# Generalization of Hydrographic Features and Automated Metric Assessment Through Bootstrapping

Center of Excellence for Geospatial Information Science (CEGIS)
United State Geological Survey (USGS)
1400 Independence Road, Rolla Missouri, 65401

Lawrence V. Stanislawski, ATA Services, Inc., lstan@usgs.gov

Barbara P. Buttenfield, Department of Geography, University of Colorado-Boulder, babs@colorado.edu

V. A. Samaranayake, Department of Mathematics and Statistics, Missouri University of Science and Technology, vsam@mst.edu

**Keywords: automated generalization, level of detail, hydrographic network, National Hydrography Dataset, bootstrapping, coefficient of line correspondence**

## 1 Introduction

Geospatial data researchers have used various techniques to assess different aspects of cartographic generalization. White (1985) and Jenks (1989) described geometric and perceptual methods to compare line generalization algorithms. Mackaness and Ruas (2007), categorized numerous ways evaluation has been applied in the map generalization process, and identified a need for additional evaluation methods to assess data after generalization, as opposed to evaluation methods for controlling or tuning generalization systems. Stoter and others (2009) developed evaluation methods to assess generalization outputs from commercial software systems based on measured or perceived satisfaction of constraints defined for generalization and mapping specifications. A large number of constraints (about 250) were identified by Stoter and others (2009), but only two measures were automated to evaluate constraint satisfaction: minimum area of buildings and minimum distance between buildings. Xiang and others (2008) developed an automated approach to assess the preservation of feature densities before and after generalization. Automated approaches that quantitatively assess the quality of cartographic generalization are not well established and require further research.

An automated method to quantitatively compare a generalized dataset to an accepted standard, or benchmark is presented in this paper. The approach generates confidence intervals for two metrics: the coefficient of line correspondence (CLC) (Stanislawski, 2009a, 2009b), Buttenfield and Stanislawski, 2010), and the coefficient of area correspondence (CAC). Both measures quantify the amount of conflation between a generalized dataset and an independent benchmark dataset. The CLC metric and the establishment of confidence intervals through bootstrapping (Fortin and others, 2002) is the subject of this paper. Bootstrapping is a nonparametric method that establishes a probability distribution for a statistical sample when conventional parametric assumptions cannot support reliable inference (Efron, 1981). Observed data are resampled with replacement to generate additional samples during bootstrapping.

A variety of climate and terrain conditions exist in the United States, and optimal cartographic generalization techniques for one area of the country may not be suitable for another, particularly when working with surface hydrographic data. Nonparametric, bootstrapped confidence intervals for the CLC and CAC can be utilized to verify the suitability of tailoring generalization procedures to one or more specific landscape conditions. Ultimately, bootstrapping could help optimize development efforts of automated generalization for the United States.

The Center of Excellence in Geospatial Information Science (CEGIS) of the U. S. Geological Survey (USGS) has been developing automated generalization tools to support display and delivery of *The National Map* (Stanislawski, 2008, 2009a, 2009b, Stanislawski and others, 2009, Brewer and others, 2009). Currently (August 2010), the approach is to transform any data theme of *The National Map* from its compiled level of detail to reduced levels of detail (LoD data versions, Cecconi and others, 2002) that can be delivered to users or used for smaller scale mapping. Each data theme is collected and maintained at the highest required level of detail, which can vary by theme. Reduced LoDs are generated through generalization processes that ideally are fully automated, thereby, minimizing data collection and database maintenance efforts. The National Land Survey of Finland has implemented such an approach to generate 1:100,000-scale (100K) level of detail for Finland from data compiled for the 1:5,000 to 1:10,000-scale level of detail (Pätynen and Ristioja, 2009).

Confidence interval processing for a subbasin of data from the United States National Hydrography Dataset (NHD) is presented in this paper. The 1:24,000-scale (24K) data were pruned and generalized to create an LoD appropriate for displays ranging from 1:50,000 to 1:200,000 (referred to as a 50K LoD) by using procedures tailored for the local climate and terrain conditions of the subbasin. The resulting 50K LoD was compared to the benchmark 100K NHD by computing measures assessing correspondence between line features at the two scales, and then developing statistical confidence intervals that specify a range of acceptable correspondence values. Development of metrics and confidence intervals is discussed in the methodology section after a description of the data and generalization process.

