9

Research Article

Xianyong Gong*, Fang Wu, Ruixing Xing, Jiawei Du, and Chengyi Liu

LCBRG: A lane-level road cluster mining algorithm with bidirectional region growing

https://doi.org/10.1515/geo-2020-0271 received January 31, 2021; accepted June 30, 2021

Abstract: Lane-level road cluster is a most representative phenomenon in road networks and is vital to spatial data mining, cartographic generalization, and data integration. In this article, a lane-level road cluster recognition method was proposed. First, the conception of lane-level road cluster and our motivation were addressed and the spatial characteristics were given. Second, a region growing cluster algorithm was defined to recognize lanelevel road clusters, where constraints including distance and orientation were used. A novel moving distance (MD) metric was proposed to measure the distance of two lines, which can effectively handle the non-uniformly distributed vertexes, heterogeneous length, inharmonious spatial alignment, and complex shape. Experiments demonstrated that the proposed method can effectively recognize lane-level road clusters with the agreement to human spatial cognition.

Keywords: cartographic generalization, spatial data mining, spatial cluster, lane-level road cluster, distance measurement

1 Introduction

Emerging geographic information system applications such as vehicle navigation, intelligent transportation systems, and self-driving technology ask for precise lane information in road networks [1–9]. Hence, road networks are more and more modeled in a realistic way, capturing all the lane-level information of the networks in datasets [10,11]. Lane-level road cluster is common in

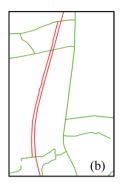
road network dataset and is of great importance to spatial data mining, cartographic generalization, data integration, road change detection, and even the urbanization process [12–15].

Lane-level road clusters treat more than one road as a unit. In data integration, lane-level road clusters can transform the most challenging many-to-many pair matching to one-to-many and one-to-one matching problems, which will greatly reduce the difficulty of data integration [12]. For example, there are three roads in the cluster in Figure 1a, while two in Figure 1b. If road clusters are identified and treated as a unit, the 3:2 multiscale matching problem will be transformed to 1:1, which is much easier. This is also termed as structure matching. In addition, object matching is used in various applications including conflation, data quality assessment, updating, and multi-scale analysis [16].

On the other hand, the highly modeled road network is one of the most important data sources for spatial data collecting and updating. However, the quality and level of details are not guaranteed when referring to the national or authoritative topographic map productions. According to the National Mapping Agency (NMA) cartographic specifications, roads are always represented using polyline or dual-polyline on map production. Lane-level road cluster with high density will bring great limitation to map legibility [17], which is defined as a combination of map readability (discerning the symbols) and map interpretation (understanding the content of the map). Hence, necessary processing such as transforming lane-level road cluster to dual-polyline or polyline is needed to make it conform to the NMA cartographic specifications. This is termed as structured generalization in map generalization [18], such as extracting the centerlines. For example, the road cluster in Figure 1a at the scale of 1:10,000 is transformed to dual-polyline in Figure 1b at the scale of 1:50,000, and single polyline in Figure 1c at the scale of 1:250,000. In the research framework, the lane clusters will be first cleaned and then used for several applications, such as lane information extraction or correction in road analysis, centerline or dual-polyline road extraction in map generalization, and so on.

^{*} Corresponding author: Xianyong Gong, Institute of Surveying and Mapping, Information Engineering University, Kexue Road 62, Zhengzhou 450001, China, e-mail: gongxygis@whu.edu.cn Fang Wu, Ruixing Xing, Jiawei Du, Chengyi Liu: Institute of Surveying and Mapping, Information Engineering University, Kexue Road 62, Zhengzhou 450001, China





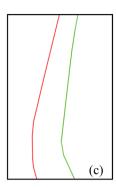


Figure 1: The road cluster in multi-scale spatial data. The scales are: (a) 1:10,000, (b) 1:50,000, and (c) 1:250,000, respectively.

In order to achieve these applications, the first step is to find which roads need to be transformed. This comes to be the focus of this article, with the aim to find the roads by the proposed method as the operational objective of such applications. Here we propose a lane-level road cluster mining algorithm with bidirectional region growing (LCBRG) to identify lane-level road clusters, based on which, map generalization algorithms including road cluster typification, centerline extraction, and simplification could be better performed.

In addition, as a lane is modeled as a line in Geographic Information Science, from a more macroscopic scientific perspective, lane-level road clustering actually contributes to the research content of line clustering problem. In spatio-temporal data mining, trajectory analysis is a typical application. By clustering the similar trajectories, their main trend can be better understood, which brings great advantages in principal component analysis, behavior pattern analysis, and trend prediction [19]. The analysis of movement behavior has been investigated for different purposes and explored in several domains. For example, by clustering the typhoon trajectories, we could forecast its direction and trend. The trajectory data record daily human mobility, such as working, shopping, and engaging in entertainment and leisure activities. Such data contain various patterns of human behavior which can be utilized to identify hotspots in urban areas [20].

Cluster analysis is the main task of exploratory data mining. Clustering algorithms can be categorized based on their cluster model [21,22]. (1) Connectivity-based clustering, also known as hierarchical clustering, is based on the core idea of objects being more related to nearby objects than to objects farther away, such as BIRCH [23]. (2) Centroid-based clustering, where clusters are represented by a central vector, which may not necessarily be a member of the dataset, such as K-means and CLARANS

[24]. (3) Distribution-based clustering, where clusters can then easily be defined as objects belonging most likely to the same distribution, such as EM [25]. (4) Density-based clustering, where clusters are defined as areas of higher density than the remainder of the dataset. Objects in these sparse areas, that are required to separate clusters, are usually considered to be noise and border points, such as DBSCAN [26].

Most algorithms are developed for points, while only few are for the line or polygon dataset; however, line clustering is a classic issue and is in urgent need in both theory and practice. Lu et al. [27] presented a clustering method to classify contour lines using wavelet analyzing and numerical statistic, which is based on contour line's geography fractal character together with its position. Contour lines are characterized by fractal dimensions that exhibit similar patterns at increasingly small scales, but lane-level roads show great spatial heterogeneity. Liu et al. [28] proposed a spatial lines clustering algorithm based on their connectivity. Based on K-means, this algorithm selects the spatial line connectivity as the distance measurement between lines to cluster spatial lines. A similar idea was employed by Zhu [29], who used DBSCAN to cluster lines to detect the outliers, where intersection and adjacent relationship are used as distance. However, the relationships of connectivity, intersection, and adjacency are qualitative and weak spatial constraints, resulting in arbitrary orientations and discretionary distances of the clusters. While lane-level road clusters are synchronized with the law of common fate in the perspective of Gestalt principles [30]. Tang et al. [31] proposed an efficient partition-and-filter model to filter trajectories with expected accuracy according to the spatial feature of high-precision GPS data and their error rule. GPS trajectory is still defined via massive zerodimensional points with high density, while the object in this article is one-dimensional line cluster.

With the scope of cartographic generalization, a lane-level road cluster recognition method was proposed in this article. First, the conception of lane-level road cluster and our motivation were addressed, and the spatial characteristics are given in Section 2. Second, a region growing cluster algorithm is defined in Section 3 to recognize lane-level road clusters, where constraints including distance and orientation were used. A novel moving distance (MD) is proposed in Section 4 to measure the distance of two lines, which can effectively handle the non-uniform length and heterogeneously distributed vertexes.

