Research Journal of Recent Sciences _____ ISSN 2277-2502
Vol. **3(11),** 39-45, November **(2014)** *Res.J.Recent Sci.*

An Information-Based Approach for Douglas-Peucker Thresholding in Polyline Generalization

Hassany Pazoky Sajjad, Saadatseresht Mohammad and Chehreghan Alireza*
[1]Department of Geomatics, Faculty of Engineering, University of Tehran, Tehran, IRAN

Available online at: www.isca.in, www.isca.me
Received 21st November 2013, revised 8th March 2014, accepted 9th May 2014

Abstract

Douglas-Peucker is the most widely used approach for line simplification. Stopping threshold in the standard method is defined in terms of a spatial distance specified by the user. However, the user, in many cases, is not able to set the threshold as s/he really wishes. Thus, the process is either performed again or is accepted while the ideal result is not provided yet. The aforementioned problems are probable, since the relationship between the spatial threshold, the number of saved vertices and the final formation of the line is unknown. To contribute in solving this problem, the terms "data" and "information" are defined in a line simplification context and line vertices are prioritized according to their relative influence on the whole line information. The result of data-information analysis is a diagram in which data and information are the two axes. The diagram reveals how the amount of information is changing as line vertices are decreasing gradually. It is shown that at a certain point in the diagram, the slope of information reduction increases tremendously like a waterfall. It reveals that at this point, for every omitted vertex, a large amount of information is lost and it's not worth discarding any other vertices. The user can either choose the point where s/he prefers not to lose any more information or the optimal point can be automatically recognized by the application. The results as shown in the paper diagrammatically prove that using such a method can heavily simplify map simplification process.

Keywords: Cartography, generalization, line simplification, information content, data-information diagram.

Introduction

There are lots of objects in the world that can be modeled into GIS applications. The modeled objects are called features. Features are made up of points, lines or polygon[1]. Point is the base element. Lines are made from joining successive points and polygons are made from joining lines[2]. Number of vertices used to present a line may vary based on line length, accuracy needed for special use, file size constraint, map scale, etc. On the other hand, some presentations of lines may be too bulky and there may be no need to store such data. Therefore, line simplification algorithms are needed to reduce the number of vertices as well as representing the line how it really is based on the user's preference[3].

Line simplification which is a subcategory of generalization is referred to a set of algorithms and approaches to simplify the lines to reduce the size of the linear data while representing the line as nearly as possible to the original shape. The goal of line simplification is to represent a line using minimum number of vertices and as accurate as possible. Omitting the superfluous vertices has following advantages[4]: i. Reduce storage space, which will result in faster data retrieval and management[5]. ii. Faster vector processing, like translation, rotation, re-scaling, cartographic analysis, etc. iii. Faster vector to raster conversion. iv. Faster plotting time. v. Faster data transfer over networks such as the Internet.

Material and Methods

Information Content of Lines: Data and information are defined many times in different fields, although they are being used interchangeably by many researchers[6]. Data is raw and has no meaning by itself; however, information is data that has been giving meaning by data processing[7-8].

In Spatial sciences, map features are means to deliver information from the real world to map users. The amount of information that a feature delivers is constrained to the following criteria[9]: i. Data gathering accuracy. ii. Scale of the map. iii. Usage of features, that means where a feature is being used. Based on the usage, the cartographer may reduce, smooth, enhance, exaggerate, simplify, eliminate or move a feature. In this case, the cartographer in devoting information to other criteria.

As nobody has yet investigated line simplification from an information aspect, data and information should be first defined. The information that a linear feature delivers has a relationship with the vertices forming it that includes the number of vertices and their locations. The more vertices retained, the more information is preserved. Therefore, this is an optimization problem.
The principle of simplifying linear features based on information content is that each vertex's contribution to the whole line information content is different. Some of the vertices have no contribution, some of them with low contribution and a

minority with a significant effect on information delivery. Here, the problem arises. How each vertex's contribution in the line information content can be calculated?

Assume a line that all the adjacent pixels in the original raster line are vectorized and all the vertices are stored as the line vertices. The line contains 100% of the information. On the other hand, if just the first and the last vertices are preserved, the resultant line would become a straight line joining the two end points. This line contains 0% of the information. The first and the last points cannot be discarded, because the minimum number of points to draw a line is two and if they are discarded, the line cannot be drawn. The position of these two points cannot be altered either, because the line would be another line and may be confused with other features. It should be restated that this is what the authors have defined as line information. Thus, the assumptions are as follows as shown in figure-1: i. All pixels forming the line are regarded as line vertices. ii. This line contains 100% of the information content of the original line. iii. If all the vertices except the first and the last points are discarded, the line contains 0% of the information content of the original line.