## 2 Methodology

### 2.1 Test data

The test subbasin (#10290107 in the 24K NHD), forms the watershed for the Pomme de Terre River, in the Midwest United States, in Missouri. The subbasin sits in the Ozark Plateau of the Interior Highlands physiographic province of the coterminous United States (Fenneman and Johnson, 1946). The subbasin covers about 2,190 square kilometers ($km^2$). The landscape is hilly but not mountainous terrain, and the climate is humid.

Landscape conditions for the subbasin are quantified from average elevation (318.5 meter (m)), standard deviation of elevation (15.41 m), and average slope (2.52 percent rise) in 5-km grid cells over the subbasin as determined from the 1:250,000-scale USGS digital elevation model (DEM). Runoff estimates for the 5-km cells were obtained from James Falcone and Dave Wolock of the USGS. Runoff estimates were developed using the water-balance model described by Wolock and McCabe (1999) that estimates mean annual watershed runoff from 1951 to 2000. Average runoff for the 5-km cells in the subbasin is 324 millimeters per year.

### 2.2 Generalization process

The high-resolution NHD data for this area were compiled from 24K source material. Deriving the LoD involves several generalization processes tailored to enrich the hydrographic attributes with catchment areas, approximate upstream drainage areas (Stanislawski 2008, 2009b), and stream channel densities that enable cartographic display at scales ranging from 1:50,000 to about 1:200,000 after pruning, selection, simplification and custom symbolization.

Enrichment processing serves multiple purposes, first to estimate local density values for each stream reach (confluence-to-confluence), which supports density stratification, also referred to as partitioning (Bobzien and others, 2008, Chaudhry and Mackaness, 2008). Second, approximate upstream drainage area values identify relative prominence for linear network features. Local density and upstream drainage areas are used to prune the network features to 50K. Pruning iteratively eliminates entire stream reaches while preserving topologic integrity of the flow network, until the subbasin channel density meets a user-specified threshold. Pruning can homogenize channel density for the entire subbasin, and, in some cases, pruning can stratify flow network features to protect geographically important density differences (for example those caused by geomorphic, volcanic, or glacial processes (Stanislawski and others, 2009)). Network features in the test data were manually stratified into normal and high density partitions; although the long term goal is to automate this step.

Appropriate tolerance thresholds for pruning streams were estimated by using a variation of the Radical Law (Töpfer and Pillewizer 1966). The variation is based on feature length rather than number of features. Preservation of feature lengths has long been accepted as an important validation of generalization quality; that is, cartographers generalize to eliminate details while preserving overall lengths. In the Pomme de Terre subbasin, pruning from 24K to the 50K LoD reduced total channel length from 1,923 km to 1,303 km in the high density partition, and from 1,507 km to 1,055 km in low density partition, yielding a total channel length of 2,358 km, which is a 31.3 percent reduction.

After pruning, the remaining linear network features were simplified (vertices removed) by using ArcGIS "Simplify Line" tool (ESRI, 2009) with the bend simplify algorithm and a 100 m tolerance (Wang 1996, Wang and Muller 1998). Simplification further reduced the pruned data to 2,304 km, an additional 2.3 percent reduction. The topologic integrity of the flow network was maintained after pruning and simplification, which is essential for NHD model applications.

A minimum area of 0.02 km$^2$ was applied as a general rule to prune polygon features. Feature type rules also were applied to differentially eliminate hydrographic areas and waterbodies; that is, ponds and lakes were treated differently than reservoirs, flood inundation areas, etc. (fig. 1).

