Minimizing uncertainty in complete automation of Douglas-Peucker Algorithm for geospatial mapping

Rong Wang  
Lanzhou Jiaotong University

Muhammad Shafeequ (✉️ shafeequ@uni-bremen.de)  
University of Bremen: Universitat Bremen  https://orcid.org/0000-0002-2152-6672

Haowen Yan  
Lanzhou Jiaotong University

Lu Xiaoming  
Lanzhou Jiaotong University

Research Article

Keywords: Optimal distance threshold, Map scale change, Automatic map generalization, Geometric accuracy, Spatial similarity relationships

Posted Date: December 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1162837/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Minimizing uncertainty in complete automation of Douglas-Peucker Algorithm for geospatial mapping

Abstract

It is qualitatively evident that the greater the map scale change, the greater the optimal distance threshold of the Douglas-Peucker Algorithm, which is used in polyline simplification. However, no specific quantitative relationships between them are known by far, causing uncertainties in complete automation of the algorithm. To fill this gap, the current paper constructs quantitative relationships based on the spatial similarity theories of polylines. A quantitative spatial similarity relationship model was proposed and evaluated by setting two groups of control experiments and taking \(<C, T>\) as coordinates. In order to realize the automatic generalization of the polyline, we verified whether these quantitative relationships could be fitted using the same function with the same coefficients. The experiments revealed that the unary quadratic function is the best, whether the polylines were derived from different or the same geographical feature area(s). The results also show that using the same optimal distance threshold is unreasonable to simplify all polylines from different geographical feature areas. On the other hand, the same geographical feature area polylines could be simplified using the same optimal distance threshold. The uncertainties were assessed by evaluating the automated generalization results for position and geometric accuracy perspectives using polylines from the same geographic feature areas. It is demonstrated that in addition to maintaining the geographical features, the proposed model maintains the shape
characteristics of polylines. Limiting the uncertainties would support the realization of completely automatic generalization of polylines and the construction of vector map geodatabases.

**Keywords:** Optimal distance threshold; Map scale change; Automatic map generalization; Geometric accuracy; Spatial similarity relationships

1. Introduction

In realistic cartography, different cartographers always select different maps as the generalization results, and in most situations, cartographers often need to input parameter(s) in the execution of an algorithm. This humans' constant interference consequently leads to the non-automation of generalized algorithms, which decreases the efficiency of map generalization and increases the uncertainty of the resulting maps (Yan 2014; Makris et al., 2021). Linear data such as contour, road network, and rivers are the main spatial data types widely used in different fields. In the process of downsizing small-scale maps or constructing vector map databases, it is indispensable to generalize polylines from one larger scale (e.g., 1:5000) to corresponding generalized counterparts at the other smaller scale (e.g., 1:50000). Therefore, it is extremely essential to realize the completely automatic generalization of polylines.

Numerous scholars have been making tremendous efforts to investigate the realization of automatic polyline generalization issues (Ai et al. 2016; Shen et al. 2018; Kronenfeld et al. 2020). Meanwhile, various methods have been presented (Mcmaster 1987), such
as the Douglas-Peucker algorithm (DP) (Douglas and Peucker 1973), Li-Openshaw (Li and Openshaw 1992), Bend Group algorithm (Qian HZ 2017). However, the problems are only partially solved. Among these approaches, the DP remains the most well-known and effective polyline simplification algorithm in map generalization (Ramer 1972; Hershberger and Snoeyink 1992; Saalfeld 1999; Yan 2014; Sandu et al., 2015). The evaluation and minimization of uncertainties in polyline simplification methods might validate these algorithms' acceptability, consistency, and advanced application for automatization.

Spatial similarity relationships are critically important measures for the accuracy assessment of map generalization results (Yan 2015, 2016; Makris et al., 2021), which have attracted the increasing attention of cartographers in recent years. However, few studies have been published on the automatic generalization of polylines based on spatial similarity relationships. The reason mainly focuses on three aspects: Firstly, no in-depth investigations on integrating spatial similarity degree into map automatic generalization are available. Only a few studies have been tested (Yan 2014), i.e., it is ambiguous about how to automatically calculate the transition conditions threshold of DP using the spatial similarity degree. As a result, it leads to constant human intervention in the execution of DP. Secondly, very little is known about the specific quantitative functional relationship between optimal distance threshold and map scale change. Is it linear or nonlinear? Is it a linear equation with one variable or a quadratic equation, or is it a logarithmic function? These questions are still being sought. Thirdly,
when referring to the effect of regional geographical features on selecting distance threshold, some cartographers believe that it should be considered, while others argue that there is no need (Li and Sui 2000). Besides, they assume that the reason why the map is generalized is only the change of map scale, while other surrounding factors remain unchanged. For example, regardless of the chosen threshold, the degree of terrain fragmentation remains unchanged before and after the generalization. Hence, the conflicting opinions adversely influence the determination of the optimal distance threshold and cause uncertainty. A computational model based on quantitative spatial similarity relationships might overcome the challenge of realizing an entirely automatic generalization of the polylines.

Constructing the quantitative functional relationships between the optimal distance threshold \( T \) and map scale change \( C \) would realize the completely automatic generalization of polylines based on the DP without manual parameterization. Meanwhile, the optimal distance threshold mainly depends on the magnitude of map scale change. Technically, when the spatial similarity degree between the intermediate polyline \( L_i \) simplified by DP and the standard target scale polyline (LM2) reaches the maximum, the corresponding threshold is obviously the optimal distance threshold. The main goal of the current research is to minimize the uncertainties in determining the optimal distance threshold for improving the complete automation of the DP algorithm.
2. Methodology

In order to automatically calculate the optimal distance threshold and realize the completely automatic generalization of the individual polyline, this paper firstly constructs and validates the spatial similarity evaluation model by quantitatively selecting the evaluation index and determining its weight, and then calculates the optimal distance threshold based on the spatial similarity relations. Afterward, the optimal distance threshold and corresponding map scale change are recorded when the similarity degree between the intermediate result and the standard target scale polyline reaches the maximum. Then taking \( <C, T> \) as the coordinates, the best quantitative empirical formula is fitted using the regression analysis and least square model. The generalization results are finally validated from the position and geometric accuracy perspectives to confirm the acceptability of the introduced methods.

