Minimising Positional Errors in Line Simplification Using Adaptive Tolerance Values

Nadia Shahriari[1] and Vincent Tao[2]

The University of Calgary, Department of Geomatics Engineering, 2500 University Dr. NW, Calgary, AB, Canada T2N 1N4, nshahria@ucalgary.ca[1]
York University, Department of Earth and Space Science & Engineering, 4700 Keele Street, Toronto, ON, Canada M3J 1P3, Tel: (416) 736-5221, Fax: (416) 736-5817, tao@yorku.ca[2]

Abstract

Line simplification is an important function in GIS and cartography and is widely used in commercial GIS software packages. Most line simplification algorithms require the user to supply a tolerance value, which is used to determine the extent to which simplification is to be applied.. All simplification algorithms induce positional errors in the data set, because they produce a discrepancy between the original line and its simplified version. The amount of this error depends on both the tolerance value and the shape of the line. This is the reason that many researchers have focused on measuring the geometric characteristics (or complexity) of lines. Using one tolerance value for all lines in the data set results in different positional errors for different lines. What is usually important for the user, is to maintain a specific level of quality, and not the tolerance value itself. The question is, 'how does one specify a tolerance value for each line based on the user specified level of accuracy'? This paper presents a solution to solve this problem. In this approach, the user supplies the target level for desired accuracy and the simplification tolerance value is calculated accordingly.
Keywords: line simplification, complexity of line, adaptive tolerance value, positional error

1 Introduction

Cartographic generalisation involves selecting the features to be maintained at the targeted scale, simplifying non-relevant characteristics, enhancing significant shapes, displacing without defacing global and local shapes, and finally harmonising the final presentation (Plazanet, 1997). Until about a decade ago,

trained cartographers usually practised map generalisation, by using rules, examples and intuition. Traditionally, map-making agencies applied generalisation using manual techniques. Many mapping organisations, especially the larger national, state or provincial establishments had their own guidelines, standards and procedures for manual map generalisation. Only recently has this situation begun to change, but it has done so dramatically and definitively, as maps – along with every other form of human communication – have become digital (Dutton, 1999).

Recent reviews of digital generalisation procedures have identified a number of distinct processes. Most of the techniques focus on either the manipulation of vector- or raster-mode data. Much of the work on raster-mode generalisation of images has come from the field of remote sensing and includes such techniques as high- and low-frequency filtering (McMaster, 1989). High-frequency filters emphasise fine details and edges by passing the higher frequencies while low-frequency filters emphasise the more generalised trends. Fourier analysis, two-dimensional convolution, linear edge detection, non-linear edge enhancement are some of the common generalisation techniques for raster imagery. Since the mid-1960s, however, most work in this area has addressed generalisation in vector-mode. Specifically, researchers have thoroughly addressed the generalisation of digital lines, while leaving other aspects of generalisation as yet unresolved. For example, many algorithms have been developed for the simplification, smoothing, enhancement, displacement and merging of linear features (McMaster, 1989).

Line simplification is an important function in cartography and GIS. In line simplification, source data are transformed to reduce data volume, to merge databases with different scales, or to maintain cartographic quality when scale is reduced. Line simplification involves the selective elimination of vertices along a cartographic line to remove unwanted information. Line simplification is only one component of cartographic generalisation. It is, however, the most commonly applied generalisation operation and is widely used in commercial GIS software packages (Veregin *et al*, 1999).

Several approaches for the simplification of digital lines have been presented in the cartographic, computer science, remote sensing and mathematics literatures. While some of these algorithms are extremely simplistic in nature, the more sophisticated ones consider geometric properties of the line. McMaster's (1987) classification of simplification algorithms includes: independent point routines, localised processing routines, constrained extended local processing routines, unconstrained local processing routines and global routines. Global routines focus on the careful selection of the critical points, or the salient geometric characteristics of the line (McMaster, 1989). Simplification is considered successful if the simplified caricature of a line closely resembles the original version. Most simplification algorithms need a tolerance value. All simplification algorithms introduce positional error. The amount of this error depends on both the tolerance value used for simplification and the shape of the line. The problem arises when the tolerance value stays the same for all lines in the data set, regardless of the variations in their shape.. The result is that simplification introduces different amounts of positional errors on different lines and the user

usually does not have a direct understanding of these effects on data quality. In other words, the user specifies a tolerance value for simplification, without knowing how much the positional accuracy of each line is affected by this tolerance value. Therefore, adaptive tolerance values should be used for line simplification of a data set.