Our contributions are summarized as follows:

- With the aim of cartographic generalization, the motivation and conception of lane-level road cluster were addressed, and the spatial characteristics were given.
- A novel MD metric was proposed to measure the distance of two lines, which can effectively handle the nonuniformly distributed vertexes, heterogeneous length, inharmonious spatial alignment, and complex shape.
- Based on the above work, a new region growing cluster algorithm was defined to recognize lane-level road clusters, where constraints including distance and orientation were used.

2 Line cluster and its characteristics

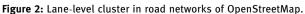
2.1 Motivation

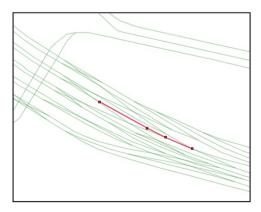
A cluster refers to a group of similar things that are close together. Lanes are always modeled as lines in the spatial database. Generally speaking, a lane-level road cluster is defined as a set of lane lines that are clustered into groups by some constraints such as distance and orientation. The lane lines in one cluster are not limited to these with strict parallel relationships. As shown in Figure 2, some short lanes are nonparallel but they are still regarded as a part of the cluster because they are located in the region of the existing cluster. However, the existing research cannot deal with these complex situations. For example, Luan and Yang [32] and Savino and Touya [10] provide methods for parallel line recognition, which are unable to handle complex line cluster with nonparallel sense.

With the development and integration of mobile communication and wireless Internet, smart mobile devices, and mobile sensors and measurement, the spatial information is massively surging. The possible reasons for line cluster include at least the following aspects [33]:

- (1) In terms of user demands, the spatial data used for driving and riding navigation require that the geographical features should be detailed as precisely as possible, and that motorway, lanes, sidewalks, and so on should be modeled in detail.
- (2) From the view of data collection, Volunteered Geographic Information (VGI) systems are now important data source for spatial data updating. Some participants of VGI are short of professional cartographic knowledge (such as the concept of map scale), so they collect data according to their personal need and experience without authoritative supervision, strictly in accordance with neither the navigation data standards nor the topographic map standards. The subjectivity of the participants leads to the appearance of multiple inharmonious shapes and attributions of the same geographic object. This may lead to the existence of data repetition and multiple and inconsistent levels of details (LoDs) [34].
- (3) The reference dataset is of great variety, including existing map productions, different high resolution







remote sensing image datasets, the US Census Bureau TIGER, Ordnance Survey OpenData, ArcGIS Open Data, open navigation data, VGI systems such as OpenStreetMap, and other open or authoritative data sources. These platforms are independent of each other, and there are no uniform data production specifications, resulting in spatial data inconsistency (in referencing system, data model, coding system, visualization, and so on). The location and geometric shape of different reference datasets are always inconsistent. In addition, the LoDs of some data sources are higher than the topographic map specifications, while others are lower or even short of.

(4) New acquisition equipment such as vehicle-based GPS and smart mobile devices are widely used [35–37], resulting in the acceleration of the speed of data acquisition and the improvement of data currency, but the data quality is not always guaranteed: spatial and temporal accuracy of different receivers and positioning methods (such as GPS/Beidou satellite navigation system, mobile wireless positioning, or integrated navigation) are inconsistent; the storage formats and attributions of data collected by different equipment lack uniform standards; unprofessional operations such as arbitrary driving lead to irregular trajectory; and acquisition equipment is also greatly affected by platforms, weather, and other natural conditions [38]. The noises may make the road trajectory irregular.

A large amount of noises in spatial data not only increase the data redundancy but also affect the overall accuracy of the map production, which will increase the burden of the cartographers, thus affecting the quality of the map updating. For example, in the emergency mapping [39,40], noise information is a serious obstacle to the acquisition of fine information in the disaster area, which will delay the relief time and will directly bring great life and property damage. Therefore, it is necessary to quickly and automatically extract high-quality road information for cartographic mapping.

2.2 Characteristics of line cluster

The abovementioned investigation shows the following characteristics: (1) most of the line clusters in various reference datasets are always with multiple inconsistent versions; (2) they do not comply with the existing NMA cartographic specifications; and (3) the topological relationships of the line primitives in the lane-level road cluster are chaotic. Although the symbolized Digital Cartographic Model (DCM) map production appears to be correct, the Digital Landscape Model (DLM) dataset behind may be chaotically organized with disordered topological relationships. Figure 3a shows the DCM result from OpenStreetMap website after symbolized rendering, while Figure 3b shows the corresponding DLM dataset. It can be seen from the DLM dataset (and the five enlarged views) that the length and orientation of the lines in the lane-level road cluster are almost inharmonious, but the one or more line primitives together contribute to the geometric representation of the corresponding geographic object.



Figure 3: DCM and DLM of lane-level roads from OpenStreetMap: (a) DCM and (b) DLM.

Examining the different conditions of the lane-level road cluster, we found that the line primitives that make up the cluster have the following characteristics or relationships:

- (1) **Condition C1: Distance.** The distance between line primitives satisfies the certain prior threshold. In general, a semantic road entity (road sections that have the same semantic name) may be represented collectively by one or more lane-level primitives, but the distance between them is generally less than the road width w plus the data error δ , namely the sum $(w + \delta)$.
- (2) Condition C2: Parallelism. The orientation between line primitives is approximately parallel. Due to the practical factors such as traffic rules, vehicle driving safety, and so on, the lane-level road primitives that represent the traffic information are generally parallel to each other.
- (3) **Condition C3: Containing relationship.** The line primitives located in the region of existing clusters are regarded as a part of the cluster. This condition mainly deals with the nonstandard lane information such as overpass and footway with the corresponding steps. These facilities are usually modeled as lines and stored together in the road database, sometimes with attributions provided by conscientious and dedicated participants.

Of the three conditions above, not all the three relationships are necessary to be a lane-level road cluster. If conditions C1 and C2, or C1 and C3, or C1, C2, and C3 are satisfied, it is possible to determine those lanes as a cluster.

As we know, the distance between two intersecting lines is zero. So not all the lines with a distance of zero constitute a cluster. Either parallelism or containing relationship or both of them are required to be satisfied to be a lane-level road cluster. Take Figure 4 as an example. The red dotted line primitive in Figure 4a is not accepted as a part of the cluster for distance constraint. Though distance constraint is satisfied, the red dotted line

primitive in Figure 4b is not accepted because neither parallelism nor containing relationship is satisfied, while the blue dotted line in Figure 4c satisfies containing relationship; therefore, it is accepted as a part of the cluster.

In addition, the composition of a cluster has no direct relationship with the length of the line primitive. Even though the lengths of line primitives are uneven, which means failing to meet the spatial alignment [41], a cluster is possible if they satisfy the above relationships. This situation will be handled by the novel MD algorithm in Section 4.

According to the characteristics of the lane-level road cluster, the constraints condition of lane r_i and r_j to be a cluster can be formalized as the following three sub constraints:

T1: the distance between r_i and r_j is less than the threshold T_d .