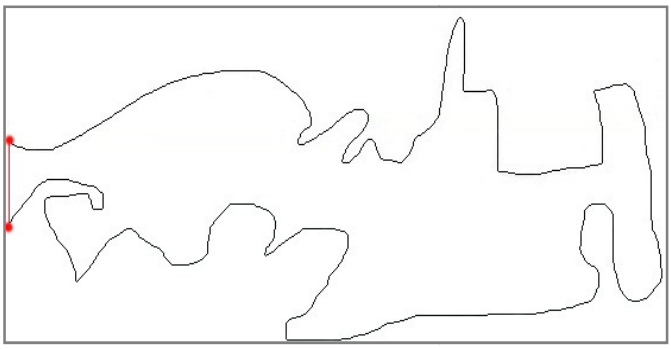


Figure-1
The black line is made up of successive pixels without any gap and is regarded as the original line. It contains 100% of the information. The red line contains just the first and the last points and has preserved 0% of the original line's information

Criteria affecting the contribution of each point in the whole line information content: To further illuminate the situation, 3 cases are shown in figure 2 (both a and b). The cartographer intends to omit the middle point. Case "a" would be the most preferable case in figure 2 (a), because the middle point is nearer to the line formed by the first and the last point and the angle is the most. In other words, in case "a", the middle point delivers the least amount of information compared to other cases. However in figure 2 (b), case "c" delivers the least amount of information; because the distance between the first and the last point is the most and therefore the angle is also the most.

The first term to discuss about is "h". This term lacks "d". Hence, as the effect of "h" and "d" are opposite, "h/d" is a far

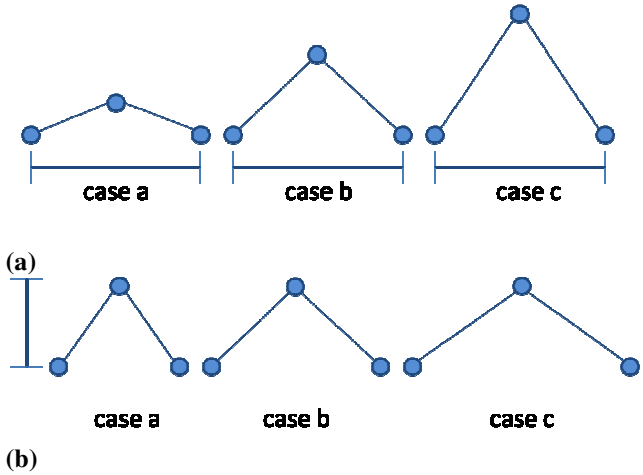


Figure-2
Different cases where the middle vertex should be discarded. (a) The distance between the first and last vertices is maintained while the height of the middle one increases. (b) The height of the middle vertex is maintained while the distance between the first and last ones increases

Let's theorize how much each vertex contributes to information delivery of the whole line. There are two criteria affecting information delivery of each vertex: Distance from the line formed by the first and the last vertices. The more distant, the more the information. Distance between the first and the last vertices. The less distant, the more the information.

Combining the two above statements, it would be: the greater distance between the middle vertex and the line formed by the first and the last ones and the less distance between the first and the last vertices, the more information would be delivered. To quantify the amount of information that each vertex adds to the line, a 3-point window is needed which is depicted in figure-3.

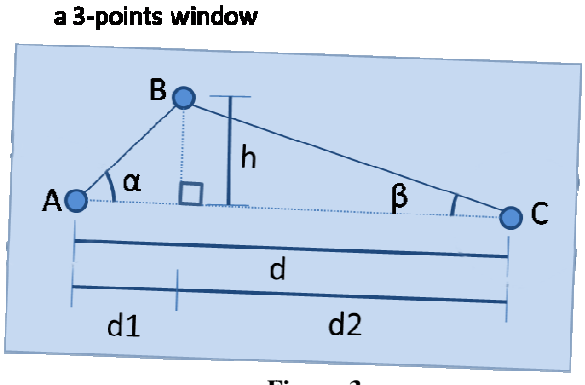


Figure-3
The variables needed to quantify each vertex's contribution to information delivery of a line

better term, although it has its own weakness that is shown in figure 4. As it is obvious, if one is to choose just one vertex in

addition to "A" and "K" in figure 4, "H" would be chosen. Therefore, the information content of "H" is the most, although "h2/d2" is far less than "h1/d1". In other words, this term is very sensitive to noise. To demonstrate, AHK and ADK are shown in figure-4(b).