2.1. Principle of parameter optimization of DP algorithm

Douglas-Peucker algorithm (DP) is a well-known and effective simplification algorithm of polylines (Douglas and Peucker 1973). DP has been widely used in map generalization to reduce the number of points on a curve (Yan 2014; Gu et al. 2016). According to its simplification fundamentals, we can clearly know that the degree that polyline is simplified mainly depends on the threshold of each simplification process, i.e., the initial threshold and step size. Moreover, in which points are deleted, the order is the opposite of that when points are reserved. However, it is consistent with the
process of DP. The optimal distance threshold determines the number of points to be
deleted and the degree to that a polyline is generalized. Hence, the current research is
aimed to find an appropriate approach for determining the quantitative relationship
between T and C. To this end, the order that points to be deleted is firstly determined
by reversing the order of points to be selected. As a result, the similarity between the
intermediate result and the standard target scale polyline reaches the maximum; the
 corresponding threshold $\lambda_k$ is the optimal distance threshold that the polyline is
generalized from scale $M_1$ to scale $M_2$. Finally, the point pairs (C, T) used to fit the
quantitative relations between them are recorded after multiple circulative iterations
(Figure 1).

![Figure 1. Flow chart of the methodology applied in this study. $\lambda_0$ is the initial distance threshold. $\Delta \lambda$ is a
gradually changing value, and the value of $\Delta \lambda$ is inversely correlated with the similarity degree between $L_{M1}$
and $L_i$. For example, suppose the polyline is generalized from 1:50000 to 1:250000, then if the similarity degree
(S) between $L_{M1}$ and $L_i$ is less than 0.6, $\Delta \lambda$ will be 0.02 km, while if S is more than 0.6 and less than 0.8, $\Delta \lambda$
will be 0.01 km; and if S is more than 0.8, $\Delta \lambda$ will be 0.005 km.](image-url)
2.2. Selection of evaluation index and determination its weights

The judgment of spatial similarity degree is the essence of map generalization. However, people often ambiguity about the similarity between two objects when judging the spatial similarity relationship between multi-scale maps because they remain uncertain about what properties should be considered in similarity assessments (Yan 2010). Therefore, among many polyline properties, including length, distance, complexity, sinuosity, etc., it is necessary to extract major properties that influence people's judgment at first and then determine its corresponding weights to construct a spatial similarity evaluation model.

According to previous research, 12 properties can be used to evaluate the spatial similarity between multi-scale polylines, including position (Olteanu-Raimond et al. 2015), length (Zhang et al. 2014), distance (Deng 2007, Tong et al. 2014), size, sinuosity (Danlel et al. 2014), complexity, orientation (Samal 2004, Olteanu-Raimond et al. 2008), area, shape, buffer-overlapped area, degree of node and attribute. These can be further divided into geometric properties and thematic attribute properties (Li et al. 2006, Zhang 2009). Due to the variety of the geometric properties, it is not clear to choose what combination of geometric properties to produce the best solution. In addition, if all properties have been considered, correlated and redundant information is likely imported, which may lead to low computational efficiency and the accuracy of
the result. This phenomenon can be explained by the curse of dimensionality or Hughes's effect (Hughes 1968). Using as few indicators as possible to express as much polyline information as possible should be solved first. Therefore, absolute distance matrix $D$ between factors is employed to overcome the above problem and measure the closeness of the indicators based on the results of genetic algorithm.

\[
D_{12 \times 12} = \begin{bmatrix}
0 & 3.22 & 0 \\
3.22 & 0 & 1.94 & 3.49 & 0 \\
4.22 & 2.46 & 3.95 & 0 & 5.14 \\
5.14 & 4.41 & 4.13 & 3.56 & 0 \\
1.23 & 3.39 & 1.99 & 4.51 & 5.24 & 0 \\
1.94 & 3.22 & 1.60 & 3.71 & 4.39 & 1.99 & 0 \\
1.45 & 3.18 & 2.39 & 4.46 & 5.37 & 0.99 & 2.72 & 0 \\
5.04 & 3.25 & 4.46 & 2.32 & 2.73 & 5.27 & 5.49 & 5.49 & 0 \\
3.16 & 2.98 & 3.73 & 4.59 & 4.60 & 3.79 & 4.79 & 3.19 & 4.79 & 0 \\
5.28 & 4.53 & 4.71 & 3.95 & 3.86 & 6.02 & 4.12 & 5.72 & 4.04 & 4.12 & 0 \\
4.21 & 4.06 & 4.02 & 4.04 & 2.62 & 4.88 & 3.42 & 4.89 & 3.45 & 3.42 & 3.42 & 0
\end{bmatrix}
\]

where, $R$ is a symmetric matrix, and $d_{ij} = \sum_{k=1}^{n} |w_{ik} - w_{jk}|$ ($i, j = 1, 2, \ldots, 12, k = 1, 2, \ldots, 9$), $w_{ik}$ represents the kth factor that affects the ith data group in Table 5 of Chehreghan's research.

