Extraction of terrain ridge lines and valley lines based on agglomeration analysis of terrain points: a cluster analysis method

Zhang Cheng
] Information Construction Department of Jiangsu Open University
Dou Wanfeng (douwanfeng@njnu.edu.cn)
Nanjing Normal University
Pang Yuan
Nanjing Normal University

Research Article

Keywords: terrain feature line extraction, DBSCAN clustering algorithm, ridge line, valley line

Posted Date: September 16th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2054890/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Earth Science Informatics on January 24th, 2023. See the published version at https://doi.org/10.1007/s12145-023-00943-7.

Abstract

Terrain feature extraction is one of the critical issues in geographic information science. As important terrain feature lines, ridge lines and valley lines, play an important role in hydrological analysis, terrain reconstruction and automatic integration of contour lines. But, the extraction of terrain feature lines is complicated and time-consuming task. In this paper, a terrain feature line extraction method is proposed based on clustering technique. The terrain feature points are automatically extracted according to the agglomeration of terrain points, and the similar points are automatically identified according to the DBSCAN clustering algorithm. The points with high similarity are clustered along the direction of ridge or valley, and the whole terrain will be clustered into multiple sub-regions. The nearest sub-regions are found by calculating the minimum distance between these sub-regions, and the adjacent sub-regions are connected orderly by their center line to obtain terrain feature lines. Compared with other methods, the cluster analysis method in this paper has simple process and high efficiency.

1. Introduction

Terrain feature extraction (Hu, 2017) is an important research topic in GIS (geographical information science). Ridge lines and valley lines, as important terrain feature lines, play an important role in the simplification of topographic model, sample-based topographic generalization, topographic and geomorphic research, hydrological analysis, terrain reconstruction, contour automatic synthesis, and so on. But, the extraction of terrain feature lines is complicated and time-consuming task, how to reduce the computing complication and increase accuracy will be a significant research problem.

There are three types of existing DEM-based ridge and valley line extraction algorithms: methods based on image processing technology to detect curve structures, algorithms based on terrain surface flow simulation, and algorithms based on surface geometry analysis. The method based on image processing techniques (Peucker and Douglas, 1975; Pang et al., 2013) applied the method of curve structure detection in images to DEM feature line detection, but it was sensitive to noise data, and branch connections of ridges and valleys without obvious linear features were prone to fracture. A term ‘structualist’ is used to develop ridge and valley line extraction from digital images based on lines drawn by moving under logical constraints in the image, starting from previously selected points (Riazanoff et al., 1998). A mathematical framework based on grey level morphological transformation is developed to extract the ridge and valley connectivity networks to understand spatial distribution (Sagar et al., 2003).

The algorithm based on the physical simulation of surface water flow (O'Callaghan and Mark, 1984) determines the water flow direction and water catchment amount at each point by simulating the surface water flow process, and connects the feature points with water catchment amount is greater than the threshold to form a valley line, and takes the boundary of the catchment area as the boundary of the watershed area. Ridge line, the-over-reality of the feature line extracted by this algorithm is a suitable method at present. However, the extracted ridge line is a closed curve, which is not in line with reality, and it is difficult to judge the direction of water flow in a flat area. The feature line extraction based on topographic surface flow analysis is more accurate comparing with other two methods. It is difficult to connect feature lines in geometric form analysis, and terrain points are often omitted when extracting feature points by surface flow simulation method. Ai (Ai and Li, 2010) develops a structured analytical approach to generalize DEM data by identifying small valleys and filling in corresponding depression locations. But the water system pattern and distance between adjacent valleys are not considered. According to their hydrological significance, unimportant river valley branches are detected and their cover is filled by raising the terrain, making the terrain surface smoother, this method can effectively preserve the main geographical features of the terrain, but they only used valley cover area as a decision-making factor, and also considered valley length, layer order, and valley density, the water system pattern and distance between adjacent valleys are not considered. Extracting topographic feature lines from point cloud is proposed based on SSV(signed surface variation) and HC-Laplacian smoothing method(Zhou et al., 2016), in which the potential feature points are segmented into different clusters by region growing based on the Euclidean distance and SSV.

Some automated algorithms and software of the extraction of ridges or ridge axes from DEMs are still not practical for the applications. Most researchers tend to design specific lineament extraction algorithms without a generic solution. The main problem is that the process of axis thinning and segment connection are still too complicated to have a universe solution (Chang and Sinha, 2007). Due to DEM solution and production errors many of problems for automated feature recognition algorithms are minimized by manual feature detection to overrule local inconsistencies for preserving the global trend to avoid false truncations and fragmented lineaments. For extracting ridge and valley profiles of mountains from amorphous point cloud data, a projection-based spatial morphological extraction framework is proposed for detecting mountain profile (Maurya et al., 2018). They consider that the membership, neighborhood, and cohesion between points are fundamental problems and should be examined to extract ridge and valley profiles from the point cloud surface data.

The traditional method based on image technology is easy to be affected by noise data, and the lines without obvious features will be broken. The feature line extracted by the method based on surface water flow is a closed curve, which is not in line with the reality. The method based on surface geometry analysis is easily affected by the size of the window, and it needs several iterations for small undulating terrain. To overcome the above problems, this paper presents a feature line extraction method based on clustering analysis. The idea of this method is based on the spatial clustering characteristics of terrain points with ridges and valleys. The method is simple, easy to operate, and has a fast extraction speed.Through the clustering of the original terrain data, the abscissa, ordinate and elevation of each terrain point are taken as its three-dimensional space, and the terrain coordinate points are clustered and analyzed. According to the clustering analysis and evaluation results, the clusters close to each other are connected to form a feature line. Finally, the elevations of classes in the feature lines are compared to extract ridge and valley lines. The experimental results show that the accuracy of our method can reach the requirements of terrain feature line extraction and computational efficiency is satisfactory.

2. The Idea And Principle Of Terrain Feature Line Extraction

The topographic ridge lines and valley lines (Chang et al., 1998; Masavi et al., 1999) are curves connected by a group of continuous ridge points or valley points, it represents a topographic boundary that intersects between slopes, which can describe the characteristics of topographic fluctuation change and pattern distribution, it is of great importance for the analysis of topographic form and trend. On the other hand, the topographic points on the slopes on both sides of a topographic feature line have local agglomeration. That is, the topographic points in the vicinity of topographic ridge points or valley points are relatively near in spatial distance, and they can form a cluster close to the ridge points or valley points, with the ridge points or valley points serving as the cluster's center. Of course, topographic feature points like peak points and depression sites have similar qualities, but they're generally isolated and can't be used to construct terrain feature lines.