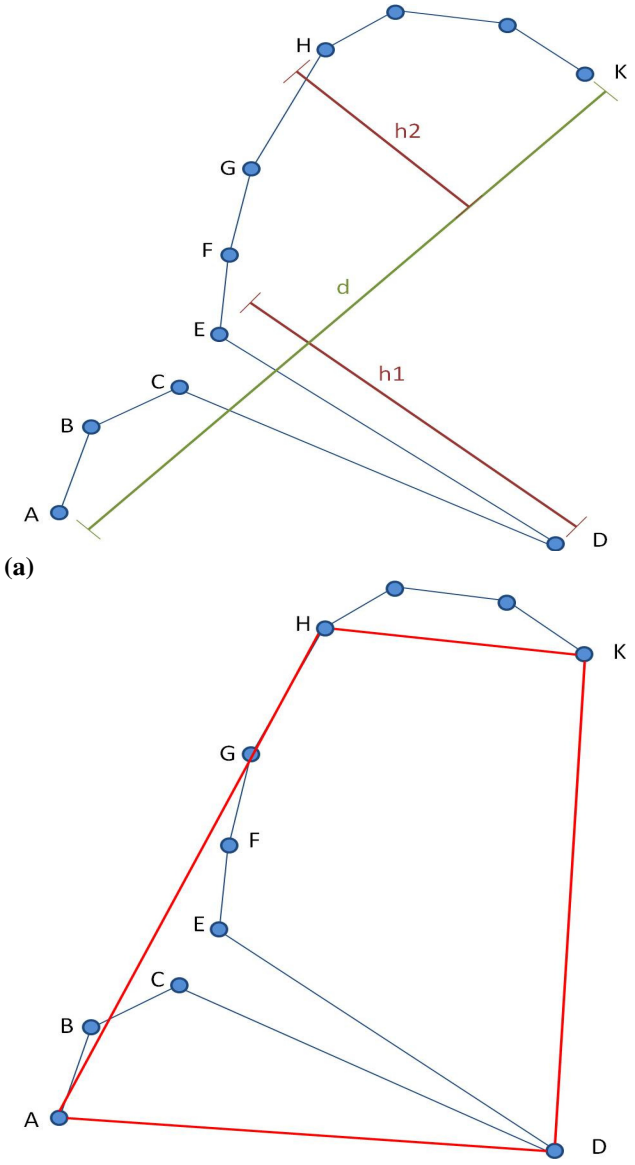


Figure-4
The weakness of "h/d" term. (a) "h1/d1" is more than "h2/d2", but information delivery of "H" is much more than that of "D". (b) AHK and ADK are depicted to help the reader understand why "H" delivers more information than "D" and "D" is a noise

The third term is "h2/d1.d2". It is a modification on h/d term. The concept behind this term is that the greater the sum of α and β (shown in figure 3), the more the information of the point. As everyone knows:

$$tan\,\alpha = \frac{h}{d1} \quad (1)$$

$$tan\,\beta = \frac{h}{d2} \quad (2)$$

Therefore,

$$\alpha = tan^{-1}\frac{h}{d1} \quad (3)$$

$$\beta = tan^{-1}\frac{h}{d2} \quad (4)$$

The Maclaurin series expansion of inverse tangent is:

$$tan^{-1}x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots \quad (5)$$

Using the first term of the above series, we have:

$$\alpha = tan^{-1}\frac{h}{d1} \approx \frac{h}{d1} \quad (6)$$

$$\beta = tan^{-1}\frac{h}{d2} \approx \frac{h}{d2} \quad (7)$$

Thus, as both α and β are effective parameters on information delivery of each point, we have:

$$\alpha.\beta \approx \frac{h}{d1}.\frac{h}{d2} \quad (8)$$

$$\alpha.\beta \approx \frac{h^2}{d1.d2} \quad (9)$$

As this term is similar to "h/d" term which both of them don't have dimension, the root of the term should be considered, since squaring the amount will further highlight the important vertices and diminishes the value of less important ones. Consequently, the term will be: $\sqrt{\frac{h^2}{d1.d2}}$.

