially constrained clustering methods. Spatial K-means is an adaptation of K-means clustering that treats the coordinates of each spatial unit as two attributes, and the weights of the coordinates can be selected. For example, if the weight of the coordinates is set to 0.2, the weights of the x and y coordinates are 0.1, and the weight of the nonspatial attributes (i.e., DEM in this experiment) is 0.8. This method has been implemented in GeoDa software. The Ward-like hierarchical clustering algorithm, including spatial/geographical constraints, modifies the measure of dissimilarity between the spatial units in hierarchical clustering. The new dissimilarity is a linear combination of the attribute dissimilarity D_0 and the spatial dissimilarity D_1 . The weight of D_1 is α , and the weight of D_0 is $1 - \alpha$. This method is implemented by the R package ClustGeo. Both methods traded off the attribute similarity and location similarity of the clustering results by adjusting the weights. The third category of non-strict spatially constrained clustering did not participate in the comparison due to its requirements for data distribution.

The proposed method provides two different data classification schemes. We used the head/tail breaks scheme for data with heavy-tailed distribution and the natural breaks for data with nonheavy-tailed. The selection of window size, range threshold R and standard deviation threshold V , and adaptation mechanisms for weak fluctuations can adjust the attribute similarity and location similarity of the clustering results. This paper compares two versions of the new algorithm with the two baseline methods described above.

Evaluation indicators

We have selected three metrics to evaluate the clustering results. The first metric is the Geodetector q statistic⁵⁵, which measures the spatial stratified heterogeneity. Spatial hierarchical heterogeneity refers to the spatial heterogeneity among strata or regions (each stratum or region consists of several units)⁵⁶, specifically, the geographical phenomenon that the variance within a stratum (i.e., 'cluster' in this paper) is smaller than the variance between

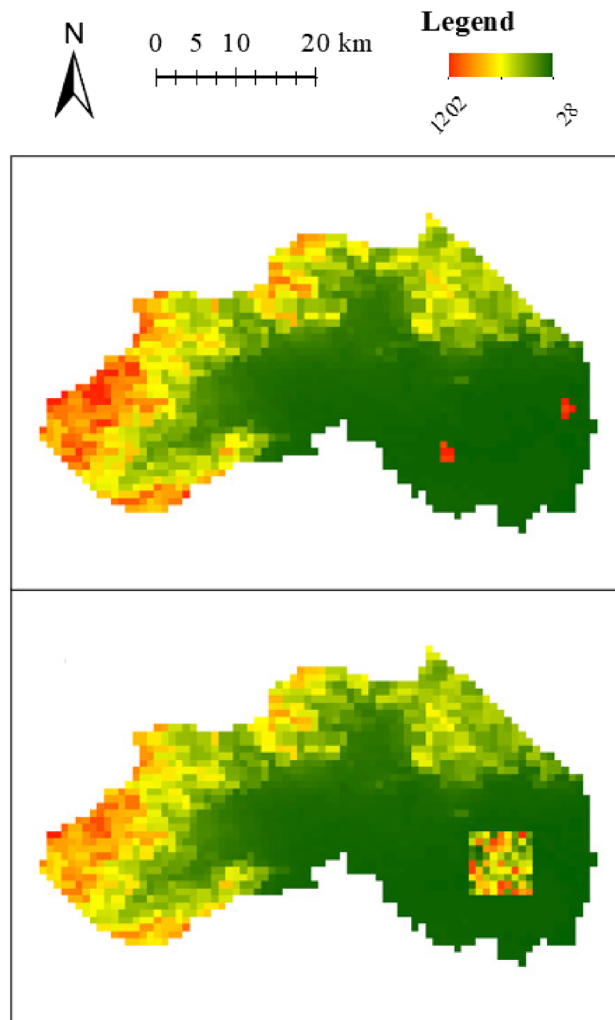


Figure 5. Extreme areas (top) and volatility area (bottom) in Changping District.

strata. Hence, the higher the requirement of spatial integration in clustering, the less significant the spatial hierarchical heterogeneity of the results. The formulation for the q statistic is

$$q = 1 - \frac{\sum_{i=1}^K N_i \sigma_i^2}{N \sigma^2} \quad (1)$$

where $i = 1, \dots, K$ are clusters; N_i and N are the number of units in cluster i and the whole region, respectively; σ_i^2 and σ are the variances in cluster i and the whole region, respectively.

The other two metrics are common clustering evaluation indexes, based on the principles of measuring intra-cluster similarity and inter-cluster dissimilarity. These two metrics are the Silhouette Index and the Davies–Bouldin Index. The formula for the Silhouette index of unit x is as follows,

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}} \quad (2)$$

where $a(x)$ is the average of the attribute distances from unit x to units of the same cluster, and $b(x)$ is the minimum value of the average distance from unit x to all units in a cluster that does not contain it. The value of the Silhouette Index is between $[-1, 1]$, and the closer it is to 1, the better the clustering effect. The formulation for the Davies–Bouldin Index is

$$DB = \frac{1}{n} \sum_{i=1}^n \max_{j \neq i} \left(\frac{d_i + d_j}{d(c_i, c_j)} \right) \quad (3)$$

Here, n represents the number of clusters, c_i is the center of the i -th cluster, d_i is the average distance from all points in cluster i to its center, and $d(c_i, c_j)$ is the distance between the center points c_i and c_j . In the following experiments, we will compare which method obtains a larger Silhouette Index for the same value of q . A smaller DB value indicates a better cluster result.