Cluster analysis (Blashfield and Aldenderfer, 1978; Busu et al., 2021) divides a given data set into several subsets, and each subset is called a cluster, so that the distance between data objects or data points in the same cluster being extremely close or the similarity being extremely high, while the distance between objects in different clusters being extremely far or the similarity being extremely low. Three-dimensional coordinates (x, y, z) can be used to represent each terrain point, where x and y are ground coordinate points and z represents the point's elevation. When the terrain is represented by a DEM data set of regular grid, the terrain points can be represented by grid points in terrain data set. The spatial distance between terrain points is depicted as a three-dimensional Euclidean geometric distance, the distance between two points is used to evaluate their similarity. Set a terrain point to define a neighborhood, calculate the agglomeration of the point and other points in the neighborhood, to determine whether they will form a cluster, or by dividing the entire terrain point according to distance, until it cannot be split, multiple clusters can be obtained. Then we can gain terrain ridge lines or valley lines through first finding the extreme point of each cluster, then calculating the point with the farthest distance from the extreme point on both sides of the cluster, forming the center line of the geometric shape of each cluster, and finally connecting these center lines by head-to-tail ligation.

Cluster analysis is a popular data mining technique. Its most notable characteristic is that the data set does not require prior information (that is, the category identifier of the object). It requires directly starting from the data set itself and assigning a category label to each object in the data set according to some similarity criterion. As a result, cluster analysis methods are also known as unsupervised classification methods. Cluster analysis is critical in assisting individuals in obtaining potential and valuable information while filtering out irrelevant data.

For DEM terrain data analysis on a regular grid, the similarity between any two terrain points can be expressed by their Euclidean distance.

Cluster analysis is not only related to the data set, but also to the similarity measure it chooses, and the choice of the similarity measure is still a challenging problem to this day. The shape of the cluster obtained by the cluster analysis is an important reference for the selection of the clustering algorithm. Usually, the clusters have two types: spherical (convex) and non-spherical. The clusters generated by the distance function are generally spherical. However, the shapes of clusters are various, especially when the shapes of clusters are irregular, it is necessary to use density-based cluster analysis.

Density-based clusters consist of relatively dense regions between objects surrounded by regions of low density. This is generally achieved by specifying the minimum number of points around any object in the cluster. When clusters are irregular in shape or coil around each other, and there are noise points and outliers, density-based cluster definitions often yield better clustering results.

The idea of density-based clustering is that, as long as the density of objects in a region is greater than a certain threshold, they will be added to the clusters that are close to it. The algorithm can overcome the shortcoming that distance-based algorithms can only find "circular-like" clusters, and it can find clusters of any shape and is not sensitive to noise data. The DBSCAN (density-based spatial clustering of applications with noise) is a representative density-based clustering algorithm, it is a partial clustering algorithm, that is, the union of all clusters obtained by clustering cannot cover the data set itself. Its characteristic is that it defines a cluster as the largest set of all density-connected symmetrical structures. That means it can cluster the regions with high enough density into a cluster and find clusters of arbitrary shapes in noisy datasets.

The shape of the clusters formed by terrain points is irregular and has various shapes, so the density-based clustering algorithm is more suitable for the clustering analysis for DEM-based terrain data. In this paper, the DBSCAN-based clustering algorithm is used to extract terrain feature points, then ridge lines and valley lines are inferred from the feature points.

3. The Dbscan Clustering Algorithm

DBSCAN clustering (Schubert et al., 2017; Khan et al., 2014) is a local clustering method. The union of all clusters cannot completely cover the data set itself. The noisy data in the data set is excluded by the algorithm and cannot form a cluster or join other clusters. The idea of the DBSCAN algorithm is to form a cluster of all objects connected by density. Before the operation, two parameters need to be input: neighborhood radius and minimum points (Minpts). A cluster can be formed only when the number of points in the neighborhood exceeds the minimum number of samples.

Given a data set S and a neighborhood radius ε, for any data object X∈S, $\epsilon\left(\mathrm{X}\right)=\left\{\mathrm{Y}\middle|\,\mathrm{Y}\in\mathrm{S},\mathrm{d}\left(\mathrm{Y},\mathrm{X}\right)\leq\epsilon\right\}$, Y is in the neighborhood with X as the core and ε as the radius, ε(X) is called the ε-neighborhood of X in S. Here d (Y, X) represents the distance between X and Y.

Given parameters (ε, Minpts) and an arbitrary data object X∈S, if |ε (X)|>= Minpts, then X is said to be a core point of S about (ε, Minpts).

Any X, Y∈S, if X is a core point in the data set S, Y∈ε(X), Y is in the neighborhood of X, then Y is directly density-reachable from X with respect to (ε, Minpts).

If there are data objects X and Y in S, there are data columns X_1, X_2, X_3, X_4, where X = X_1, and from X_1 to X_2, X_2 to X_3, X_3 to X_4, X_4 to Y are directly density reachable, then X to Y are density reachable with respect to (ε, Minpts).

For any data object Z, Y∈S, if there is a data object X∈S, so that both X and Z, Y is density-reachable, then Z and Y are density-connected with respect to (ε, Minpts).

For any Y∈S, if Y is not a core point about (ε, Minpts), but there is a core point X, let Y∈ε(X), Y is said to be the boundary point of S.

For any Y ∈ S, if Y is neither about (ε, Minpts) core points nor boundary points, then Y is the noise point of S about (ε, Minpts).

The basic idea of the DBSCAN algorithm is to find a cluster generated through examining the ε-neighborhood of each point X in the data set S. For each X ∈ S, if ε(X) > = Minpts, create a new cluster with X as the core point and merge all objects density-reachable from X into this new cluster until there are no new ones. When a point can be added to this cluster, the next point in S is checked until every point in S has been checked. In this process, if some points checked later already belong to a previously generated cluster, there is no need to generate a new cluster for this point.

The process of the DBSCAN algorithm is to search for clusters by checking the neighborhood of each point in the data set. During the operation, all data needs to be traversed. First, each point in the data set is marked as unvisited, when a point is visited, it is marked as visited, then how many points are calculated in its neighborhood according to the spatial distance calculation formula (1). If the number of points is greater than the minimum number of sample points, the point is a core point. If it is not the core point, the algorithm continues to traverse the next point unvisited in the data set. If it is a core point, the points in its neighborhood are classified and a cluster with the core point is created as the core region, and this cluster is classified as a new data gathered region. The points in its neighborhood are accessed until all points are visited.

4. Topographic Feature Line Extraction Method

4.1 Topographic feature line extraction process

To extract ridge lines and valley lines based on the DBSCAN algorithm, it is necessary to rasterize the terrain data and convert it into text data that can be read by the algorithm. We select the two parameters of the DBSCAN algorithm through a certain strategy for clustering analysis and study the terrain features according to the clustering results further.

The specific steps of terrain feature line extraction are as follows:

Step1: Terrain data preparation.

The DEM data for the terrain is obtained, and the raster data is converted into text format and read by the algorithm through a series of processing steps in ArcGIS software.

Step 2: Terrain data cluster analysis.

With the help of the clustering algorithm, the terrain is divided into multiple separated sub-regions $\mathrm{SV_i}\ (\mathrm{i} = 1,2,\cdots,\mathrm{K})$, K is the number of sub-regions.

The DBSCAN clustering algorithm is used for terrain clustering division, and its similarity is expressed by Euclidean distance. The Euclidean distance between terrain grid elements can be calculated by the formula (1). The judgment rule is: when $d \leq \varepsilon$ (ε is the set threshold), and P_i is a point in sub-region V_i, P_j does not belong to the point in sub-region V_i, $V_i=V_i \cup \{P_j\}$, where d is the distance between two points.

