Automated Metric Assessment of Line Simplification in Humid Landscapes

Lawrence V. Stanislawski[1], Paulo Raposo[2], Michael Howard[1], and Barbara P. Buttenfield[3]

[1] Center for Excellence in Geospatial Information Science (CEGIS), United States Geological Survey (USGS), Rolla, Missouri USA
Email: lstan@usgs.gov

[2] Department of Geography, Pennsylvania State University, Pennsylvania USA
Email: paulo.raposo@psu.edu

[3] Department of Geography, University of Colorado, Boulder Colorado USA
Email: babs@colorado.edu

ABSTRACT: Automated cartographic generalization involves several operations, which includes simplification of line features. Line simplification affects the geometric characteristics of represented features and can affect associated analyses. As the use of digital map information progresses, appropriate representation of features becomes increasingly important. Terrain conditions affect geometric characteristics of linear features for several map feature types, notably road and water networks. Proper line simplification should account for effects of local landscape characteristics. Focusing on the bend simplify algorithm, this paper describes the use of several metrics that quantify geometric characteristics of linear features and help assess effects of line simplification within humid climate regions. Linear road and hydrographic features from *The National Map* databases of the U. S. Geological Survey (USGS) are simplified using the bend simplify algorithm and evaluated. Subsets of features from flat, hilly, and mountainous topographies are evaluated to assess geometric feature characteristics across humid landscapes of the coterminous United States. Geometric feature characteristics before and after five levels of simplification are measured and evaluated through the following metrics: segment length, sinuosity, areal displacement, vector displacement, and Hausdorff distance. These metrics may guide the automated selection of tolerance values for feature simplification that is appropriate for cartographic and analysis purposes over a range of reduced map scales and geographic conditions.

KEYWORDS: automated generalization, line simplification, sinuosity, Hausdorff distance, terrain

Introduction

Vector feature simplification is an important topic in cartography and geographic information science because it affects map representation and data analysis. The design of automated line or polygon simplification procedures is an ongoing research topic for several reasons. One reason stems from the challenge of devising algorithms that produce simplified features that are cartographically acceptable. Another reason lies in the facility with which simplified features can be measured and compared in digital environments.

Most literature on cartographic simplification either presents, compares, or describes requirements of new or existing algorithms. This paper seeks to contribute to a smaller group of pieces that considers measurements, or metrics, of simplification. In particular, the work presented determines if differences caused by two landscape characteristics, namely slope and local line density, can be found among an array of simplification metrics computed for linear networks representing hydrographic and roadway features. In addition to addressing the question of differences observable over diverse topographies, this study considers metrics of line displacement with regard to map accuracy standards, such as those maintained by national topographic mapping agencies.

Focusing on the bend simplify (BS) algorithm (Wang and Muller, 1998), which enjoys widespread availability as part of Esri's ArcGIS® Simplify Line tool, this study provides indication that simplification tolerances for this algorithm can be tailored to both landscape characteristics and positional accuracy standards for a range of scales.

A Brief Chronology of Advances in Line Simplification Metrics

Active research in cartographic line simplification precedes the early digitization of mapmaking (e.g., Perkal, 1966). Line simplification is one element of the larger field of cartographic generalization described by Harrie and Weibel (2007) as having evolved from scenarios in which cartographers respond to problems encountered during processing, through semi-automated processes in which human interaction is crucial, to constraint-based modeling, where fully automated processes operate to satisfy user-defined requirements. Most line simplification to date, even among the research community, involves human interaction, but until recently has lacked evaluation methods. Some recent advances have utilized fully automated agent-based models with embedded conflict resolution and inclusion of some self-evaluation metrics (Duchêne et al., 2012; Ruas, 1999).

Influential early work by Peucker (1976) modeled digital polylines as potentially noisy representations of linear features made up of varying vertex frequencies along and on either side of a line, where certain "high frequency" vertices could be removed to simplify lines. Several later studies asserted the importance of vertex reduction (Jenks, 1989; McMaster, 1987), thereby suggesting an obvious metric of simplification, namely the proportion of lost information. More recent research, however, has shifted focus to metrics of displacement since vertex reduction does not reflect representation accuracy (Dutton, 1999; Li, 1993; Raposo, 2010), whereas measures of deviation can be related to map accuracy constraints.

Many different algorithms for line simplification have been developed, both in cartography and cognate fields such as signal processing and pattern recognition. A relatively small set of algorithms have been made available in commercial geographic information system (GIS) software. Notable among these are the two algorithms available in Esri's ArcGIS® Simplify Line geoprocessing tool: the Douglas-Peucker point-remove algorithm (Douglas and Peucker, 1973; also independently designed in computer science by Ramer, 1972), and the BS algorithm (Wang, 1996; Wang and Muller, 1998). The point-remove algorithm uses a user-specified bandwidth tolerance measured perpendicular to the line to determine which vertices should be deleted, referencing Peucker's frequency-based conception of a cartographic line described above. In contrast to this, the BS algorithm iteratively considers the angles formed at vertices along the line for elimination based on whether the "bend" they form in the line is cartographically significant. Significance is established with reference to a user-provided "baseline" polyline segment length. When an angle formed by three sequential vertices is eliminated, the algorithm connects the remaining two endpoints that form the "baseline." Both algorithms define simplified lines by a subset of the input line vertices. Cartographers generally use the point-remove algorithm to process lines with orthogonal angles and the BS algorithm to simplify more curvilinear features.

