# HYDROGRAPHIC FEATURE GENERALIZATION IN DRY MOUNTAINOUS TERRAIN

Lawrence V. Stanislawski [a], Barbara P. Buttenfield [b]

[a] ATA Services, Center for Excellence in Geospatial Information Science (CEGIS), United States Geological Survey (USGS), Rolla, Missouri, USA – lstan@usgs.gov
[b] Department of Geography, University of Colorado, Boulder, Colorado, USA – babs@colorado.edu

**KEY WORDS**: automated generalization, National Hydrography Dataset, coefficient of line correspondence

## ABSTRACT:

A wide 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. This paper presents generalization and data modelling to produce reduced scale versions of hydrographic data for a multi-resolution national data set, *The National Map*, of the United States Geological Survey (USGS). The approach distinguishes regional differences in geographic factors to demonstrate that knowledge about varying terrain and climate conditions can support the design of tailored generalization operations that preserve distinct hydrographic patterns. Hydrographic generalization procedures are being tailored for different terrain (mountainous, hilly, and flat) and climate (humid and dry) conditions within the United States. We demonstrate using a sequence of automated generalization operations tailored for a dry mountainous subbasin watershed of the United States National Hydrography Dataset (NHD). NHD data for the subbasin, compiled from 1:24,000-scale source material, were generalized to create hydrographic data that are appropriate for cartographic mapping at scales between about 1:50,000 and 1:200,000. Generalization results are metrically compared to a 1:100,000-scale NHD benchmark through the Coefficient of Line Correspondence (CLC) and the Coefficient of Area Correspondence (CAC). Confidence intervals for the CLC and CAC are generated through a non-parametric bootstrapping approach. These metrics and associated confidence intervals can help establish the geographic extents that are suitable for each set of tailored generalization procedures.

## 1. INTRODUCTION

A framework for automated cartographic generalization is being developed by the U.S. Geological Survey (USGS) to support the display and delivery of *The National Map* products. The framework must be able to handle the wide variety of climate and terrain conditions that exist in the United States. For instance, within the conterminous United States, elevations range from about 85 meters below sea level in Death Valley, California, to greater than 4000 meters above sea level in some parts of the Rocky and Cascade-Sierra Mountain ranges (USGS, 2001), and average annual precipitation between 1961 and 1990 ranged from 0 to 200 inches (in) (0 to 508 centimetres (cm)) (Daly and Taylor, 2000). Given the range of conditions, various natural landscapes have evolved and must be adequately represented in the multiple-scales of USGS cartographic products through the generalization framework.

The Center of Excellence for Geospatial Information Science (CEGIS) of the USGS is collaborating with the University of Colorado-Boulder and the Pennsylvania State University to develop and test automated processes that enrich, prune, and simplify high-resolution (1:24,000-scale or larger) data from the National Hydrography Dataset (NHD) to produce 1:50,000-scale level of detail (50K LoD) NHD data (Brewer and others, 2009; Stanislawski, 2009; Stanislawski and others, 2009; Buttenfield and others, 2010). Through symbol redesign and elimination operations, the 50K LoD is used to represent hydrographic features on topographic maps ranging in scale from 1:50,000 to 1:200,000 (Brewer and others, 2009). Sequences of enrichment, pruning, and simplification operations that produce 50K LoDs have been manually tailored for six NHD subbasins, which were selected from six different climate (dry, humid) and terrain (flat, hilly, mountainous) regimes that span much of the conterminous United States (Buttenfield and others, 2010). Upon developing the generalization knowledge base for the NHD, subsequent research will investigate terrestrial classification systems that can be coordinated with the tailored generalization sequences to furnish a database of blended generalization parameters, constraints, and operations, which produce 50K LoDs with smooth transitions over landscape boundaries. This strategy follows that of Burghardt and Neun (2006), who proposed a collaborative filtering approach to predict the best sequence of generalization operations for features from a knowledge base of successfully generalized features.

Building a generalization framework that applies operations, which are tailored for different geographic conditions has been proposed by Touya (2008) and is being developed for France (Touya, 2010). Touya’s sophisticated framework simultaneously handles multiple data themes and coordinates collaboration among several generalization models or processes (Touya, 2008). The Mapping Information Branch (MIB) of Canada is developing a database of generalization knowledge that stores integrity constraints and patterns, which, through implementation, orchestrates the automated application of generalization algorithms over the various geographic conditions in Canada (Pilon and others, 2010). The MIB is applying this approach to generate data at 1:250,000-scale from 1:50,000-scale data.