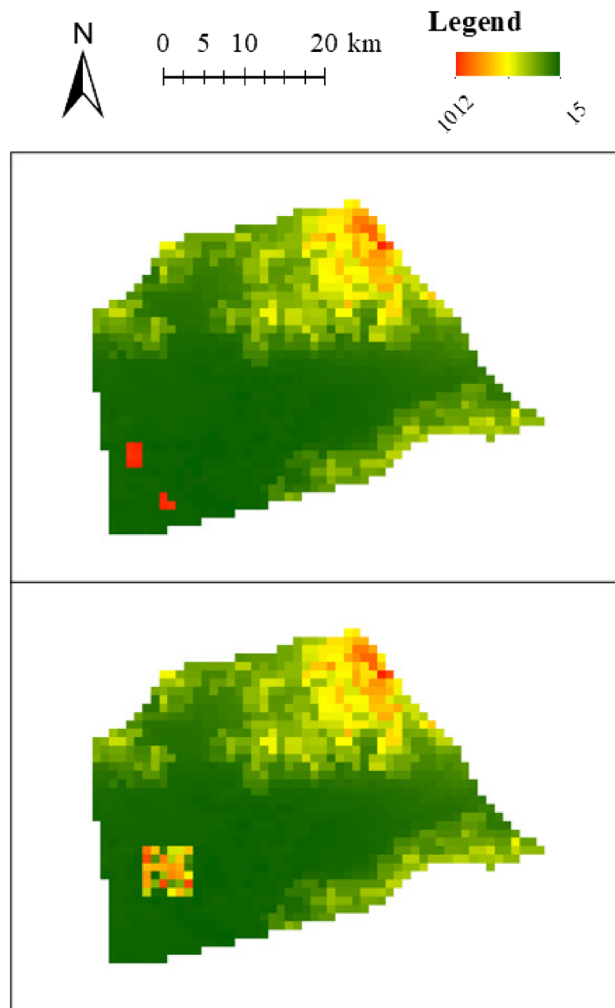


Figure 6. Extreme areas (top) and volatility area (bottom) in Pinggu District.

For non-strict spatially constrained clustering, multiple clustering results can be achieved under various spatial-attribute tradeoffs. To ensure comparability between different clustering results, we aim to obtain results with the same q -value using distinct methods and, in this situation, compare the values of the remaining metrics.

Experimental results

Results of clustering extreme areas

The first experiment is carried out on the DEM of Changping District. Since the DEM data in Changping District have a heavy-tailed distribution, the natural breaks scheme was used for data classification. The number of classes is set to 5 (the default number of classes in ArcGIS). As a result, the number of clusters generated by spatial clustering is also 5. The numbers of clusters for SKM and WHS are also set to 5. For SKM and WHS, we set their location weights α to 0.2, 0.4, 0.6 and 0.8, respectively, and calculate the q of the spatial clustering results under each weight. The q s of the SKM clustering results are 0.908, 0.893, 0.796 and 0.720 in order. For WHS, when the location weight is 0.8, the q is too small, and the extreme areas cannot be divided, so no comparison is made. The q s of the WHS are 0.836, 0.732 and 0.677, respectively. By adjusting the window size, range threshold, and standard deviation threshold of the proposed method, the q s obtained are consistent with the above results. The clustering results of the three methods are shown in Fig. 7 and Supplementary Figure S3.

The robustness to spatial outliers of the proposed method can be seen in the map. For SKM, when the location weight is relatively large (in this experiment the weight is 0.8), it cannot correctly divide the extreme areas. For WHS, the two extreme regions are divided into the same cluster with the northeast. Whereas, according to the DEM values, it is known that their attributes are not similar, which is caused by the location joining the clustering. Instead of improving the spatial integration, this clustering result also weakens the similarity of the attributes. While the proposed method can divide extreme areas in all cases.

The numerical results also show the reasonable clustering of spatial outliers by the proposed method. The metrics of all the clustering results are calculated. The Silhouette Indexes and Davies–Bouldin Indexes corresponding to a q s are shown in Supplementary Figure S4 (for simplicity, NM is used to represent the proposed method, the same as below). The figure shows that the Silhouette Indexes of the proposed method are larger