Several approaches to metric assessment of line simplification have been undertaken. McMaster (1986, 1987) detailed 6 objective measures of simplification, having determined these as significant from a principle components analysis of 30 measures. White (1985) compared algorithms based on areal displacement, whereas Jenks (1989) considered vector deviation as an objective approximation of how well a simplification retained the "characteristic points" (Attneave, 1954; Marino, 1979) inherent to the input line. Measures of geometric difference from mathematics also can be used to evaluate map line simplification (Hangouët, 1995; Raposo, 2011).

In addition to assessing the performance of various algorithms, researchers have sought to use measures of simplification to guide the parameterization of simplification algorithms. Buttenfield (1991) introduced the idea of "structure signature", created by partitioning a cartographic line into segments and recording diverse simplification parameters for successive levels of resolution. Veregin (1999) suggested that measures of deviation could be used to optimize simplification algorithms, and sought to make distinctions in the meaningfulness of various metrics taken on products of the Douglas-Peucker algorithm across both rivers and roadways. Veregrin (1999) further suggested that positional errors introduced by the Douglas-Peucker algorithm can be related to the input tolerance when deviation values are considered in aggregate, and that this information can be used to govern the selection of tolerances when a given accuracy standard must be met.

Virtually all line simplification algorithms thus far used in cartography involve a user-defined parameter which determines the degree to which the algorithm reduces linear detail. Several authors have advocated for the calibration of these parameters to quantifiable properties such as visual resolution (Li and Openshaw, 1990; Raposo, 2010; Tobler, 1988), accuracy standards (Veregin, 2000), and geometric and scale-specific difference along a linear feature (Buttenfield, 1986, 1989). This research seeks to extend these ideas by beginning to consider the effect that a line's local topography may have on the degree of positional deviation caused by a given tolerance value.

Methods

Test Data from Humid Regions of the Coterminous United States

Ten test subbasins from the National Hydrography Dataset (NHD) are selected from humid regions that lie within flat, hilly and mountainous landscape topographies of the coterminous United States (Figure 1). For this work, humid regions experience more than 140 millimeters per year (mm/year) of runoff as depicted by a 5-kilometer (km) cell mean annual runoff model (McCabe and Wolock 2008; McCabe and Markstrom 2007) for 1951 to 2000. Landscape categories are estimated from average slope values determined for 5-km cells of USGS 1:250,000-scale 3-arc-second digital elevation models, which have a nominal resolution of 90 meters (m). Slope categories range from 0.0 to 1.5; 1.5 to 7.0; and greater than 7.0 percent rise for the flat (low-slope), hilly (mid-slope), and mountainous (high-slope) landscapes, respectively.

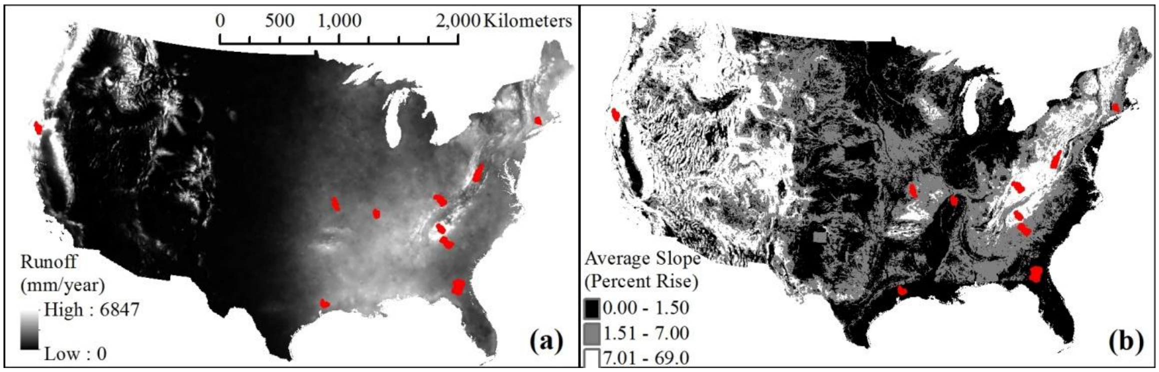

Figure 1. Distribution of 10 test subbasins displayed over (a) mean annual runoff, and (b) average slope.

Linear features from the hydrography and roadway networks of *The National Map* databases are evaluated for each subbasin. With respect to hydrography, one goal of this study is to measure geometric relations of linear features with landscape; therefore, only naturally-occurring stream line features are tested. Tested NHD subbasin data are compiled from 1:24,000-scale (24K) source material. Linear road networks from TomTom North America, Inc. compiled in 2011 were clipped to each subbasin boundary and evaluated. USGS displays TomTom road data on US Topo, 24K topographic maps.

To study the impact of local density, hydrography and road network lines are assigned to local line-density partitions having density class breaks of less than 1.0, between 1.0 and less than 2.5, and greater than 2.5 kilometers per square kilometer (km/km^2) for the low-, medium-, and high-density partitions, respectively. Line-density partitions were generated through a raster-based approach with a minimum mapping unit of 15 square kilometers (km^2) (Stanislawski and Buttenfield, 2011). For hydrography, the total number of features in the low-, medium-, and high-density partitions range from 11 to 3,287; 578 to 6,632; and 0 to 13,460, respectively. For transportation, the total number of road features in the low-, medium-, and high-density partitions range from 77 to 1,387; 1,894 to 12,721; and 351 to 6,682, respectively.

Line Simplification and Assessment Metrics

Linear network features for hydrography and roads were simplified with the BS algorithm (Wang and Muller, 1998) available in Esri's ArcGIS® Simplify Line tool. Features were simplified with five tolerance values: 15, 25, 50, 100, and 200 meters. Metrics computed for original and simplified line features are average sinuosity, percent change in average sinuosity, average displacement area, maximum Hausdorff distance (MHD), average vector displacement per kilometer, mean of average segment length per feature, and average segment length per class. Because of the page limits, only average sinuosity, MHD, average vector displacement per kilometer, and average segment length per feature are presented for hydrography, and MHD and average vector displacement per kilometer are presented for roads. Metrics separately are computed for each subbasin and density class. Processing is automated through Python programming and ArcGIS® geoprocessing functions.