And then, the direct cluster genealogy diagram (Figure 2) is derived based on the matrix $R$. 
According to the F statistical analysis of cluster analysis, $F > F_\alpha(f_1, f_2) = 2.494$ ($\alpha = 0.05$, $f_1 = r - 1 = 5$, $f_2 = n - r = 3$), which indicates that the classes in Figure 2 are significantly different, the classification strategy is reasonable. It can be determined that $Z$ belongs to $(2.46, 2.62)$ because the attribute as the thematic property can be extracted as an independent class, meanwhile, the clustering degree of other geometric properties reaches maximum. Finally, the impact indices are clustered into six categories: 

\{Attribute\}, \{Sinuosity\}, \{Degree of node\}, \{Position, Complexity, Area, Distance, Orientation\}, \{Size, Shape, Length\}, \{Buffer-overlapped area\}. In this paper, only geometric properties were considered because, in most situations, cartographers pay more attention to the geometric attributes of polylines and ignore their thematic attribute information (Deng 2007). It should be noted that \{degree of node\}, \{Position, Complexity, Area\} and \{Size, Length\} are excluded from the selected impact factors
since they impact only a few part groups of the dataset. Finally, the chosen geometric factors are \{Sinuosity\}, \{Distance, Orientation\}, \{Shape\} and \{Buffer-overlapped area\}. Except for distance and orientation, the correlation coefficient $|R_{xy}|$ between the selected geometric factors is less or equal to 0.3, which is consistent with the basis for the exclusivity of index selection. Although distance and orientation were clustered in one class, and their correlation coefficient $R_{xy}$ is 0.76, they impact all groups of datasets in stability test, which indicates that both factors are very stable. Thus, Sinuosity, Distance, Orientation, Shape, and Buffer-overlapped area are chosen. It is generally not allowed to move polylines on maps in the process of map generalization, i.e., orientation between polylines is not changed after map generalization. Therefore, Sinuosity, Distance, Shape, and Buffer-overlapped area are finally chosen as evaluation indices.

**Distance** is a fundamental concept in geospatial science. According to the previous research, it can be known that compared with Euclidian distance (Peuquet 1992), Bottleneck distance (Efrat et al. 2001), and Fréchet distance (Nayyeri et al. 2015), Hausdorff distance is one of the most used distances for spatial objects in GIS, but it is sensitive to the shape of the objects, especially to the outliers. Besides, it does not satisfy the change law of similarity (Li et al., 2018). Therefore, Mean- Hausdorff distance (MHD) is selected as the distance similarity metric to obtain more stable and accurate result (Deng et al. 2007). On the other hand, the shape is viewed as the most crucial geometric factor that describes planar curves. The essence of shape similarity is to judge
the coincidence degree between polylines. Therefore, Li (2018) comparatively assessed the advantages and disadvantages of different measurement indices through experiments, e.g., Vertex Reserved, Length Ratio, Turning Function, and Dual-side Bend Forest Shape Similarity. The results indicated that the cumulative turning function (Fan et al. 2014) (Equation (1)) could eliminate the interference of detail difference, which is more suitable for the comparison between multi-scale polylines of the same target.

\[
\overline{\text{Sim}_{L_1,L_2}^{\text{shp}}} = (\int f(\theta_{L_1} - \theta_{L_2}) ds)^{\frac{1}{2}} = (\int_0^1 T_A(L_1) ds - \int_0^1 T_A(L_2) ds)^{\frac{1}{2}}
\]  

(1)

where, \( f(\theta_{L_1} - \theta_{L_2}) = \begin{cases} |\theta_{L_1} - \theta_{L_2}| & (|\theta_{L_1} - \theta_{L_2}| \leq 180^\circ) \\ 360^\circ - |\theta_{L_1} - \theta_{L_2}| & (|\theta_{L_1} - \theta_{L_2}| > 180^\circ) \end{cases} \). The \( \overline{\text{Sim}_{L_1,L_2}^{\text{shp}}} \) is inversely correlated with the similarity degree between \( L_1 \) and \( L_2 \). If the angle is clockwise, \( \alpha_i \) is positive, and if the angle is counterclockwise, \( \alpha_i' \) is negative. The relation between \( \alpha_i \) and \( \alpha_i' \) is that \( \alpha_i = 360^\circ - |\alpha_i'| \). For instance, if \( \alpha_i' = -18^\circ \), then \( \alpha_i = 360^\circ - 18^\circ = 342^\circ \).

![Figure 3. Cumulative angle function of the multi-scale polylines L_1 and L_2.](image)
Buffer-overlapped area can reflect the horizontal distance and vertical distance of two
delineation entities. However, it is worth noting that the threshold of buffer radius is the
key to solving various research problems. According to the visual perception degree of
position deviation on the figure, buffer radius of the buffer-overlapped area is generally
set as 1~2 times of the minimum distance (0.2mm) between two points on the target
scale map, e.g., if the target scale is 1:250000, then the buffer radius is 50m.

However, suppose we want to study the mechanism of spatial similarity varying with
map scale change. In that case, map scale change must be the unique independent
variable, and the same set of weights for different groups of datasets should be used to
eliminate the influence of weights variation, which may adversely influence the results.
Therefore, nine groups of weights of four stability factors are normalized to the same

group of weights (Table 1) by using the Sum-Product method.

| Impact factor       | Distance | Sinuosity | Shape  | Buffer-overlapped area |
|---------------------|----------|-----------|--------|------------------------|
| Weight              | 0.32     | 0.23      | 0.11   | 0.34                   |

The weights of sinuosity, distance, shape, buffer-overlapped area are 0.23, 0.32, 0.11,
and 0.34, the critical ranking of which is consistent with the weights obtained from the
opinions of experts based on spatial cognition experiment or the well-adopted criteria
(Chehreghan et al. 2016). The large weight value indicates that the corresponding factor
strongly influences the judgment of spatial similarity relations and vice versa. Thus,
according to the value of the adopted weights, the buffer-overlapped area has the
greatest impact, followed by distance.