Another important criteria affecting information delivery of each vertex is the surface of the triangle that each vertex makes with its immediate neighbor vertices. As everybody knows, the surface of triangles can be calculated with the following formula:

$$S = \frac{h.d}{2} \quad (10)$$

The term is not dimensionless. Therefore, to be comparable to "h" term, the root of the surface is taken into account, so that the dimension of both terms is equal to "distance". Furthermore, dividing by two has no effect on the final result, thus, it is omitted to decrease the computational cost. Therefore, this term is $\sqrt{h.d}$. The strange point about this term is that opposed to what is mentioned earlier about the opposite effect of "h" and "d", they are multiplied in this term. Investigation to candidate the most appropriate term is performed in subsequent sections.

Proposed Approach: To choose the best term, a window of 3points is considered and it moves along the line (as shown in Figure-3). For each step of the window, information term is calculated and stored. After calculating the term for all the vertices of the line, the ones with minimum information delivery are discarded. Then the procedure of calculating the term for remaining points is performed again. The reason of recalculating the term is that when some points are omitted, the

information content of other points may change– if there are omitted points in their vicinity. This loop will continue until just two end points remain. To have a better understanding of the information values, they are normalized and stated in terms of percentage. The values are then sorted ascending.

This process can be divided into two parts: lossless and lossy data reduction[10]. In lossless data reduction, the vertices that convey no information are detected and omitted. The variable "h" for these vertices is equal to zero. In this case, the 3 points in the window are collinear. When all the collinear vertices are omitted, all the remaining ones deliver some information and each vertex that is omitted, some information is lost.

Having the information delivery of each vertex in the line, a diagram can be drawn which the x-axis is the percent of number of vertices and y-axis shows the percent of maximum cumulative amount of information content of the line.

Results and Discussion

Information term Test: To investigate different information terms, three lines are put to test and the data-information diagrams are produced. The first line contains smooth curves, the second one is made up of high-frequency curves and the third one is a combination of low-frequency and high-frequency curves. These lines are shown in figure-5.

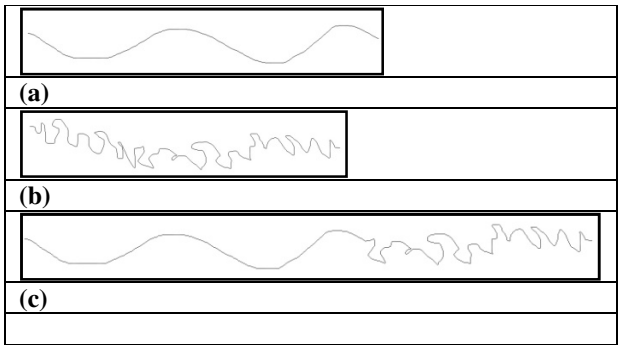
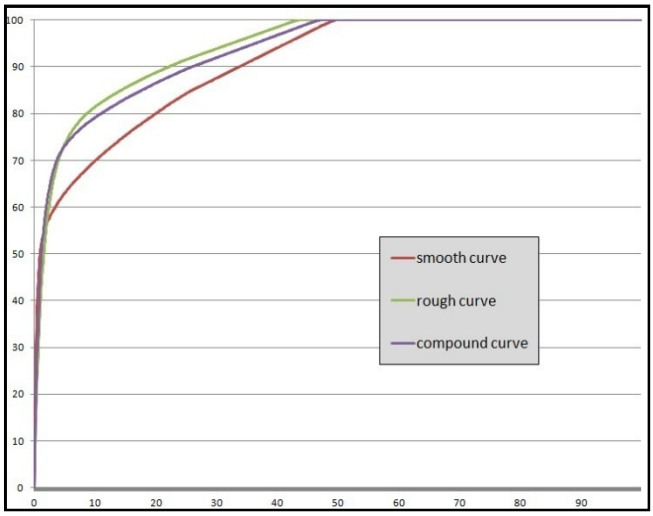


Figure-5
Three test lines (a) smoothly curved, (b) roughly curved and (c) a combination of the previous ones