Average sinuosity. The sinuosity of a line feature is defined as the total length of all polyline line segments divided by the distance between the polyline endpoints. Average sinuosity is computed for all features in each density class.

Maximum Hausdorff distance (MHD). The Hausdorff distance (HD) between two geometric shapes can be defined as the largest minimum distance between any point on one shape to any point on the other (Hangouët, 1995; Rucklidge, 1996; Nutanong, 2011). This is an important measure to compare along with area displacement; a feature could have a large displacement area without the lines significantly deviating from their source features. An HD is computed between the original geometry and a simplified version of the feature's geometry. An MHD is determined from the HDs for the set of features in each density class.

Checking the distance from every point in a source feature to every point in the generalized version of the feature is computationally intensive; therefore, an alternate, approximate method was used that constructs the minimum bounding rectangle (MBR) around each displacement polygon for a feature. The maximum vector displacement (MVD) for a polygon is determined by comparing the length and width of the MBR with the length of the intersection of the simplified line with the displacement polygon. If the difference between the MBR width and the intersection length is greater than the difference of the MBR length and the intersection length, then MVD is estimated as the width of the MBR, otherwise it is the length of the MBR. The HD for a feature is the maximum MVD for all displacement polygons on the feature.

To further reduce processing time, HDs were computed for a 5-percent sample of features in any density class having more than 1,000 features, otherwise all features were used. Sample features are selected by sorted length to preserve a length distribution similar to the distribution of all features in the associated density class.

Average vector displacement per kilometer. Vector displacement for a feature is the average MVD for all displacement polygons for the feature. Vector displacement per kilometer for a feature is the vector displacement divided by the feature length in kilometers. The average vector displacement is the sum of the vector displacement per kilometer values for all features in a density class divided by the number of features in the set. Sampling for this computation is completed in the same manner as used for the MHD.

Mean of average segment length per feature. Average segment length per feature is the total length of a feature divided by the number of line segments in the feature. The mean value is the mean of all average segment lengths per feature in a density class.

Results

Results are discussed for each feature type and metric type below.

Hydrography

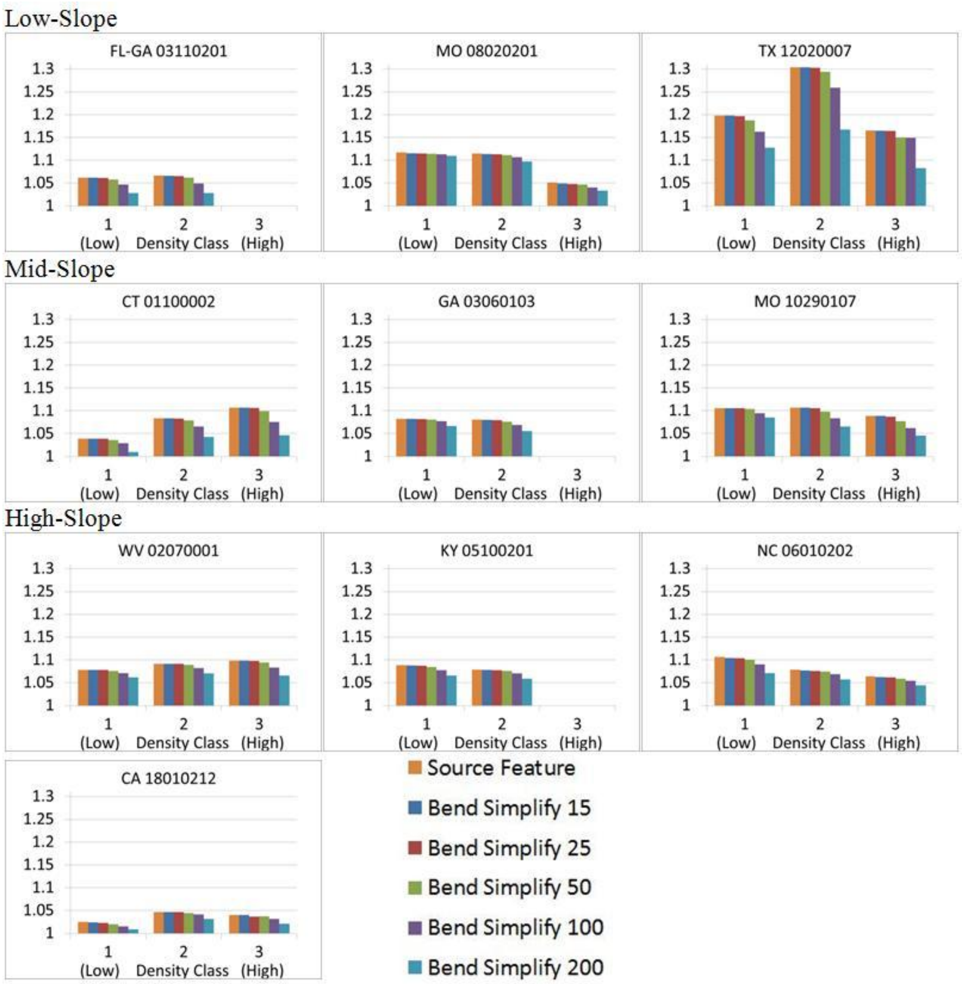

Figure 2. Average sinuosity for original and simplified stream features in three density classes for 10 test subbasins in low-, middle- (mid-), and high-slope landscapes in humid regions of the coterminous United States.