T2: the orientation difference between r_i and r_j is less than the threshold T_{dir} .

T3: though T2 is not satisfied, r_i (r_j) is located in the region of any existing cluster.

A bidirectional region growing algorithm is proposed to lane-level road clustering in Section 3, and the involved quantitative constraints is descripted in detail in Section 4.

3 The bidirectional region growing algorithm to lane-level road cluster

3.1 Method's basic strategy

The lane-level road cluster can be extracted by geometric and semantic methods. The semantic information such as the street names is very efficient when extracting the

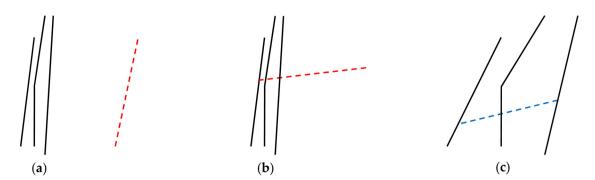


Figure 4: Three cases of cluster conditions.

lane-level clusters. In practical project, we prefer the semantic information in the extraction work when it is available and with high quality. However, there are still many difficulties when using the semantic information. For example, the semantic information may be omitted or uncompleted during the data acquisition. Another example is that, the semantic information may have different encoding methods and recording formats, which leads to more measurement difficulties in semantic similarity.

From the perspective of scientific research, geometric methods and semantic methods are two different categories, so the research objects are greatly different. In this article, we focus on the geometric information. From the perspective of practical application, the geometric method and semantic method complement each other and need to work together.

By analyzing the characteristics of the lane-level road cluster, the following strategy is proposed to identify the cluster: First, according to condition C1, the distance constraint is used to identify the initial cluster. Second, line primitives that do not meet the parallelism relationship in condition C2 are removed from the recognized initial clusters. Finally, for the remaining line primitive which has not been assigned to any cluster, check condition C3. If the line primitive is contained by the region of any existing cluster, assign it to the corresponding cluster.

Based on this strategy, a lane-level road cluster mining algorithm with bidirectional region growing (LCBRG) is proposed, which extracts clusters both horizontally and vertically. The algorithm steps are as follows:

- Step 1: Calculate the quantitative constraints of lines and capture the proximity relationships among line primitives.
- Step 2: Region growing cluster on the vertical (RGCV) examines the neighboring lines of the initial seed line along the line orientation and determines whether the line neighbors should be assigned to the existing cluster. Here the neighboring relationship means the connection between neighboring roads.
- Step 3: Region growing cluster on the horizontal (RGCH) examines the neighboring lines of the initial seed line which are perpendicular and adjacent to the seed line and determines whether the neighbors should be assigned to the existing cluster. Here the neighboring relationship means the proximity.
- Step 4: Traverse all the remaining ungrouped line primitives and examine whether the line primitive is contained by any existing cluster. If yes, assign this line primitive to the corresponding cluster.

The complexity of LCBRG is $O(n^2)$, where n denotes the lane number. In order to reduce the computational complexity, strokes are first constructed according to the principle of good continuation [42]. Line primitives with similar orientation at the touching endpoint are assigned to the same stroke. In fact, stroke construction is in some sense a sub-step of RGCV.

The two key sub-steps of LCBRG algorithm, RGCV and RGCH, are conducted in the Sections 3.2 and 3.3, respectively to recognize the lane-level road cluster.

3.2 RGCV

RGCV examines neighboring lines of the initial seed line along the line orientation and determines whether the line neighbors should be added to the existing cluster. Here the neighboring relationship means the connection between the neighboring roads. Considering the quality characteristics of lane-level road cluster, lines may have dangles or pseudo-nodes. As shown in Figure 5, the endpoints should be touched, but they are not due to the lack of professional knowledge of the participants. Hence, a buffer circle with the radius α is offered to cover the dangles or pseudo-nodes that are not connected to the seed line. That is, if two lines are connected by the same buffer circle, they are regarded as neighbors and will be assigned to the same cluster.

Here the buffer radius α is critical. Figure 5a shows the case that buffer radius α is too small to detect the candidate neighbors, where line B is ignored which should be assigned to the cluster of line A. While Figure 5b shows the case that buffer radius α is too large that too many candidate neighbors are detected, where lines C and D are detected which should not be assigned to the cluster of line A. Considering the fact that the maximum distance between lanes is generally less than the road width, half of the road width is suggested for buffer radius α . Even in the case that candidate neighbors are excessively detected, the detected candidates are still located in the region of the road polygon, which will not affect the identification results of lane-level clusters.

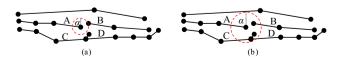


Figure 5: The buffer circle in RGCV.

Thus, the maximum value of buffer radius α is suggested to be half of the road width.

Let the integer K denote the cluster number, the integer array cls() denote the cluster that a stroke belongs to, the list array clu(k) denote the objects that a kth cluster contains, and the Boolean array bVisited() denote whether the stroke has been visited. The RGCV is conducted by the following steps:

- Step 1: For the interested dataset, strokes are first identified using the method in research [32]. Initialize all the notation variables to 0.
- Step 2: For the identified strokes, select one of them (the first in this article) as seed stroke r_0 .
- Step 3: If $bVisited(r_0)$ = false, query the candidate neighboring strokes set $Cnt(r_0)$ which are not disjoint with the buffer circles of the endpoints of stroke r_0 . If $cls(r_0) = 0$, then K++; set $cls(r_0) = K$, and add r_0 to $clu(cls(r_0))$. Let $bVisited(r_0)$ = true.
- Step 4: For each stroke $cr_i \in Cnt(r_O)$, if $cls(cr_i) = 0$, test the constraints condition $c(r_O, cr_i)$. If $c(r_O, cr_i)$ is satisfied to T2, then $cls(cr_i) = cls(r_O)$, add cr_i to $clu(cls(r_O))$. Let $bVisited(cr_i) = true$.
- Step 5: Let $r_0 = cr_i$, recursively carry out step3 to step5 until all the strokes are visited.

article uses the buffer analysis with fixed radius β to capture the neighbors. The buffer radius β needs to agree with traffic rules, driving safety, road design principles, and other related factors.

Using the notations in RGCV, the RGCH is conducted by the following steps:

- Step 1: Perform RGCV. Reinitialize the Boolean array bVisited() to zero. Set variable k = 0.
- Step 2: While $k \le K$, visit the kth result clu(k) identified in RGCV. For each stroke $r_i \in clu(k)$, go to step 3.
- Step 3: If $bVisited(r_i)$ = false, query the candidate neighboring strokes set $Near(r_i)$ which are not disjoint with the buffer of r_i . Let $bVisited(r_i)$ = true.
- Step 4: For each stroke $nr_j \in Near(r_i)$, if $cls(nr_j) \neq cls(r_i)$, test the constraints condition $c(r_i, nr_j)$. If $c(r_i, nr_j)$ is satisfied to both T1 and T2, then add the list array $clu(cls(nr_j))$ to $clu(cls(r_i))$. For each stroke $r_q \in clu(cls(nr_j))$, set $cls(r_q) = cls(r_i)$ and $bVisited(r_q) =$ true. Let $bVisited(nr_i) =$ true.
- Step 5: Let $r_i = nr_i$, go to step 3.
- Step 6: Let k = k + 1, recursively carry out step 2 to step 5 until k > K or all the strokes are visited. The list array clu(k) with more than one object inside is regarded as a candidate cluster.