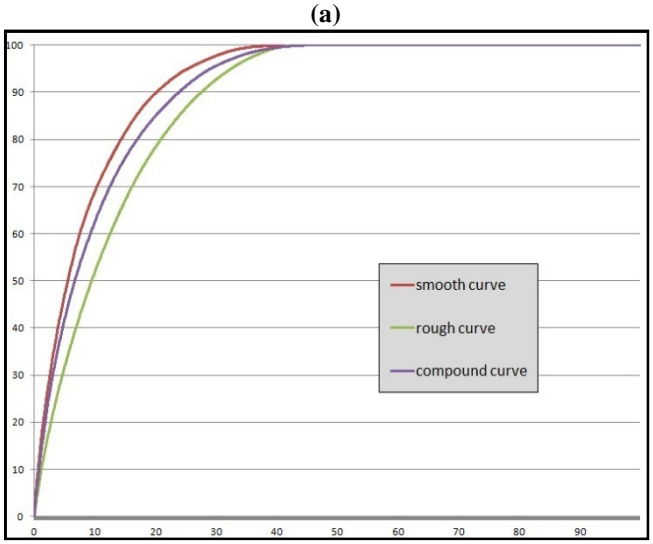
The data-information diagrams of the three lines with different information terms are shown in figure-6.
As one can see, the difference of the diagrams in figure-6(a) and (b) are more than that of figure (c) and (d). This fact shows that the term $\sqrt{\frac{h^2}{d1.d2}}$ and "$\frac{h}{d}$" are less sensitive to line type.

An interesting point is that the rough curved line in figure 6(b) rises and reaches the top earlier, but in figure 6(a) and (d) it is vice versa. It is obvious that, for a certain amount of information, more vertices are needed to reach the information level of a high-frequency line than that of the low-frequency one. Therefore, it is deduced that "$\frac{h}{d}$" information term is more realistic than information terms which have dimension ("h" and "$\sqrt{h.d}$"). Also in figure 6(c), the compound line is not moving between the two ends of line frequency spectrum, therefore, it is not acceptable either.

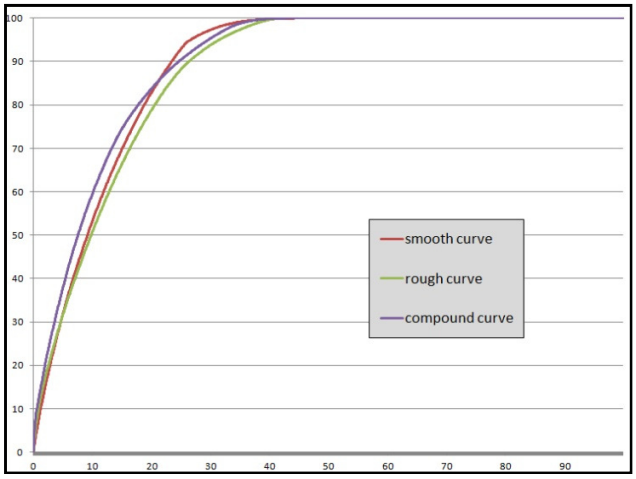
Analysis of the diagram shows that more than half of the vertices are omitted in lossless data reduction procedure. After that, some vertices have a very little effect on information delivery and as more vertices are omitted, the information values of the remaining ones rise tremendously. It means that there should be a point where vertices are worth being omitted, since they don't convey much information. The question is how to find the point.



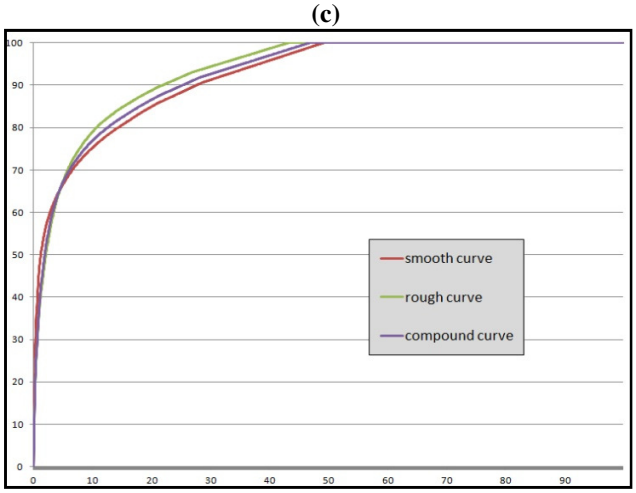
(a)



(b)



(c)



(d)

Figure-6

Data-information diagram drawn for the lines of figure 4

The term in (a) is "h", in (b) is "$\frac{h}{d}$", in (c) is "$\sqrt{\frac{h^2}{d1.d2}}$" and in (d) is "$\sqrt{h.d}$"