Average sinuosity. Average sinuosity values for hydrography range from nearly 1.0 for the California (CA) subbasin up to as much as 1.3 for the Texas (TX) subbasin (Figure 2). The length of the CA subbasin stream features average about 0.41 km, whereas features in the TX subbasin average about 1.0 km. In addition the TX subbasin includes a relatively large proportion of braided channels within flood plain areas compared to other subbasins. As expected, average sinuosity decreases with increasing BS tolerance within each density class for each subbasin, and the percent change in sinuosity increases with increasing tolerance. The low-slope subbasins seem to have larger range and average sinuosity values than the mid- and high-slope categories. Local density does not have a consistent effect on sinuosity.

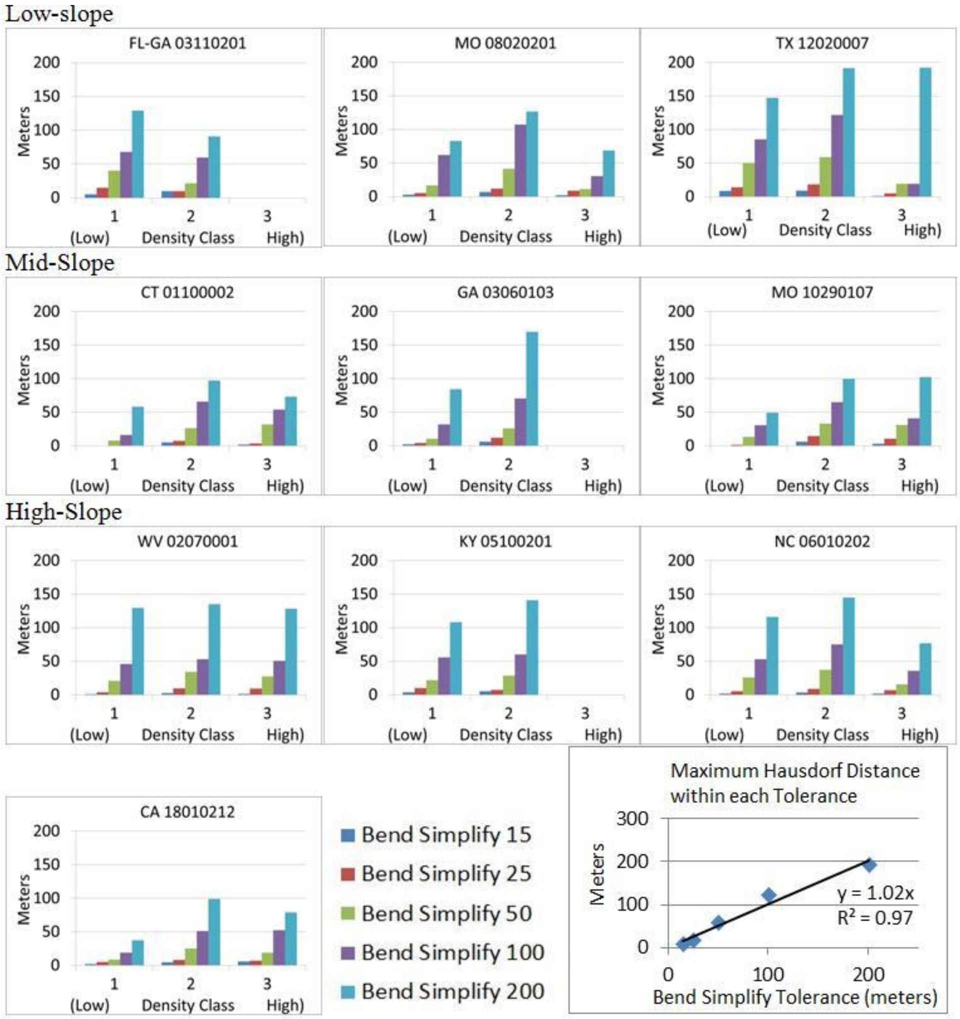

Figure 3. Maximum Hausdorff distance between original and simplified stream features in three density classes for 10 test subbasins in low-, middle- (mid-), and high-slope landscapes in humid regions of the coterminous United States. Regression of maximum within each bend-simplify tolerance also is shown.

Maximum Hausdorff distance (MHD). MHDs range from about 0 to 193 m for all density, slope, and tolerance values (Figure 3). As with sinuosity, low-slope subbasins seem to have larger MHD values than mid- and high-slope subbasins; therefore, there may be some interaction between sinuosity and MHD for stream features. Again, no consistent relation is apparent between MHD and local density, but there seems to be a consistent positive relation between MHD and the BS algorithm tolerance. United States National Map Accuracy Standards (NMAS) identify scale-based limits of acceptable horizontal displacements for well-defined points (US Bureau of Budget, 1947). The regression line predicting MHD from the algorithm tolerance could assist in constraining simplification displacement within NMAS, although further analysis is needed.