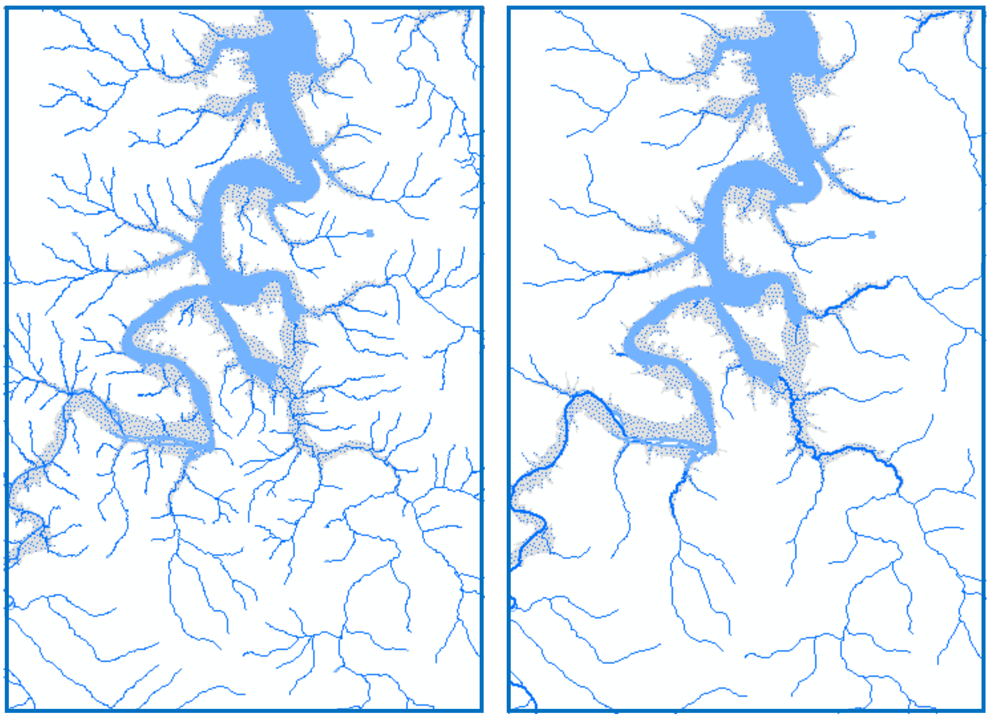

**Figure 1.** A part of the Pomme de Terre subbasin showing original 1:24,000-scale hydrography (left), and hydrography at the 1:50,000-scale level of detail after enrichment, pruning, and simplification (right).

### 2.3 Calculating the coefficient of line correspondence (CLC)

Through an automated process, the CLC measures the conflation of linear features in a generalized LoD with a standard benchmark, which in this case is the medium-resolution, 100K NHD database. Medium-resolution NHD data were originally compiled through a manual generalization of photo-reduced mosaics of 24K hydrography. Over the years, the 24K NHD largely has been updated to more recent representations, whereas the 100K has maintained original content. The CLC metric is based on channel length because pruning and simplification are intended to preserve feature length. Stream features are matched for the conflation assessment by confluence-to-confluence segments and by feature position; a small buffer compensates for slight displacements caused by simplification. The coefficient is computed as follows:

$$\text{CLC} = \frac{\sum \text{conflations}}{\sum \text{conflations} + \sum (\text{omissions} + \text{commissions})}.$$

where

conflations → length of conflated 100K benchmark channels
omissions → length of channels in 100K benchmark but not in LoD, and
comissions → length of channels in LoD but not in 100K benchmark.

Commission lengths are divided by the benchmark-to-LoD length expansion factor, which compensates for higher granularity in the LoD feature representations (Stanislawski, 2009b). The CLC metric ranges between 1 (perfect correspondence between LoD and benchmark) and values approaching 0 (as increasing discrepancies appear in the generalized data).

The subbasin was covered by a grid of 200 square cells with 3,670 m sides. A cell weight, equal to the area of the cell in the subbasin divided by the subbasin area, was determined for each cell to avoid edge bias. Weighted CLC values were computed for each cell by applying a 1.04 length expansion factor to reduce the length of the 50K LoD commission errors to the granularity of the benchmark lines, and then multiplying the cell weight by the cell CLC (Stanislawski, 2009b).