The DBSCAN clustering algorithm needs to input two parameters, the minimum number of sample points (represented by Minpts) and the neighborhood radius (represented by ε), which can be determined according to the formation principle of terrain feature lines and the method of traversal analysis. The details are introduced in Section 4.2.

Step 3: Extraction of ridge points and valley points.

The terrain feature points are determined by calculating the extreme points of the sub-area. At this time, two points will be selected in the sub-area, that is, the maximum value point and the minimum value point. The point closest to the center of the sub-area is retained, and the other point is deleted, and take the reserved extreme points as the feature points of this sub-region.

The calculated feature points are the ridge points and valley points in the terrain, the ridge point is the feature point of the ridge line, which is at a high altitude, and the valley point is the feature point of the valley line, which is at a low altitude.

Step 4: Feature line extraction to determine ridge lines or valley lines.

Usually the sub-area is distributed in a certain direction. Calculate the points on both sides of the sub-area that are farthest from the extreme point, draw the center line of the sub-area, and pass through these three points. The simulated line can reflect the characteristics of the sub-area line distribution.

Calculate the minimum distance between each sub-region and find adjacent sub-regions. The minimum distance between sub-regions is the smallest of any two grid distances in the two sub-regions. Connect the two points with the minimum distance between adjacent sub-regions, and the connecting line can represent the feature line between the sub-regions. Combining the connecting lines within the sub-regions and between the sub-regions, a completed feature line is formed.

The minimum distance between sub-regions can be expressed as formula (1). A_i and A_j are two sub-regions, and the distance between any two grid points, $X \in A_i$, $Y \in A_j$, is d(x, y). The smallest distance between any two grid points is defined as the distance measured between the two sub-regions,

$$\mathrm{ds}\left(\mathrm{A}_i, \mathrm{A}_j\right) = \min\left\{\mathrm{d}\left(\mathrm{X}, \mathrm{Y}\right) \middle| \mathrm{X} \in \mathrm{A}_i, \mathrm{Y} \in \mathrm{A}_j\right\} \quad (1)$$

4.2 Parameter determination of the DBSCAN algorithm

In the DBSCAN algorithm, whether or not the object in the data set *S* is a core point depends on the density parameter (ε, *Minpts*) composed of the neighborhood radius and the minimum number of samples. The DBSCAN algorithm divides the regions that reach or exceed the density generation into clusters, it can find the clusters of arbitrary shape in the data set with "noisy", and then form a partial cluster of *S*. It is worth noting that the DBSCAN algorithm is sensitive to the user-defined density of (ε, *Minpts*). That is, the settings of the parameters of ε and *Minpts* will directly affect the clustering results.

For regular grid-based DEM terrain datasets, *Minpts* represents the minimum number of grid cells within the neighborhood radius that form a core point. Here, we set the minimum number of grid cells to form a core point to 2, that is to say, if there are 2 sample points in the neighborhood, a core point will be formed.

After the *Minpts* are determined, the value of the neighborhood radius ε can be determined by the method of dynamic traversal. First, the initial value of ε is determined, the lower limit of ε, which is calculated as *MinEPS*. Assuming that the resolution of the terrain is R×R, the initial value of ε is determined to be R meters, then *MinEPS* = R. Next, the upper limit of the neighborhood radius ε needs to be determined. In the DBSCAN algorithm, the ε is used to determine the threshold for whether a grid becomes a cluster, which is essentially a measure of the degree of smoothness between adjacent grids. When adjacent grids can gather together to form a cluster, it means that they should be relatively smooth, which means the slope between the grids is not very steep. Generally, a slope exceeding 45 degrees is considered to be very steep, so it is

determined that the upper limit of the slope is 45 degrees, the center distance between two adjacent grids can be calculated by the formula (2).

D = R/cos(π/4)=$\sqrt{2}$R (2)

Therefore, the upper limit of ε is $\sqrt{2}$R. In this way, we have determined that the traversal range of ε is [R, $\sqrt{2}$R].

After determining the value range of the neighborhood radius, we can determine the optimal minimum number of sample points by traversing the value of the neighborhood radius. The method of determination is that when the minimum number of sample points reaches a certain value, the number of noise data and the number of clusters do not change with the change of the neighborhood, then the minimum number of sample points is the current value minus 1.

The minimum number of sample points is determined, the number of clusters within the fixed domain radius is calculated, and the neighborhood radius is determined by calculating the change rate of the cluster with the adjacent domain. The neighborhood radius with the absolute value of the smallest change rate is selected as the best value. At this time, the influence on the formation of the feature line is minimal. The calculation method of the change rate is shown in formula (3). Assuming that the adjacent domain radii are e_i and e_j, the number of corresponding clusters are C_i and C_j, and the absolute value of the change rate is expressed by T_{i-j}, then the calculation formula is as follows:

$$T_{i-j} = \left| \frac{C_j - C_i}{e_j - e_i} \right|$$

3

In the process of terrain feature line extraction, to exclude the useless data of the hillside position, the DBSCAN algorithm can also be used to complete. The specific method is to set a relatively large minimum sample point according to the optimal neighborhood radius value to ensure that the points at the top of the mountain can be gathered, and the points at the hillside are identified as noise data.

5. Experimental Analysis And Evaluation

5.1 DBSCAN extracts feature lines

Figure 2 is a mountainous terrain area adopted in this experiment. The resolution of the terrain based on DEM is 30 × 30 meters, and the elevation range is between 1327 meters to 3596 meters. It is preprocessed by ArcGIS software and converted into grid text format. The grid text data were subjected to the DBSCAN algorithm for clustering analysis. The experiments are executed on a computer of CPU @2.3GHZ, Intel(R) Core (TM) i5-6300HQ, and main memory of 12GB.

The first is the choice of parameters. Since the resolution of the terrain grid is 30 × 30 meters, the minimum value of the neighborhood radius is 30 meters. According to the upper limit of the maximum slope of π/4, the upper limit of the neighborhood radius can be determined as 42 meters according to formula (3), then the interval is [30, 42].

Figure 3 shows the number of clusters and noise points under different minimum sample points. When selecting the value of *Minpts*, a relatively large value should be selected to ensure that the non-feature lines of the hillside area are excluded. As shown in Fig. 3 (a), when *Minpts* = 6, the number of noisy data is constant throughout the neighborhood and equal to the total number of the data set. At this time, when comparing *Minpts* equal to 4 or 5, selecting a relatively large minimum number of sample points can exclude more non-feature points, and reduce the influence of non-feature points on the extraction of feature lines. Within a smaller radius of a neighborhood the rate of change of the cluster is also small. Finally *Minpts* = 5 is determined to be the appropriate minimum number of sample points according to Fig. 3.

Taking 5 as the minimum sample point, calculate the number of noise data with different neighborhood radius values. The selected neighborhood radius should enable the points at the top of the mountain to be clustered into multiple clusters, while the points at the hillside are identified as noise data.

Fixing the number of minimum points to 5, we can change the value of neighborhood radius and observe the clustering results under different neighborhood radii. The change rates between adjacent neighborhood values according to formula (3) are shown in Table 1.