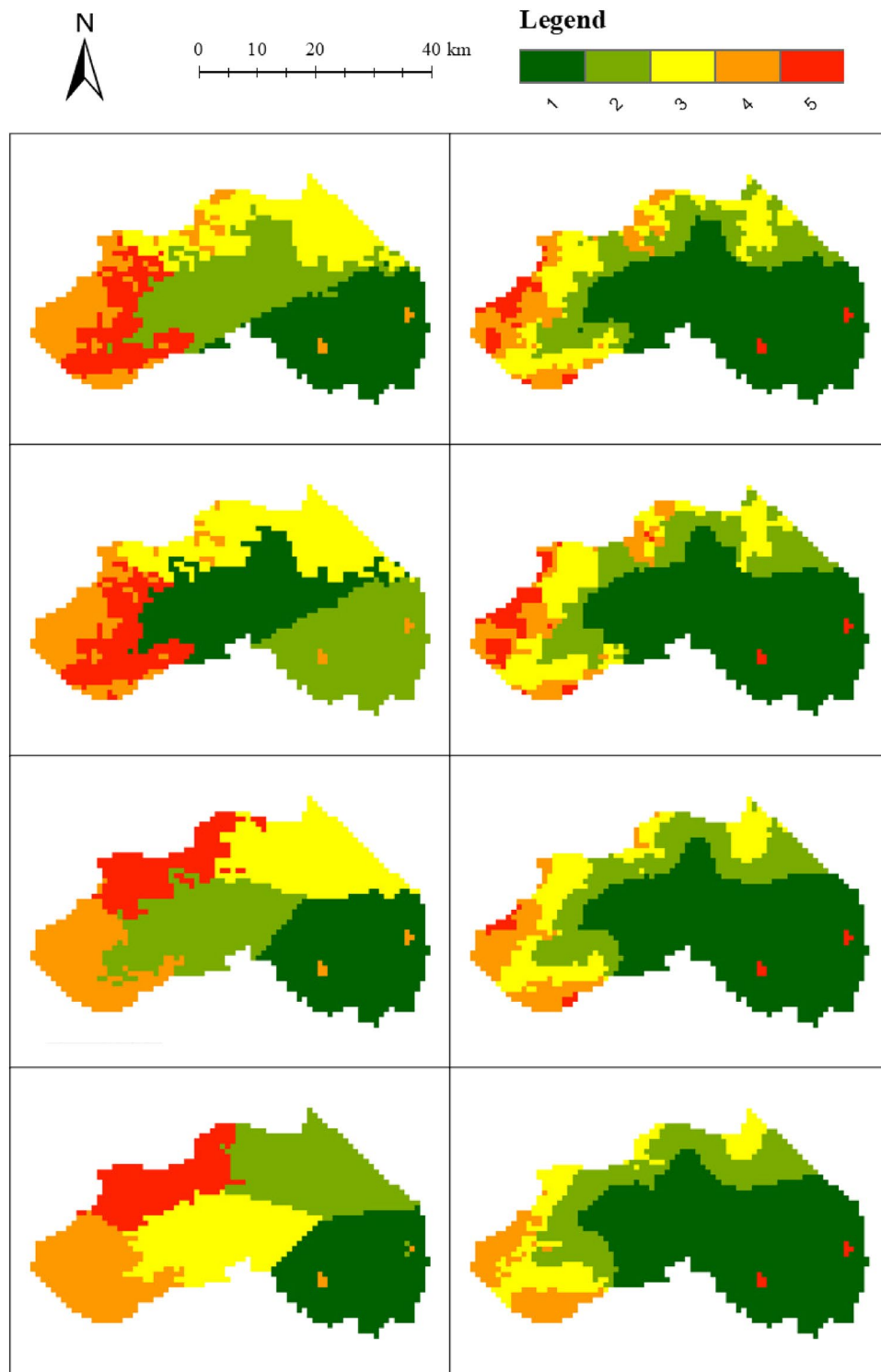


Figure 7. Results of SKM (left side) and NM (right side) for Changping District with extreme areas when q_s are the same (same row).

than those of the two baseline methods, regardless of the q . Additionally, for all values of q , the Davies–Bouldin Indexes of the proposed method are greater than that of the two baselines.

The following experiment was carried out on the DEM data of Pinggu District. Since the DEM in Pinggu District has a heavy-tailed distribution, the head/tail breaks scheme was used for data classification. The number of classes obtained by the head/tail breaks was 6, and the number of clusters obtained after clustering was 6. The numbers of clusters for SKM and WHS are also set to 5. For SKM and WHS, we set their location weights α to

0.2, 0.4, 0.6 and 0.8, respectively, and calculate the q of the spatial clustering results under each weight. The q s of the SKM clustering results were 0.919, 0.897, 0.835 and 0.715, respectively. The q s of the WHS clustering results are 0.857, 0.816, 0.803 and 0.528 in order. By adjusting the window size, range threshold, and standard deviation threshold of the proposed method, the q s obtained are consistent with the above results. The clustering results of the three methods are shown in Figs. 8 and Supplementary Figure S5.