3.3 RGCH

RGCH examines neighboring lines of the initial seed line which are perpendicular and adjacent to the seed line and determines whether the neighbors should be added to the existing cluster. Here the neighboring relationship means the proximity relationship. The methods to measure proximity include the topological method (such as Delaunay triangle) and metric method (such as buffer analysis). The topological method using Delaunay triangle captures proximity relationship without the limitation of distance, and two features are regarded as neighbors if they are connected by the same Delaunay triangle and are not separated by other features, which means that it is a qualitative approach. It can be seen that even if there is a street block between two lanes from different roads (the distance between them is larger than road width), they are considered adjacent. This is against our motivation. In fact, only the lanes with specific distance can be treated as a cluster. So the topological method using Delaunay triangle may fail to detect the correct candidate neighbors, thereby making the subsequent steps timeconsuming. Taking the road width into account, this

4 Quantitative constraints

According to the characteristics and three conditions of lane-level road cluster mentioned above, the quantitative constraints for cluster recognition mainly include distance and orientation.

- (1) Distance. At present, the distance measurement method between lines includes Euclidean distance (ED), Hausdorff distance (HD), and Fréchet distance (FD). However, ED is unable to measure the distance between complex lines. The HD and FD are generally used to measure the matching degree of point sets. When the vertexes are non-uniformly distributed, the local shape mutation has a great influence on the HD and FD. That is, the stability of the HD and FD is poor [43–48].
- (2) Orientation. The orientation of a line not only describes its own orientation characteristics but also can be used to reflect the relative orientation relationship between two lines. The orientation of a line can be divided into two categories, namely global and local orientation. The global orientation refers to the azimuth determined by the first and last vertex of the line. The local orientation refers to the azimuth determined by

every segment of the line, usually weighted by the segment length. For a lane cluster, the line with angle 0 and the line with angle 180 have the same contributions. Hence, the angle of a line ranges from 0 to 180 in our algorithm. Here the orientation difference is used to describe the relative orientation relationship between two candidate lines. Considering the situation in this article, the orientations of line primitives in a lane-level cluster suffer a slight change. So we employ the global orientation to approximately measure the orientation difference of two lines, denoted by:

$$D(\theta) = |\theta_A - \theta_B|, \tag{1}$$

where θ_A and θ_B are the global orientations of the two lines, respectively.

It should be noted that the cluster is a fuzzy spatial concept, and line primitives in a lane-level cluster may be heterogeneous in length and inharmonious in spatial alignment, as shown in Figure 6. For the purposes of this study, the composition of a cluster has no direct relationship with the length of the line primitive, so the dotted lines in Figure 6a and b should be identified as a part of the cluster. Hence, a new distance metric is needed which is compatible with the situations in Figure 6a and b.

Taking the advantages of HD and FD, here a novel MD metric method is provided to avoid the effects of non-uniformly distributed vertexes, heterogeneous length, inharmonious spatial alignment, and complex shape. The MD method consists of two stages, namely distance metric strategy and moving strategy.

Considering the heterogeneous geometry and the inharmonious spatial alignment of lanes, a novel moving metric method is proposed to calculate the distance and orientation between lanes.

4.1 Distance metric strategy

First, the facing projection distance (PD) metric strategy is achieved by the following:

For the given two lines, noted as $A = \{a_0, a_1, ..., a_p\}$ and $B = \{b_0, b_1, ..., b_q\}$, interpolation is conducted. For line A, for each vertex b_i on B, the travel distance between



Figure 6: Heterogeneous length and spatial alignment.

the b_i and b_0 is noted as $subLen(b_0, b_i)$ and the length of B is noted as Len(B). Create a new vertex on A at the location with the length $subLen(b_0, b_i) \times Len(A)/Len(B)$. Traverse all the vertexes on B, so that we get the interpolated A noted as $A' = \{a'_0, a'_1, ..., a'_t\}$, where t = p + q. Similarly, using the same strategy, interpolation on B is conducted and the interpolated B is noted as $B' = \{b'_0, b'_1, ..., b'_t\}$. The aim of interpolation is to reduce the effect of the unequal number and non-uniform distribution of vertexes.

With this, the distance between *A* and *B* is defined as

$$PD(A, B) = MAX\{PD(A', B'), PD(B', A')\},$$
 (2)

$$PD(A', B') = \underset{a \in A'}{\mathsf{AVG}}(\|a - B'\|),\tag{3}$$

$$PD(B', A') = \underset{b \in B'}{\text{AVG}}(\|b - A'\|),$$
 (4)

where MAX[·] means the maximum function, AVG[·] means the average function, and PD(A', B') denotes the average of facing PDs between each vertex of lines A' and B'. $\|a - B'\|$ means the facing PDs between the vertex a and the line B'. The facing PD between a vertex and a line is the distance between the vertex and its projection point on the target line.

The projection point is calculated as follows. As shown in Figure 7, L_1 and L_2 are two lines and $C(x_1, y_1)$, $D(x_2, y_2)$, and $B(x_3, y_3)$ are the vertexes. The projection point $z_t(x_t, y_t)$ of source vertex B projected on the segment linking C and D is defined as:

$$\begin{cases} x_t = x_2 + \frac{\lambda}{\mu} \Delta x_{21} \\ y_t = y_2 + \frac{\lambda}{\mu} \Delta y_{21} \end{cases}$$
 (5)

where $\lambda = \Delta x_{21} \Delta x_{32} + \Delta y_{21} \Delta y_{32} (\Delta x_{pq} = x_p - x_q, \Delta y_{pq} = y_p - y_q)$ and $\mu = \Delta x_{21}^2 + \Delta y_{21}^2$.

Here it is critical to find which segment the projection point is projected on. However, except the two endpoints, every vertex is shared by two segments. When the angle

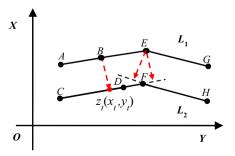


Figure 7: The calculation of projection point.

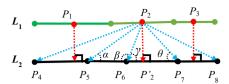


Figure 8: The position judgment of facing projection points.

of the two touching segments is sharper, it is difficult to find which segment the projection point is on. For example, the projection point of vertex E in Figure 7 will fall beyond any segments of L_2 but on the extension of the segment. Xing et al. [47] proposed the midpoint of the source segment as reference point to calculate the facing PD.

The method to find the projected segment is as follows: Taking Figure 8 as an example, midpoints of every segments of L_1 are calculated first, and they are treated as reference points. Taking P_2 as an example reference point, its projection point projected on L_2 is noted as P $'_2$. Among the segments linking P_2 and the vertexes of L_2 , the distance of P_2P_6 is the minimum. Hence, the projection point certainly falls on either P_5P_6 or P_6P_7 . According to the geometry characteristics of the triangle, if the perpendicular segment starting from one vertex of a triangle falls inside the triangle, the interior angles at the other two vertexes cannot be obtuse angle. As shown in Figure 8, both the interior angles at P_6 and P_7 of the triangle $\Delta P_2 P_6 P_7$ are not obtuse angle, so the segment linking P_6 and P_7 is suggested as the segment that the projection point P'_2 is projected on.