Whereas Buttenfield and others (2010) describe NHD generalization for a humid hilly subbasin, this paper describes generalization procedures developed by CEGIS and the Colorado team to produce a 50k LoD for a dry mountainous NHD subbasin. To validate results, the 50k LoD is compared to a benchmark dataset, the 1:100,000-scale (100K) NHD, through the coefficient of line correspondence (CLC) and the coefficient of area correspondence (CAC).

## 2. METHODS

### 2.1 Piceance-Yellow Subbasin

Features in the high-resolution (HR) Piceance-Yellow NHD subbasin (14050006) were generalized to 50k LoD for this study. The subbasin covers about 2,370 square kilometres (km$^2$) in the Colorado Rocky Mountains and is the watershed for the Piceance and Yellow Creek stream channels. Fenneman and Johnson (1946) identified eight physiographic divisions in the coterminous United States, with each having similar topography, rock types and structure, and geologic and geomorphic history (Figure 1). The Piceance-Yellow subbasin lies in the Intermontane Plateaus physiographic division (Figure 1).

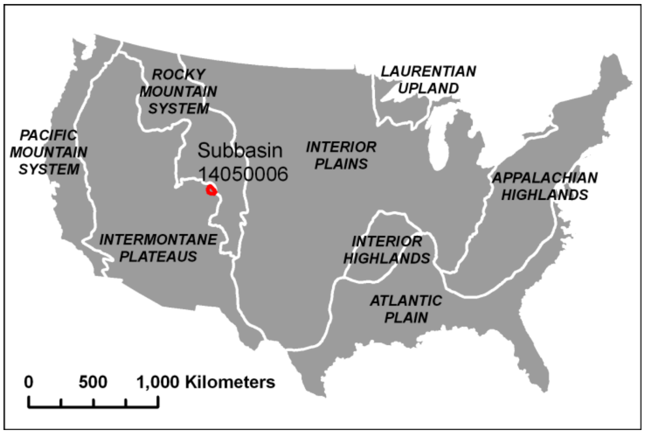

Figure 1. Piceance-Yellow Creeks subbasin (14050006) outlined in red within the physiographic division of the coterminous United States (Fenneman and Johnson, 1946).

There are 3,768 flow-directed (FD) features in the subbasin HR NHD flowline feature class, some of which were updated on 13 January 2010. FD flowlines comprise about 3,297 kilometres (km) of stream channels, which is 1.39 kilometres per km$^2$ (km/km$^2$) overall drainage density for the subbasin. About 32 km of flowline features without flow direction exist in the subbasin and these, largely composed of ephemeral streams, were removed from subsequent analysis. The remaining FD flowlines were compiled from 1:24,000-scale (24K) source material. A total of 196 polygonal hydrography features—lakes, ponds, streams, rivers, washes, and swamps--exist in the subbasin and cover about 1.82 km$^2$.

### 2.2 Generalization Procedures

The Piceane-Yellow HR NHD subbasin was enriched and differentially generalized, which employs pruning and additional generalization operations that consider feature types and local densities (Buttenfield and others, 2010). NHD enrichment consists of estimating catchment area, upstream drainage area (UDA), a density partition, and symbolization attributes for each FD flowline feature. Through automation, catchment areas were estimated using Thiessen polygons (Stanislawski and others, 2007), and UDA values were estimated by accumulating catchment areas through an augmented-directed graph traversal algorithm (Stanislawski and others, 2006). Flowline features were manually partitioned into three (low, medium, high) density classes, having 1.07, 1.43, and 2.03 km/km$^2$ drainage densities, respectively (Figure 2). Automated partitioning techniques are under development, but are not fully implemented for this analysis.

Following the described enrichment processes, flowline and polygon features were automatically pruned to remove less prominent features. A variation of the Radical law (Töpfer and Pillewizer, 1966) was separately applied to compute a target 50K density for each 24K partition. Computed 50K target densities were 0.74, 0.99, and 1.40 km/km$^2$ for the low, medium, and high density partitions, respectively. Each partition was iteratively pruned to a minimum UDA tolerance, which was incrementally increased until the associated 50K target density was achieved for the partition. Topological connectivity was maintained among remaining flowline features.