It can be seen from the figures that when the location weight is greater than 0.6, the two baseline algorithms cannot correctly divide the extreme areas. The Silhouette Indexes corresponding to q s are shown in

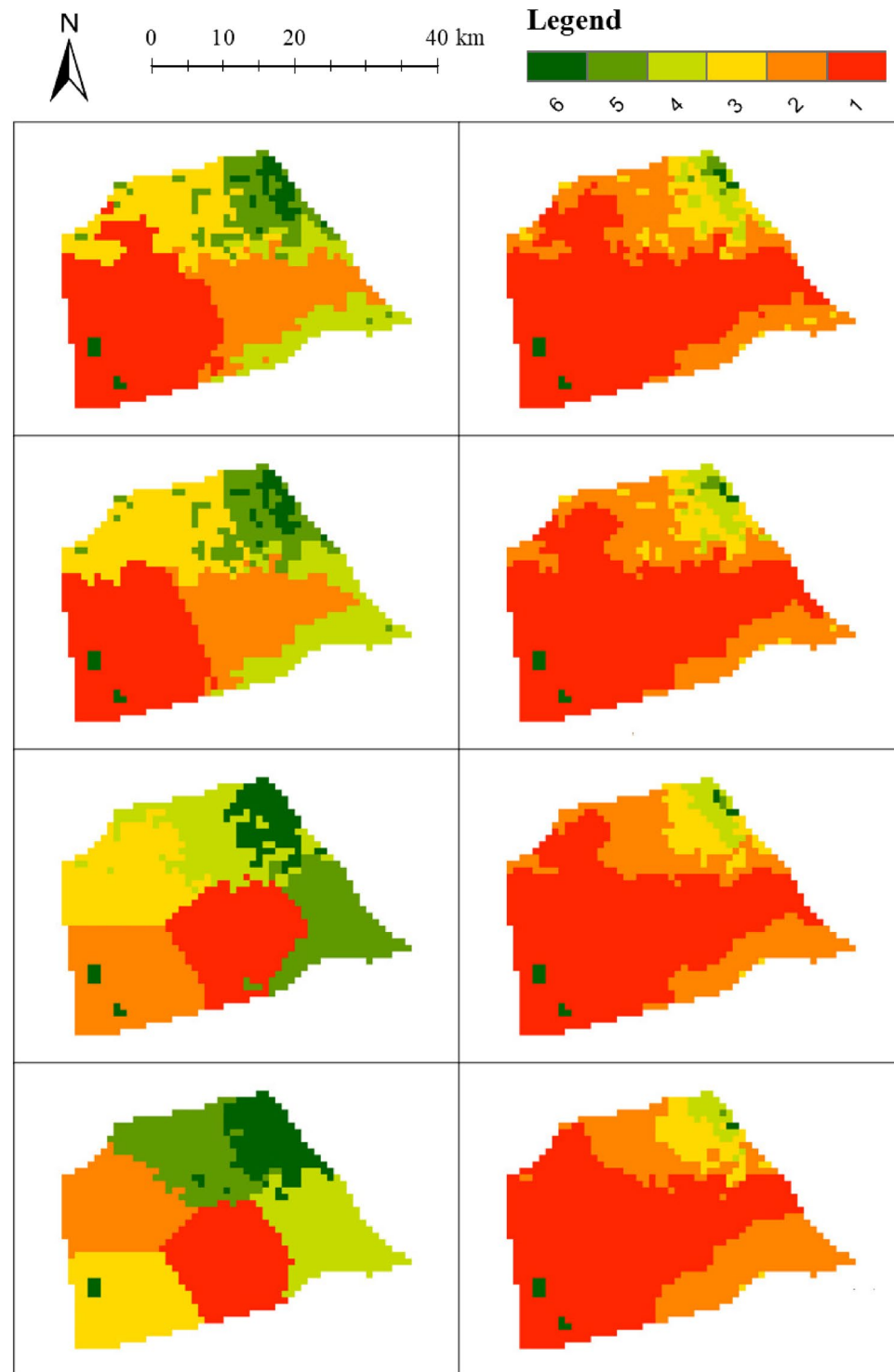


Figure 8. Results of SKM (left side) and NM (right side) for Pinggu District with extreme areas when q s are the same (same row).

Supplementary Figure S6, which shows that the Silhouette Indexes of the proposed method are always larger than those of the two baseline methods. The Davies–Bouldin Indexes of the proposed method are also the best.

Results of clustering volatility areas

Similarly, the DEM of Changping District is tested first. The natural breaks scheme was also used for data classification. The number of classes is set to 5. Additionally, the numbers of classes generated from SKM and WHS were set to 5. For SKM and WHS, we set their location weights to 0.2, 0.4, 0.6 and 0.8, respectively, and calculated the qs of the spatial clustering results under each weight. The qs of SKM are 0.912, 0.890, 0.776 and 0.582 in order. The qs of the WHS are 0.875, 0.819, 0.817 and 0.623, respectively. By adjusting the window size, range threshold, and standard deviation threshold of the proposed method, the qs obtained are consistent with the above results. The clustering results of the three methods are shown in Supplementary Figure S7 and Supplementary Figure S8.

Both baseline algorithms cannot reasonably cluster the volatility areas. When location weight of SKM reaches 0.6, about half of the volatility area merge with the surrounding cluster, and the remaining half are clustered into the same cluster. When the location weight reaches 0.8, the volatility area is completely merged with the surrounding units into one cluster. The result of WHS clusters the volatility area into almost the same cluster at all location weights. And similar to the extreme value region, the volatility area is divided into the same cluster with the northeast. And the proposed method divides the volatility areas to different degrees under various spatial stratified heterogeneity. The Silhouette Indexes shown in Supplementary Figure S9 also reveal that the effect of the proposed method are better. At the same time, in most cases, the Davies–Bouldin Indexes of the proposed method are the best, with only two scenarios where the Davies–Bouldin Indexes of the proposed method are slightly inferior, which correspond to the two worst cases in terms of q .