More details about facing PD can be found in our previous research by Xing et al. [47]. Existing research about the line distance is based on the set of the line vertexes, while the PD metric is based on every vertex of the line to another line, which makes it outstanding in dealing with inharmonious spatial alignment [47].

4.2 Moving strategy

The spatial alignment is of great importance in measuring distance and orientation between lines. When the spatial

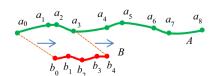


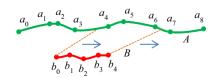
Figure 9: The moving strategy schematic of MD metric.

alignment suffers a greater difference (twice as suggested), the true distance and orientation may become greater than they are. Hence, we give more attention to spatial alignment by employing the proposed moving strategy and define a novel MD metric. The intended purpose is to get the minimum orientation and distance between two lines with large differences. The moving strategy is performed when the length of the longer line is greater than or equal to twice the length of the shorter one [49]. The moving step length, noted as *stepLen*, is set by the user. It is suggested as 1 m in this work. The comparison in Section 5.1 shows that the MD metric is more capable than PD with the situation of heterogeneous length.

Let *Len*() denote the length function. Referring to Figure 9, the moving strategy is carried out by the following steps to get the MD and orientation:

- Step 1: Let *A* be the longer line and *B* the shorter. Set interpolation step length as *stepLen* and interpolate *A* and *B*.
- Step 2: Calculate the ending condition N, N = (Len(A)-Len(B))/stepLen. Set variable i = 0.
- Step 3: Select a_0 as the starting seed vertex (i = 0), get the sub-curve A'_i with the distance of Len(B), and calculate the distance PD_i and orientation θ_i between A'_i and B.
- Step 4: If i < N, then move A'_i along A with the distance of stepLen. That is, move A'_i to the next interval location, so that A'_{i+1} is obtained. Calculate the distance PD_{i+1} and orientation θ_{i+1} between A'_{i+1} and B. i = i+1. Execute step 4 until $i \ge N$. If $i \ge N$, then go to Step 5.
- Step 5: Find the minimum value of PD_i . The corresponding PD_i and θ_i are regarded as the moving distance and orientation difference, respectively, between A and B. That is, the moving distance $MD(A, B) = MIN(PD_i)$.

The computational processes of MD mainly include point distance, point interpolation, facing PD, and the moving strategy. The computational complexities are O (1), O(t), O(t²), and O(t³), respectively. So the computational complexity of MD is O(1 + t + t² + t³) = O(t³).



5 Experiment results and discussion

5.1 Validation of MD

To check the validation and compatibility of the proposed MD for line clustering, five examples are provided, as shown in Figure 10. The dashed line is the buffer region with a distance of 5 m. From the perspective of human visual perception, it can be qualitatively found that the distance between lines in the experimental examples is approximately 10 m. Four sets of comparison experiments are designed. Compared with Figure 10a, the length of the line below in Figure 10b is smaller, the shape complexity is less in Figure 10c, the vertex number of the line below in Figure 10d is less, and the spatial alignment in Figure 10e is more heterogeneous. With the line distance, these examples include the aspects of length, shape complexity, vertex distribution, vertex number, and spatial alignment. Actually, Figure 10b-d collaboratively contribute to the shape complexity to some extent.

Besides the classic ED, HD, and FD, other two recently introduced distance metrics from Huang et al. [46] and

Xing et al. [47] are employed to compare with the MD method proposed in this article. More details of these five distance methods can be found in refs [43–47]. For short, these five metrics are noted as ED, HD, FD, HBHD, and PD. For the five examples in Figure 10, the distance results calculated by these five methods are shown in Table 1.

To understand Table 1, the readers have to compare both different rows and columns. Different columns demonstrate the different metric capabilities with the same factors. Different rows demonstrate the different metric stabilities with different factors. For example, the HD and MD in Figure 10a are 25.487063 and 10.745835. In Figure 10e, the line below only moves right side, and the spatial alignment is the more heterogeneous, While the HD and MD are 54.021876 and 9.679917. The rates of value change are |25.487063 – 54.021876|/25.487063 and |10.745835 – 9.679917|/10.745835, that is, 111.958027% and 9.919359%, respectively. This means that when measuring distance, MD has a better stability and is more robust to the heterogeneous spatial alignment.

For further analysis, statistical indicators, i.e., maximum (Max), minimum (Min), average (Avg), median (Med), and standard deviation (SD), are calculated, as shown in Table 2. Box plot is also employed to visualize the range, distribution, central value, and variability of

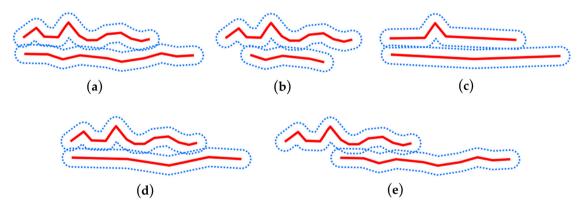


Figure 10: Example for the validation of MD.

Table 1: Distance results by different methods

| | ED | HD | FD | HBHD | PD | MD |
|---|----------|-----------|-----------|-----------|-----------|-----------|
| a | 8.564134 | 25.487063 | 19.279053 | 12.052824 | 10.745835 | 10.745835 |
| b | 8.564134 | 19.482878 | 14.942706 | 12.284965 | 13.138293 | 11.818602 |
| С | 8.739612 | 28.556984 | 18.628247 | 12.531829 | 12.314051 | 10.292981 |
| d | 8.674569 | 25.487063 | 19.011023 | 11.853981 | 11.682789 | 10.390045 |
| e | 7.393849 | 54.021876 | 45.149334 | 18.215809 | 9.679917 | 9.679917 |

| Table 2: | Statistics | of different | distance | methods |
|----------|------------|--------------|----------|---------|
|----------|------------|--------------|----------|---------|

| | ED | HD | FD | НВНD | PD | MD |
|--------|-----------|-----------|-----------|-----------|-----------|-----------|
| Max | 8.739612 | 54.021876 | 45.149334 | 18.215809 | 13.138293 | 11.818602 |
| Min | 7.393849 | 19.482878 | 14.942706 | 11.853981 | 9.679917 | 9.679917 |
| Avg | 8.387260 | 30.607173 | 23.402073 | 13.387882 | 11.512177 | 10.585476 |
| Med | 8.564134 | 25.487063 | 19.011023 | 12.284965 | 11.682789 | 10.390045 |
| Avg-10 | -1.612740 | 20.607173 | 13.402073 | 3.387882 | 1.512177 | 0.585476 |
| SD | 0.501225 | 12.071724 | 10.987031 | 2.424597 | 1.205058 | 0.705705 |

each metric, as shown in Figure 11. The box ranges from 25 to 75%. The whisker indicates variability outside the upper and lower quartiles such as the most extreme values in the dataset (maximum and minimum values).