Table 1
Absolute value of change rate between neighborhood values

Eps	Absolute value of change rate	Eps	Absolute value of change rate
31–32	50	37–38	16
32–33	3	38–39	1
33–34	21	39–40	12
34–35	9	40–41	10
35–36	16	41–42	15
36–37	1	42–43	32

As shown in Fig. 3 (b), when the neighborhood radius is 38 meters between adjacent neighborhood, the rate of change of clusters is the smallest, as when the neighborhood radius is 37 meters, the number of clusters is 176. When the radius is 38 meters, the number of clusters is 160. When the radius of 39 meters, clusters in the field of number is 159. By calculating the rate of change of the number of clusters, find out the most accurate neighborhood radius. According to formula (3), the neighborhood change rate between 37 and 38 is 16, and the change rate of the domain between 38 and 39 is 1. The radius of the neighborhood with a smaller change rate is selected, indicating that the change of the neighborhood has little influence on the cluster when the neighborhood changes. The purpose of choosing 38 meters instead of 39 meters is to ensure that the data in the hillside area can be identified as noise data and that the extracted ridge lines are more accurate. Therefore, the clustering parameter of comprehensive selection is (38 meters, 5).

As shown in Table 1, it can be seen from the table that the change rate of the number of clusters in the neighborhood range is first smaller, then larger, then smaller when the neighborhood is 36m, then larger when the neighborhood is 37m, and then smaller when the neighborhood is 38m, and then it has been increasing. From 36 m to 37 m, the change rate of this process is 1. Although it is very small, there is still a process with the change rate of 1 at 38 m, which means that when the neighborhood radius is 36 m, it is not the best neighborhood. When the neighborhood radius is 37 m, it is the time when the number of clusters is the most. At this time, many clusters are clustered on the ridge, and at the same time, many areas are clustered on the hillside. When the neighborhood radius is 38 m, the number of clusters starts to decrease, that is, some clusters on the ridge are merged together, because the elevation change of the points on the hillside is greater than that on the ridge, 38 is the critical value for forming the ridge line without clustering the classes on the hillside, and is also the best value for feature extraction. Therefore, we select (38m, 5) as the clustering parameters of DBSCAN algorithm .

In order to verify the accuracy of the results, parameters (34m, 5) (35m, 5) (36m, 5) (37m, 5) (38m, 5) (39m, 5) were selected for clustering. The results are shown in Fig. 4. When the radius of the neighborhood is set from 34 to 36 meters, the clustered class cannot represent the ridge and valley contour of the terrain, and the clustering effect is obviously poor. When neighborhood radius is 37 meters, the main outline of the ridge and valley lines can get out, but some of the branch parts have not been successful identification. When the neighborhood radius of 39 m, the ridge branch part was out, the silhouette is also better, but the class is too thick, will also count in the area around the ridge line, extract ridge line, valley line effect is not ideal.

To sum up, the density (ε, Minpts) parameter of the DBSCAN algorithm is selected as (38m, 5), the DBSCAN clustering is performed on the entire terrain, and the sub-region division results are shown in Fig. 5(a).

Find the point with the largest distance between the feature points in the sub-region and its two sides, the center line of the sub-region starts from these two points, and passes through the feature points of the sub-region. The closest points are connected to form the characteristic line between the sub-regions, and finally, the characteristic line graph as shown in Fig. 5(b) is formed.

Find the highest or lowest point in the sub-area, find the two points farthest from the point in the sub-area, connect the two points with the extreme point, and simulate the trend of the characteristic line of the sub-area. Calculate the minimum distance between two adjacent sub-regions, use the minimum distance between the sub-regions as the weight, construct the spatial topology structure between the sub-regions, randomly select several consecutive or systematically spaced points on each feature line, and calculate Its average value, and the average value can represent the average elevation of the characteristic line, as shown in Table 2. If the elevation magnitude is alternating, it is a ridge line or a valley line. If the three consecutive lines are increasing or decreasing, the middle line is at a hillside location, and points of similar elevation along the hillside location are grouped together.

Table 2
Elevation distribution of grid points on feature lines

feature line	Point 1	Point 2	Point 3	Point 4	Point 5	Average elevation
1	1020,3210,3118	1530,3210,3159	3120,4260,3596	3480,4320,3173	3870,4830,2457	3100.6
2	1620,5160,1781	1650,5340,1651	1710,4140,2714	1860,4320,2457.	1590,4710,2358	2192.2
3	2040,4800,1928	2010,4860,1884	2220,4380,2277	1890,5100,1751	2070,5310,1815.	1931
4	2070,5310,1815	2940,3810,3457.	2520,4800,2574	2550,4680,2648	2700,5340,2056	2510
5	2700,5340,2056	2760,5310,2156	3300,4470,2400	3330,4500,2700.	3570,5280,2450	2352.4
6	3870,4230,3067	4020,4230,2954	4170,3990,2697.	4380,4170,2862.	4410,4140,2834.	2882.8
7	4770,3240,2148.	4800,3210,2127	4830.3180,2113	4920. 3150. 2107.	5070. 2940. 2029.	2104.8
8	3630,4080,2884	3810,3390,2746	3810,2220,2958.	3810,2340,3037	3990,3660,2496.	2824.2
9	3660,2070,3024	3630,2370,3202.	3270,3060,3179	3600,2430,3193	3810,2820,2747	3069
10	2370,2610,2229	2580,2490,2289	2580,2250,2077.	2670,2250,2118	2790,1800,2074.	2157.4
11	1710,2940,2884.	2070,1860,2084.	1770,2940,2845	1950,2010,2043	1980,2610,2502	2471.6
12	1560,1950,2114	1710,2580,2485	1770,2550,2437	1860,2430,2331	1800,2490,2388	2351
13	1110,2280,2528	1200,2160,2425	1200,3120,3291	1230,2790,3010.	1230,1890,2277	2706.2
14	180,3030,1868.	1050,4530,1806	1080,4620,1815.	120,3540,2172	120,3150,1925	1917.2
15	2940,1650,2046.	3000,2490,2514.	3030,2550,2564	3000,1950,2302	3390,1860,2504.	2386

We show the clustering result in the distribution map by using ArcGIS 10.5 software. Connecting these distributions of the clusters form several lines as shown in Fig. 6.

Figure 6(a) is a top view above the terrain, and Fig. 6(b) is a side view of the terrain. The average elevation of the No. 1 feature line is 3100.6 meters, and there are many feature lines on both sides, which is a main ridge line, and other feature

lines are its branches. The average elevation of the No. 2 feature line is 2192.2 meters, the average elevation of the No. 3 feature line is 1931 meters, and the average elevation of feature line 14 is 1917.2 meters, indicating that the No. 2 feature line is a ridge line, and the feature line elevations on both adjacent sides are smaller than it. If the elevation of the point on the feature line 4 is greater than that of the feature line 3, the feature line 3 is a valley line. The elevation of the feature line 5 is smaller than that of the feature line 4, and the line 4 is a ridge line, because No. 5 feature line is clustered on the adjacent edge, the No. 5 feature line can be judged to be mostly on a hillside or a valley line. The average elevation of the No. 6 feature line is 2882.8 meters, and the average elevation of the adjacent No. 7 feature line is 2104.4 meters, the elevation of the large cluster is also smaller, it is judged that the No. 7 line is a valley line. the No. 7 feature line is a valley line, the No. 8 feature line is 2824.2 meters high. The elevations of the No. 9 feature line is 3069 meters, and the elevation value these three characteristic lines is increasing, so the No. 8 line is located on the hillside, not a ridge line or a valley line. This is because the elevations in a certain direction on the hillside are similar and can be grouped together, but the elevation difference on the slope is large and cannot be clustered together, resulting in a "line-like" clustering on the hillside position, then No. 9 feature line is a ridge line. Through analysis and comparison in turn, it is concluded that the No. 10 feature line is a valley line, the No. 11 feature line is a ridge line, the No. 12 feature line is a valley line, the No. 13 feature line is a ridge line, the No. 14 feature line is a valley line. The elevation of No.15 is 2386 meters, it has an incremental change with feature lines 9 and 10, which is the convergence of hillside positions, not ridge lines or valley lines.