After pruning flowlines of all partitions, the 24K polygon features smaller than one square inch (6.45 square cm) at map scale were pruned. First, polygons that had been stripped of all flowlines passing through them during the flowline pruning process were removed. Minimum area constraints compiled by feature type from 100K NHD standards were then used to remove smaller polygons from the remaining polygons.

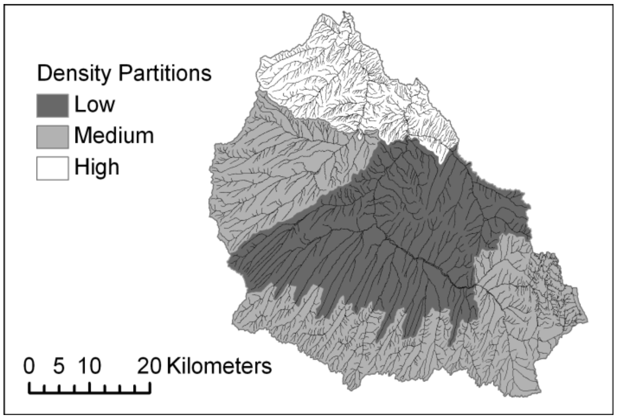

Figure 2. Density partitions manually assigned for National Hydrography Dataset flowlines in the Piceance-Yellow subbasin (14050006) in Colorado.

Generalization operations subsequent to pruning were tailored to preserve geometric characteristics (e.g. vertex spacing, curve shape) similar to the 100K NHD features in the subbasin. These operations consisted of removal of all overlapping polygons of less prominent feature types—such as, submerged streams or rapids—that may be present, and simplification of retained polygon edges and flowlines. Simplification of polygon boundaries and flowlines was completed with the Bend Simplify algorithm (Wang, 1996; Wang and Muller, 1998) using tolerance thresholds of 200 m and 150 m, respectively.

Enrichment, pruning, differential generalization, and validation (described in next section) procedures have been logically subdivided into tools and incorporated into an NHD generalization toolbox implemented largely through Python and the ESRI ArcGIS geoprocessing environment. The toolbox allows an interactive approach for evaluating various generalization options and provides the ability to string together a series of tools for batch processing in a production setting.

Harrie and Weibel (2007) indicate three categories of models for batch processing the overall generalization process: condition-action, human interaction, and constraint based. Although the generalization procedures described above are a work in progress and have yet to incorporate interactions between other data themes, the procedures fall within the range of the condition-action and human interaction models described by Harrie and Weibel (2007). The constraint-based approach could be achieved through future enhancements that incorporate formalized constraints into the use of the tools.

### 2.3 Validation: Comparison to Benchmark

The 50K LoD was compared to the 100K NHD through CLC and CAC computations to validate the generalized data to a benchmark. The CLC and CAC estimates how well the lines and polygons match between the two sets of data, respectively. A spatial distribution of the CLC and CAC was generated by comparing features within a grid of 201, 3.66 by 3.66 km cells that cover the subbasin. Using a scale-based buffer of 152.4 m to estimate the accuracy of the data, omission and commission errors were determined for line and polygon features in weighted CLC and CAC values for each cell, respectively (Stanislawski and others, 2010; Buttenfield and others, 2010). Weighted cell values for CLC and CAC were summarized to compute the subbasin CLC and CAC values, and confidence intervals for the subbasin values were estimated through a bootstrap resampling approach (Stanislawski and others, 2010).

## 3. RESULTS AND DISCUSSION

### 3.1 Generalization

The HR NHD data before and after pruning and differential generalization for the Piceance-Yellow subbasin is shown in figure 3. Densities achieved after automated pruning for the low, medium, and high density partitions are 0.74, 0.99, and 1.40 km/km$^2$, respectively, which are precisely equivalent to the desired target densities estimated for 50K by the radical law. Thus, local density variations are maintained in the 50K