2.3. Construction of spatial similarity evaluation model

According to the definition and description of each index in Section 2.2, it is known to all that some factors are positive indicators, such as distance, shape, sinuosity, while others are negative, i.e., the larger $\text{Sim}^{\text{Dis}}_{L_iL_j}$, indicates the less similarity between polylines, while the $\text{Sim}^{\text{Pref}}_{L_iL_j}$ is an opposite factor. Moreover, the units and dimensions of factors are different, and their value varies greatly, which also has a great impact on the results. Therefore, to eliminate the effects of negative factors and different units on results, the similarity degree of these factors should be normalized to $[0, 1]$ before determining the spatial similarity evaluation model, which is achieved by using the most applicable technique range standardization as follows:

$$
\text{Sim}^P_{L_iL_j} = \begin{cases} 
\frac{\text{Sim}^P_k - \text{MIN}_{1 \leq i \leq N} \text{Sim}^P_{L_iL_j}}{\text{MAX}_{1 \leq i \leq N} \text{Sim}^P_{L_iL_j} - \text{MIN}_{1 \leq i \leq N} \text{Sim}^P_{L_iL_j}} & \text{when } \text{Sim}^P_k \text{ is a positive index.} \\
\frac{\text{MAX}_{1 \leq i \leq N} \text{Sim}^P_{L_iL_j} - \text{Sim}^P_k}{\text{MAX}_{1 \leq i \leq N} \text{Sim}^P_{L_iL_j} - \text{MIN}_{1 \leq i \leq N} \text{Sim}^P_{L_iL_j}} & \text{when } \text{Sim}^P_k \text{ is a negative index.}
\end{cases}
$$

(2)

$\text{Sim}^P_{L_iL_j}$ is the similarity degree of the factor $P_k$ after normalized, where the large $\text{Sim}^P_{L_iL_j}$ indicates the higher spatial similarity. $N$ is the number of groups of datasets.

Finally, Equation (3) (Yan 2014) can be employed to calculate the spatial similarity degree between polylines based on four factors.

$$
\text{Sim}(L_i, L_j) = \sum_{k=1}^{n} W_p \text{Sim}^P_{L_iL_j}
$$

(3)
where, $\text{Sim}(L_1, L_2) \in (0,1]$, $W_{pk} \in [0,1]$, $n$ is the total number of impact factors, which is 4.

### 2.4. Geometric accuracy assessment

In the generalization of an individual polyline, the influence of simplification on the geometric accuracy is mainly reflected in the change of geometric features and the position of points on the polyline (Wu F et al. 2008, Chen et al. 2016). The variations of the former are mainly represented by the changes in curve length and sinuosity. The variations of the latter are mainly manifested as the global or local displacement before and after the generalization of the polyline. Such variations are evaluated using the standardized deviation of displacement (SMD), position error (PE), vector displacement, area displacement, and buffer tolerance. Among these indices, the SMD mainly evaluates the local maximum, and it is the most suitable index for comparing the different simplification results of the same linear features (Wu et al. 2008).

Furthermore, PE is often used to evaluate the overall displacement (White 1985; McMaster 1987). In this research work, the geometric accuracy of the simplified polyline is evaluated based on the length ratio (LR) (Chehreghan and Abbaspour 2018), sinuosity degree (SD) (Danlel and Yang 2014), SMD (Chen 2016), and PE.

By comparing the length ratio of the polyline, it can be seen how much the simplification algorithm compresses the points on the polyline, and the ratio is inversely correlated with the amount of compression (Wu 2008). The degree of curvature can be obtained by sum the included angles of adjacent straight lines on the polyline. If the
length of polylines changes more and the sinuosity degree changes less, this indicates that the geometric features of the simplified polyline are better maintained. The smaller value of SMD represents a smaller displacement of a feature on the map. This implies a lesser likeliness of intersection, which highlights that the spatial relationship is maintained very well. Since the SMD can only be used to obtain the maximum value of the polyline's local displacement, it cannot properly describe the overall displacement. So, the PE is used to evaluate the positional accuracy of a polyline. The smaller the PE value is, the better the accuracy of the point position on the curve is maintained. Furthermore, the simplification result can effectively maintain the overall shape characteristics of polylines.

3. Experiments and Results

According to different geographical features, the plain of China is divided into the Northeast China Plain, the North China Plain, and the Yangtze Plain. Thirty-eight groups of multi-scale datasets of rivers and roads, which include 1:10k, 1:25k, 1:50k, 1:100k, 1:250k, and 1:500k, were selected to calculate the quantitative relationships between S and C. These data are divided into two major group categories: control and experimental. The experimental group is composed of the multi-scale polylines datasets from the same geographical feature area, each group of the dataset includes river and road from the same sample area, and the former is composed of the multi-scale polylines datasets from the different geographical feature areas, each group of the
dataset includes river or road from the different sample areas. These multi-scale datasets include 12 groups of river samples from the Huaibei Plain (Sample area I), which belong to a part of the North China Plain, seven groups of river samples from the Northeast China Plain (Sample area II). There are 19 groups of samples from the lower Yangtze Plain (Sample area III), including 12 groups of the river and 7 groups of roads, which are separately sampled from the Jianghuai Plain and the Yangtze River Delta Plain, because both of which are part of the lower Yangtze Plain, we assume they share the same geographic features. These multi-scale vector data, as mentioned above, were derived from the data vectorization results of GF-2 images under different scales using ArcGIS10.6. The GF-2 image is obtained by the fusion of Pan and multi-spectral image, whose spatial resolution is 1m and 4m, respectively. Multi-scale datasets of rivers and roads, which include 1:5w, 1:10w, 1:25w, 1:50w, and 1:100w, were derived from the current results of the National Geomatics Center of China (NGCC).
Figure 4. Sampling area of experimental data and sampling data of multi-scale river and road in the plain area.