5.2 Experimental comparison between hydrological analysis and clustering analysis

To compare the extraction results with other methods, we implement the hydrological analysis method. The extraction of ridge lines and valley lines are carried out by using the method based on hydrological analysis (Stanislawski et al., 2021; Zhou et al., 2021; Testa et al., 2011; Hinnell et al., 2010) in ArcGIS 10.5. First, we use the window to calculate the average value of the original DEM. This process is called focal mean. In the raster calculator, the original DEM is subtracted by the results of focus statistics. Then we reclassify the results and filter out the positive and negative terrains by using 0 as the grading boundary.

The following is the process of the depression filling. First, the original DEM is filled with the depression filling tool of ArcGIS and the flow direction data is obtained by the flow direction processing of the filled data. Then the confluence accumulation is obtained by the flow processing. After the corresponding processing, the ridge lines are obtained.

We use the raster calculator to calculate the inverse terrain DEM and perform the flow direction processing on the inverse terrain DEM to obtain the flow direction data with ArcGIS. Then we use the raster calculator to extract the value of the confluence accumulation amount of 0 to obtain the flow direction data, then reclassify the data whose confluence accumulation is 0 and divide it into two levels, and adjust the critical point. Finally, the part with the attribute value close to 0 is the valley line.

The ridge and valley lines extracted by hydrology analysis are at the same position as that extracted by the DBSCAN algorithm. Hydrology analysis is also a traditional method of feature line extraction. It calculates the flow direction and water catchment through anti-topography and analyzes the feature lines according to the water catchment. Finally, the valley lines are directly obtained. The ridge lines are extracted by filling in the original terrain, then we calculate the flow direction and water catchment to obtain the ridge line.

Figure 7 is a comparison of the results of feature lines extracted by the two methods. From the comparison between Fig. 7 (a) and Fig. 7 (b), for the same terrain, they can both extract the ridge line and valley line. In the ridge line and valley line extracted by DBSCAN, there are a large number of points in some areas; that is, the gathered clusters are long and wide, and it reflects that the terrain similarity of grid points in this area is high, and it indicates the gentle terrain slope of the target area. However, the clusters are small, or some areas on a feature line are not clustered, it indicates that the slope of the area

is steep, the elevation of the points along with the feature line changes, and the elevation of the feature line increases or decreases in a certain direction.

In terms of extraction steps, DBSCAN clustering extraction of terrain feature lines is relatively simple, it can extract ridge lines and valley lines at the same time. But in the hydrological analysis, ridge lines and valley lines are extracted by different steps or sub-methods and the extraction process is much more complex than the DBSCAN algorithm. Ridge line extraction needs to obtain the filling data of the original terrain first. But valley line extraction needs to obtain the anti-terrain data of the original terrain first. After obtaining two different datasets, it is necessary to combine the positive and negative terrain data to obtain the feature line, and the water collecting line and water collecting line in the branch area of water extraction were not gathered. For getting accurate analysis time, we make several times of experiments to run clustering analysis and the hydrological analysis respectively. The average time of the DBSCAN clustering method is 429 seconds and the hydrological analysis method is 780 seconds. Obviously, the method in this paper reduces 351 seconds and provides 1.81 times in comparison to the hydrological analysis method.

There is not much difference in the accuracy of extraction, but in terms of extraction time and method, the proposed DBSCAN method has a greater improvement in time than hydrological analysis, and the method is also simpler.

6. Conclusion

In this paper, the DBSCAN clustering algorithm is used to extract ridge lines and valley lines in a new way. According to the DBSCAN algorithm automatically identifies similar points, the points with high similarity are gathered together along the direction of a ridge or valley, and the whole terrain will be gathered into multiple regions. By calculating the minimum distance between regions, a sub-region map is constructed, and the lines with larger distances between sub-regions are deleted to form multiple feature lines. By comparing the elevation values of the points on these feature lines, it is easy to determine whether they are ridge lines or valley lines, or a line formed by points on the hillside. If there are several adjacent lines with increasing or decreasing elevation, the feature line with the largest average elevation is the ridge line and the lowest is the valley line. The method presented in this paper is an effective application of the combination of cluster analysis and terrain feature extraction. Through the DBSCAN parameter selection strategy proposed in this paper, the feature lines in the terrain can be well extracted with higher efficiency, the feature line extraction method in this paper is a novel method and can reflect the slope change of the feature line.

The future work is on how to determine the density parameters of the DBSCAN algorithm based on terrain features, and study further the relationship between the selection of density parameters in the DBSCAN clustering method and terrain features such as resolution, different types of terrain, and so on. The terrain distribution and morphological characteristics are also further analyzed by the shape and size of the clusters formed by the DBSCAN algorithm.

Declarations

Data and codes availability:

The terrain data and codes that support the findings of this study are available in figshare with the identifiifier doi:10.6084/m9.figshare.21078646.

Disclosure statement:

No potential conflict of interest was reported by the authors.

Funding:

This work was partly supported by the National Natural Science Foundation of China (No. 41930102, 41771411).