How To Use the Diagram: Data-information diagram can be used in three ways. The first possibility is to specify the amount of data user wishes to preserve. Using the diagram, the maximum amount of information that can be preserved by these vertices is specified. The other possibility is that user specifies the amount of information and the minimum number of vertices to deliver such amount of information is retrieved through the diagram. For example as it is depictedin figure-7, using 10% of the vertices, the generalized line preserves 65% of the information and if 90% of the information is needed, 25% of the vertices should be preserved. The other possibility is to automatically detect the optimum point which is further discussed in the next section. It should be mentioned that the line which the data-information diagrams of figure-7 are drawn with is the one shown in figure 1 and the information term is "$\frac{h}{d}$".

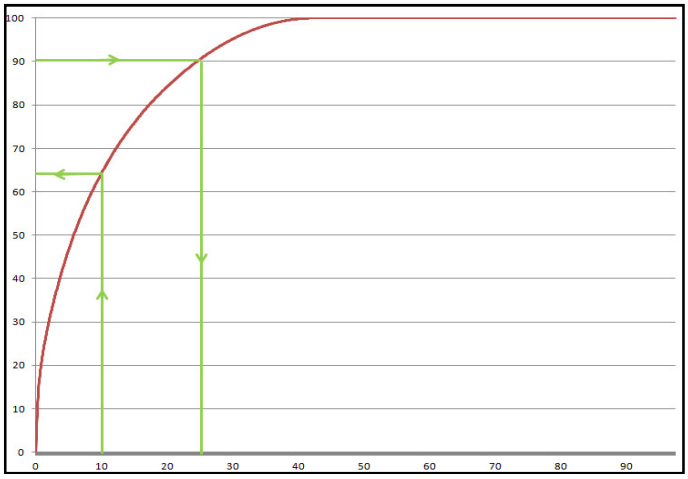


Figure-7

Using the diagram to simplify lines. If the data is specified by the user, maximum information is retrievable from the diagram and if the information is specified by the user, minimum amount of data is retrievable from the diagram

Detection of the Optimum Threshold: To find the optimum threshold, three parameters can be used: compression ratio, redundancy and graph slope. Compression ratio is defined as follows:

$$C = \frac{100}{D} \quad (11)$$

Where "C" is the compression ratio and D is the amount of data. This parameter shows the amount of compression instead of data itself. Redundancy is defined as follows:

$$R = 1 - \frac{1}{C} \quad (12)$$

Where "R" is redundancy and "C" represents compression ratio. This parameter shows how redundant data are.

Slope of data-information graph shows that by adding a very small amount of data, how much information is added to the generalized line. For example in figure-7, at zero, the slope is 67, at 10% is 2.8 and at 25% is 1.1. When the slope equals 1 (in this case at 26%), it shows that data and information increase equally. Similarly when the slope equals 2 (in this case at 14%), it means that for every percent of data, two percent of information is preserved. The slope can also be stated in terms of α which are the angle between the tangent to the graph and x-axis.

Therefore, data-information diagram can evolve to more efficient diagrams that areslope-information, compression ratio-information and redundancy-information diagrams. The aforementioned diagrams related to figure-7 are shown in figure 8. It should be noted that in figure-8, vertices with no interference in information content of the lines are discarded to make a better view.