In order to measure the deviation between the distance of different metrics and that of human visual perception, the deviation between the average and 10 m (visual distance) is calculated, noted as avg-10 in Table 2. The dashed line standing for the distance of 10 m is also illustrated in Figure 11 to visualize the deviation. It can be found that the deviation (avg-10) of MD is the smallest in Table 2, which means it is closest to the distance of human spatial cognition as shown in Figure 11.

On one hand, as stated in refs [46,47], ED measures the minimum distance between point sets without the consideration of shape, size, order, and alignment of object groups. On the other hand, though the SD of ED

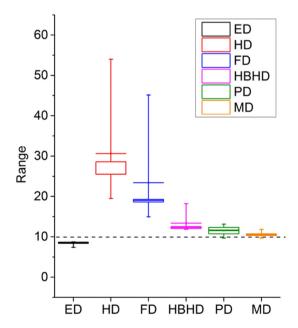


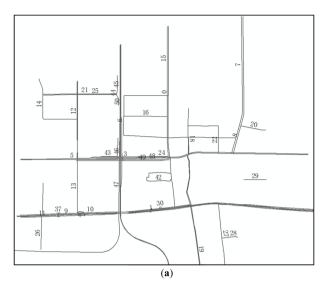
Figure 11: Box plot to display the range of each metric.

is the smallest, the avg-10 deviation is larger than that of MD. Among the other four, the variation range of MD is the smallest, which means that MD has better stability when dealing with the situation of non-uniformly distributed vertexes, heterogeneous length, inharmonious spatial alignment, and complex shape.

Actually, HBHD and PD are improved based on HD and FD. The results also reflect the fact that they perform better than the latter. On the state of the art, the applicability for line clustering from large to small is MD, PD, HBHD, FD, HD, and ED. The theory analysis, measuring stability, and visual reliability jointly proved that the proposed MD method is suitable for line clustering.

5.2 Lane-level road clustering

To verify the effectiveness of the proposed method, a real dataset from OpenStreetMap is used in the experiments. The experimental region is located in the east of Beijing City, between 39°56'25.30" N to 39°54'24.40" N latitude and 116°29'3.22" E to 116°32'14.60" E longitude, as shown in Figure 12. The semantic information of the region is uncompleted and fails to conduct the extraction work, so lane-level road clusters are extracted using the proposed LCBRG method. The distance threshold is critical and it is necessary to consider traffic planning, road design, and other factors. The lane width is the width required for safe and comfortable driving on the road, which considers the vehicle width and the extra width that is necessary for the overtaking or parallel driving. In general, the width of motorway in the city's main road is 3.5-3.75 m, and the width of the sidewalk varies from 3 to 10 m [50,51]. The width of the central isolation zone ranges from 1 to 10 m [50,51]. There may also be overpasses, toll stations, and so on. Accordingly, the distance threshold $T_{\rm d}$ of lane width in this article is set to 25 m. Therefore, the buffer radius α in RGCV is 12.5 m and the radius β in RGCH is 25 m. The orientation difference threshold $T_{\rm dir}$ is set to 30°. The parameter table is shown in Table 3.



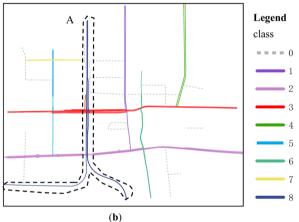


Figure 12: Lane-level road cluster input data and result. (a) The input strokes labelled with strokes IDs; (b) Lane-level road cluster result.

Table 3: The parameter table

| <i>T</i> _d (m) | T _{dir} (°) | α (m) | β (m) |
|---------------------------|----------------------|-------|--------------|
| 25 | 30 | 12.5 | 25 |

There are 119 road lanes in the experimental area, as shown in Figure 12. Because of the good organization of road networks, we use the method in ref. [42] to identify the strokes. Fifty-two strokes are recognized including the stroke-structured road group and single road lane. Using the proposed LCBRG method to mine the lane-level clusters from Figure 12, eight lane-level clusters are obtained, where 38 strokes are involved, as the colored solid class 1–8 shown in Figure 12. The details of the lane-level road clusters are listed in Table 4. Also, the statistics of lane-level road cluster result are shown in Table 5. It

Table 4: Details of lane-level road clusters

| Cluster ID | Stroke ID | Stroke number |
|------------|--|------------------|
| C1 | 0, 15 | 2 |
| C2 | 1, 2, 4, 9, 10, 11, 17, 30, 31, 32, 33, 34, 35, 36, 37, 38 | 16 |
| C3 | 3, 5, 23, 24, 43, 48, 49 | 7 |
| C4 | 7, 8 | 2 |
| C5 | 12, 13 | 2 |
| C6 | 18, 19 | 2 |
| C7 | 21, 25 | 2 |
| C8 | 45, 46, 47, 50, 51 | 5 |

Table 5: Statistics of lane-level road cluster result

| RN | SN | CN | SiC | SiC/SN | SSN |
|-----|----|----|-----|--------|-----|
| 119 | 52 | 8 | 38 | 73.08% | 14 |

RN: road number, SN: stroke number, CN: cluster number, SiC: stroke number in cluster, SiC/SN: SiC divided by SN, SSN: single stroke number.

can be seen that 73.08% strokes are recognized as lanelevel clusters. There are still 14 (52-38, as shown in Table 5) single road strokes that are not involved in the lane-level clusters, shown as the dashed class 0 in Figure 12.

In addition, there is one undesirable phenomenon in the recognized result, the cluster A shown in Figure 12, which is surrounded by the dotted line. Cluster A contains strokes {#45, #46, #47, #50, #51}. The overlap proportion of stroke {#45} and strokes {#46 and #47} is consistent with the constraint of lane-level constraint condition and therefore they are identified as part of the same cluster. There are at least the following two reasons responsible for this undesirable phenomenon:

- (1) Stroke construction algorithm. The experimental dataset does not have semantic information, so strokes are geometrically identified mainly according to the principle of good continuation. Hence, two lanes may be regarded as parts of a stroke because of their good local continuation, leading to the fact that the lanes may have large curvatures, such as the strokes #45, #46, and #47. In other applications, users should take into account both semantic information and good continuation principle to identify a stroke.
- (2) Lane-level road cluster recognition. In order to be compatible with the situation of heterogeneous length and inharmonious spatial alignment, the radius β in RGCH is relaxed. The advantage is that short lanes, such as those in the region C in Figure 13, are also

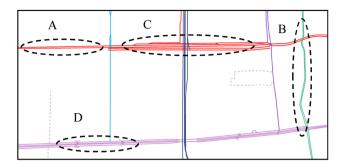


Figure 13: Enlarged view of lane-level road cluster result.

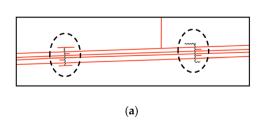
identified as parts of a cluster, which will greatly increase the robustness and universality. But the drawback is also clear that there will be some undesirable lanes such as those in the dotted region A in Figure 12. On the whole, this relaxation is worthy, because the number of short lanes is far more than the number of that with large curvature (at least 10 times in general). In other applications, the threshold of the length constraint should be set based on user demands, data characteristics, and actual situation.