3.1 Validation of the proposed spatial similarity evaluation model

It is qualitatively evident that the greater the map scale change, the smaller the spatial similarity degree between multi-scale polylines. However, no specific quantitative relationships between them are known by far, which hampers the complete automation of the algorithm. There are five potential candidate functions to map this gap that can describe such changing trends, including linear function, polynomials, power functions,
logarithmic functions, and exponential function. The three or more order polynomials have \( n-2 \) inflection point(s). Hence the curve has more than one monotonicity decreasing interval. Ultimately, only second-order polynomials should be considered.

\[
\begin{align*}
  y &= a_2x^2 + a_1x + a_0 \quad (a_2 > 0, x \in \left[1, -\frac{a_1}{2a_2}\right]) \\
  y &= a\ln(x) + a_0 \quad (a < 0, x \geq 1) \\
  y &= ax^b \quad (a > 0, -1 < b < 0, x \geq 1) \\
  y &= ae^{bx} \quad (a > 0, b < 0, x \geq 1) \\
  y &= ax + b \quad (a < 0, x \geq 1)
\end{align*}
\]

Taking \( <C_{S_i,S_j}, \text{Sim}(L_i,L_j)> \) as coordinates, the unknown coefficients of the Equation can be determined using the curve fitting approach. Here, for simplification, \( \text{Sim}(L_i,L_j) = f(C_{S_i,S_j}) \) is replaced with \( S = f(C) \). Statistical analysis is conducted based on Microsoft Excel (V10.0) and the least square method \( R^2 \) (\( R^2 \in [0,1] \)) is employed to choose the best-fitted function (Yang et al. 2015). Considering that the five groups of datasets have similar variation tendencies, multi-scale polylines from the Huaibei plain are taken as an example (Figure 5).

![Figure 5. Fitting results of five functions and its R^2 of multi-scale polylines from the Huaibei plain.](attachment:image.png)

\( R^2 \), as a good indicator, often is used to compare the candidate functions. The larger value of \( R^2 \) always indicates the better fitting curve. As can be clearly seen from Figure 5, the \( R^2 \) of the power function is closest to 1\((R^2=0.8152)\), which achieves the best
fitting relationship of the curve among all candidates. Therefore, Equation (5) is chosen as a quantitative model for describing the relationships between spatial similarity degree and map scale change of individual polylines. This conclusion is consistent with Yan's conclusion (Yan 2014).

\[
\text{Sim}(L_i, L_j) = \begin{cases} 
1, & \text{if } L_1 \text{ and } L_2 \text{ are straight line, otherwise} \\
(aC_{L_i, L_j})^b
\end{cases}
\]

(5)

where, \( a > 0, \ -1 < b < 0; \ \text{Sim}(L_i, L_j) \in (0,1], \ C_{L_i, L_j} \in [1, +\infty) \).

To validate the reliability of the proposed model, this paper selected five groups of vector sampling multi-scale polylines from previous research of Yan (2015) and Chehreghan (2016). Figure 6 shows the fitting results of the proposed model, and Table 2 shows the accuracy comparison result with the existing evaluation model (Yan 2014).

Figure 6. Quantitative relationships fitting results of validation data between S and C.

Table 2. Accuracy comparison result with the existing evaluation model.

| Group | Previous evaluation | Proposed evaluation | Accuracy improvement |
|-------|---------------------|---------------------|---------------------|
| (A)   | \( S=1.0164C^{-0.3427} \), \( R^2=0.8770 \) | \( S=0.903C^{-0.085}, R^2=0.9135 \) | 4.16% |
| (C)   | \( R^2=0.8770 \) | \( S=0.6716C^{-0.213}, R^2=0.9470 \) | 7.98% |
| (E)   | \( R^2=0.8770 \) | \( S=0.7165C^{-0.28}, R^2=0.9781 \) | 11.52% |
Compared with the previous model, the fitting accuracy of the evaluation model proposed in this paper is improved by 4.16%~11.52%. However, compared with Chehreghan's conclusion, there is a nonlinear power function relationship between spatial similarity and map scale change (Figure 6 (B), (D)). Although the trend of three groups of the multi-scale polyline (Figure 6(A),(C),(E)) is the same, they cannot be described using the same power function whether they are the same ground objects of different types or the different ground objects of same type, which is some different with Yan's and Chehreghan's conclusion.

In order to further verify whether these quantitative relationships can be fitted using the same function with the same coefficients, this paper selects 38 groups of the dataset of different plains, and the curve fitting results are shown in Figure 7.
Figure 7. Fitting results between S and C of the multi-scale polylines.

Figure 7(A) shows the fitting results between S and C of multi-scale individual polylines from the same geographical feature area, and Figure 7(B) shows the fitting result of multi-scale individual polylines from the different geographical feature areas.
Through the comparative analysis of the above fitting results in Figure 7, it can be seen clearly that the relationship between spatial similarity degree and map scale change of different groups of the dataset from the same geographical feature plain can be described using the same power function curve with the same parameters, e.g., in Figure 7(A1), 12 groups of multi-scale polylines of the river are all from the Huaibei Plain, the quantitative relationship between S and C can be fitted using the same power function $S=0.9245C^{-0.224}$, and the fitting accuracy $R^2$ is up to 0.8152. Moreover, 19 groups of multi-scale polylines of river and road are all from the Lower Yangtze Plain in Figure 7(A4), although they are the different types of the polyline, and some groups of polylines of which are sampled from the Jianghuai Plain, while others are sampled from the Yangtze River Delta Plain, the fitting accuracy is also up to 0.7966. However, Figure 7(B) shows that no matter whether the polylines are the same or different type(s) of features, the coefficients of their power function are entirely different for polylines from the different geographical feature areas. For example, as shown in Figure 7(B1), 12 groups of multi-scale polylines are from the Huaibei Plain, and the other 12 groups of multi-scale polylines are from the Lower Yangtze Plain, both of them are the same type of ground objects, but the fitting accuracy of power function is only 0.5005. In Figure 7(B3), 12 groups of multi-scale polyline of the river are from the Huaibei Plain, and seven groups of multi-scale polyline of the road are from the Lower Yangtze Plain; both are the different types of ground objects, the fitting accuracy is also only 0.4155.
3.2. Determination of distance threshold

The essence of the DP algorithm is to compare the straight-line distance connecting the start point and endpoint with the optimal distance threshold. Therefore, this part aims to find the optimal distance threshold, which determines the degree of the polyline to be simplified, corresponding to map scale change to realize the completely automatic generalization of polylines.