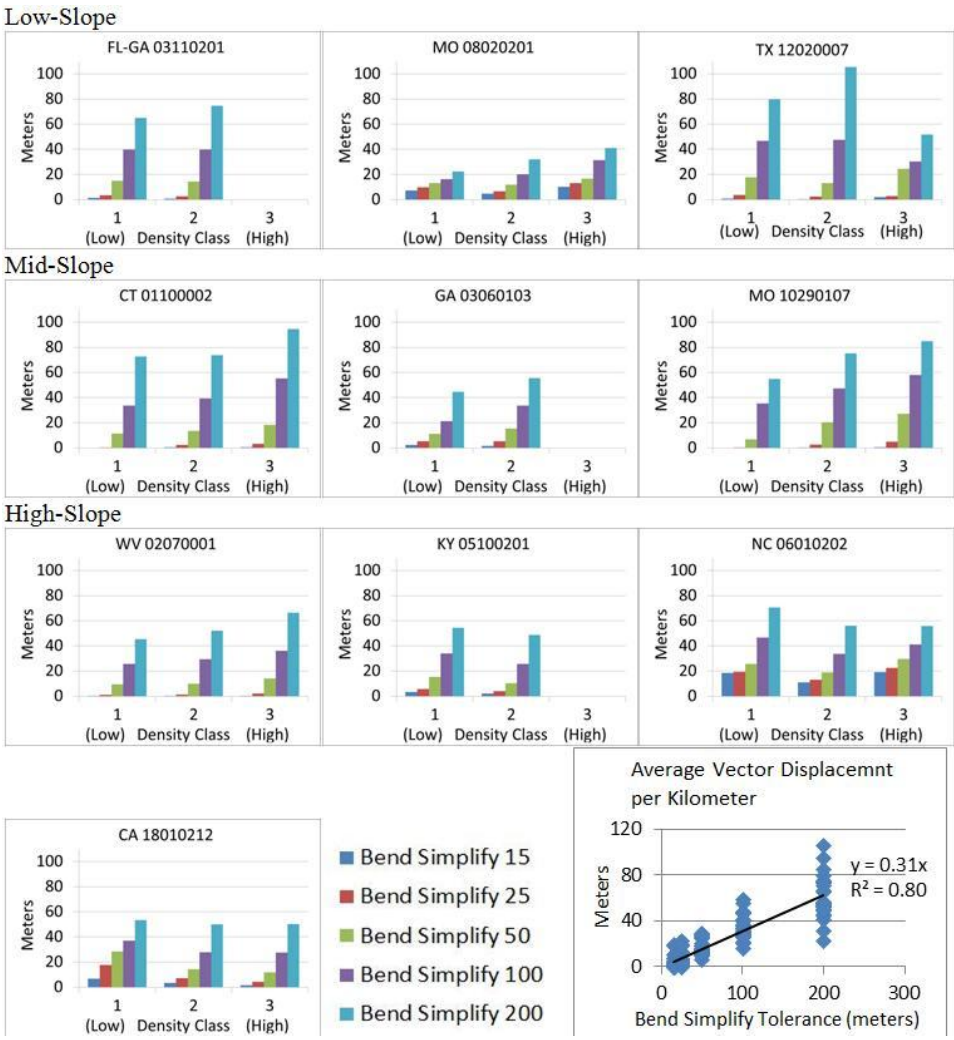

Figure 4. Average vector displacement per kilometer between original and simplified stream features in three density classes for 10 test subbasins in low-, middle- (mid-), and high-slope landscapes in humid regions of the coterminous United States. Regression line with bend-simplify tolerance also is shown.

Average vector displacement per kilometer. Average vector displacement per kilometer values range from about 0 to 106 m for all subbasins and BS tolerance values (Figure 4). No consistent relation visually is apparent between average displacement per kilometer with slope, or with density; however, average displacement per kilometer has a good positive relation with algorithm tolerance, similar to MHD.

Mean of average segment length per feature. This metric ranges from about 6 to 165 m for all subbasins and BS tolerance values (Figure 5). The low-slope Missouri (MO) subbasin has relatively high mean values compared to other subbasins. This heavily farmed MO subbasin is in the coastal plains section of the Mississippi River watershed, and it has a drainage network composed largely of man-made, channelized drainage features with few natural streams. Man-made or canal features usually are straight with few vertices in their representation, and consequently have relatively large average segment lengths. No consistent relation visually is apparent between the mean of average feature segment length with slope, or with density. Overall, there is a very weak positive relation between the mean of average segment length per feature and BS tolerance (Figure 5, R^2=0.19); however, visual inspection indicates a consistent positive relation between this metric and BS tolerance within each density class of each subbasin.