The following experiment was carried out on the DEM data of Pinggu District. The number of classes obtained by the head/tail breaks was 4. Additionally, the numbers of classes generated from SKM and WHS were set to 4. For SKM and WHS, we set their location weights to 0.2, 0.4, 0.6 and 0.8, respectively, and calculated the qs of the spatial clustering results under each weight. The qs of SKM are 0.890, 0.808, 0.758 and 0.517, respectively. For WHS, when the location weight is 0.8, the q is too large, and the volatility areas cannot be divided, so no comparison is made. For WHS, when the location weight is 0.8, the q is too small, and the extreme areas cannot be divided, so no comparison is made. The qs of the WHS are 0.806, 0.785, 0.527 in order. By adjusting the window size, range threshold, and standard deviation threshold of the proposed method, the qs obtained are consistent with the above results. The clustering results of the three methods are shown in Supplementary Figure S10 and Supplementary Figure S11.

Both baseline algorithms are only able to delineate the volatility areas when the location weights are low. However, when the location weight of SKM reaches 0.8, it almost loses the ability to cluster volatility area, while when the location weight of WHS reaches 0.6, the volatility area are divided into the same cluster with Northwest which is closer rather than with more similar attributes. The Silhouette Indexes and Davies–Bouldin Indexes of the proposed method shown in Supplementary Figure S12 are also better than two baseline algorithms.

Discussion

Regarding the experimental results, we note that the clustering evaluation indexes of the results obtained by the two baseline methods is always worse than that of the proposed method. On the one hand, as previously analysed, the proposed method can better delineate extreme and volatile regions. On the other hand, these two baseline methods tend to divide similar spatial units with similar attributes into different clusters, which is also one of the reasons for the poor performance of the clustering evaluation indexes. The results showed that although non-strict spatially constrained clustering relaxes the enforcement of spatial constraints when compared with strict spatially constrained clustering, the result that similar regions are separated sometimes cannot be avoided. This may be due to the fact that the inclusion of location makes the algorithm tend to generate ‘circular’ clusters, which have the highest similarity in the location of spatial units. In contrast, the shape of the cluster generated by the proposed method is only related to the characteristics of the data itself, and there is no tendency to cluster into a certain shape.

During the clustering process, it is worth noting that the adaptation mechanisms indirectly affect the clustering results. The two adaptation mechanisms have different advantages. From the numerical results, the averaging of q is better than the occupying of q . It is easy to understand because averaging turns the grids in the window into a compromised cluster so that the attribute values of the grids do not differ too much from that cluster. As a result, this method is mathematically optimal. However, occupying expands the area of the cluster in the local area and optimizes the spatial structure of the clusters. Therefore, when performing spatial clustering, averaging is more suitable if it focuses on the mathematical optimization; if it focuses on the optimization of the spatial structure, occupying is more suitable.

The influence of window size on clustering results is also a worthwhile topic to explore. To compare the effects of different window sizes, we designed an experiment. While keeping the range threshold and standard deviation threshold constant, we compared the clustering results under different window sizes. We conducted experiments on the Changping District DEM data with extreme areas and volatility area, respectively. Figures S13 depict the spatial distribution of clustering results. It can be observed that the spatial contiguity of the clustering results initially increases and then decreases. As seen on Figures S13, when the window size is between 4 and 6, spatial continuity is at its highest, and subsequently, clustering results gradually become more fragmented. This is because with a smaller window size, increasing it results in an increase in the number of grids that need to change their categories. However, as the window size continues to increase, the number of grids covered by the sliding window also increases, potentially leading to higher range and standard deviation values. With the range and standard deviation thresholds held constant, the number of grids with changed classes decreases, resulting

in a more fragmented spatial outcome. Therefore, in practical usage, window size, range threshold, and standard deviation threshold need to be carefully coordinated to achieve the desired clustering effect.

Conclusion

This study proposes a non-strict spatial clustering method for univariate raster data. This method uses data classification schemes to classify the research area initially and then obtains each locality of the research area through a sliding window. For all grids in the sliding window, the spatial outliers are delineated by judging their attribute values' range and standard deviation. For windows with a range and standard deviation that are too large, the classes of the initial classification are retained to preserve the spatial outliers. Otherwise, the grids with relatively close distances are classified into one cluster to improve the spatial integration. To prove the effectiveness of the proposed method, we designed a comparative experiment involving the proposed method and two other representative methods. The research areas were selected to facilitate the comparison by including two types of outlier areas, namely, extreme areas and volatility areas. The results demonstrated the advantage of the proposed method; it kept outlier areas in the large-scale contiguous clusters and avoided the problem of similar units being divided into different clusters, whereas the other representative methods did not. Because the proposed method involves the specification of several parameters, pre-experiment debugging is usually required to obtain clustering results that meet users' needs. In future research, we can try to establish the correlation between parameter changes and clustering results to improve the efficiency of parameter setting. By extending the definitions of the range and standard deviation to higher dimensions, the proposed method can be applied to multivariate data.