It is qualitatively evident that the greater the map scale change, the greater the optimal distance threshold of DP. Hence, three potential candidate functions can be employed to fit the changing trend between λ and C: polynomial, linear equation, and logarithmic function. Since the other polynomials have n-1 inflection point(s), e.g., cubic polynomial, which indicates that the curve is not monotonic. Hence, only the second-order polynomial \( (x \in [0, -b/2a]) \) satisfies the variation tendency that the dependent variable increases with the independent variable. Ultimately, only quadratic polynomial, linear equation, and logarithmic function are considered (Equation (6)).

\[
\begin{align*}
  y &= ax^2 + bx + d \quad (x \in [1, -\frac{b}{2a}]) \\
  y &= a\ln(x) + b \quad (x \geq 1) \\
  y &= ax + b \quad (x \geq 1)
\end{align*}
\]

According to the principle described in section 2.1, the point pairs (C, T) of the optimal distance threshold and map scale change are continuously recorded with the gradual generalization of the polyline. The fitting results between them are shown in Figure 8.
As can be clearly seen from Figure 8, the $R^2$ of the unary quadratic function is closest to 1 ($R^2 = 0.9585$), which achieves the best fitting relationship of the curve among all the candidates. Therefore, the unary quadratic function (Equation (7)) is chosen as the quantitative model for describing the relationships between T and C.

$$T = \begin{cases} 0 & \text{if } C = 1 \\ aC^2 + bC + d & \text{if } a < 0, C \in (1, \max\{C_i\}] \end{cases}$$ (7)

In order to further verify whether the quantitative relationships between T and C can be fitted using the same function with the same coefficients, the experimental datasets of this part are divided into two parts: control and experimental. The former consists of multi-scale datasets from the different geographical feature areas, and the quadratic functions are shown in Figure 9(a). The latter consists of multi-scale datasets from the different geographical feature areas, and the quadratic functions are shown in Figure 9(b).
Figure 9. Quantitative relationship fitting results between T and C of multi-scale polylines.

It can be clearly seen from Figure 9(a) that the $R^2$ of each dataset in the control group is no less than 0.9585, and the maximum is up to 0.9917, indicating a very high fitting accuracy. However, Figure 9(a1) shows that no matter whether the polylines are the same or different type(s) of features (e.g., B and C both are roads; D and E are the different types of features of the real world), the coefficients of their unary quadratic function are entirely different for polylines from the different geographical feature area. Consequently, these results convincingly demonstrate that it is unreasonable to describe all five groups of datasets using the same single quadratic Equation with the same coefficients; besides, $R^2$ is only 0.4636. Therefore, it is impossible to simultaneously realize the completely automatic generalization of all polylines using the same optimal distance threshold.
Based on the results presented in Figure 9(b), the first group of experimental data which is consisted of four groups of multi-scale polyline from Shanxi Mountain (Figure 9(b1)), the other consists of 19 groups of multi-scale polylines from the Lower Yangtze River Plain (Figure 9(b2)). The $R^2$ values of the two groups of datasets are not less than 0.8521, with the best in the first groups ($R^2 = 0.9012$). Hence, the fitting accuracy is satisfactory. i.e., it is affirmative to realize the completely automatic generalization of all two groups of sampling datasets, respectively. Therefore, these results convincingly demonstrate that it is reasonable to realize the complete automation of the DP generalization algorithm for the polylines from the same geographical feature area. Therefore, polylines from the same geographical feature area are the objects of the following research.

3.3. Automated generalization of a polyline

Taking the multi-scale polylines from the Shanxi Mountain as an example, the theoretical optimal distance threshold is determined if the original and target map scales are given (Table 3). For example, suppose the original and target map scales are 1:50000, 1:250000, respectively, and then the map scale change $C$ is 5. Hence, the optimal distance threshold $T$ will be 0.0095 km. Afterward, six groups of experimental datasets are automatically generalized using DP based on the corresponding theoretical optimal distance threshold.

Table 3. Parameter $T$ optimally results of multi-scale polylines.
Figure 10 shows the results of the experimental datasets based on the proposed method.

![Comparison graphs of river and road automatic generalization results.](image)

Figure 10. Experiment results of the different datasets from the same geographical feature area.

### 3.4. Accuracy evaluation

The proposed model is validated for its acceptability and consistency. In this regard, length ratio (LR), Sinuosity ratio (SR), PE, and SMD are employed to evaluate the geometric and position accuracy of generalization results (Wu et al. 2008).

**Table 4. Geometric accuracy comparison between simplification results and available data.**

| Features | Target scale reference Data (L_{MC}) | Simplification results of proposed method (L) |
|----------|-------------------------------------|-----------------------------------------------|
|          | LR  | SR  | PE/km | SMD (%) | LR'  | SR'  | PE'/km | SMD'   |
| Road1    | 0.9901 | 0.9145 | 0.0211 | 0.4661 | 0.9915 | 0.9004 | 0.0578 | 0.4565 |
Table 4 also reveals that comparing with $L_{M2}$, $\Delta LR$ of samples is less than zero, and $\Delta SR$ is almost greater than zero, the error of the compression amount ($\Delta LR$), but the sinuosity degree change rate ($\Delta SR$) of the road and river are the smallest. It highlights that the compression amount (LR) and the sinuosity degree change rates (SR) of $L$ are consistent with $L_{M2}$, and the proposed approaches can better maintain the geometric features of road and river; the data compression effect is superior.