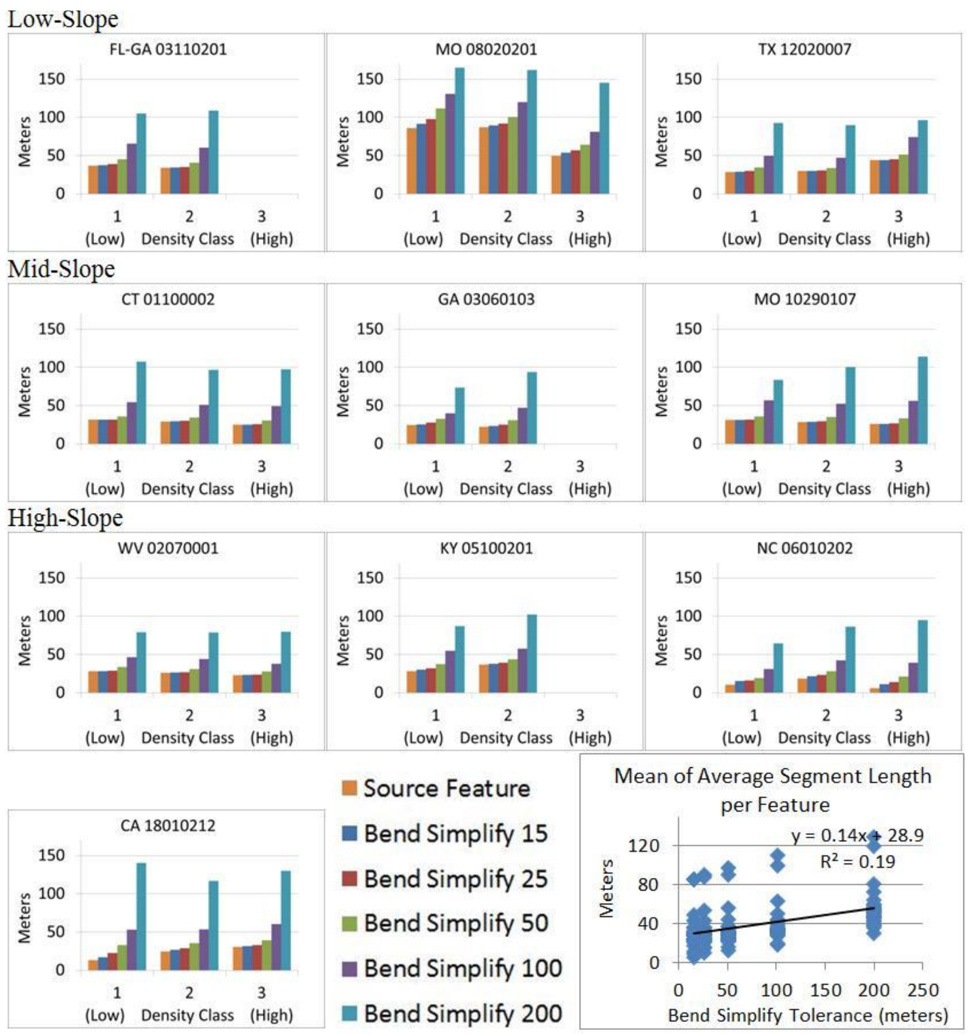

Figure 5. Mean of average segment length per feature for original and simplified stream features in three density classes for 10 test subbasins in low-, middle- (mid-), and high-slope landscapes in humid regions of the coterminous United States. Regression line with bend-simplify tolerance also is shown.

Transportation

Maximum Hausdorff distance. MHDs for road features range from about 0.5 to 396.5 m for all density, slope, and tolerance values (Figure 6); however, the highest value for the California (CA) subbasin is overestimated with the MBR method. The correct MHD for the 200 BS tolerance of high-density features for the CA subbasin is 306.2 m. Opposite to results for hydrography features, MHDs seem to increase with slope. No consistent relation is apparent between MHD and local density. As with hydrography, a consistent positive relation exists between MHD and the BS tolerance, but the slope of this relation is about 50 percent higher than the hydrography relation. This result again suggests a regression between MHD and BS tolerance may be used to help constrain simplification displacement within NMAS.

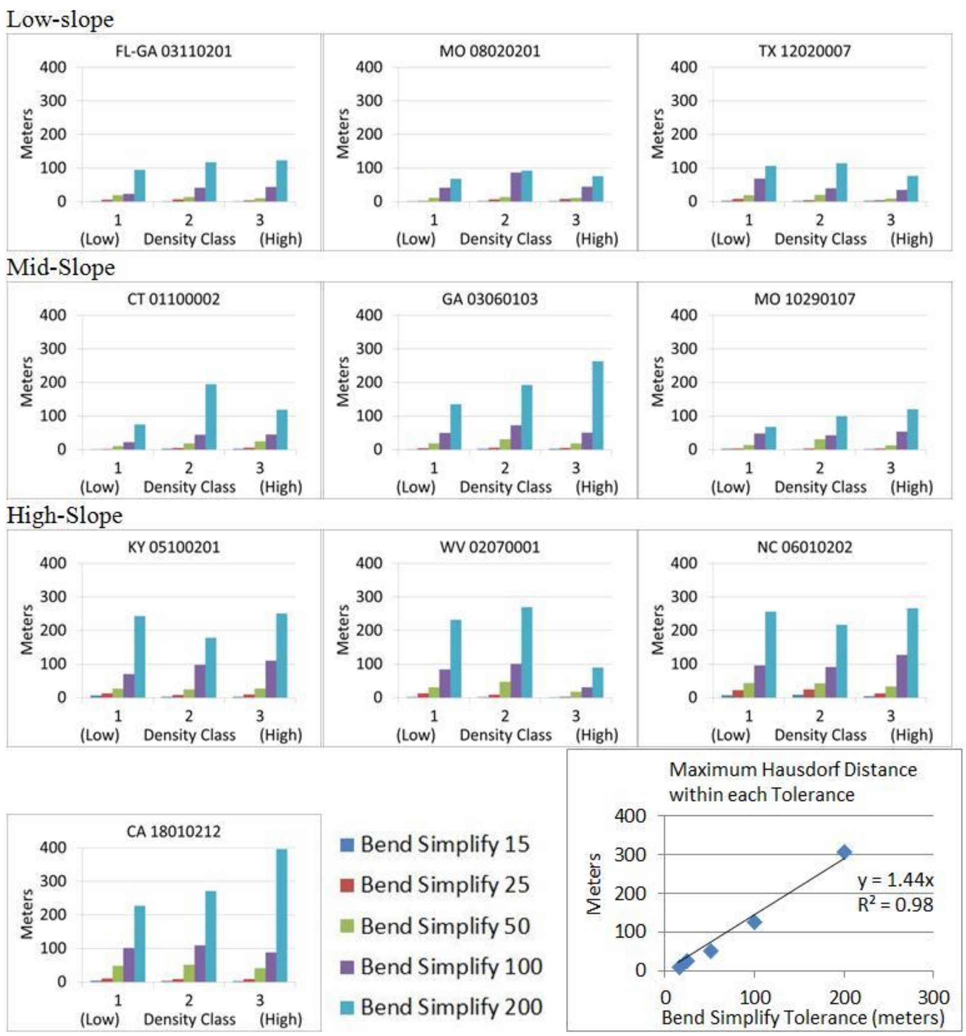

Figure 6. Maximum Hausdorff distance between original and simplified road features in three density classes for 10 test subbasins in low-, middle- (mid-), and high-slope landscapes in humid regions of the coterminous United States. Regression of maximum within each bend-simplify tolerance also is shown.