References

1. Hu QM (2017) Research progress of topographic feature line extraction. Surveying and Mapping and Spatial Geographic Information 40(4):47–50
2. Peucker TK, Douglas DH (1975) Detection of surface-specific points by local parallel processing of discrete terrain elevation data. Comput Graphics image Process 4(4):375–387
3. Wood J (1996) The geomorphological characterization of digital elevation models. University of Leicester (United Kingdom)
4. O'Callaghan JF, Mark DM (1984) The extraction of drainage networks from digital elevation data. Computer vision, graphics, and image processing, 28(3): 323–344
5. Milligan GW, Cooper MC (1987) Methodology review: Clustering methods. Appl Psychol Meas 11(4):329–354
6. Khan K, Rehman SU, Aziz K et al (2014) DBSCAN: Past, present and future. The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014). IEEE, pages, 232–238
7. Li M, Bi X, Wang L et al (2021) A method of two-stage clustering learning based on improved DBSCAN and density peak algorithm[J]. Comput Commun 167:75–84
8. Chang YC, Song GS, Hsu SK (1998) Automatic extraction of ridge and valley axes using the profile recognition and polygon-breaking algorithm. Comput Geosci 24(1):83–93
9. Masavi M, Natarajan P, Binello S et al (1999) Knowledge based extraction of ridge lines from digital terrain elevation data[C]. IEEE International Geoscience and Remote Sensing Symposium, IGARSS'99 (Cat. No. 99CH36293). IEEE, 5: 2492–2494
10. Blashfield RK, Aldenderfer MS (1978) The literature on cluster analysis. Multivar Behav Res 13(3):271–295
11. Busu M, Nedelcu C, Cadis A (2021) An overview of the academic level among EU countries. A cluster analysis approach. //Proceedings of the International Conference on Business Excellence. 15(1): 210–217
12. Schubert E, Sander J, Ester M et al (2017) DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Trans Database Syst (TODS) 42(3):1–21
13. Khan K, Rehman SU, Aziz K et al (2014) DBSCAN: Past, present and future. The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014). IEEE, pages, 232–238
14. Stanislawski LV, Shavers EJ, Wang S et al (2021) Extensibility of U-Net neural network model for hydrographic feature extraction and implications for hydrologic modeling. Remote Sens 13(12):2368
15. Zhou Q, Su J, Arnbjerg-Nielsen K et al (2021) A GIS-Based Hydrological Modeling Approach for Rapid Urban Flood Hazard Assessment. Water 13(11):1483
16. Tesfa TK, Tarboton DG, Watson DW et al (2011) Extraction of hydrological proximity measures from DEMs using parallel processing. Environ Model Softw 26(12):1696–1709
17. Hinnell AC, Ferré TPA, Vrugt JA et al (2010) Improved extraction of hydrologic information from geophysical data through coupled hydrogeophysical inversion. Water Resour Res 46(4):1–14
18. Ai TH, Li JZ (2010) A DEM generalization by minor valley branch detection and grid filling. ISPRS J Photogrammetry Remote Sens 65(2):198–207
19. Pang XF, Song Z, Xie WY (2013) Extract valley-ridge lines from point-cloud-based 3D fingerprint models. IEEE Comput Graphics Application 33(4):73–81
20. Zhou W, Peng RC, Zhang LH, Wang WJ (2016) Topographic feature line extraction from point cloud based on SSV and HC-laplacian smoothing. Proceedings of International Conference on Advances in Energy and Environment Research(ICAEER), Guangdong, China, Aug. 12–13, 2016. Eds:Bachir Achour and Qiyan Wu, pages, 246–253. CRC Press
21. Zhou W, Peng RC, Dong J, Wang T (2018) Automated extraction of 3D vector topographic feature line from terrain point cloud. Geocarto Int 33(10):1036–1047

22. Riazanoff S, Cervelle B, Chorowicz J (1988) Ridge and valley line extraction from digital terrain models. Int J Remote Sens 9(6):1175–1183
23. Daya Sagar BS, Murthy MBR, Babu Rao C, Raj B (2003) Morphological approach to extract ridge and valley connectivity networks from Digital Elevation Models. Int J Remote Sens 24(3):573–581
24. Maurya RK, Kulkarni ST, Kalita N (2018) Projection-based spatial morphology for extracting ridge and valley profiles of mountains from 3D amorphous data. IEEE International Conference on Computing Communication and Automation(ICCCA), Greater Noida, India, Dec. 14–15, pages 1–6
25. Romero BE, Clarke KC (2018) Exploring uncertainties in terrain feature extraction across multi-scale, multi-feature, and multi-method approaches for variable terrain. Cartography and Geographic Information Science 45(5):381–399
26. Chang Y-C, Sinha G (2007) A visual basic program for ridge axis picking on DEM data using the profile-recogintion and polygon-breaking algorithm. Comput Geosci 33(2):229–237

Figures

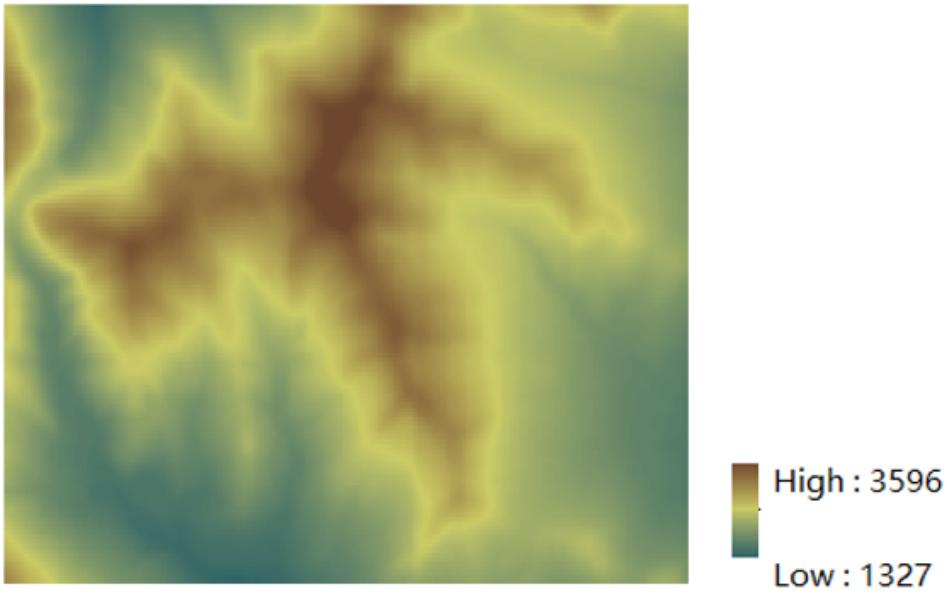

Figure 1

Mountainous terrain with the size of 44 × 120

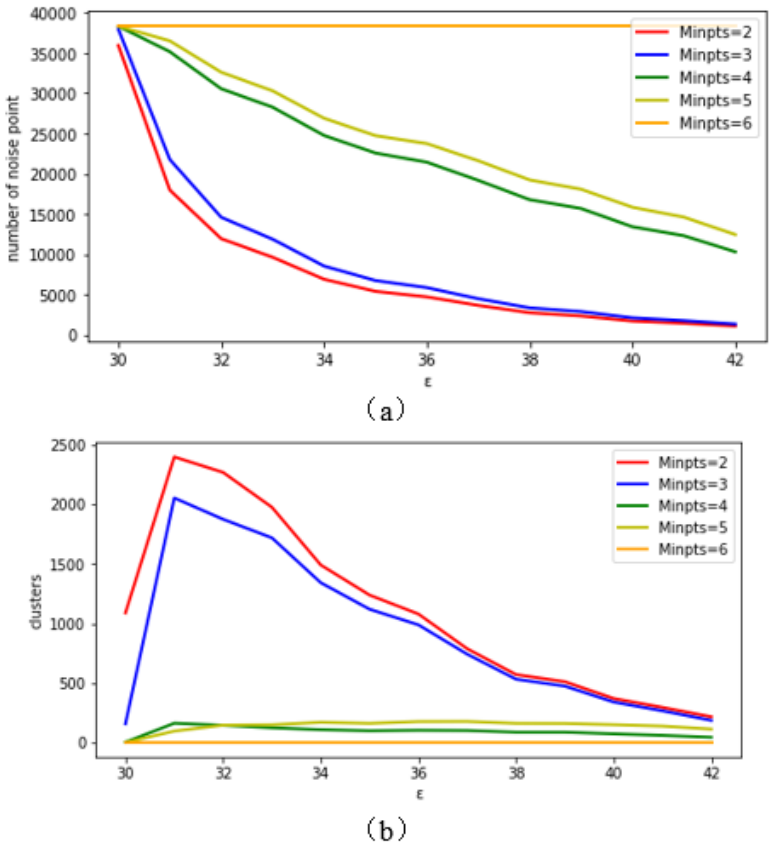

Figure 2

Variation of the number of noise points and the number of clusters with the neighborhood under different minimum sample points: (a) number of noise points; (b) number of clusters