![Figure 11. Error rate comparison of geometric and position accuracy of simplification results](image)

The road and river simplification results are comparable to the standard target scale polyline in terms of maintaining the position accuracy (Figure 11). Compared with the standard target scale polylines, the position errors of road and river are significantly superior to the error-index values of LM2 except for two groups of samplings, but the maximum position errors are only 0.012. It highlights that the simplification results of the proposed method could better maintain the global position accuracy of polylines.
As shown in Table 4, the SMD is almost greater than zero, but the maximum of SMD of L only is up to 0.0408, i.e., the index SMD also indicates that the simplification results of the proposed method could better maintain the local position accuracy of polylines. In summary, the proposed method effectively maintains the local and global shape characteristics of roads and rivers.

4. Discussions

Compared with special subjective parameter determination algorithms (Douglas and Peucker 1973; Li and Openshaw 1992; Ai et al. 2016; Qian 2017), which require humans' constant interference, this paper constructed the quantitative relationships between the optimal distance threshold and map scale change using the spatial similarity relations of polylines based on two groups of control experiments, which consider polylines from the same or different geographical feature areas. Finally, based on the theoretical optimal distance, the polylines from the same geographical feature areas are completely automatically generalized using the Douglas-Peucker algorithm.

First, compared with the linear Equation and logarithmic function, the $R^2$ of the unary quadratic function is no less than 0.8521. Therefore, the unary quadratic functions are the best as the quantitative model for describing the relationships between the optimal distance threshold and map scale change.

Second, no matter whether the polylines are from the same or different geographical feature area(s), the unary quadratic function can be employed to describe the
quantitative relationship between the optimal distance threshold and map scale change. However, coefficients a, b, and c are the same for the polylines from the same geographical feature area, whereas the coefficients are different for the polylines from the different geographical feature areas. Similar results were also found between the similarity degree and map scale change. Thus, there are two main objective factors affecting map generalization: scale and regional geographic characteristics. This conclusion illustrates those regional geographic characteristics influence the determination of the quantitative relationship between the optimal distance threshold and map scale change. Therefore, it is reasonable to generalize the polylines from the same geographical area using the same single optimal distance threshold.

Third, according to the quantitative functional relation of the polylines from the same geographical feature area, if the map scale change is given, the theoretical optimal distance threshold can be calculated, and vice versa. Compared with semi-automatic or non-automatic polyline simplification algorithm (Gu 2016; Qian HZ 2017; Li CM 2017), this not only can facilitate the automation of map generalization software but avoid human intervention in the process of map generalization, which will promote the efficiency of the algorithm and the accuracy of map generalized results.

Last, this paper validated the acceptability and consistency of the geometric and position accuracy by comparing them with the standard target scale polyline. The experiment results indicate that the SMD of the road and the river simplified by the proposed method is better than the standard target scale polyline, i.e., geometric
features retention of the proposed method is more superior. This conclusion is consistent with the conclusions drawn by Wu (2008), who compared the Douglas-Peucker, Li-Openshaw, Circle, and asymptotic algorithms. Therefore, it can be concluded that the proposed method can better maintain the geometric features and shape characteristics of polylines on the whole in terms of geometric and position accuracy. In this paper, multi-scale individual polylines are taken as research objects. However, the generalization of polyline group objects is not only dependent on the features of ground features; it is closely related to the surrounding ground features. Therefore, due to the defects of the DP algorithm itself, the simplification results may intersect or intersect with other adjacent features that are close in distance, e.g., two adjacent contours with a smaller distance between them simplified by the DP algorithm may intersect. It is recommended that future work should conduct extensive experiments over polyline group objects, e.g., contour cluster, road network, reticulate drainage.

5. Conclusions

This paper proposed a model, which describes the quantitative relationships between the optimal distance threshold and map scale change, to minimize the uncertainties in the complete automation of the Douglas-Peucker algorithm. It can calculate the theoretical optimal distance threshold of the polylines from the same geographical feature area taking map scale change as the only independent variable and vice versa.
It is indicated that realizing the complete automation of the Douglas-Peucker algorithm is affirmative for the polylines from the same geographical feature area. The findings of the current study not merely provide an idea and method for realizing the complete automation of map generalization by looking for the quantitative relationship between the parameter of algorithm and map scale change but also facilitate the automation of map generalization algorithms and system of polyline and improve the accuracy of map generalized results. The proposed model optimizes algorithm threshold settings, eventually reducing the inevitable differences between automated-simplified and cartographers' ratified simplification results. Reduced uncertainties in realizing the Douglas-Peucker Algorithm automation would support improved spatial data matching and establish a vector map database.

References

Ai, T.H., Shu, K., Yang, M., et al., 2016. Envelope generation and simplification of polylines using Delaunay triangulation. International Journal of Geographical Information Science, 31(2), 297-319.

Chen, J. N., Qian, H. Z., Wang, X., et al. 2016. Geometric accuracy assessment of linear features' simplification algorithm. Geomatics and Information Science of Wuhan University, 45(4):486-493. [in Chinese]

Chehreghan, A., Abbaspour, R.A., 2018. A geometric-based approach for road matching on multi-scale datasets using a genetic algorithm. Cartography and Geographic Information Science, 45(3), 155-269.
Chehreghan, A., Abbaspour, R.A., 2016. An assessment of spatial similarity degree between polylines on multi-scale, multi-source maps. Geocarto International, 32(5), 471-487.

Danlel, P.A., Yang, P., 2014. Quantitative methods for comparing different polyline models. Journal of Geographic Information System, 6, 88-98.

Deng, M., Li, Z.L., Chen, X.Y., 2007. Extended Hausdorff distance for spatial objects in GIS. International Journal of Geographical Information Science, 21(4):459–475.