Average vector displacement per kilometer. Average vector displacement per kilometer values of transportation features range from about 0 to 102 m (Figure 7). As with MHD, average displacement per kilometer seems to increase with increasing slope categories. No consistent relation is apparent between vector displacement per kilometer and density. Between subbasins, a weak positive linear relation exists that predicts average displacement per kilometer from BS tolerance (Figure 7). Average displacement consistently increases with increasing BS tolerance within each density class of each subbasin.

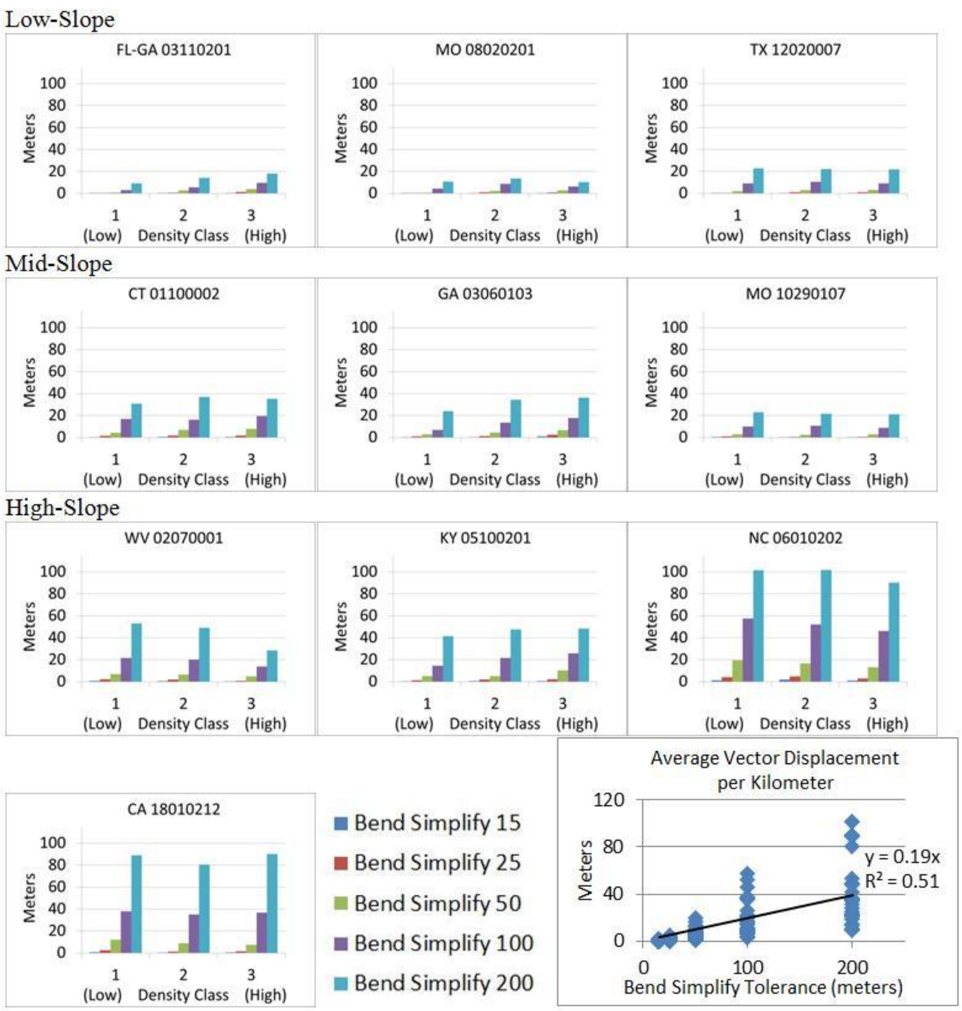

Figure 7. Average vector displacement per kilometer between original and simplified road features in three density classes for 10 test subbasins in low-, middle- (mid-), and high-slope landscapes in humid regions of the coterminous United States. Regression line with bend-simplify tolerance is also shown.

In summary, initial results of line simplification through the BS algorithm suggest a consistent positive relation exists between displacement and segment length metrics with BS algorithm tolerance, but results also imply that more aggressive algorithm tolerance values might modify the slope of regression models. Expanding the range of algorithm tolerance values forms a component of ongoing research. Also, relations between BS tolerance and the various metrics seem to vary with local conditions. Landscape slope appears to consistently affect metrics for original and simplified road features, which seems intuitive given the effect of increasing slope on road and highway construction. Additional testing in other landscape conditions and through more rigorous statistical analysis should better establish patterns in the geometric characteristics of linear stream and road features with relation to local landscape conditions and simplification operations. Ultimately, relations established through this work will help devise rules and constraints to automate the selection of tolerance values for feature simplification that is appropriate for cartographic and analysis purposes over a range of reduced map scales and geographic conditions.

Acknowledgements

Paulo Raposo's work is supported by CEGIS research grants to Penn State and by the Gould Center in the Department of Geography. The work of Dr. Buttenfield is supported by USGS-CEGIS grant #04121HS029, “Generalization and Data Modeling for New Generation Topographic Mapping.”

References

Attneave, F. (1954) Some Informational Aspects of Visual Perception. *Psychological Review*, 61, 3, pp. 183-193.

Brassel, K. and Weibel, R. (1988). A Review and Conceptual Framework of Automated Map Generalization, *International Journal of Geographical Information Systems*, 2, 3, pp. 229-244.

Buttenfield, B. P. (1986) Digital Definitions of Scale-Dependent Line Structure. *AutoCarto* London Conference, London, England.