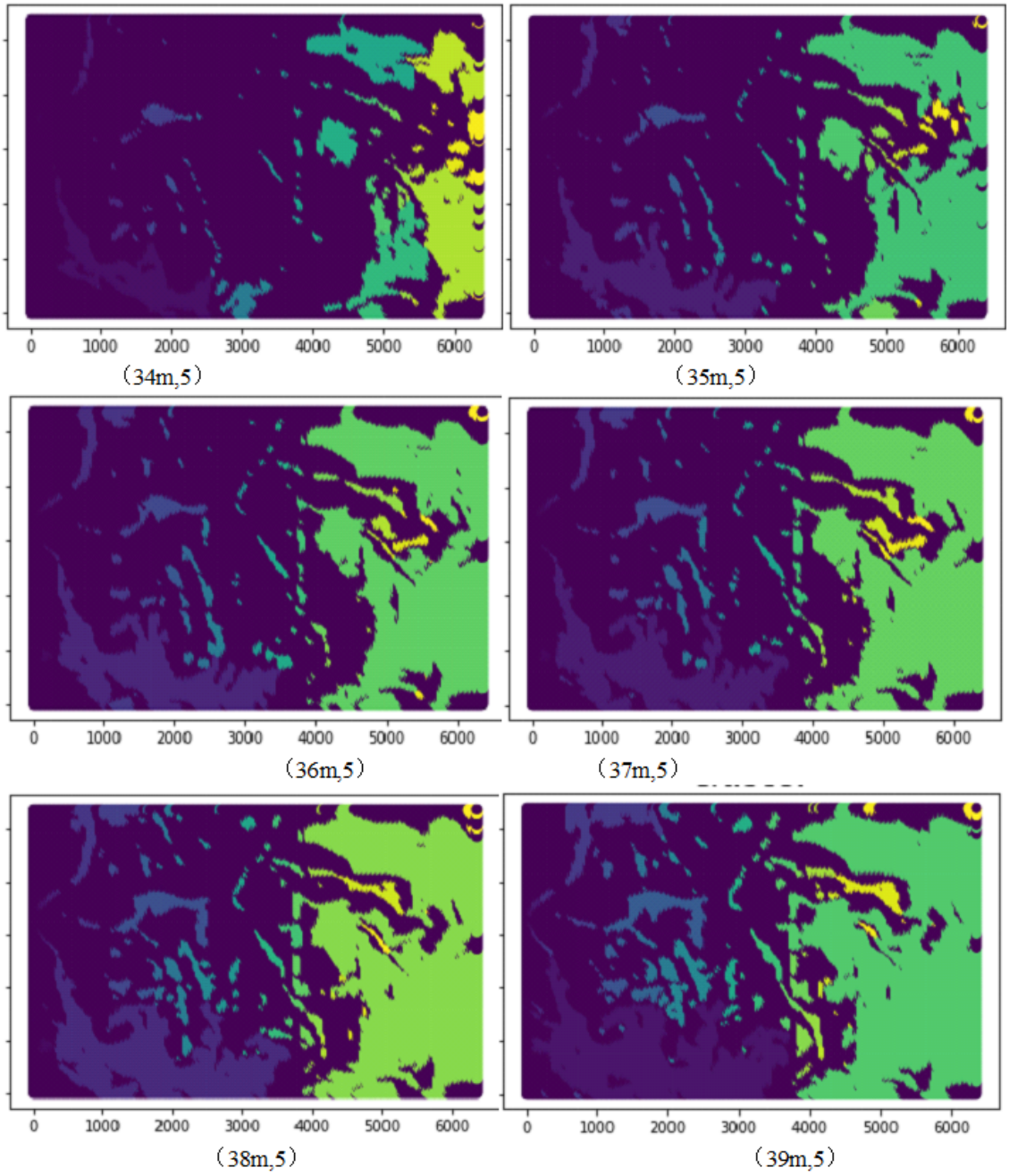

Figure 3

Clustering results of neighborhood changes when Minpts=5

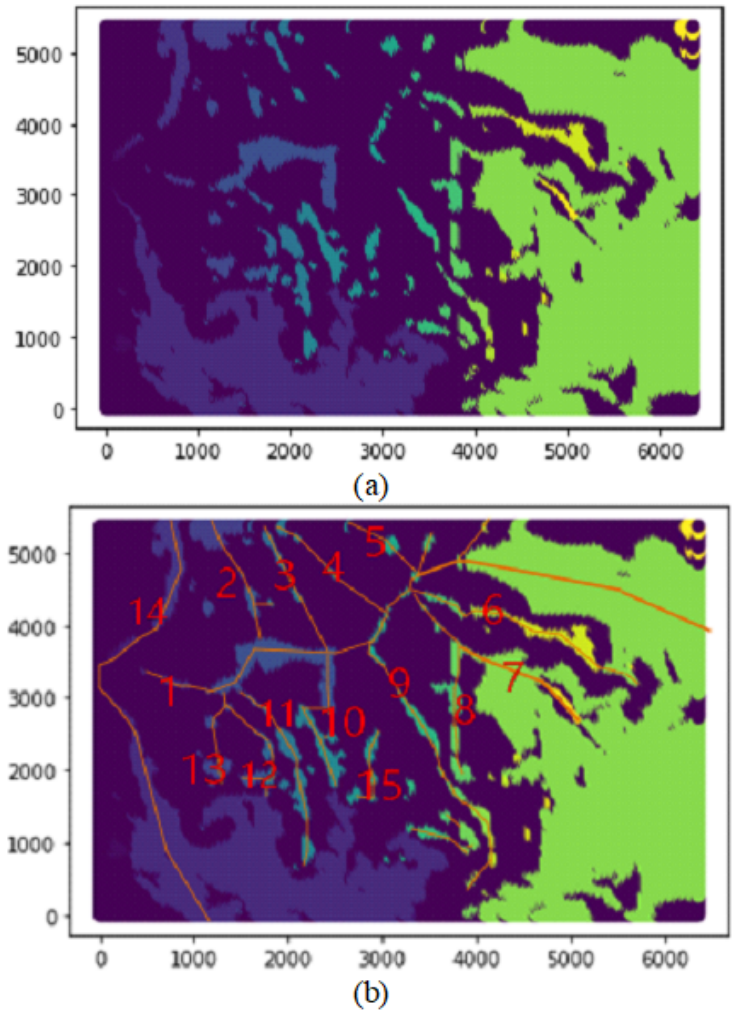

Figure 4

DBSCAN clustering results: (a)the clustering results of the given terrain; (b) the extracted feature lines.