This paper presents a solution to this problem. First some measures of line complexity will be discussed and applied to a data set to illustrate the differences in geometry of different linear features. Then, the paper describes a new approach by which the user specifies the maximum positional error instead of a tolerance value. The tolerance value will be specified for each individual line based on this maximum valid positional error. In fact, the tolerance value will be adapted to each line. Although applying one tolerance value for each line is not logically correct, if the line is not geometrically homogenous, however, the assumption here is that the lines are homogenous. This new approach may be extended to line sections if the lines are not homogenous and the result will be different tolerance values for different parts of a line.

2 Complexity Measures of Lines

Researchers have introduced a variety of mathematical measures to evaluate the complexity of a line. Some of these measures may be applied to a single line while the others are applied to compare the geometry of a line before and after simplification. McMaster (1986) called these categories of measures as "single attribute measurements" and "displacement or comparative measurements" respectively. Most of the research in the field of complexity measures focuses on four categories of measures: *length*, *density*, *angularity* and *curvilinearity*. These categories are discussed in the following subsections.

2.1 Length Measures

The length of a line is probably the simplest geometric attribute of it. The length, however, cannot singularly represent the complexity of the line. Nevertheless, it can be used to represent the effects of the simplification.. Because much of the sinuosity is eliminated, a line becomes shorter, as it undergoes the simplification process. McMaster (1986) introduced "the percentage change in line length" as a length measure. This measure has been expressed as the total length of the simplified line divided by the length of the original line and presented as a percentage. This ratio is expected to decrease as the line is more simplified.

Jasinski (1990) introduced another measure, "error variance", which is an average perpendicular displacement of every point in a line to the anchor line. Error variance measures the deviation of the line from its straight-line approximation, its anchor line. Buttenfield (1984) defined the anchor line as a straight segment linking the first and the last point of the coordinate string. Error

variance is expected to increase as the line is more simplified. As the mean value of a distribution does not say anything about dispersion of values in the data set, Jasinski (1990) introduced another measure called "coefficient of variation of error variance" which gives one the relative variability of a distribution as a ratio of the standard deviation to the mean.

2.2 Density Measures

Density of a line represents the frequency of detail that exists in it. Number of points of a line, as a measure of density, solely cannot represent the complexity of the line. It can be used, however, to represent the effect of simplification on a line by comparing the density of the line before and after simplification. All simplification algorithms, principally, reduce the number of points along a line. Therefore, density of the line is expected to decrease as the line is more simplified.

McMaster (1986) introduced three coordinate measures to compare the density of the simplified line with the density of the original line. These measures are "ratio of change in the number of coordinates", "difference in average number of coordinates per inch" and "ratio of change in the standard deviation of the number of coordinates per inch".

Jasinski (1990) introduced another density measure called "average segment length" which is the average length of all segments between the points of the line. The value of this measure is expected to increase as the line is more simplified because fewer points are retained and segments automatically get longer. Jasinski (1990) also introduced another density measure called "coefficient of variation of average segment length" because the average segment length does not give much information whether most of the segments are similar or whether the line consists of some very short and some very long ones.

2.3 Angularity Measures

Angularity of a line is one of its primary geometric characteristics. Angularity measures evaluate specifically the individual angular changes along a line. McMaster (1986) defined nine measures for angularity of a line to compare its angularity before and after simplification. The "percentage change in angularity" is one of these measures, which was expressed as the sum of the angles between consecutive vectors on the simplified line divided by this sum on the original line. He defined the "absolute angle of change between each pair of consecutive vectors" as a basic angularity measure. Jasinski (1990) defined another formula to calculate the "average angularity", which ranges from 0 (straight line) to 1 (the line backs on itself) and another measure, which is called "Coefficient of variation of average angularity". The latter measure is expected to decrease as the line is more simplified.