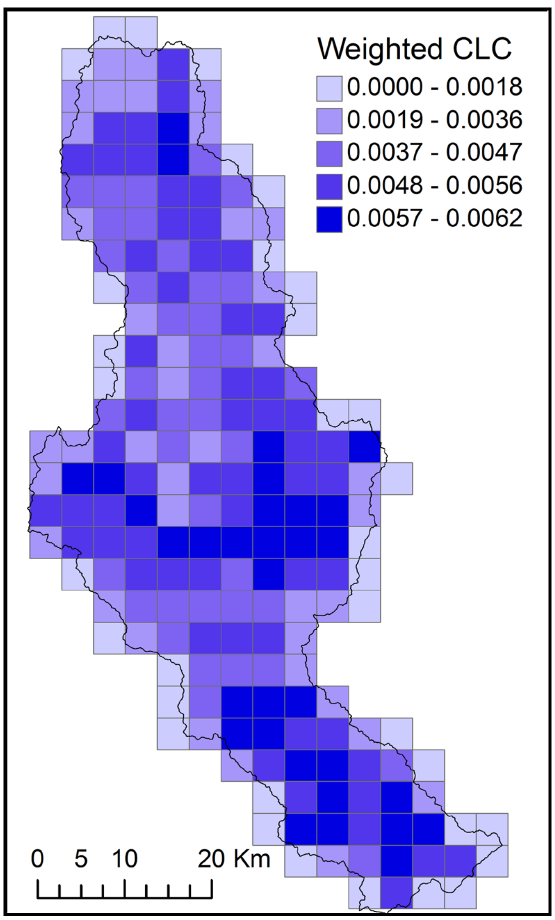


**Figure 2.** Weighted coefficient of line correspondence (CLC) values determined for the 200 cells in the Pomme de Terre subbasin (boundary shown in black).

Figure 2 shows the spatial distribution of weighted CLC values across the subbasin. Lowest values appear along the edge of the subbasin because relatively small proportions of these edge cells are inside the subbasin, which produces small cell weights. Furthermore, very few LoD flowline features are in the small sections of the subbasin inside these edge cells, where a perfect match or mismatch with the benchmark features is more likely. Small weights for the edge cells help reduce affects of sample bias for edge cells.

The sum of the cell weights is 1.0000, and the sum of the 200 weighted CLC cell values is 0.8020, which represents the CLC value for the subbasin. The sums of the weighted proportions of commission and omission errors are 0.1358 and 0.0622, respectively, indicating that a greater proportion of features appear to be included in the generalized dataset than the proportion of benchmark features that are missing from the generalized data, which is expected for a 50K LoD comparison to a 100K benchmark.

## 2.4 Confidence intervals for CLC

A confidence interval for a parameter not only provides approximate lower and upper bounds for the unknown value, but also states how confident one is about capturing the true value of the parameter within the bounds. A narrow confidence interval with a high confidence level indicates a reliable estimate, whereas a wider interval may signal that there is high variability in the data thus making any estimate based on that data less precise. For example the subbasin CLC (in this case 0.8020) measures how well the generalized LoD matches the standard benchmark dataset, and a confidence interval for a subbasin CLC provides a reliability measure for that particular CLC estimate. Furthermore, a CLC confidence interval can be used as a measure of the precision, or variability, of the CLC over the subbasin, and it can be furnished at standard levels of confidence, such as at the 90, 95, or 99 percent confidence levels.

Hence, CLC confidence intervals can be used to compare multiple CLC values for the same subbasin, which may be generated to test several generalization alternatives that use slightly different selection parameters or simplification tolerances. Such CLC confidence interval comparisons may indicate whether the patterns of conflation error from alternate generalizations are slightly or dramatically different. Additionally, results from the same generalization process applied to nearby subbasins may be compared through confidence intervals for the subbasin CLCs. Consequently, with regard to the evaluation subtypes defined by Mackaness and Ruas (2007), subbasin CLC values and associated confidence intervals may be considered evaluations for tuning generalization parameters, or for assessing quality after generalization.

Confidence intervals around the CLC metric were generated through a bootstrap resampling procedure (Fortin and others, 2002). Given a random sample from a population, the bootstrap procedure operates by resampling this original sample many times, treating it as if it were a population. Resampling is done with replacement, so each resample can be different from the original sample. Each resample is called a bootstrap sample and each bootstrap sample can be used to estimate the parameter of interest (eg., the population mean). These repeated estimates yield a frequency distribution for the estimate. This distribution enables one to compute sampling variability of the estimate and obtain statistics such as confidence intervals. When the resampling rate is very high, the frequency distribution can be assumed to replicate the probability distribution underlying the population from which the sample was initially drawn. In the case of the generalized LoD, individual cell CLC values can be resampled from the set of 200 cells to generate a bootstrap. To ensure that the resampling process reflects all types of weighted CLC values, a stratified random sampling strategy was used.