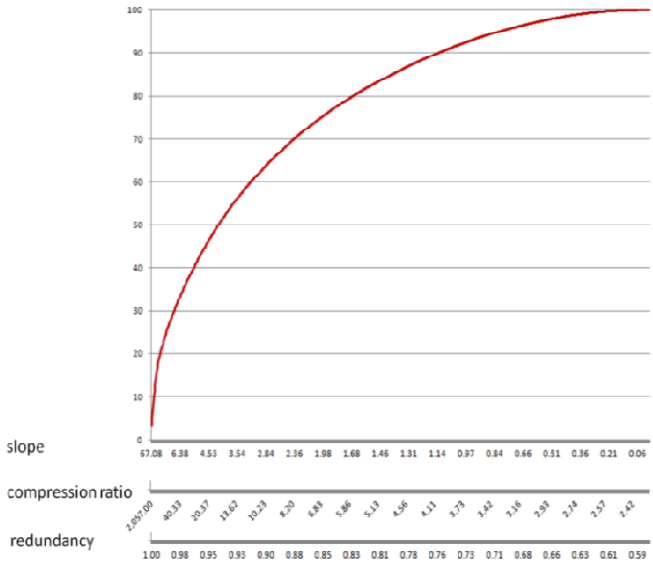
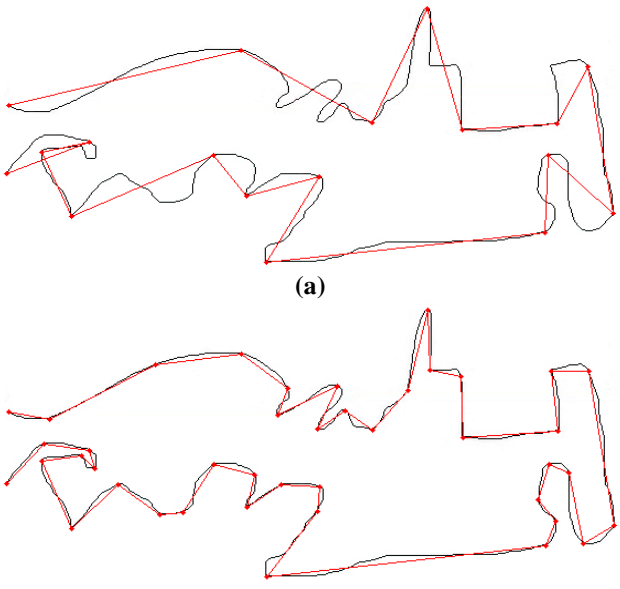


Figure-8
Slope-Information, compression ratio-information and redundancy-information diagrams related to figure-7

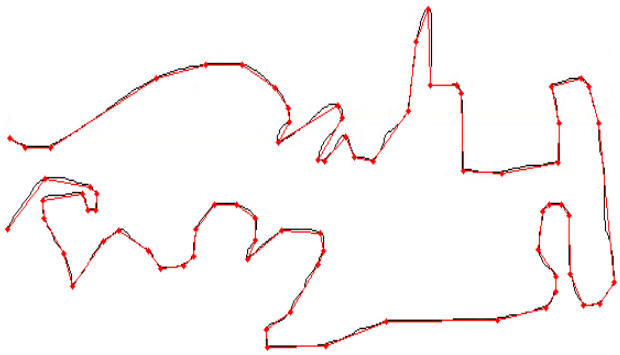
Using one of the diagrams mentioned above, user can choose the threshold on the basis of the significance of vertices on information. For example using slope-information diagram, choosing 2 as the threshold means that vertices are saved to where the rate of information increase is two times more than that of data. To make this process automatic, a default value for slope can be set.

To show the results, the simplified lines of different information and data are shown in figure-9. Different Parameters of each simplified line is mentioned is table-1.

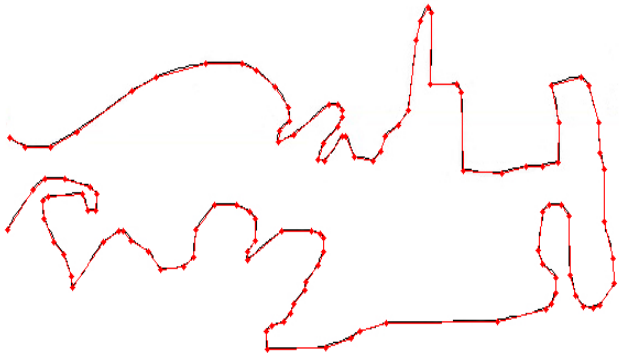


(a)

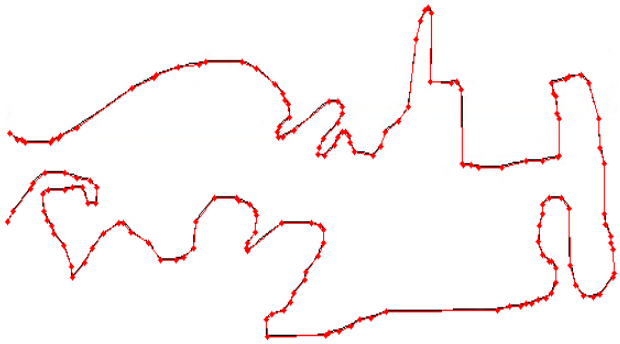
(b)