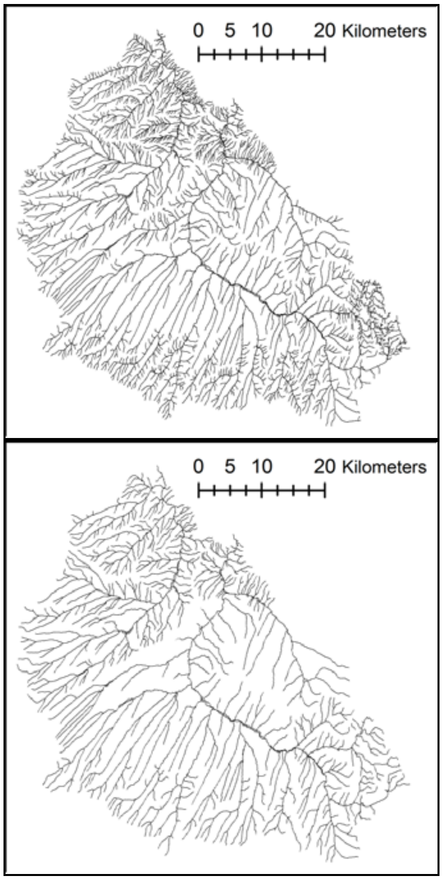

Figure 3. High-resolution NHD data before (above) and after (below) pruning and differential generalization to produce a 1:50,000-scale level of detail for the Piceance-Yellow subbasin.

LoD, although this does not appear as distinct as depicted in the HR data. Though not visible in figure 3, the number of hydrography polygons was reduced from 196 to 123 through pruning, with lakes, ponds, and washes maintained in the 50K LoD.

Line simplification decreased the total length of the pruned flowline features by about 2.5 percent; however, the total area of the 123 hydrography polygons after pruning was increased by about 15 percent through the simplification process of the polygon boundaries. The reduction of flowlines because of simplification was an expected result. The effect of boundary simplification increasing the polygon areas will be further investigated on other prototype subbasins.

### 3.2 NHD polygons at 1:100,000-scale

The HR NHD includes 26 polygon feature types. 100K NHD polygon feature types within each subbasin of the 100K NHD were summarized by area and frequency for the conterminous United States. This was completed to identify less significant feature types that may be removed from 50K LoD representations at 100K and smaller scales, and to visibly inspect the distribution of 100K hydrography polygons.

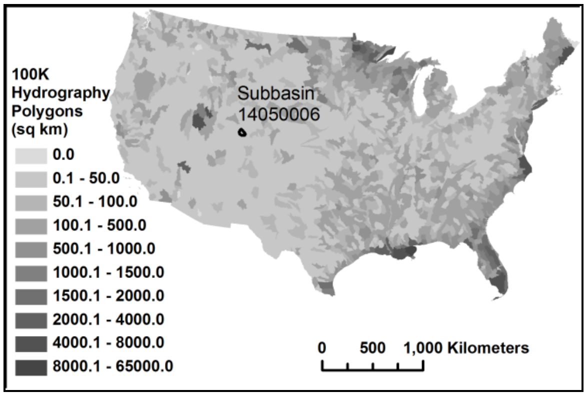

Figure 4. Distribution of square kilometres of 1:100,000-scale NHD hydrography polygons, excluding playas and washes, within each subbasin within the conterminous United States.

The summary of the 100K NHD polygons indicates that the 100K subbasins include from 0 to 4,692 hydrography polygons, which range from 0 km$^2$ to more than 64,000 km$^2$. Removing the ephemeral and dry feature types of wash and playa from the 15 polygon feature types present in the 100K NHD produces the distribution of polygonal features within the conterminous United States as shown in figure 4. The dry-to-wet, west-to-east pattern in the distribution resembles the pattern displayed by mean annual runoff and precipitation estimates for this area (Daly and Taylor, 2000). Aside from only displaying the 13 of 15 polygon feature types found in the 100K NHD, this pattern of feature types can be replicated when drawing 50K LoD features on maps with scales of 100K or smaller.

Washes exist within the 100K NHD, but not in the Piceance-Yellow subbasin; however, the 50K LoD includes 16 wash polygons because they were larger than the minimum size criterion. Along with other conditional rules, wash features can be masked by other features, or removed when displaying 50K LoD features on maps having scales of 100K or smaller (Brewer and others, 2009).

### 3.1 Generalization validation

The distributions of weighted CLC and CAC values are displayed in figure 5. Subbasin CLC and CAC values are 0.80 and 0.68, respectively, indicating about 80 and 68 percent of the 50K LoD features match the line and polygon features in the 100K NHD. About 13 and 7 percent of the line mismatches are due to commissions and omissions, respectively. None of the 100K polygons were determined to be omitted from the 50K LoD, with all polygon mismatches in the CAC due to extra polygon features being included in the 50K LoD and not in the 100K NHD benchmark. However, removing from the 50K LoD the polygon features of type wash, which are not typical of 100K features in this part of the country, improves the subbasin CAC to 0.78 (figure 6).