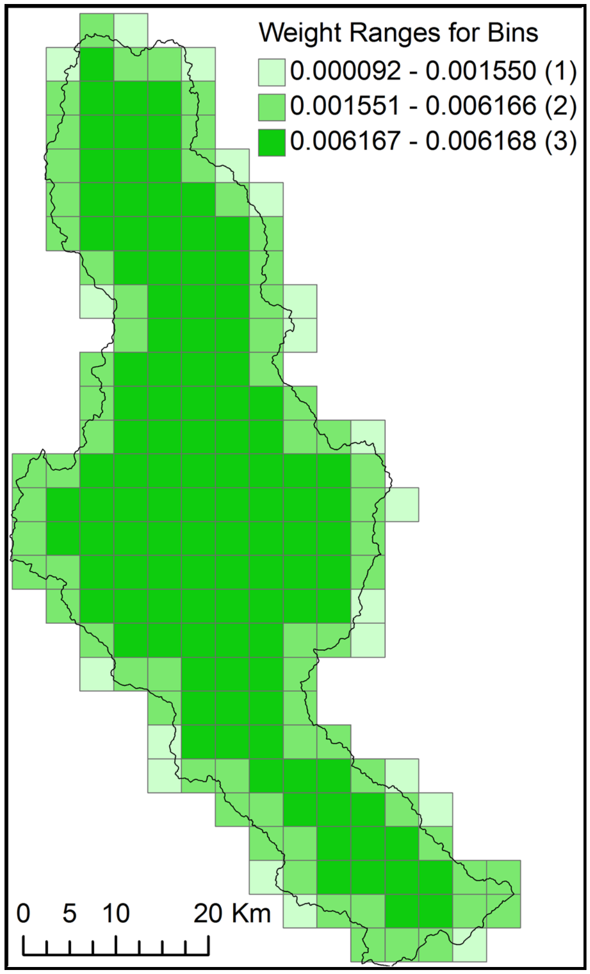


**Figure 3.** Spatial distribution and range of weights assigned to stratification bins (1, 2, and 3) for the bootstrap resampling process for Pomme de Terre subbasin (boundary shown in black).

The cells were assigned into three bins based on CLC weight, where edge cells with the lowest weights were assigned to bin 1, interior cells with the largest weights were assigned to bin 3, and remaining edge cells were assigned to bin 2. A total of 20, 66, and 114 cells were assigned to bins 1, 2, and 3, which is 10, 33, and 57 percent of the 200 cells, respectively. These percentages stratified each 200-cell random sample for the bootstrap process to produce a similar weighting structure in each sample. Figure 3 displays the spatial distribution of the bin stratification in the Pomme de Terre subbasin and the range of weights assigned to each bin.

Through automation, one thousand stratified random samples of 200 cells were resampled with replacement from the original set of 200 cells. Each resampled set of 200 weighted CLC values included the same number of cells from each bin as in the original set. The resulting 1,000 average CLC values created the bootstrap frequency distribution from which non-parametric confidence intervals were obtained.

The resulting bootstrapped interval at the 90 percent level of confidence for the weighted CLC value ranges from 0.7832 to 0.8205, with a bootstrapped average of 0.8018 (figure 4). The bootstrap distribution has a median of 0.8014 and mode of 0.801 from 0.001 bins. The mean, median, and mode are equal in a normal distribution. Furthermore, the normally distributed bootstrapped values, centered on the weighted CLC indicate the bootstrapping process is well adapted for the dataset (fig. 4). More specifically, the number of cells (200) used for subdividing the subbasin and the bin assignment percentages used for maintaining the relative diversity in each sample appear to produce an acceptable distribution. Additional testing should be done to verify this observation.

The bootstrapped distribution indicates the generalized results are within an acceptable level of reliability when assessed against the medium resolution (100K) NHD benchmark. The boostrapped average (0.8018) and original subbasin CLC value (0.8020) suggest that, loosely speaking, about 80 percent of the linear features match between the two tested datasets. The range between the upper and lower bounds of the 90 percent confidence interval is 0.0373, which is less than five percent of the bootstrapped average CLC, indicating that the average CLC estimate is fairly precise.