2.4 Curvilinearity Measures

Curvilinearity of a line is defined by the number of inflection points in it. Inflection points divide the line into curvilinear segments, which are the portions of a line in which all angles are in the same direction, either positive or negative. Curvilinearity is a measure of direction of angular changes while angularity is the measure of magnitude of angular changes. McMaster (1986) defined four curvilinearity measures. These measures compare the curvilinearity of a line before and after simplification. McMaster (1986) defined the "ratio of the change in the number of curvilinear segments" as the basic curvilinearity measure, which is the number of curvilinear segments of the simplified line divided by the number of curvilinear segments of the original line. Jasinski (1990) defined the ratio of total curvilinearity to the number of all turns as the "curvilinearity ratio".

3 Discussion of Data and Measurements

A set of four lines was selected for the project. Each of these lines is a section of a river in the USA and provided by ESRI data which has come with ArcView software. For simplicity, the rivers have been named river1, river2, river3 and river4 which are parts of the rivers "Platte", "Arkansas", "Brazos" and "Red River of the North" respectively. In order for the set of lines to be representative of a wide range of shapes, it was decided to select the rivers with different irregularities. River1, which is less irregular, has 459.46 km length and includes 19 points. The length of river2 is 714.99 km and includes 40 points. River3 is 962.17 km and has 87 points. River4 which is 443.72 km, has almost the same length as river1, but has more irregularities and includes 137 points. Fig. 1 illustrates the selected rivers.

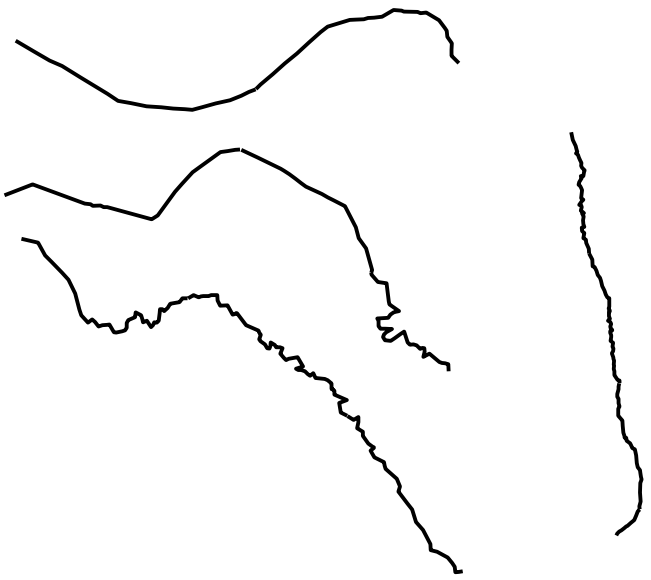

Fig. 1. Data set used for complexity measures

The Douglas-Peucker's (1973) line simplification algorithm has been selected in this project. There are several reasons for this selection. This algorithm is the most accurate at selecting critical points (White, 1983). Additionally, the algorithm's selection of these critical points generates simplifications that mimic those generated by manual generalisation, and retains details critical for map-reader recognition (Buttenfield, 1991).

Visually, it is obvious that these four lines have different shapes. The difference can be shown mathematically using the complexity measures. For each aspect of line complexity, namely, *length*, *density*, *angularity* and *curvilinearity*, different measures have been applied on the test data set, however, only some of them will be discussed in this paper. The results have been plotted as graphs, with the measurement value against the tolerance value set for simplification.

"Error variance" (Jasinski, 1990) has been applied on the data set as a length measure. Fig. 2 illustrates the results. As could be expected, the error variance increases as the lines are more simplified. This is logical, since the progressive simplification routine by Douglas-Peucker retains characteristic points which tend to have a high error variance (Jasinski, 1990).