Douglas, D.H, Peucker, T. K., 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. The Canadian Cartographer, 2(10):112-122.

Efrat, A. et al. Geometry helps in bottleneck matching and related problems. Algorithmica, 2001, 31 (1):1–28.

Gu, T., Chen, X. Y., Liu, C.Q., 2016. A modified line simplification method combined Douglas-Peucker with Li-Openshaw. Journal of East China University of Technology, 39(4):396-400. [in Chinese]

Hershberger, J., Snoeyink, J., 1992. Speeding up the Douglas-Peucker line-simplification algorithm. In: Proceedings of the fifth symposium on data handling UBC Technical report TR-92-07, pp:134-143.

Hughes, G.F., 1968. On the mean accuracy of statistical pattern recognizers. IEEE Trans. Inf. Theory, 14, 55-63.

Kronenfeld, B. J., Stanislawsuki, L.V., Buttenfield, B.P., et al, .2020. Simplification of
polylines by segment collapse: minimizing areal displacement while preserving area. International Journal of Cartography, 6(1):22-46.

Li, Z. L., Openshaw, S., 1992. Algorithms for automated line generalization based on a natural principle of objective generalization. International Journal of Geographic Information Systems, 16(5):373-389.

Li, C. M., Guo, P.P., Yin, Y., et al., 2017. A line simplification algorithm considering spatial relations between two lines. Acta Geodaetica et Cartographica Sinica, 46(4):498-506. [in Chinese]

Li, Z. L., Sui, H.G., 2000. An integrated technique for automated generalization of contour maps. The Cartographic Journal. 37(1):29-37.

Li, ZX, Zhai, J.S., Wu, F., 2018. Shape similarity assessment method for coastline generalization. ISPRS International Journal of GeoInformation, 7(7):283.

Li, B., Fonseca, F.T., 2006. TDD: A comprehensive model for qualitative spatial similarity assessment. Spatial Cognition and Computation, 6(1):31-61.

Makris, A., Silva, C.L.d., Bogorny, V. et al. Evaluating the effect of compressing algorithms for trajectory similarity and classification problems. Geoinformatica 25, 679–711 (2021). https://doi.org/10.1007/s10707-021-00434-1.

Mcmaster, R. B., 1986. A statistical analysis of mathematical measures for linear simplification. The American Cartographer, 13(2):103-116.

Mcmaster, R. B., 1987. Automated line generalization. Cartographica, 24(2):74-111.

Nayyeri, A. et al., Computing the Fréchet Distance between polygons with holes. In M.
Olteanu-Raimond, A., Mustière, S., Ruas, A., 2015. Knowledge formalisation for vector data matching using belief theory. Journal of Spatial Information Science, 10:21–46.

Olteanu-Raimond, A., et al., 2008. Data matching - a matter of belief. In: Ruas A, Gold C, editors. Headway in spatial data handling. Lecture notes in geoinformation and cartography. Berlin: Springer, p. 501-519.

Peuquet, D.J., 1992. An algorithm for calculating minimum euclidean distance between two geographic features. Computers & Geosciences, 18 (8), 989–1001.

Qian, H. Z., He, H. W., Wang, X., et al. Line feature simplification method based on Bend group division[J]. Geomatics and Information Science of Wuhan University, 2017, 42(8):1096-1103. [in Chinese]

Ramer, U., 1972. An iterative procedure for the polygonal approximation of plane curves. Computer Graph Image Process, 1(3):244-256.

Saalfeld, A. (1999). Topologically consistent line simplification with the Douglas-Peucker algorithm. Cartography and Geographic Information Science, 26(1), 7-18.

Sama,l A., Seth, S., Cueto, K., 2004. A feature-based approach to conflation of geospatial sources. International Journal of Geographical Information Science,
Sandu Popa, I., Zeitouni, K., Oria, V. et al. Spatio-temporal compression of trajectories in road networks. Geoinformatica 19, 117–145 (2015). https://doi.org/10.1007/s10707-014-0208-4.

Tong, X.H., Liang, D., Jin, Y.M., 2014. A linear road object matching method for conflation based on optimization and logistic regression. International Journal of Geographical Information Science, 28(4):824-846.

Wu, F., Zhu, K. P.. 2008. Geometric accuracy assessment of linear features' simplification algorithm. Geomatics and Information Science of Wuhan University, 33(6):600-603. [in Chinese]

White, E. R., 1985. Assessment of line-generalization algorithms using characteristic points. The American Cartographer, 12(1):17-28.

Yan, H. W., 2010. Fundamental Theories of Spatial Similarity Relations in Multi-scale Map Spaces. Chinese Geographical Science, 20(1):18–22.

Yan, H. W., Li, J., 2014. Spatial similarity relations in multi-scale map spaces [M]. Springer International Publishing Switzerland.

Yan, H. W., Shen, Y. Z., Li, J., 2016. Approach to calculating spatial similarity degrees of the same river basin networks on multi-scale maps. Geocarto International, 31(7):765-782.

Yan, H. W., 2015. Quantitative relations between spatial similarity degree and map scale change of individual linear objects in multi-scale map spaces. Geocarto
Yang, W. F., Yan, H. W., Jonathan, L., 2015. Formula for calculating spatial similarity degrees between point clouds on multi-scale maps taking map scale change as the only independent variable. Geodesy and Geodynamics, 2(6):113-125.

Shen, Y. L., Ai, T. H., Wang, L., et al. 2018. A new approach to simplifying polygonal and linear features using superpixel segmentation. International Journal of Geographical Information Science, 32(10): 2023-2054.

Zhang, M., Yao, W., Meng, L., 2014. Enrichment of topographic road database for the purpose of routing and navigation. International Journal of Digital Earth, 7(5): 411–431.

Zhang, M., 2009. Methods and implementations of road-network matching. Munich: Technical University of Munich