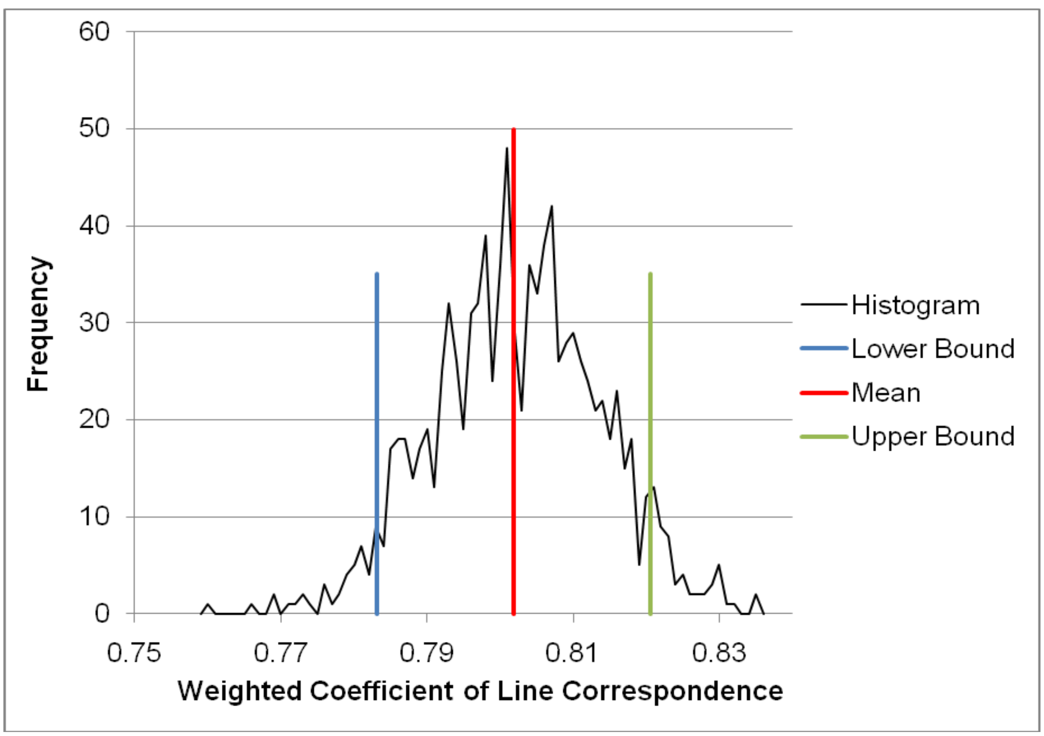


**Figure 4.** Histogram and 90 percent confidence interval for the 1000 weighted coefficient of line correspondence (CLC) values bootstrapped from the 200 cells in the Pomme de Terre subbasin.

## 3 Discussion

Although additional metrics could be used to assess overall adequacy of generalized data, the CLC and associated bootstrapped confidence interval can be a valuable tool for optimizing development of automated generalization procedures. Fully automated procedures for generating weighted CLC values for a subbasin of NHD data and the associated bootstrapped confidence intervals require 2 to 3 hours. The spatial distribution of weighted CLC values identifies areas that have relatively more unmatched linear features between the 50K LoD and the benchmark dataset, and highlights parts of the generalized dataset where further processing refinements may be focused. In addition, the confidence interval for the CLC can be used to establish the spatial extent where tested hydrographic generalization procedures may be successfully applied, and where variations of existing approaches should be applied or alternative approaches developed. While further validation and interpretation is needed, bootstrap-based confidence intervals for the CLC as described, and similarly for the CAC, appear to be an original and viable contribution for metric assessment of generalized cartographic data.

Assessment based on CLC and CAC estimates hinges on the existence of an acceptable benchmark dataset. The 100K NHD has been used as the benchmark to compare to the 50K LoD for this study. Strictly speaking, the assessment described in this paper only applies to maps that display the tested 50K LoD NHD at the 100K scale. Furthermore, compilation standards of generalized and benchmark datasets must be considered when making CLC and CAC based comparisons. Are the generalized and benchmark datasets compiled with similar standards from similar time periods? Given these limitations, variability in CLC and CAC based comparisons should be expected over large regions, such as the United States. However, the CLC and CAC appear to furnish consistent, reliable automated approaches for comparing how well two datasets, representing similar features, match.

## References

Bobzien, M., Burghardt, D., Petzold, I., Neun, M., and Weibel, R., 2008, Multi-representation databases with explicitly modeled horizontal, vertical, and update relations, *Cartography and Geographic Information Systems,* 35(1):3-16.

Brewer, C.A., Buttenfield, B.P, and Usery, E.L., 2009, Evaluating generalizations of hydrography in differing terrains for *The National Map* of the United States, *24th International Cartography Conference*, November 15-21, 2009, Santiago, Chile.