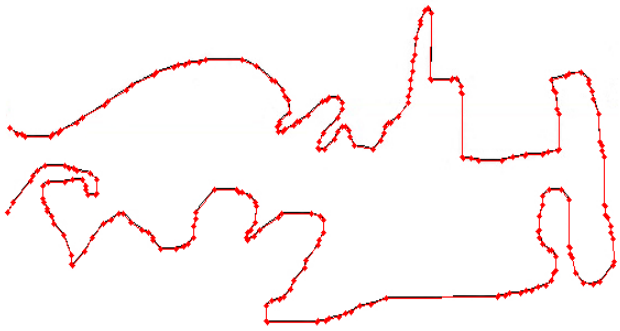
(c)



(d)



(e)



(f)

Figure-9
Results of the proposed line simplification algorithm with different data and information content shown in table 1

Table-1
Different Parameters of simplified lines depicted in figure 9

Simplified Line	Data Percentage	Information Percentage	slope	Compression Ratio	Redundancy
a	0.8	20	10.2	121	.99
b	2	30	6.9	50.2	.98
c	3.6	40	5.4	27.4	.96
d	5.7	50	4.2	17.4	.94
e	8.5	60	3.1	11.8	.91
f	12	70	2.4	8.2	.89

Conclusion

In this research, anovel approach based on information content for Douglas-Peuckerthresholdingis presented for line simplification. Most of current algorithms of line simplification are based on a spatial threshold. The problem is that user may not be able to set the threshold as s/he really requires. Therefore, a process of trial and error occurs that is time-consuming and not efficient.

In the approach discussed in this research, user can specify the amount of information needed for any specific purpose.

This approach is based on the assumption that different vertices have unequal parts in information delivery of lines. Hence, several terms for information delivery of each vertex in a line are proposed and put to test. Having the information of each point, a data-information diagram can be drawn. Using this diagram, the optimum threshold can be automatically specified as well as direct user preference on how much data or information s/he requires to preserve. Furthermore, data-information diagram is evolved to other diagrams to further facilitate line simplification process such as: slope-information**,** compression ratio-information and redundancy-information diagrams.

Although there is a long way in front of the authors to work on different aspects and weaknesses of this approach, the results are satisfying.

Specifying an acceptable information term of each vertex is the most significant requirement of developing this approach. The best information term may not be linear and this can make the processes more time-consuming. A data-information diagram based on human cognition should be produced and compared with the proposed information terms. This could be very helpful, since line simplification is based on human cognition and by no means can be written in programming languages. Therefore, by comparing algorithms by what human really thinks, the best algorithm can be recognized. Shifting the approach from bottom-to-top to top-to-bottom is another aspect of this problem that should be corrected.

References

1. Golledge R.G., Rice M.T. and Jacobson R., Multimodal Interfaces for Representing and Accessing Geospatial Information, in Frontiers of Geographic Information Technology, S. Rana and J. Sharma, Editors, Springer Berlin Heidelberg., 181-208 **(2006)**

2. Krygier J. and Wood D., Making Maps: A Visual Guide to Map Design For GIS, *New York: Guilford,* **2, (2011)**

3. Shi W. and Cheung. C., Performance Evaluation of Line Simplification Algorithms for Vector Generalization, *Cartographic Journal*., **43(1)**, 27-44 **(2006)**

4. Taylor G., Line Simplification Algorithms. Online available (January 5, 2012): www. comp. glam. ac.uk/ pages/staff/ getaylor/papers/lcwin.PDF., **(2005)**

5. Jenks G., Geographic Logic in Line Generalization, Cartographica., **26(1)**, 27-42 **(1989)**

6. Boisot M. and Canals A., Data, information, and Knowledge: Have We Got it Right ? *Journal of Evolutionary Economics.,* **14**, 1–25 **(2004)**

7. Bellinger G., Castro D. and Mills A., Data, Information, Knowledge, and Wisdom, Available from: Online available (January 5, 2012): http://www.systems-thinking.org/dikw/dikw.htm.**(2004)**

8. Davis G. and Olson M., Management Information Systems: Conceptual Foundations, Structure and Development, 2nd ed, *New York: McGraw Hill* **(1985)**

9. Robinson A.H. et al., Elements of Cartography, 6th ed, *New York : John Wiley and Sons,* **(1995)**

10. Li Z., Zhu C. and Gold C., Digital Terrain Modeling: Principles and Methodology, *Taylor and Francis* **(2010)**