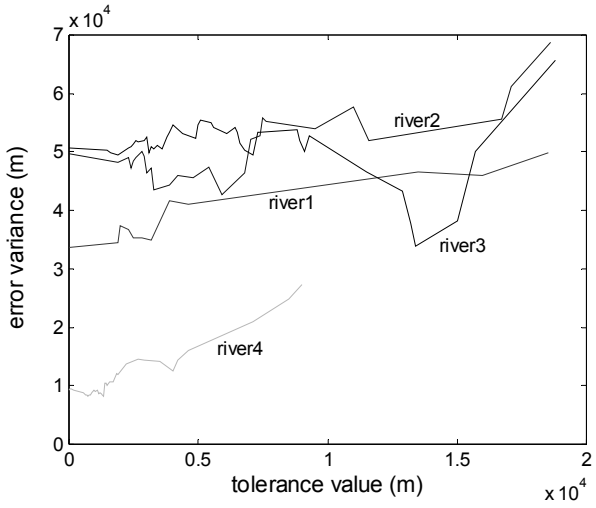

Fig. 2. Error variance

"Percentage change in number of coordinates" (McMaster, 1986) has been used to illustrate the density differences of four lines. Fig. 3 illustrates the result of this measure. As could be expected, the percentage change in number of points is decreasing as the lines are further simplified. Despite it's greater irregularity, the decrease is faster in the case of river4, because its points are close to its anchor line. Therefore even with a small tolerance value, many points are removed by the simplification.

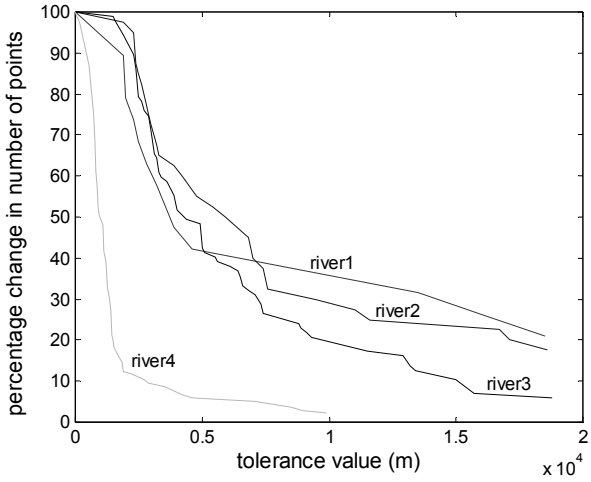

Fig. 3. Percentage change in number of points

The third measure applied to the data set was the "percentage change in angularity" (McMaster, 1986). Fig. 4 illustrates the results of this measure. As expected, the percentage change in angularity is decreasing as the lines are further simplified. Again, this decreasing is faster in the case of river4, because the line is mostly simplified with low tolerance values. Therefore, it loses its angularity much faster than the other three lines.

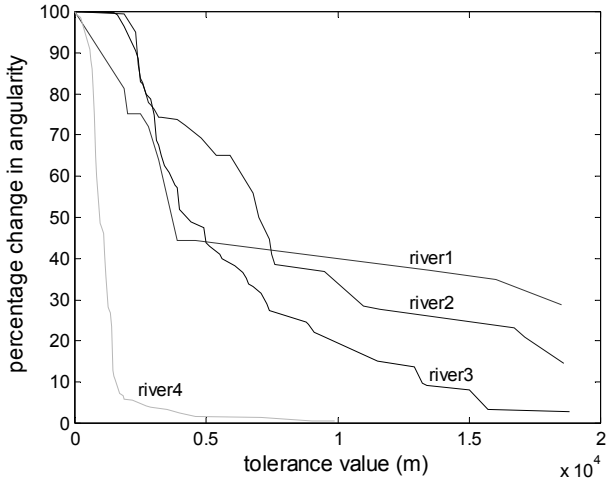

Fig. 4. Percentage change in angularity

Fig. 5 illustrates the results of the last measure, "ratio of the change in the number of curvilinear segments" (McMaster, 1986).