Buttenfield, B.P., and Stanislawski, L.V., 2010 (submitting), Metric assessment of generalized hydrography, *Sixth International Conference on Geographic Information Science*, Sept. 14-17, 2010, Zurich, Switzerland.

Cecconi, A. Weibel, R, and Barrault, M., 2002, Improving automated generalization for on-demand web mapping by multiscale databases, *International Symposium on Geospatial Theory, Processing, and Applications*, July 2002, Ottawa, Canada.

Chaudhry, O. and Mackaness, W.A., 2008, Partitioning to make manageable the generalization of national spatial datasets, *11th ICA Workshop on Generalization and Multiple Representation*, June 20-21, 2008, Montpellier, France.

Efron, B., 1981, Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods, *Biometrika*, 68:589–599. doi:10.1093/biomet/68.3.589.

ESRI, 2009. ArcGIS Desktop Help (9.3.1). 380 New York Street, Redlands, CA, USA.

Fenneman, and Johnson, 1946, Physical Divisions of the United States: Washington, D.C., U.S. Geological Survey special map series, 1:7,000,000-scale.

Fortin, M.J., Jacquez, G.M., and Shipley, B., 2002, Computer-intensive methods, In El-Shaarawi, A.H., and Piegorsch, W.W. (Eds.), *Encyclopedia of Environmetrics*, vol. 1, pp. 339-402, John Wiley & Sons, Ltd., Chichester.

Jenks, G.F., 1989, Geographic logic in line generalization, *Cartographica: Numerical Generalization in Cartography*, 26(1):27-42.

Mackaness, W.A., and Ruas, A., 2007, Evaluation in the map generalization process, In: W.A. Mackaness, A. Ruas, L.T. Sarjakoski (Eds.), *Generalization of Geographic Information: Cartographic Modelling and Applications*, Elsevier for International Cartographic Association, 2007, pp. 89-111.

Pätynen, V., and Ristioja, J., 2009, Vector based generalization application in the production of small scale databases of the NLS of Finland, *24th International Cartography Conference*, November 15-21, 2009, Santiago, Chile.

Stanislawski, L.V., 2009a, An automated approach to compare generalized hydrographic network features to a benchmark dataset, in: S. Mustiere, M. Sester, F. van Harmelen, and P. van Oosterom (eds.), (2009), *Generalization of spatial information, Dagstuhl Seminar,* April 13-17, 2009. Schloss-Dagstuhl, Germany. Online: http://www.dagstuhl.de/Materials/index.en.phtml?09161

Stanislawski, L.V., 2009b, Feature pruning by upstream drainage area to support automated generalization of the United States National Hydrography Dataset, *Computers, Environment and Urban Systems*, 33(5):325-333.

Stanislawski, L.V., 2008, Development of a knowledge-based network pruning strategy for automated generalisation of the United States National Hydrography Dataset, *11th ICA Workshop on Generalisation and Multiple Representation*. June 20-21, 2008, Montpellier, France.

Stanislawski, L.V., Buttenfield, B.P, Finn, M.P., and Roth, K., 2009, Stratified database pruning to support local density variations in automated generalization of the United States National Hydrography Dataset, *24th International Cartography Conference*, November 15-21, 2009, Santiago, Chile.

Stoter, J., Burghardt, D., Duchêne, C., Baella, B., Bakker, N., Blok, C., Pla, M., Regnauld, N., Touya, G., and Shmid, S., 2009, Methodology for evaluating automated map generalization in commercial software, *Computers, Environment and Urban Systems*, 33(5):325-333.

Töpfer, F., and Pillewizer, W., 1966, The principles of selection, *The Cartographic Journal*, vol. 3, pp. 10-16.

Wolock, D.M. and McCabe, G.J., 1999, Estimates of runoff using water-balance and atmospheric general circulation models, *Journal of the American Water Resources Association,* 35(6):1341-1350.

Wang, Z. 1996, Manual versus Automated Line Generalization, *GIS/LIS '96 Proceedings*, pp. 94–106.

Wang, Z., and Muller, J.C., 1998, Line generalization based on analysis of shape characteristics, *Cartography and Geographical Information Systems,* 25(1):3 - 15.

White, E.R., 1985, Assessment of line-generalization algorithms using characteristic points, *The American Cartographer*, 12(1):17-27.

Xiang, Z., Tinghua, A., and Stoter, J., 2008, The evaluation of spatial distribution density in map generalization, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. 37, Part B2, 2008, Beijing.