Data availability

The DEM data used in this study can be obtained from the Resource and Environment Science and Data Center (<https://www.resdc.cn/data.aspx?DATAID=123>).

Received: 16 June 2023; Accepted: 27 January 2024

Published online: 19 February 2024

References

- Fang, C., Liu, H., Luo, K. & Yu, X. Process and proposal for comprehensive regionalization of Chinese human geography. *J. Geog. Sci.* **27**, 1155–1168. <https://doi.org/10.1007/s11442-017-1428-y> (2017).
- Assunção, R. M., Neves, M. C., Câmara, G. & Freitas, C. D. C. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *Int. J. Geogr. Inf. Sci.* **20**, 797–811. <https://doi.org/10.1080/13658810600665111> (2006).
- Grubestic, T. H., Wei, R. & Murray, A. T. Spatial clustering overview and comparison: Accuracy, sensitivity, and computational expense. *Ann. Assoc. Am. Geogr.* **104**, 1134–1156. <https://doi.org/10.1080/00045608.2014.958389> (2014).
- Nigussie, T. Z., Zewotir, T. T. & Muluneh, E. K. Detection of temporal, spatial and spatiotemporal clustering of malaria incidence in Northwest Ethiopia, 2012–2020. *Sci. Rep.* **12**, 3635. <https://doi.org/10.1038/s41598-022-07713-3> (2022).
- Cuadros, D. F. *et al.* Spatial structure of depression in south Africa: A longitudinal panel survey of a nationally representative sample of households. *Sci. Rep.* **9**, 979. <https://doi.org/10.1038/s41598-018-37791-1> (2019).
- Chu, J. *et al.* Female breast cancer mortality clusters in Shandong Province, China: A spatial analysis. *Sci. Rep.* **7**, 1–8. <https://doi.org/10.1038/s41598-017-00179-8> (2017).
- Jiang, F. *et al.* Spatial distribution and clusters of pancreatic cancer mortality in Shandong Province, China. *Sci. Rep.* **9**, 1–7. <https://doi.org/10.1038/s41598-019-49357-w> (2019).
- Chen, D., Jin, G., Zhang, Q., Arowolo, A. O. & Li, Y. Water ecological function zoning in Heihe River Basin, Northwest China. *Phys. Chem. Earth A/B/C* **96**, 74–83. <https://doi.org/10.1016/j.pce.2016.08.005> (2016).
- Xie, D. *et al.* Ecological function zoning of Poyang Lake wetland: A RAMSAR site in China. *Water Policy* **15**, 922–935. <https://doi.org/10.2166/wp.2013.181> (2013).
- Fang, Q., Zhang, L., Hong, H., Zhang, L. & Bristow, F. Ecological function zoning for environmental planning at different levels. *Environ. Dev. Sustain.* **10**, 41–49. <https://doi.org/10.1007/s10668-006-9037-4> (2008).
- Zhang, B. *et al.* A geomorphological regionalization using the upscaled dem: The Beijing-Tianjin-Hebei Area, China case study. *Sci. Rep.* **10**, 10532. <https://doi.org/10.1038/s41598-020-66993-9> (2020).
- Wang, N., Cheng, W., Wang, B., Liu, Q. & Zhou, C. Geomorphological regionalization theory system and division methodology of China. *J. Geog. Sci.* **30**, 212–232. <https://doi.org/10.1007/s11442-020-1724-9> (2020).
- Grubestic, T. H., Wei, R. & Murray, A. T. Spatial clustering overview and comparison: Accuracy, sensitivity, and computational expense. *Ann. Assoc. Am. Geogr.* **104**, 1134–1156. <https://doi.org/10.1080/00045608.2014.958389> (2014).
- Tobler, W. R. A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* **46**, 234–240. <https://doi.org/10.2307/143141> (1970).
- Kou, Y., Lu, C.-T. & Chen, D. Spatial weighted outlier detection. In *Proceedings of the 2006 SIAM International Conference on Data Mining (SDM)*, 614–618. <https://doi.org/10.1137/1.9781611972764.71> (2006).
- Achtert, E. *et al.* Spatial outlier detection: Data, algorithms, visualizations. In *Advances in Spatial and Temporal Databases*, 512–516 (Springer, Berlin, 2011).
- de Bernardo, G., Álvarez-García, S., Brisaboa, N. R., Navarro, G. & Pedreira, O. Compact queriable representations of raster data. In *String Processing and Information Retrieval* (eds Kurland, O. *et al.*) 96–108 (Springer, Cham, 2013). https://doi.org/10.1007/978-3-319-02432-5_14.
- Guyon, I., Von Luxburg, U., Williamson, R. C. & Clustering: Science or art. In *NIPS, Workshop on Clustering Theory*, 1–11 (NIPS Vancouver, BC, 2009).
- Jiang, X., Ma, J., Jiang, J. & Guo, X. Robust feature matching using spatial clustering with heavy outliers. *IEEE Trans. Image Process.* **29**, 736–746. <https://doi.org/10.1109/TIP.2019.2934572> (2020).
- Ren, K., Ye, Y., Gu, G. & Chen, Q. Feature matching based on spatial clustering for aerial image registration with large view differences. *Optik* **259**, 169033. <https://doi.org/10.1016/j.ijleo.2022.169033> (2022).
- Peter, S. J. Minimum spanning tree based clustering for outlier detection. *J. Discrete Math. Sci. Cryptogr.* **14**, 149–166. <https://doi.org/10.1080/09720529.2011.10698329> (2011).
- Galán, S. F. Comparative evaluation of region query strategies for DBSCAN clustering. *Inf. Sci.* **502**, 76–90. <https://doi.org/10.1016/j.ins.2019.06.036> (2019).
- Kim, K. & Cha, S. K. Sibling clustering of tree-based spatial indexes for efficient spatial query processing. In *Proceedings of the Seventh International Conference on Information and Knowledge Management*, 398–405 (1998).