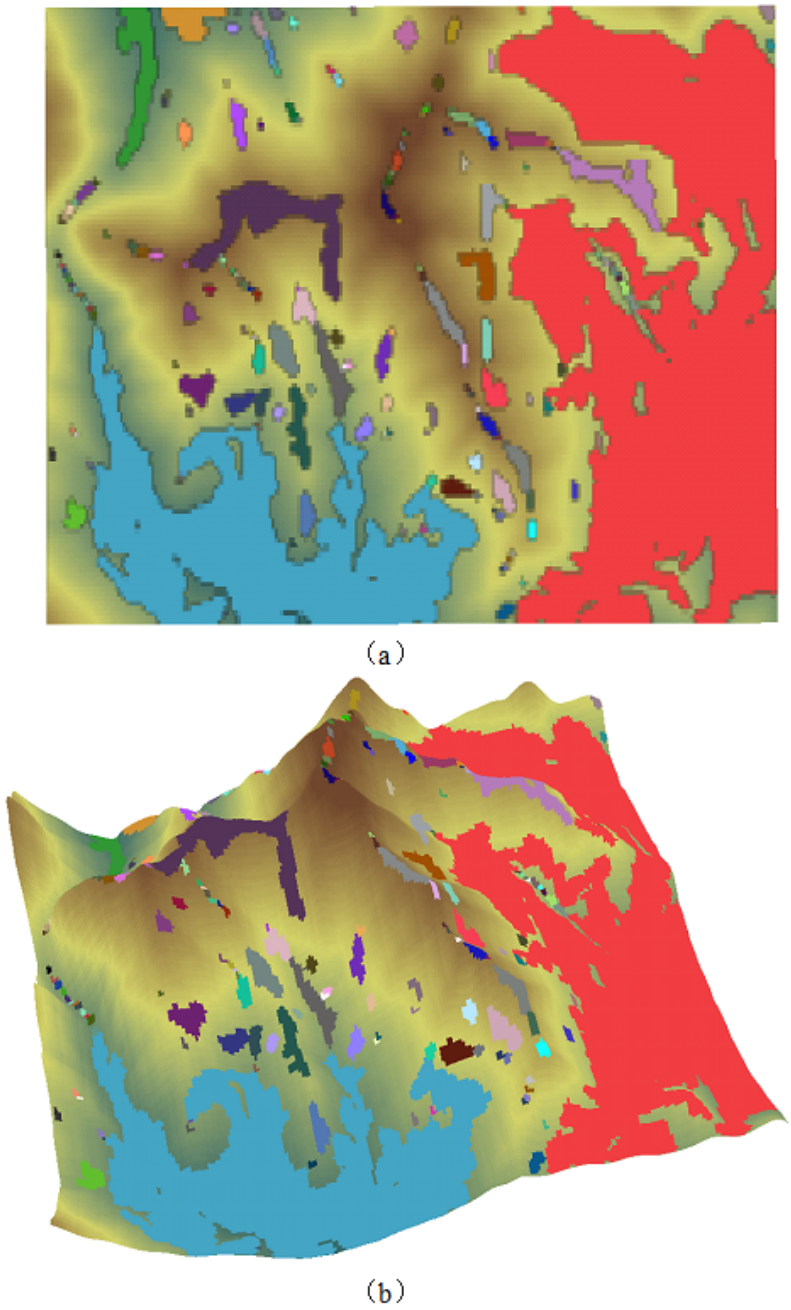

Figure 5

mapping of terrain subregions. (a) lines on the top view above the terrain; (b) lines on side view of the view;

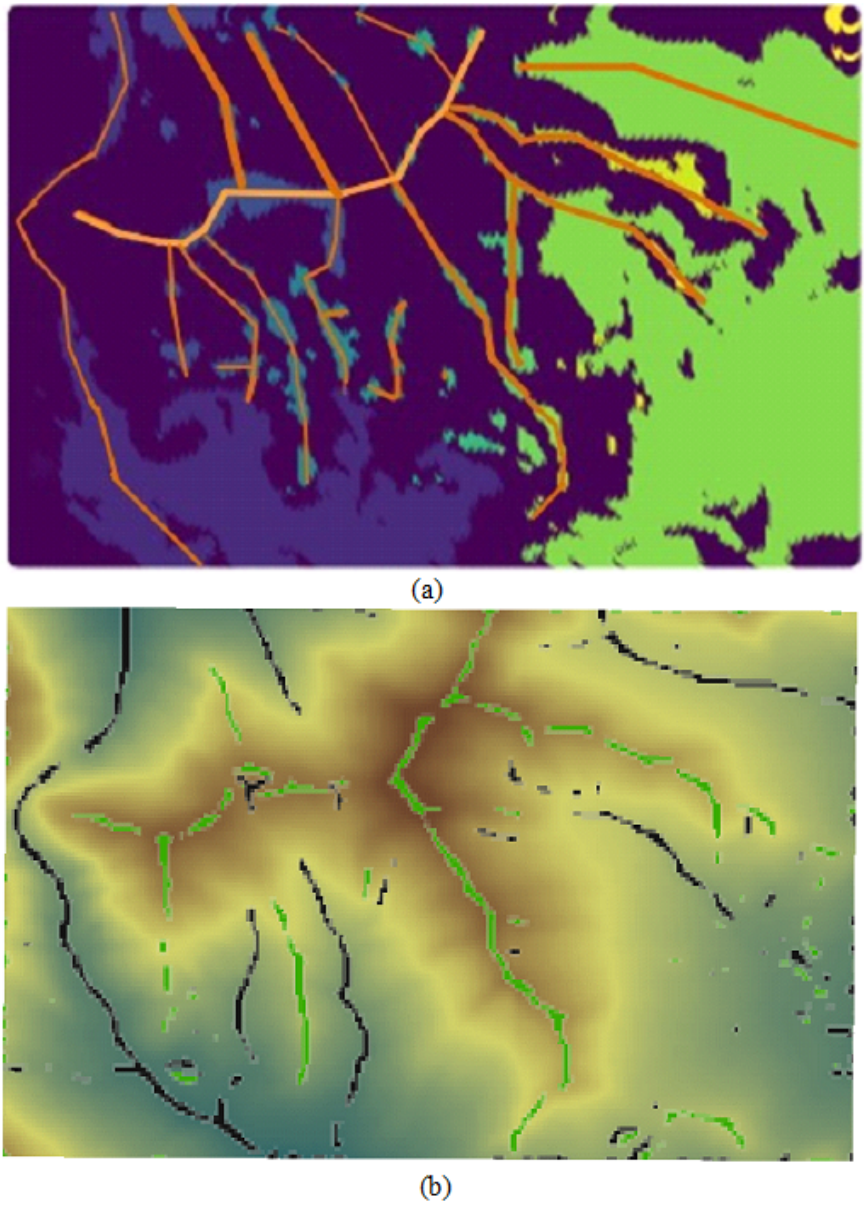

Figure 6

Results comparison of different extraction methods: (a) DBSCAN; (b) hydrological analysis.