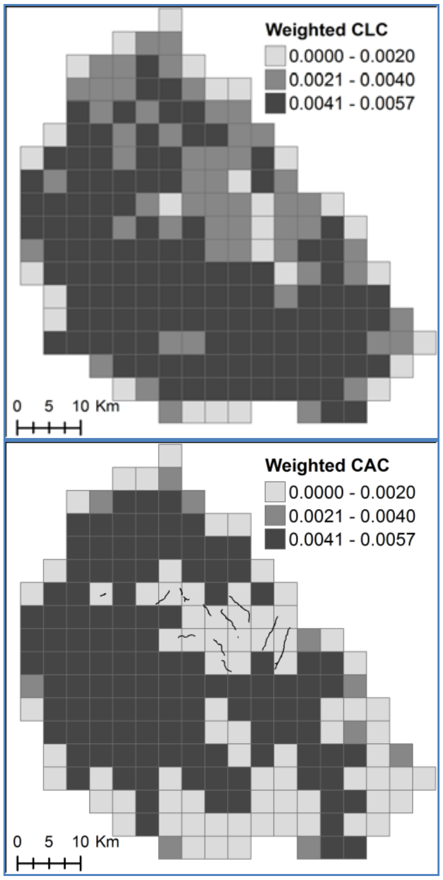

Figure 5. Distribution of weighted coefficient of line correspondence (CLC) (above) and weighted coefficient of area correspondence (CAC) (below) values for the 201, 3.66 by 3.66 km cells in the Piceance-Yellow subbasin. Weighted CLC and CAC measure the amount of match between line and polygon features of the 1:50,000-scale level of detail (50K LoD) and 1:100,000-scale NHD datasets, respectively. Higher values (dark shade) represent better match between

features in the two data sets. Wash features from the 50K LoD are overlain on the CAC grid.

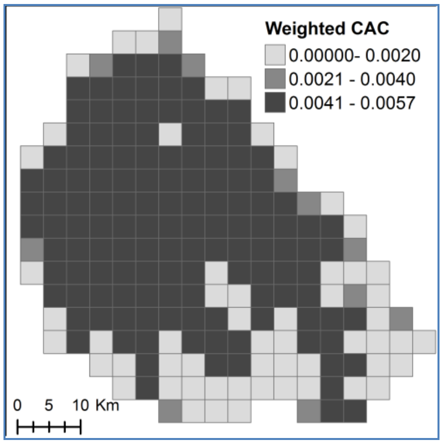

Figure 6. Distribution of weighted coefficient of area correspondence (CAC) values for the 201, 3.66 by 3.66 km cells in the Piceance-Yellow subbasin. Weighted CAC measures the amount of match between polygon features of the 1:50,000-scale level of detail (50K LoD) **after** removal of wash polygons and 1:100,000-scale NHD datasets. Higher values (dark shade) represent better match between features in the two data sets.

The lower and upper bounds of the 90 percent confidence interval estimated for the CLC are 0.775 and 0.819. Lower and upper bounds of the 90 percent confidence interval estimated for the CAC with washes removed from the 50K LoD are 0.735 and 0.831. These precision estimates help clarify whether significant changes are generated from different generalizations of the same subbasin, or when comparing generalizations of different subbasins. CAC and CLC and associated precision estimates can be used to define generalization constraints for subbasins.

## 4. CONCLUSIONS

This paper demonstrates the use of the NHD generalization toolbox to generate a 50K LoD of NHD data from the HR NHD data, which is compiled at 24K or larger scale. The test subbasin is situated in a dry mountainous environment. Feature simplification operations were tailored for the subbasin based upon physiographic characteristics (dry landscape, mountainous terrain). Subbasin processing from enrichment through simplification took about one hour for this subbasin consisting of roughly 4000 features. An additional two and a half hours were required to complete validation procedures, but alternative processing techniques could vastly reduce validation time.

Advantages of fully automated generalization include reduced workloads and greater consistency of results. The described automated validation techniques are valuable for refining generalization procedures. Metric assessment of the quality of generalization has been a longstanding goal for mapping agencies and researchers alike (Mackaness and others, 2007). The CAC and CLC along with associated precision estimates suggest the resulting 50K LoD matches about 75 percent of the 100K NHD representations, with the majority of mismatches due to commission errors, as logically expected. Disadvantages of fully automated generalization include additional work to formalize pruning and simplification parameters. Limitations of generalization algorithms also occur, but these problems would be incurred with manual processing as well.  For example, simplification of polygon feature boundaries increased polygon areas. Furthermore, simplification introduces some degree of channel displacement, which cannot be avoided and may impact the CLC metric. Implications of these effects will be investigated further on other prototype subbasins.