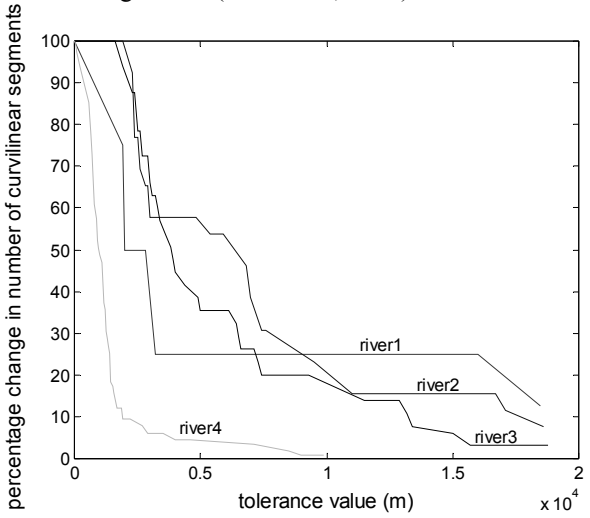

Fig. 5. Percentage change in curvilinearity

The "ratio of the change in the number of curvilinear segments" is a curvilinearity measure. As expected, the percentage change in curvilinearity is decreasing with more simplification. Fig. 5 shows how the simplification affects the curvilinearity of lines with different shapes. The straight lines in Fig. 5 means that the associated line is more simplified, however, the number of curvilinear segments does not change.

4 Data Quality

Line simplification algorithms simplify the lines by selective elimination of vertices (shape points) along the line. This process results in positional error because it produces a discrepancy between the location of the original line and that of the simplified line (Veregin *et al*, 1999). McMaster (1986) defined three groups of displacement or comparative measurements, which are used to evaluate differences between the base line and the simplified line. These measurements are "vector displacement", "polygonal displacement" and "perimeter displacement" measures. The vector displacement measure has been used in this project to investigate the effect of simplification in inducing positional errors in the data set. As a result of coordinate elimination, a vector difference is produced each time part of the simplified line is displaced. This distance may be measured as the perpendicular distance from the eliminated coordinates on the base line to the new vector on the simplified line. Sum of the length of all vectors between the two lines divided by the length of the original line is a measure of vector displacement (McMaster, 1986). This measure is expected to increase when the line is more simplified. This is logical, because the discrepancy between the location of the original line and that of the simplified line increases as the line is more simplified. Fig. 6 illustrates the results of applying vector displacement measure on the four lines of the data set.

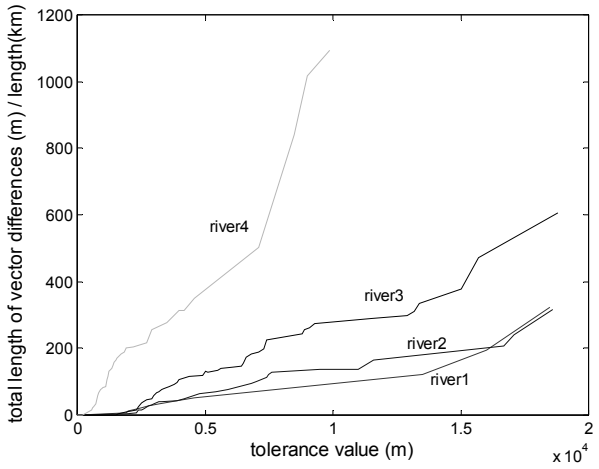

Fig. 6. Vector displacement

Figure 6 shows that the amount of vector displacement, which represents the induced positional error, is different for each line when one tolerance value is used for all four lines. In most parts of the graph for each specific tolerance value, there are four different values for vector displacement. For example, if the user specifies 7000m as a tolerance value, the vector displacements will be 53.05, 101.25, 181.25 and 350.59 m per km of line length for river1, river2, river3 and river4 respectively. This results due to the difference in the complexity of the lines, as discussed in the previous section. When using one tolerance value for all lines in the data set, simplification induces more positional errors in river4 then the other rivers in the data set. This occurs because river4 has more density, angularity and curvilinearity than the others.

5 Adaptive Tolerance Value

Line simplification, as discussed above, can have significant effects on the data quality; however, the user does not have any control on these effects. When using a simplification algorithm, the user generally specifies a tolerance value to control the degree of simplification but is unable to specify a target level of quality that must be attained. What is missing is a way for users to select a weed tolerance value that ensures a certain level of positional accuracy is maintained (Veregin *et al*, 1999). Using one tolerance value for all lines in the data set induces different degrees of positional errors for different lines. Therefore, in order to keep a certain level of positional accuracy for all features in the data set, the different tolerance values should be used for different features. The question to explore is

how should one specify a tolerance value for each feature based on the user-specified level of accuracy.