Buttenfield, B. P. (1989) Scale-Dependence and Self-Similarity in Cartographic Lines. *Cartographica*, 26, 1, pp. 79-100.

Buttenfield, B. P. (1991) A Rule for Describing Line Feature Geometry. *in* Buttenfield, B. P. and McMaster, R. B., eds., *Map Generalization: Making Rules for Knowledge Representation*, Essex: Longman Scientific and Technical, pp. 150-239.

Douglas, D. H. and Peucker, T. K. (1973) Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature. *Cartographica*, 10, 2, pp. 112-122.

Duchêne, C. Ruas A. and Cambier, C. (2012) The CartACom Model: Transforming Cartographic Features Into Communicating Agengs for Cartographic Generalization. *International Journal of Geographical Information Science*, (advance online publication), 1-30. doi: 10.1080/13658816.2011.639302

Dutton, G. (1999) Scale, Sinuosity, and Point Selection in Digital Line Generalization. *Cartography and Geographic Information Science*, 26, 1, pp. 33-53.

Hangouët, J. F. (1995) Computation of the Hausdorff distance between plane vector polylines. AutoCarto12 Conference, Charlotte, North Carolina.

Harrie, L. and Weibel, R. (2007) Modelling the Overall Process of Generalisation. *in* Mackaness, W. A. Ruas, A. and Sarjakoski, L. T., eds., *Generalisation of Geographic Information: Cartographic Modelling and Applications*, Elsevier, pp. 67-87.

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

Li, Z. (1993) Some Observations on the Issue of Line Generalization. *The Cartographic Journal*, 30, 1, pp. 68-71.

Li, Z. and Openshaw, S. (1990) A Natural Principle of Objective Generalization of Digital Map Data and Other Spatial Data. *RRL Research Report: CURDS*, University of Newcastle upon Tyne.

Marino, J. (1979) Identification of Characteristic Points Along Naturally Occurring Lines: An Empirical Study. *The Canadian Cartographer*, 16, 1, pp. 70-80.

McCabe, G. J. and Markstrom, S. L. (2007) A Monthly Water-balance Model Driven by a Graphical User Interface. *U.S. Geological Survey, Open-File Report 2007-1088*, 6pp.

McCabe G. J. and Wolock, D. M. (2008) Joint Variability of Global Runoff and Global Sea Surface Temperatures. *Journal of Hydrometeorology*, 9, pp. 816-824.

McMaster, R. B. (1986) A Statistical Analysis of Mathematical Measures for Linear Simplifcation. *The American Cartographer*, 13, 2, pp. 103-116.

McMaster, R. B. (1987) Automated Line Generalization. *Cartographica*, 24, 2, pp. 74-111.

Nutanong S. Jacox E. H. and Samet H., (2011) An incremental Hausdorff distance calculation algorithm. 37th International Conference on Very Large Data Bases, Seattle, Washington, August 29-September 3.

Perkal, J. (1966) An Attempt at Objective Generalization. Translated by W. Jakowski from Perkal, J., *Proba obiektywnej generalizacji. Geodezia es Kartografia*, Tom VII, Zeszyt 2, 1958, pp. 130-142 *in* Nystuen, J., ed., *Michigan Inter-University Community of Mathematical Geographers, Discussion Paper* 10, Ann Arbor, Michigan.

Peucker, T. (1976) A Theory of the Cartographic Line. *International Yearbook of Cartography*, 16, pp. 134-143.

Ramer, U. (1972) An Iterative Procedure for the Polygonal Approximation of Plane Curves. *Computer Graphics and Image Processing*, 1, pp. 244-256.

Raposo, P. (2010) Piece by piece: a method of cartographic line generalization using regular hexagonal tessellation. ASPRS/CaGIS 2010 Fall Specialty Conference, AutoCarto 2010, Orlando, Florida.

Raposo, P. (2011) Scale-Specific Automated Map Line Simplification by Vertex Clustering on a Hexagonal Tessellation. *Master's Thesis*, The Pennsylvania State University, University Park, PA.

Ruas, A. (1999) Modèle de Généralisation de Données Géographiques à Base de Contraintes et d'Autonomie. *Ph.D.*, University de Marne-la-Vallée, Marne-la-Vallée, France.

Rucklidge, W. (1996). *Efficient Visual Recognition Using the Hausdorff Distance*. Berlin: Springer Verlag.

Stanislawski, L. V. and Buttenfield, B. P. (2011) A raster alternative for partitioning line densities to support automated cartographic generalization. 25th International Cartography Conference, Paris, July 3-8.

Tobler, W. R. (1988) Resolution, resampling, and all that. *in* Mounsey, H. and Tomlinson, R. F., eds., Building Databases for Global Science: Proceedings of 1st Meeting of the International Geographical Union Global Database Planning Project, Hampshire, U.K.: Taylor and Francis, pp. 129-137.

US Bureau of the Budget (1947) United States National Map Accuracy Standards, [revised June 17 1947].

Veregin, H. (1999) Line Simplification, Geometric Distortion, and Positional Error. *Cartographica*, 36, 1, pp. 25-39.

Veregin, H. (2000) Quantifying Positional Error Induced by Line Simplification. *International Journal of Geographical Information Science*, 14, 2, pp. 113-130.

Wang, Z. (1996) Manual versus automated line generalization. GIS/LIS '96 Conference, Denver, Colorado.

Wang, Z. and Muller, J. C. (1998) Line Generalization Based on Analysis of Shape Characteristics. *Cartography and Geographic Information Science*, 25, 1, pp. 3-15.

White, E. R. (1985) Assessment of Line-Generalization Algorithms Using Characteristic Points. *Cartography and Geographic Information Science*, 12, 1, pp. 17-28.