24. Wang, D., Wang, X., Liu, L., Wang, D. & Zeng, Z. Urban signatures in the spatial clustering of precipitation extremes over mainland china. *J. Hydrometeorol.* **22**, 639–656 (2021).
25. Yu, B. *et al.* Object-based spatial cluster analysis of urban landscape pattern using nighttime light satellite images: A case study of China. *Int. J. Geogr. Inf. Sci.* **28**, 2328–2355. <https://doi.org/10.1080/13658816.2014.922186> (2014).
26. Liu, M. *et al.* The spatial clustering analysis of Covid-19 and its associated factors in mainland china at the prefecture level. *Sci. Total Environ.* **777**, 145992. <https://doi.org/10.1016/j.scitotenv.2021.145992> (2021).
27. Gelaw, Y. A., Magalhães, R. J. S., Assefa, Y. & Williams, G. Spatial clustering and socio-demographic determinants of HIV infection in Ethiopia, 2015–2017. *Int. J. Infect. Dis.* **82**, 33–39. <https://doi.org/10.1016/j.ijid.2019.02.046> (2019).
28. Gao, H. Agricultural soil data analysis using spatial clustering data mining techniques. In *2021 IEEE 13th International Conference on Computer Research and Development (ICCRD)*, 83–90. <https://doi.org/10.1109/ICCRD51685.2021.9386553> (2021).
29. Xu, X., Li, J., Wu, C. & Plaza, A. Regional clustering-based spatial preprocessing for hyperspectral unmixing. *Remote Sens. Environ.* **204**, 333–346. <https://doi.org/10.1016/j.rse.2017.10.020> (2018).
30. Yele, V. P., Alegavi, S. & Sedamkar, R. Hybrid hesitant fuzzy linguistic bi-objective binary coyote clustering based segmentation and classification for land use land cover in hyperspectral image. *Int. J. Inf. Technol.* 1–10 (2023).
31. Yuan, S., Tan, P.-N., Cheruvilil, K. S., Collins, S. M. & Soranno, P. A. Constrained spectral clustering for regionalization: Exploring the trade-off between spatial contiguity and landscape homogeneity. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 1–10. <https://doi.org/10.1109/DSAA.2015.7344878> (2015).
32. Chavent, M., Kuertz-Simonet, V., Labenne, A. & Saracco, J. Clustgeo: An R package for hierarchical clustering with spatial constraints. *Comput. Stat.* **33**, 1799–1822. <https://doi.org/10.1007/s00180-018-0791-1> (2018).
33. Duque, J. C., Ramos, R. & Suriñach, J. Supervised regionalization methods: A survey. *Int. Reg. Sci. Rev.* **30**, 195–220. <https://doi.org/10.1177/0160017607301605> (2007).
34. Openshaw, S. A regionalisation program for large data sets. *Comput. Appl.* **3**, 136–147 (1973).
35. Recchia, A. Contiguity-constrained hierarchical agglomerative clustering using SAS. *J. Stat. Softw. Code Snipp.* **33**:1–12. <https://doi.org/10.18637/jss.v033.c02> (2010).
36. Ambroise, C., Dang, M. & Govaert, G. Clustering of spatial data by the EM algorithm. In *geoENV I - Geostatistics for Environmental Applications* (eds Soares, A. *et al.*) 493–504 (Springer, Dordrecht, 1997).
37. Fouedjio, F. A hierarchical clustering method for multivariate geostatistical data. *Spatial Stat.* **18**, 333–351. <https://doi.org/10.1016/j.spasta.2016.07.003> (2016).
38. Gould, P. Essays on commodity flows and the spatial structure of the Indian economy. *Geogr. Rev.* **58**, 158–161. <https://doi.org/10.2307/212844> (1968).
39. Jain, A. K. & Farrokhnia, F. Unsupervised texture segmentation using Gabor filters. *Pattern Recogn.* **24**, 1167–1186. [https://doi.org/10.1016/0031-3203\(91\)90143-S](https://doi.org/10.1016/0031-3203(91)90143-S) (1991).
40. Murray, A. T. & Shyy, T.-K. Integrating attribute and space characteristics in choropleth display and spatial data mining. *Int. J. Geogr. Inf. Sci.* **14**, 649–667. <https://doi.org/10.1080/136588100424954> (2000).
41. Oliver, M. A. & Webster, R. A geostatistical basis for spatial weighting in multivariate classification. *Math. Geol.* **21**, 15–35. <https://doi.org/10.1007/BF00897238> (1989).
42. Distefano, V., Mameli, V. & Poli, I. Identifying spatial patterns with the bootstrap clustgeo technique. *Spatial Stat.* **38**, 100441. <https://doi.org/10.1016/j.spasta.2020.100441> (2020).
43. Vignes, M., Blanchet, J., Leroux, D. & Forbes, F. SpaCEM3: A software for biological module detection when data is incomplete, high dimensional and dependent. *Bioinformatics* **27**, 881–882. <https://doi.org/10.1093/bioinformatics/btr034> (2011).