An enlarged sub-view of Figure 12 is shown in Figure 13. It can be seen that the proposed method can not only extract the regular two-lane road, but also the irregular lane-level road. However, the method by Zhang et al. [48] uses a strict parallel coefficient as a quantitative indicator to extract only two-lane roads, while the proposed method uses the distance, proximity, and orientation relationships; hence, it is more advantageous in robustness. Consequently, some noise lanes that satisfy T1 or T2 are assigned to clusters, for example, the region C in Figure 13. This means that the proposed method is more general that Zhang et al. [48]. Assigning these noise lanes to clusters makes the concept of semantic road entity more sufficient.

It should be noted that the overpasses are usually modeled as lines and stored together in the road database, as the region D shown in Figure 13, which is enlarged and shown in Figure 14a. This may lead to the fact that parts of the facilities such as steps or escalators along the road may be identified as parts of lane-level clusters. Although it is reasonable in geometry and social function, it destroys the semantic integrity of overpass, and will affect the actual width of the detected road clusters. An overpass is a bridge, road, railway, or similar structure that crosses over another road. The steps or escalators are components of overpass structure and cannot be individually assigned to the cross road. Hence, steps or escalators should be treated as a whole according to their semantic information or other more excellent methods.

In addition, due to data collection, data quality, and other reasons, some lanes may be visually involved in the composition of multiple lane-level road clusters. In the "L" shaped #43 lane shown in Figure 14b, the corner vertex at the right angle visually turns the lane into two parts, the horizontal part of #43 lane is visually adjacent to the cluster C3, while the vertical part is visually adjacent to the cluster C8. The length proportion that each part falls into the corresponding cluster is approximately the same. Since the distance between the road and the cluster C3 is smaller, it is recognized as a part of cluster C3. In other applications, distance, length proportion of each cluster, and other factors may be taken into account.

The validation of the proposed method is verified using the local region of Beijing City. The undesirable phenomenon and possible affecting factors are also discussed. Furthermore, experiments are carried out on the whole city of Beijing. In the result shown in Figure 15, the red lines are the clusters recognized using the proposed method, while the gray ones are unassigned roads. It can be seen that the recognized lane-level clusters act as the skeleton and framework of road networks.



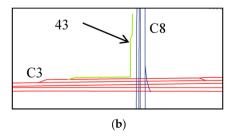


Figure 14: Two undesirable examples: (a) overpass in lane-level road cluster dataset and (b) lane involved in the composition of multiple clusters.

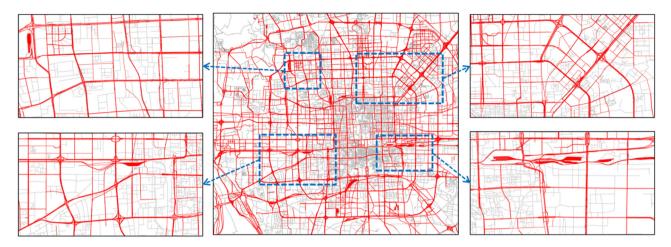


Figure 15: The lane-level road cluster result of Beijing City.

In urban planning, the lane-level roads are always the ones that are important due to the amount of traffic flow and the aim is to relieve the traffic pressure. It means that compared with single road, the lane-level roads reflect the importance and capacity to some extent. The characteristics of lanes measure the hierarchy of the road networks. In general, the more the number of the lanes is, the greater the capacity of the road is, and the more important the road is. Hence, the lane-level road always has high priority in cartographic generalization and multi-scale representation.

The experiment is conducted on a laptop equipped with a Microsoft Windows 10 64-bit operating system. The central processing unit (CPU) is an Intel Core i7-8750H, and the memory (RAM) is 32 GB in size. The total computation time of the Beijing case is 311.22 s. The proposed algorithm with high complexity is still time-consuming and may affect the scalability. For example, when the data volume is very large, it may run timeout and fail to get the result. Spatial partitioning and indexing strategy such as hashing, trees, and Morton index may be referred to speed up the searching process and reduce the number of candidates for testing.

6 Conclusion

The collective map generalization of object group is one of the difficulties of map generalization. Lane-level road clusters are common in the road network dataset. However, there is few if any research on the line group generalization. In addition, line group identification is one of the most difficult fields. This article analyzes the concept of lane-level road cluster and its causes, offers the effective

spatial constraints, and provides the basic strategy for line cluster recognition, which provides strong support for map generalization of the line group. Further research includes: the identification of high-level semantic structures (such as road roundabout and stack interchange) of road networks through geometric features and group-based line generalization operations such as simplification, typification, and so on.

Acknowledgments: This study was supported by the National Natural Science Foundation of China (Grant No. 41801396).

Author contributions: Xianyong Gong and Fang Wu conceived and designed the research; Xianyong Gong and Chengyi Liu helped in experiments and data analysis; Ruixing Xing and Jiawei Du helped in paper organization and language correction; and Xianyong Gong wrote the paper.

Conflict of interest: Authors state no conflict of interest.

References

- [1] Tang L, Yang X, Kan Z, Li Q. Lane-level road information mining from vehicle GPS trajectories based on naïve Bayesian classification. ISPRS Int J Geo-Inf. 2015;4:2660–80.
- [2] Tang L, Xue Y, Zhen D, Li Q. LRIC: collecting lane-based road information via crowdsourcing. IEEE T Intell TranspSyst. 2016;17:2552-62.
- [3] Turk MA, Morgenthaler DG, Gremban KD, Marra MM. VITS a vision system for automated land vehicle navigation. IEEE T Pattern Anal Mach Intell. 1988;10:342–61.