This paper presents a new way to solve this problem. In the approach presented, the user supplies a certain level of positional accuracy (either vector displacement or polygonal displacement or perimeter displacement or any other measure of positional accuracy depending on the user's application) instead of supplying a tolerance value. As a result, the tolerance value will be specified automatically for each line in the data set. In order to obtain the tolerance value for each line, the data set has to be pre-analysed. In this pre-analysis, vector displacement (or any other positional displacement) measure is applied on all lines in the data set using a set of tolerance values for each line, which can start from 0 to maximum tolerance value for the line. Then the proper tolerance value for each line is specified using the set of tolerance values and associated vector displacement from the pre-analysis phase. The proper tolerance value for each line is a value by which maximum simplification is achieved while maintaining the user specified positional accuracy. An example of this process is shown in Fig. 7. In this example the user specifies 150 m per km as the maximum vector displacement and the values of 14438.86 m, 11271.90 m, 6436.25 m and 1398.41 m is calculated as tolerance values for river1, river2, river3 and river4 respectively. Therefore, four rivers will be simplified using four different tolerance values and maximum vector displacement will remain under 150m per km for all four features.

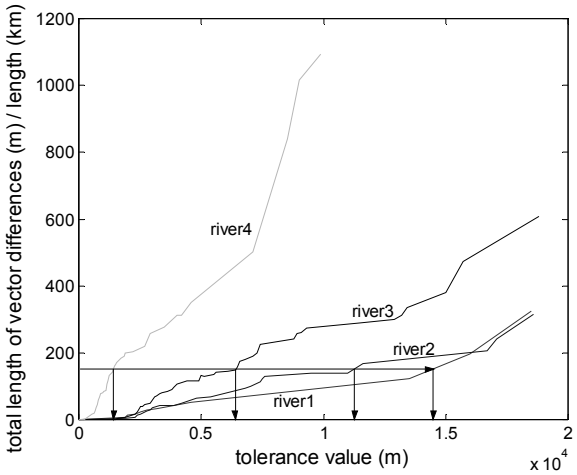

Fig. 7. Adaptive tolerance values.

This algorithm has also been tested on a larger data set. The test data set is an Arcview shape file including Canadian rivers, located in UTM zone12 (NAD83) at the scale of 1:7.5 million (National Atlas of Canada, 1997). The size of the shape file is 467KB and includes 1353 lines. The first step is the pre-analysis in which the vector displacements are calculated for each line using different

tolerance values. In order to specify a set of tolerance values for each line, first maximum possible tolerance value (T_{max}) for that line (by which the line will be simplified to a straight line connecting the start point and end point of the line) was calculated. Next a set of tolerance values was selected between zero and T_{max}. After assigning tolerance values for each line, vector displacement of the line is calculated for each tolerance value in the set of assigned tolerance values for that line. All these tolerance values and corresponding vector displacements were automatically recorded in a text file, along with feature identifiers indicating corresponding lines. The next step was simplification using adaptive tolerance values. For this step, the value of 100m/km was specified as maximum valid vector displacement for the lines of the map (as user specified level of positional accuracy). Then the proper tolerance value for each line was automatically selected using the set of tolerance values and vector displacement values stored in the text file in the previous step. Next, each line of the map was simplified using the selected tolerance value. Fig. 8 shows a section of the data set (top), together with the results of using constant tolerance value (1500 m) and the result of using adaptive tolerance value (100m/km). As was expected, the amount of vector displacement induced by line simplification is completely different for different lines when using a constant tolerance value (Fig. 8, left). The vector displacement, however, has been kept under 100m/km for all lines in the simplification test using adaptive tolerance values (Fig. 8, right). This means the user has control on the positional errors induced by line simplification. The processing time for the case of a constant tolerance value was 0.601 s while for the adaptive case was 0.911 s plus 13.309 s for pre-analysis phase. The reason that more processing time is needed for the adaptive approach is partly because of the need to search the text file to find the proper tolerance value for a line. As the order of the lines in the text file is the same as their order in the map file, the program does not have to search the whole text file to find the proper tolerance value for one line but only the section of the file that starts with the identifier of that line. Although the time difference is not that much, however, some methods can be investigated to improve the efficiency such as a binary search tree to find the appropriate tolerance value.