44. Vignes, M. & Forbes, F. Gene clustering via integrated Markov models combining individual and pairwise features. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **6**, 260–270. <https://doi.org/10.1109/TCBB.2007.70248> (2009).
45. Ngunjiri, M. W., Libohova, Z., Owens, P. R. & Schulze, D. G. Landform pattern recognition and classification for predicting soil types of the Uasin Gishu Plateau, Kenya. *CATENA* **188**, 104390. <https://doi.org/10.1016/j.catena.2019.104390> (2020).
46. Xu, C. *et al.* Developing a quantitative landscape regionalization framework integrating driving factors and response attributes of landscapes. *Landscape Ecol. Eng.* **10**, 295–307. <https://doi.org/10.1007/s11355-013-0225-8> (2014).
47. Lyu, B., Wu, W. & Hu, Z. A novel bidirectional clustering algorithm based on local density. *Sci. Rep.* **11**, 14214. <https://doi.org/10.1038/s41598-021-93244-2> (2021).
48. Schubert, E., Zimek, A. & Kriegel, H.-P. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. *Data Min. Knowl. Disc.* **28**, 190–237. <https://doi.org/10.1007/s10618-012-0300-z> (2014).
49. Basofi, A., Fariza, A., Ahsan, A. S. & Kamal, I. M. A comparison between natural and head/tail breaks in LSI (landslide susceptibility index) classification for landslide susceptibility mapping: A case study in Ponorogo, East Java, Indonesia. In *2015 International Conference on Science in Information Technology (ICSITech)*, 337–342. <https://doi.org/10.1109/ICSITech.2015.7407828> (2015).
50. Khamis, N., Sin, T. C. & Hock, G. C. Segmentation of residential customer load profile in peninsular Malaysia using Jenks natural breaks. In *2018 IEEE 7th International Conference on Power and Energy (PECon)*, 128–131. <https://doi.org/10.1109/PECON.2018.8684113> (2018).
51. Jiang, B. Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution. *Prof. Geogr.* **65**, 482–494. <https://doi.org/10.1080/00330124.2012.700499> (2013).
52. Fariza, A., Rusydi, I., Hasim, J. A. N., Basofi, A. Spatial. & flood risk mapping in East Java, Indonesia, using analytic hierarchy process—natural breaks classification. In *2nd International conferences on Information Technology. Information Systems and Electrical Engineering (ICITISEE)*, 406–411 (2017). <https://doi.org/10.1109/ICITISEE.2017.8285539> (2017).
53. Lu, Y., He, T., Xu, X. & Qiao, Z. Investigation the robustness of standard classification methods for defining urban heat islands. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **14**, 11386–11394. <https://doi.org/10.1109/JSTARS.2021.3124558> (2021).
54. Zhang, H. & Wu, Z. A head/tail breaks-based method for efficiently estimating the absolute Boltzmann entropy of numerical raster data. *ISPRS Int. J. Geo Inf.* **9**, 103. <https://doi.org/10.3390/ijgi9020103> (2020).
55. Wang, J., Zhang, T. & Fu, B. A measure of spatial stratified heterogeneity. *Ecol. Ind.* **67**, 250–256. <https://doi.org/10.1016/j.ecolind.2016.02.052> (2016).
56. Wang, J. *et al.* Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region, China. *Int. J. Geogr. Inf. Sci.* **24**, 107–127. <https://doi.org/10.1080/13658810802443457> (2010).

Author contributions

C.S. conceived the study, H.W. performed the study and wrote the manuscript, and W.J. and P.G. analyzed the results. All authors reviewed the manuscript.

Funding

Funding was provided by National Natural Science Foundation of China (Grant Nos. 42230106, 42271418), Second Tibetan Plateau Scientific Expedition and Research Program (Grant No. 2019QZKK0608), and State Key Laboratory of Earth Surface Processes and Resource Ecology (Grant Nos. 2022-ZD-04 and 2023-WT-02).

Competing interests

The authors declare no competing interest. The corresponding author is responsible for submitting a competing interests statement on behalf of all authors of the paper.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-53066-4>.

Correspondence and requests for materials should be addressed to C.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024