- [4] Cao D, Jiang Y, Wang J, Ji B, Liu Y. ARNS: adaptive relay-node selection method for message broadcasting in the Internet of vehicles. Sensors. 2020;20:1338.
- [5] Gao Z, Long K, Li C, Wu W, Han LD. Bus priority control for dynamic exclusive bus lane. Comput Mater Con. 2019;61:345-61.
- [6] Liu W, Tang Y, Yang F, Dou Y, Wang J. A multi-objective decision-making approach for the optimal location of electric vehicle charging facilities. Comput Mater Con. 2019;60:813–34.
- [7] Liu W, Tang Y, Yang F, Zhang C, Cao D, Kim GJ. Internet of Things-based solutions for transport network vulnerability assessment in Intelligent Transportation Systems. Comput Mater Con. 2020:65:2511–27.
- [8] Wang J, Tang Y, He S, Zhao C, Sharma PK, Alfarraj O, et al. LogEvent2vec: LogEvent-to-Vector-based anomaly detection for large-scale logs in Internet of Things. Sensors. 2020;20:2451.
- [9] Waqas M, Tu S, Rehman SU, Halim Z, Anwar S, Abbas G, et al. Authentication of vehicles and road side units in intelligent transportation system. Comput Mater Con. 2020;64:359-71.
- [10] Savino S, Touya G. Automatic structure detection and generalization of railway networks. Int Cartogr Conf 2015, ICA. Rio de Janeiro, Brazil: Aug, 2015.
- [11] Yeh AGO, Zhong T, Yue Y. Hierarchical polygonization for generating and updating lane-based road network information for navigation from road markings. Int J Geogr Inf Sci. 2015;29:1509–33.
- [12] Li Q, Fan H, Luan X, Yang B, Liu L. Polygon-based approach for extracting multilane roads from OpenStreetMap urban road networks. Int J Geogr Inf Sci. 2014;28:2200–19.
- [13] Yang B, Zhang Y, Lu F. Geometric-based approach for integrating VGI POIs and road networks. Int J Geogr Inf Sci. 2014:28:126–47.
- [14] Liu W, Wang J. Evaluation of coupling coordination degree between urban rail transit and land use. Int J Commun Syst. 2019;34(2):e4015.
- [15] Chen Q, Gan X, Huang W, Feng J, Shim H. Road damage detection and classification using mask R-CNN with DenseNet backbone. Comput Mater Con. 2020;65:2201–15.
- [16] Chehreghan A, Abbaspour AR. A geometric-based approach for road matching on multi-scale datasets using a genetic algorithm. Cartogr Geogr Inf Sci. 2018;45:255–69.
- [17] Guo W, Liu T, Dai F, Xu P. An improved whale optimization algorithm for feature selection. Comput Mater Con. 2020;62:337-54.
- [18] Gong X, Wu F. A typification method for linear pattern in urban building generalisation. Geocarto Int. 2018;33:189–207.
- [19] Izakian Z, Mesgari MS, Abraham A. Automated clustering of trajectory data using a particle swarm optimization. Comput Env Urban Syst. 2016;55:55–65.
- [20] Zhao P, Qin K, Ye X, Wang Y, Chen Y. A trajectory clustering approach based on decision graph and data field for detecting hotspots. Int J Geogr Inf Sci. 2016;31:1101–27.
- [21] Deng M, Liu Q, Cheng T, Shi Y. An adaptive spatial clustering algorithm based on delaunay triangulation. Comput Env Urban Syst. 2011;35:320–32.
- [22] Liu Q, Deng M, Shi Y, Wang J. A density-based spatial clustering algorithm considering both spatial proximity

- and attribute similarity. Comput Geosci-UK. 2012;46:296–309.
- [23] Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases. Proceedings of the ACM SIGMOD International Conference on Management of Data. Montreal, Quebec, Canada: Association for Computing Machinery. 1996. p. 103–14.
- [24] Ng R, Han J. Efficient and effective clustering method for spatial data mining. Proceedings of the 20th International Conference on Very Large Data Bases. Santiago, Chile: 1994. p. 144-55
- [25] Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B. 1977;39:1–38.
- [26] Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. International Conference on Knowledge Discovery in Databases and Data Mining (KDD-96), Portland; 1996. p. 226–31.
- [27] Lu L, Wu J, Liu Z. Data clustering of contour lines based on shape characteristics. Acta Geod Cartogr Sin. 2005;34:138-41.
- [28] Liu S, Ji G, Li W. Spatial line clustering algorithm based on connectivity. Comput Sci. 2011;38:179–81.
- [29] Zhu J. Line outliers detection based on topological relationships. Master Dissertation. Nanjing China: Nanjing Normal University; 2011.
- [30] Gong X. Research on settlement generalization methods considering spatial pattern and road networks. PhD thesis. Zhengzhou China: Information Engineering University; 2017.
- [31] Tang L, Ren C, Liu Z, Li Q. A road map refinement method using delaunay triangulation for big trace data. ISPRS Int J Geo-Inf. 2017:6:45.
- [32] Luan X, Yang B. Generating strokes from city road networks. Geogr Geo-Inf Sci. 2009;25:49-52.
- [33] Antoniou V, Skopeliti A. Measures and indicators of VGI quality: an overview. ISPRS Ann Photogramm Remote Sens Spatial Inf Sci, Vol II-3/W5. La Grand Motte, France; 2015. p. 345–351.
- [34] Raimond AMO, Hart G, Touya G, Kellenberger T, Foody GM, Demetriou D. The scale of VGI in map production: a perspective on European National Mapping Agencies. T GIS. 2017;21:74–90.
- [35] Qian H, Lu Y. Simplifying GPS trajectory data with enhanced spatial-temporal constraints. ISPRS Int J Geo-Inf. 2017;6:329.
- [36] Biagioni J, Eriksson J. Inferring road maps from global positioning system traces: survey and comparative evaluation. Transp Res Rec J Transp Res Board. 2012;2291:61–71.
- [37] Reinoso JF, Ariza-López FJ, Barrera D, Gómez-Blanco A, Romero-Zaliz R. A fitted B-spline method to derive a representative 3D axis from a set of multiple road traces. Geocarto Int. 2016;31:832–44.
- [38] Zhang J, Xie Z, Sun J, Zou X, Wang J. A cascaded R-CNN with multiscale attention and imbalanced samples for traffic sign detection. IEEE Access. 2020;8:29742-54.
- [39] Horita FEA, Degrossi LC, Assis LFFG, Zipf A, Albuquerque JPD. The use of Volunteered Geographic Information (VGI) and crowdsourcing in disaster management: a systematic literature review. Chicago, Illinois, USA: AMCIS 2013 Proceedings; 2013.

- [40] Camponovo ME, Freundschuh SM. Assessing uncertainty in VGI for emergency response. Cartogr Geogr Inf Sci. 2014;41:440–55.
- [41] Gong X, Xing R, Li J. Spatial alignment relationship and its quantitative description. Eng Surv Mapp. 2017;26(7–11):17.
- [42] Thomson R, Richardson D. The 'good continuation' principle of perceptual organization applied to the generalization of road networks. Proceedings of the 19th ICA, Ottawa; 1999. p. 1215–23
- [43] Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing images using the Hausdorff distance. IEEE T Pattern Anal Mach Intell. 1993;15:850–62.
- [44] Mascret A, Devogele T, Le Berre I, Hénaff A. Coastline matching process based on the discrete Fréchet distance. Proceedings of the 12th International Symposium on Spatial Data Handling. Vienna, Austria: Springer; 2006. p. 383–400
- [45] Mustière S, Devogele T. Matching networks with different levels of detail. Geoinformatica. 2008;12:435-53.

- [46] Huang B, Wu F, Xu J, Zhai R, Gong X. A method of distance measurement for corresponding linear feature. Geomat Inf Sci Wuhan Univ. 2017;42:398–401.
- [47] Xing R, Wu F, Zhang H, Gong X. Dual-carriageway road extraction based on facing project distance. Geomat Inf Sci Wuhan Univ. 2018;43:152–8.
- [48] Zhang H, Wu F, Gong X, Xu J, Zhang J. A parallel factor-based method of arterial two-lane roads recognition. Geomat Inf Sci Wuhan Univ. 2017;42:1124-30.
- [49] Fu Z, Yang Y, Gao X, Zhao X, Lu Y, Chen S. Road networks matching using multiple logistic regression. Geomat Inf Sci Wuhan Univ. 2016;41(2):171-7.
- [50] American Association of State Highway and Transportation Officials. A policy on geometric design of highways and streets; 2001.
- [51] Ministry of Housing and Urban-Rural Development of the People's Republic of China. Road traffic signs and markings part 3: road traffic markings. GB 5768.3-2009; 2009.