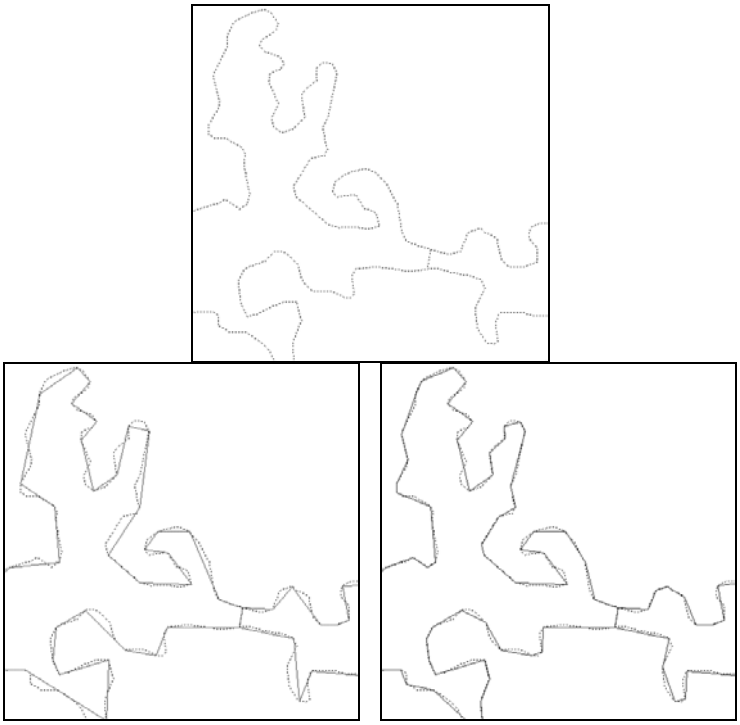

Fig. 8. A section of the test data set (top), the results of using constant tolerance value (left) and the result of using adaptive tolerance value (right)

6 Conclusion

The focus of this research has been the development of a method to calculate the adaptive tolerance values. In the proposed approach, the user supplies the target level of positional accuracy instead of a simplification tolerance value. The tolerance value is then calculated based on the given level of accuracy. The result is that each line in the data set is simplified using the calculated tolerance value specific for that line. In this paper, vector displacement has been used to define a target level of positional accuracy. However, based on a user's application, areal displacement, perimeter displacement or any other measure of discrepancy between the original line and the simplified line can be used to define the target level of accuracy. It is assumed, in this paper, that the lines are homogenous. However, the reality is that there are many linear features that are composed of non-homogenous sections. Further work may focus on extending and evaluating the implementation of the proposed approach to segments of lines in the data set.

References

Buttenfield B (1991) A rule for describing line feature geometry. In: Buttenfield B, McMaster RB (eds) Map generalization. Longman Scientific & Technical, London, pp 150-171

Buttenfield B (1984) Line structure in graphic and geographic space. Unpublished Ph.D. dissertation, University of Washington

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

Dutton GH (1999) Introduction and problem statement. In: A hierarchical coordinate system for geoprocessing and cartography. Lecture notes in Earth Science 79. Springer-Verlag, Berlin

Jasinski MJ (1990) The Comparison of Complexity Measures for Cartographic Lines. NCGIA Technical report 90-1. National Center for Geographic Information & Analysis, Department of Geography, State University of New York at Buffalo, Buffalo, New York 14260

McMaster RB (1989) The integration of simplification and smoothing algorithms in line generalization. Cartographica 26:101-121

McMaster RB (1987) Automatic Line generalization. Cartographica 24(2):74-111

McMaster RB (1986) A statistical analysis of mathematical measures of linear simplification. The American Cartographer 13(2):103-116

National Atlas of Canada (1997) Canadian Rivers [online]. GeoGratis website. Available from: geogratis.cgdi.gc.ca

Plazanet C (1997) Modeling Geometry for Linear Feature Generalization. In: ESF/NSF 1995 Summer Institute in Geographic Information Research: Bridging the Atlantic, in Portland Maine (USA) Part 3. Taylor & Francis, pp 264-279

Veregin H, Dai X (1999) Minimizing positional error induced by line simplification. Proceedings of the International symposium on spatial data quality 1999. The HongKong polytechnic University, HongKong

White ER (1983) Perceptual Evaluation of line generalization algorithm. Unpublished Master's thesis, University of Oklahoma.