The NHD generalization toolbox being developed through this research provides automated processes that generalize HR NHD to reduced levels of detail for cartographic purposes. Additional refinements are necessary, particularly to generate LoD data with the full integrity of the NHD model that can be distributed for applications. The toolbox includes tools for automated data enrichment with density partitioning, generalization operation sequences tailored for specific regimes in the country, and validation procedures. Additional research will transform existing operations into a series of operation sequences that blend generalized data over terrain gradients between physiographic regimes.

## ACKNOWLEDGEMENTS

The authors thank Christopher Anderson-Tarver of University of Colorado-Boulder, and Ryan Haney and Jeremy Koontz of the U.S. Geological Survey for assistance with data processing and programming.

## REFERENCES

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 Cartographic Conference*, 15-21 Nov. 2009, Santiago, Chile.

Burghardt, D., and Neun, M. 2006. Automated sequencing of generalization services based on collaborative filtering. In: Raubal, M., Miller, H.J., Frank, A.U., and Goodchild, M. (eds.), *GIScience 2006*, pp. 41-46.

Buttenfield, B.P., Stanislawski, L.V., Brewer, C.A. 2010. Multiscale representations of water: Tailoring generalization sequences to specific physiographic regimes. *GIScience 2010, 14-17* Sep. 2010, Zurich, Switzerland.

Daly, C., and Taylor, G. 2000. United States average annual precipitation, 1961-1990. http://www.nationalatlas.gov/mld/prism0p.html. (accessed 31 Aug. 2010).

Fenneman, N.M., and Johnson, D.W. 1946. Physical Divisions of the United States: Washington, D.C., USGS special map series, scale 1:7,000,000. http://water.usgs.gov/GIS/metadata/usgswrd/XML/physio.xml (accessed 1 Sep. 2010).

Harrie, L., and Weibel, R. 2007. Modelling the overall process of generalization. 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. 67-87.

Mackaness, W. A., Ruas, A., and Sarjakoski, L.T. (eds.) 2007. *Generalization of Geographic Information: Cartographic Modelling and Applications.* Elsevier for International Cartographic Association, 370 pp.

Pilon, D., Beaulieu, A., and Sabo, N. 2010. The generalization of the Canadian landmass: a Federal perspective. *13th ICA Workshop on generalization and multiple representation*, 12-13 Sep. 2010. Zurich, Switzerland.

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

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*, 15-21 Nov. 2009, Santiago, Chile.

Stanislawski, L.V., Buttenfield, B.P, and Samaranayake, V.A. 2010. Generalization of hydrographic features and automated metric assessment through bootstrapping. *13th ICA Workshop on generalization and multiple representation.* 12-13 Sep. 2010, Zurich, Switzerland.

Stanislawski, L.V., Finn, M., Starbuck, M., Usery, E.L., and Turley, P. 2006. Estimation of accumulated upstream drainage values in braided streams using augmented directed graphs. *AutoCarto 2006*, 26-28 June 2006, Vancouver, Washington.

Stanislawski, L.V., Finn, M., Usery, E.L., and Barnes, M. 2007. Assessment of a rapid approach for estimating catchment areas for surface drainage lines. *Proceedings ACSM-IPLSA-MSPS*, 9-12 March 2007, St. Louis, Missouri.

Töpfer, F., and W. Pillewizer. 1966. The principles of selection: a means of cartographic generalization. *The Cartographic Journal* 3(1), pp. 10-16.

Touya, G. 2008. First thoughts for the orchestration of generalization methods on heterogeneous landscapes. *12th ICA Workshop on generalization and multiple representation.* 20-21 June 2008, Montpellier, France.

Touya, G. 2010. Relevant space partitioning for collaborative generalization. *13th ICA Workshop on generalization and multiple representation.* 12-13 Sep. 2010, Zurich, Switzerland.

USGS. 2001. Elevations and distances in the United States. http://egsc.usgs.gov/isb/pubs/booklets/elvadist/elvadist.html#toc. (accessed 31 Aug. 2010).

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), pp. 3 - 15.